YOUR LARGE VISION-LANGUAGE MODEL ALREADY KNOWS HOW TO LOCALIZE CHANGES

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual change understanding necessitates both *seeing* and *saying*. However, existing approaches typically realize only one of these capabilities: LVLMs excel at *saying* changes but struggle to precisely *see* them, whereas localization-only methods can accurately *see* changes but lack the semantic capacity to *say* what those changes represent. To fill this gap, we endow LVLMs with the ability to reliably *see* changes. By investigating the attention head representations in frozen LVLMs, we find that a small subset of attention heads ($\leq 0.7\%$ of all attention heads) already encodes change-informative spatial evidence. We excavate these heads by measuring their attention sum deviation from an unchanged self-reference baseline and their spatial consistency across both image directions. Despite being selected without pixel-level supervision, these heads produce responses that closely align with the ground truth. Furthermore, we diffuse their attention through semantic affinities extracted from the frozen vision encoder—rendering stable maps for effective SAM refinements. Extensive experiments across diverse change localization benchmarks demonstrate that our training-free framework substantially surpasses LVLM-based baselines and matches dedicated localization-only approaches. Our results suggest that LVLMs encode an innate spatial understanding of changes via distinct attention patterns, which, combined with their world knowledge and strong descriptive capabilities, enables comprehensive and generalizable change understanding.

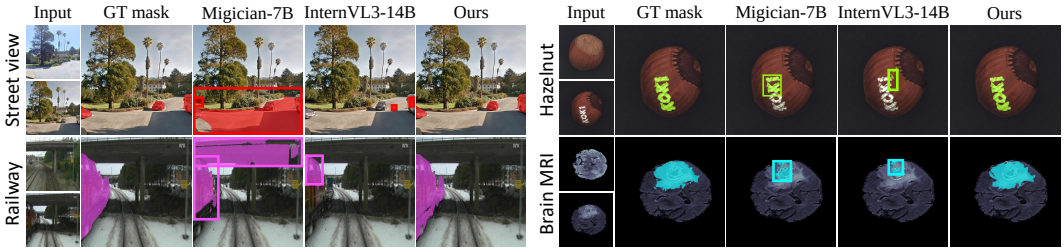


Figure 1: **Change localization results.** Existing LVLMs with bounding-box-guided SAM struggle to localize changes, whereas ours—built on InternVL3-14B—accurately identifies the changed regions.

1 INTRODUCTION

Humans effortlessly understand visual discrepancies (Gleitman et al., 2007; Hollingworth, 2007): they concurrently identify where a change occurred, articulate what has changed, and express its context and significance in natural language (Park et al., 2019; Zou et al., 2022). This seamless integration of perceptual recognition and linguistic reasoning transforms low-level visual differences into actionable understanding—enabling adaptive decision-making in ever-evolving environments (Kim et al., 2026b; Yoon & Kim, 2026). Replicating this capability to naturally *see* and *say* changes therefore has long been a central goal for AI, with applications spanning incident monitoring, autonomous driving, anomaly inspection, media forensics, and medical diagnosis (Sakurada et al., 2018; Bergmann et al., 2019; Baid et al., 2021; Wang et al., 2022a; Tu et al., 2024; Liu et al., 2024; Kim & Kim, 2025).

The recent emergence of Large Vision-Language Models (LVLMs) offers a compelling path toward this goal (Yang et al., 2025; Zhu et al., 2025b; An et al., 2026). By jointly processing image pairs and reasoning over their relationships, modern LVLMs demonstrate a remarkable ability to describe

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.

complex changes across diverse scenarios (Wang et al., 2024a; Xu et al., 2024; Kim et al., 2026a). However, as illustrated in Figure 1, their ability to *say* changes does not translate into precise *seeing*. When asked to localize changes, LVLMs often produce inaccurate bounding boxes—resulting in poor masks even after SAM refinement. This reveals a fundamental output gap: LVLMs can articulate what has changed but cannot reliably externalize the corresponding spatial evidence through coordinate generation. Conversely, localization-only methods built upon vision foundation models (Kirillov et al., 2023; Oquab et al., 2023; Lin et al., 2024; Kim & Kim, 2025) can accurately delineate changed regions but cannot explain what those changes mean; precise localization and semantic description remain disconnected. This asymmetry raises our central question: How can we unlock the spatial knowledge underlying what LVLMs already know how to *say*, enabling them to reliably *see* changes?

One possible alternative is to fine-tune LVLMs with grounding supervision; however, extending this paradigm to change localization requires costly bitemporal annotations (Zheng et al., 2021; Kim & Kim, 2025)—constraining the scale and diversity of supervision required for broad generalization. Thus, to overcome this limitation, we design two key innovations: *attention head filtering criteria* and *semantic-guided attention diffusion*. By investigating the attention head representation of the frozen LVLMs, we discover that change-informative spatial evidence is already encoded in a small subset of attention heads. We identify these heads using two principled criteria: *Attention Sum Deviation* selects heads whose bidirectional responses deviate strongly from a semantically unchanged self-reference, while *Attention Spatial Consistency* retains heads that localize the same changed regions under input-order reversal. Surprisingly, despite being selected without pixel-level supervision, these heads produce spatial responses that consistently align with the ground-truth change masks. We then propagate their peak-dominated responses through semantic affinities from the frozen vision encoder—producing coherent maps for effective SAM refinement. Finally, as described in Figure 1, our method endows LVLMs with the ability to *see* changes—all without annotation-intensive training data—by unlocking the full potential inherited in their hidden visuospatial representation.

We validate our framework across ten LVLMs spanning diverse parameter scales, architectures, and training corpora, as well as across a broad range of change recognition benchmarks—encompassing urban and suburban environments, rural railway scenes, manufactured products, and brain MRI. Our framework outperforms the existing LVLM-based methods by significant margins. Furthermore, our method performs comparably to dedicated localization-only methods that leverage vision foundation models. These results show that modern LVLMs can effectively serve as change localizers, intrinsically encoding their understanding of changes as a spatially meaningful pattern within their attention maps. To the best of our knowledge, this is the first effort to reveal this capability and amplify it into a solid framework for comprehensive and generalizable change understanding.

To summarize, we highlight the main contributions of this paper below:

- We identify that change localization signals are already encoded in a sparse set of LVLM's attention heads and present two principled criteria for systematically selecting them.
- We design a simple training-free framework that leverages selected attention heads and the LVLM's vision encoder—enabling LVLMs to *see* changes.
- We evaluate our method across a diverse set of LVLMs. Our framework significantly outperforms previous LVLM-based approaches and even matches dedicated localization-only methods.

2 Related Work

We discuss the most relevant literature here and provide more scope and discussion in Appendix E.

Visual change understanding has progressed through two complementary fields: change captioning, which describes what has changed in natural language (Park et al., 2019; Forbes et al., 2019; Shi et al., 2020); and change localization, which identifies where changes occurred in pixel-wise change masks (Alcantarilla et al., 2016; Sakurada et al., 2017)—the focus of this study. While change captioning has traditionally relied on specialized architectures and paired-image annotations, the advance of LVLMs has substantially reshaped this landscape. Building upon large-scale vision-language pretraining and increasingly capable multi-image reasoning (Xu et al., 2024; Wang et al., 2024a; Meng et al., 2024), LVLMs now leverage their broad world knowledge to directly compare image pairs and describe their differences without task-specific training (Tu et al., 2023a; 2024). As demonstrated in C3-Bench (Kim et al., 2026a), modern LVLMs substantially outperform conventional change captioning models

across a broad spectrum of real-world change contexts—marking a remarkable leap in open-domain change understanding. Therefore, the primary bottleneck is no longer verbalizing the semantics of changes, but whether this understanding can be reified visually. Accordingly, this paper focuses on change localization as the missing piece towards more comprehensive change understanding.

Early change localization methods primarily used CNNs to learn pairwise representations (Alcantarilla et al., 2016; Wang et al., 2022a). With the emergence of vision foundation models, this paradigm has shifted toward leveraging robust pretrained features rather than learning representations from scratch. Methods based on SAM and DINOv2 accomplish this by introducing lightweight adapters (Li et al., 2025c) or task-specific decoders for change mask prediction (Lin et al., 2024). This direction has also enabled training-free approaches—such as GeSCF (Kim & Kim, 2025)—which repurpose SAM's latent features for zero-shot change localization. However, despite their robustness, these approaches can only delineate where changes occurred, not explain what those changes mean in natural language.

Complementing these vision-only approaches, modern LVLM-based localization methods provide a strong baseline for change localization. Several studies have explored localization in single-image settings (Rasheed et al., 2023), where models localize a text-specified target through coordinate generation (Ren et al., 2024), specialized `[seg]` tokens (Lai et al., 2023), or internal attention signals (Wu et al., 2024a; Kang et al., 2025). More recently, their capability has been extended to multi-image grounding; for example, Migician performs free-form grounding across multiple images through a dedicated multi-image grounding dataset (Li et al., 2025a). However, these approaches assume that the localization target is readily identifiable from the given instruction (e.g., "*where did I park the car from Image 1 in Image 2?*"). Our change localization poses a distinct challenge: As established by (Kim et al., 2026a), trustworthy change reasoning requires LVLMs to jointly compare *two images* under a *prompt* that defines which differences constitute meaningful changes and which should be dismissed as pseudo-changes—the localization target is not given a priori. Therefore, reliable change localization must also be grounded in the same context-conditioned reasoning—a setting that remains difficult to address with existing LVLM-based localization methods (see results in Section 5).

3 PRELIMINARIES

Since our method employs LVLMs' internal representation, we first recap their relevant architecture.

Terminology. LVLMs typically consist of a vision encoder, a modality projector, and a large language model (Liu et al., 2023). Given an input image pair (I_a, I_b), the vision encoder $\mathcal{E}_v$ first independently extracts their visual token sequences $F_a{=}\mathcal{E}_v(I_a)$ and $F_b{=}\mathcal{E}_v(I_b)$, where $F_a, F_b \in \mathbb{R}^{P^2\times d_v}$, P^2 denotes the number of visual patch tokens, and d_v is the visual feature dimension. The projector $\mathcal{P}$ then maps the visual tokens into the language embedding space, yielding $Z_a^v{=}\mathcal{P}(F_a)$ and $Z_b^v{=}\mathcal{P}(F_b)$, while the text embedding layer $\mathcal{E}_t$ encodes the prompt T as $Z^t{=}\mathcal{E}_t(T)$. Here, $Z_a^v, Z_b^v \in \mathbb{R}^{P^2\times d_t}$ and $Z^t \in \mathbb{R}^{L\times d_t}$, where d_t denotes the language embedding dimension and L is the number of text tokens. Finally, the visual and textual embeddings are concatenated as $Z^0{=}[Z_a^v; Z_b^v; Z^t] \in \mathbb{R}^{(2P^2+L)\times d_t}$ to form the input sequence to the language model.

Multi-Head Self Attention. The input embeddings Z^0 are passed through a sequence of decoder blocks, each containing multi-head self attention and a feed-forward network (Vaswani et al., 2017). We focus on the self-attention heads, as they operate at the finest granularity for token-level interactions. At layer ℓ and head h, the hidden state from the previous layer $Z^{\ell-1}$ is projected into query Q, key K, and value $V \in \mathbb{R}^{(2P^2+L)\times d_h}$ matrices, where d_h is the hidden dimension of the attention head. Then, the attention head computes the attention weights as:

$$\mathbf{A}^{\ell,h}(Z^{\ell-1}) = \text{Softmax}(QK^T/\sqrt{d_h}) \in \mathbb{R}^{(2P^2+L)\times(2P^2+L)}. \tag{1}$$

Here, attention weights quantify the similarity between the query Q and key K matrices.

Text-to-Image Attention Maps. Given the autoregressive nature of large language models, information accumulates from preceding tokens to subsequent ones—causing the final token to encapsulate the semantics of the entire sentence (Kang et al., 2025; Kim et al., 2025). Therefore, to *reify where an LVLM focuses on* under a given instruction, we examine how the query embedding of the final prompt token q_t attends to the visual tokens of each input image. Specifically, we define the attention weight vector as:

$$\mathbf{a}^{\ell,h} = \text{Softmax}(q_tK^T/\sqrt{d_h}) \in \mathbb{R}^{1\times(2P^2+L)}. \tag{2}$$

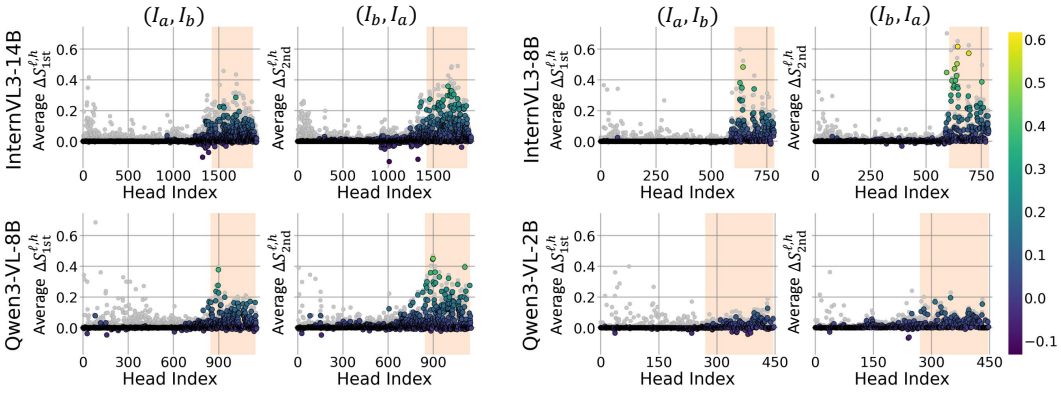


Figure 2: **Average $\Delta S_{1st}^{\ell,h}$ and $\Delta S_{2nd}^{\ell,h}$ values for each attention head**. Across different LVLMs, attention heads exhibit similar patterns, with the heads at middle-to-late layers consistently showing strong responses. The gray points denote heads that were highly activated before the subtraction.

We then extract the visual token segments $\mathbf{a}^{\ell,h}[:, 0:P^2]$ and $\mathbf{a}^{\ell,h}[:, P^2:2P^2]$, which correspond to the text-to-image attention maps for the first and second images—denoting them as $\mathbf{a}_{\text{1st}}^{\ell,h}$ and $\mathbf{a}_{\text{2nd}}^{\ell,h}$. Importantly, these maps are defined *positionally*: for (I_a, I_b), they correspond to I_a and I_b, whereas swapping the input image order to (I_b, I_a) reverses the correspondence.

4 Methodology

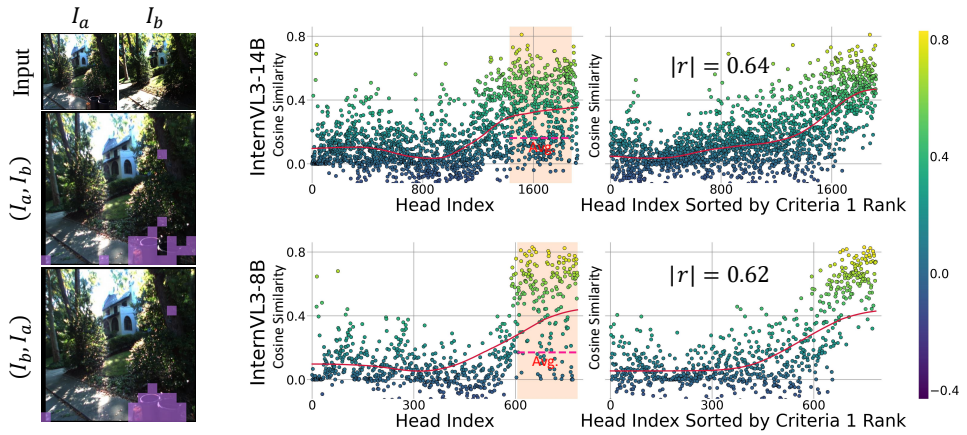
As demonstrated in prior studies (Voita et al., 2019; Clark et al., 2019; Raghu et al., 2021; Walmer et al., 2022; Li et al., 2023; Gould et al., 2023; Kumar et al., 2023; Cabannes et al., 2024; Zheng et al., 2024; Bi et al., 2024; Kang et al., 2025; Rawlekar et al., 2026), attention heads exhibit distinct functional specializations. Here, our objective is to discover change-informative attention heads and translate their spatial signals into final change masks.

For the following analyses, we randomly sample 1,000 image pairs from the VL-CMU-CD training set (Alcantarilla et al., 2016) and investigate LVLMs' text-to-image maps. This choice offers two key advantages: *(i)* As changed instances in VL-CMU-CD are in only one image (I_a) within each pair, it isolates attention head responses without confounding potential effects from simultaneous changes in both images; and *(ii)* Severe illumination, weather, and seasonal differences introduce challenging pseudo-changes, retaining attention heads that respond to meaningful changes rather than superficial appearance shifts. Despite being investigated under this one-sided setting, our method generalizes to diverse settings where changes may occur in both images—such as ChangeVPR (Kim & Kim, 2025). Appendix C.2 further confirms robust performance across head-analysis datasets.

4.1 Attention Head Filtering Criteria

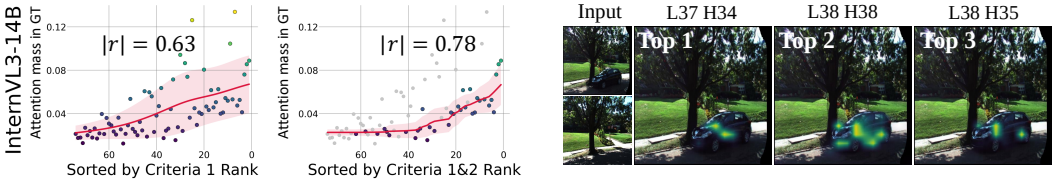
Inspired by prior studies (Kang et al., 2025; Kim & Kim, 2025), we posit that effective localization requires *change-responsive* and *order-consistent* attention heads. For this, we present two key criteria.

Criterion 1: Attention Sum Deviation. In general, attention sum directly measures how strongly a head responds to each image in the changed input pair (I_a, I_b). However, a high activation does not necessarily indicate strong change responsiveness. This is because attention values may also be driven by salient objects, distinctive textures, or other image-specific artifacts unrelated to the change (Siméoni et al., 2021; Wang et al., 2022b; Rawlekar et al., 2026). To address this issue, we measure the *deviation* from a self-augmented, unchanged baseline $(I_a, \hat{I}_a)$ rather than investigating the changed pair alone. Specifically, we build $\hat{I}_a$ by applying pseudo-change transformations (Ramkumar et al., 2021)—such as photometric perturbations—to the change-containing image I_a. We then subtract the resulting baseline attention map element-wise from that of (I_a, I_b) and sum the residual responses to obtain $\Delta S_{\text{1st}}^{\ell,h} = \sum_{i=1}^{P^2} (\mathbf{a}_{\text{1st}}^{\ell,h}[i] - \hat{\mathbf{a}}_{\text{1st}}^{\ell,h}[i])$. Applying the same procedure under the reversed input order (I_b, I_a) yields $\Delta S_{\text{2nd}}^{\ell,h}$. Figure 2 visualizes these deviations averaged over 1,000 samples. Surprisingly, strong deviations concentrate in a sparse subset of middle-to-late layer heads, while early-layer



(a) Position bias in LVLM (b) Attention spatial consistency analysis

Figure 3: **(a) Position bias in LVLM**. When the input order is reversed, the LVLM's attention focus shifts even though it refers to the same changed instances. **(b) Attention spatial consistency analysis**. As heads rank higher under criterion 1, spatial consistency increases, with a Pearson correlation $|r|>0.6$. Moreover, the corresponding heads exhibit cosine similarities far above the overall average.



(a) Attention mass in GT across criteria (b) Top-3 selected heads

Figure 4: **(a) Attention mass in GT across criteria**. We report the average $\Delta_{\mathrm{GT}} S^{\ell,h}$ for bidirectional deviation maps (1st and 2nd). The gray points indicate filtered heads. **(b) Top-3 selected heads**. Here, L37 H34 refers to the 34th attention head in the 37th layer of the LVLM (Kang et al., 2025).

activations are largely suppressed through baseline subtraction. Based on this insight, we consider heads with large deviations under both input directions highly responsive to visual changes.

Criterion 2: Attention Spatial Consistency. While criterion 1 coarsely excavates heads that activate strongly under both input orders, it does not guarantee that these responses correspond to the same spatial regions. As depicted in Figure 3(a), position biases (Wu et al., 2024b; Tian et al., 2025; Kim et al., 2026a) may cause a head to respond differently depending on whether the changed instance appears first or second in the input sequence. Reliable change localization, therefore, requires heads that remain robust to such biases. We therefore examine the cosine similarity between the patch-level deviation maps $(\mathbf{a}^{\ell,h}_{\mathrm{1st}} - \hat{\mathbf{a}}^{\ell,h}_{\mathrm{1st}})$ and $(\mathbf{a}^{\ell,h}_{\mathrm{2nd}} - \hat{\mathbf{a}}^{\ell,h}_{\mathrm{2nd}})$—both pointing to the changed instance in I_a—and retain the top-k heads whose similarity exceeds τ, preserving only those change-responsive heads that consistently localize the same changed regions. Interestingly, as illustrated in Figure 3(b), the rank of attention sum deviation and spatial consistency exhibit a strong Pearson correlation ($|r|>0.6$). Moreover, layers with heads highly ranked by criterion 1 exhibit cosine similarities well above the overall average. These findings highlight that both criteria provide complementary yet closely related signals, enabling us to identify change-responsive heads with reliable spatial consistency.

4.2 Identifying Change-Informative Heads via Two Criteria

First, we rank all heads independently according to the average $\Delta S^{\ell,h}_{1st}$ and $\Delta S^{\ell,h}_{2nd}$, retaining the top 100 heads from each ordering. We then take their intersection to identify heads that respond strongly under both directions (criterion 1). For InternVL3-14B (Zhu et al., 2025b), this results in 74 heads, approximately 4% of total attention heads. Applying a cosine similarity threshold τ of 0.5 further reduces this set to 33 heads, from which we select the top-k heads with the highest spatial consistency (criterion 2). Figure 4(a) shows that the two criteria effectively identify heads responsive to actual changes. The Pearson correlation between the criterion-induced rank and GT attention mass $\Delta_{\mathrm{GT}} S^{\ell,h}$ increases from 0.63 after criterion 1 to 0.78 after criterion 2—indicating strong correlations.

Similar trends are also observed in other LVLMs in Appendix C.1, demonstrating their generality. This result indicates that our criteria not only select heads that are responsive and spatially consistent, but also favor heads that focus on the actual changed regions, aligned with the GT change mask.

Meanwhile, as illustrated in Figure 4(b), the selected heads exhibit two key characteristics: (*i*) They attend to a shared core of the changed instance, while each attention head covers slightly different surrounding regions—suggesting that aggregating multiple heads can be beneficial; (*ii*) Their attention is highly peak-dominated, causing direct binarization to capture only fragmented changes—motivating *semantic-guided attention diffusion* to propagate sparse responses across coherent changed regions.

4.3 Semantic-guided Attention Diffusion (SAD)

Motivated by the personalized PageRank algorithm (Page et al., 1999), we employ semantic affinities derived from the LVLM's frozen vision encoder (ViT) to diffuse attention values—in a self-contained manner—as ViTs tend to capture semantic relationships across patches (Caron et al., 2021). First, we construct a semantic affinity graph $\mathbf{S}$ with normalized ViT features (lines 1–2). We then retain the top-n neighbors per patch (line 3), symmetrize it, and convert it into a diffusion matrix $\mathbf{W}$ (line 4). This matrix controls attention diffusion to semantic neighbors—propagating higher attention values to more similar patches. Finally, peak-dominated attention $\mathbf{y}$ is diffused to its convergence—yielding the stabilized map $\mathbf{y}_{\text{sol}}$ (lines 5–9); here, we use direct linear solve without iterative updates.

Algorithm 1: Semantic-guided Attention Diffusion (SAD)

Input: Peak attention $\mathbf{y}$, visual features F, n, and α
Output: Refined attention $\mathbf{y}_{\text{sol}} = \lim_{T\to\infty} \mathbf{y}^{(T)}$

1: $\widehat{F}_i \leftarrow F_i / \|F_i\|_2$
2: $\mathbf{S} \leftarrow \widehat{F}\widehat{F}^\top$
3: $\mathbf{A} \leftarrow \text{TopN}(\mathbf{S}, n)$
4: $\mathbf{W} \leftarrow \text{RowNorm}\big(\max(\mathbf{A}, \mathbf{A}^\top)\big)$
5: $\mathbf{y}^{(0)} \leftarrow \mathbf{y}$
6: **for** $t = 0, \ldots, T-1$ **do** ▷ Conceptual recurrence
7: $\quad \mathbf{y}^{(t+1)} \leftarrow \alpha \mathbf{W}\mathbf{y}^{(t)} + (1-\alpha)\mathbf{y}$
8: **end for**
9: **return** $\mathbf{y}_{\text{sol}} = \lim_{T\to\infty} \mathbf{y}^{(T)} = (1-\alpha)(\mathbf{I} - \alpha\mathbf{W})^{-1}\mathbf{y}$

4.4 Localizing Changes with Your Large Vision-Language Model

In the previous subsection, we showed that our two criteria adequately find change-informative heads and presented a straightforward algorithm to stabilize them. Based on this insight, we design a training-free framework for localizing changes (see Figure 5). Specifically, following the procedure in Section 4.2, we select the top-k heads that satisfy both criteria. Next, the image pair and text prompt are fed into the LVLM, and bidirectional deviation maps are extracted from the selected heads. The resulting maps are interpolated and assembled via element-wise summation to form the combined map, which is then stabilized using the SAD algorithm (Section 4.3) to produce the change response map. Finally, we binarize the obtained change response map using its mean value and select sufficiently overlapping SAM or MedSAM (Ma et al., 2025) mask proposals—following the paradigm in GeSCF (Kim & Kim, 2025)—as the final change mask. Note that our framework can readily adapt to diverse change scenarios by exploring the LVLM's knowledge through simple prompt modification (`[context]`). Here, `[context]` specifies the relevant changes to detect and the irrelevant differences to ignore for each change scenario (Kim et al., 2026a). Further details—including full equations, hyperparameters, and prompts—are offered in Appendix B.

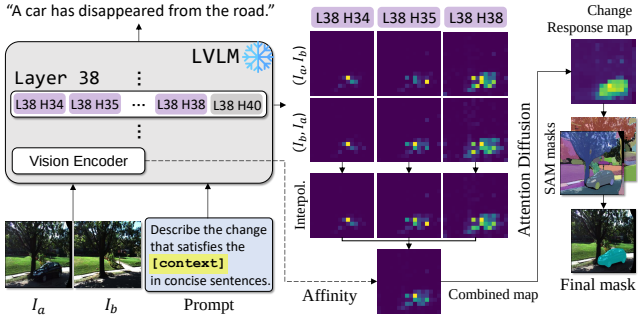


Figure 5: **Our training-free change localization method.** Here, `[context]` can be modified to adapt to different change scenarios. See Appendix B.3 for the full `[context]` prompts.

5 Experiments and Results

5.1 Experimental Setup

Benchmarks. We evaluate the effectiveness of our framework based on two localization benchmarks: *(i)* We adopt the *GeSCD benchmark* of (Kim & Kim, 2025), which evaluates the generalizability of

change localization algorithms on the ChangeVPR dataset. The dataset comprises three challenging splits—SF-XL, St Lucia, and Nordland—spanning urban, suburban, and rural railway scenes under substantial appearance shifts; *(ii)* We further evaluate on the *zero-shot anomaly detection benchmark*—using MVTecAD (Bergmann et al., 2019) and Brain AD (Baid et al., 2021)—to localize anomalies in industrial inspection and brain MRI. Following (Gao, 2025), we formulate anomaly detection as a paired-image setting by matching each query image with a normal reference. We measure F1-Score as our primary metric. Additional details of our benchmark settings are provided in Appendix A.1.

Models. We validate our approach on ten frozen LVLMs to demonstrate its broad applicability. Our main experiments include: Qwen-2-VL (Yang et al., 2024), LLaVA-OneVision families (Li et al., 2024b;a; An et al., 2025; 2026), Qwen-3-VL-Instruct (Yang et al., 2025), and InternVL3 (Zhu et al., 2025b), with model parameters spanning from 2B to 14B. These models natively support multi-image inputs. To broaden our evaluation, we additionally include LLaVA-1.5-7B (Liu et al., 2023), which is originally designed for single-image inputs. Following prior works (Nguyen et al., 2024b; Xu et al., 2024), we extend it to process multiple images and apply the same protocol described in Section 4.4. We select and *fix* the top-k heads from the VL-CMU-CD training set, with k=3 for all LVLMs.

Baselines. To position our method within the broader landscape of change localization, we compare our method against a diverse set of approaches capable of localizing changes: *(i) Previous localization-only methods*, which are explicitly designed to predict pixel-wise change masks (Chen et al., 2021; Wang et al., 2022a; Chen et al., 2023; Lin et al., 2024; Cao et al., 2024; Kim & Kim, 2025; Gao, 2025); and *(ii) Previous LVLM-based methods*. For the frozen LVLMs, following (Ren et al., 2024), we instruct each model to predict bounding boxes and convert them into change masks using SAM. We ensure prediction fidelity by retrying if responses are not aligned with the desired parsing format. We further compare against single-image grounding-specialized LVLMs, including LISA (Lai et al., 2023), GLaMM (Rasheed et al., 2023), LocalizationHeads (Kang et al., 2025), and EGG (Gröpl et al., 2026). As these methods are built on LLaVA-1.5, we use their multi-image-adapted variants mentioned above. Finally, we include Migician-7B (Li et al., 2025a), a Qwen2-VL-7B-based model fine-tuned for multi-image grounding, also paired with SAM. For the localization-only methods, all models in Table 1 except GeSCF are trained on the VL-CMU-CD dataset, while those in Table 2 are evaluated using pretrained weights in their official papers. More details are offered in Appendix A.2.

5.2 Main Results

Tables 1 and 2 present the main results—highlighting the key advantages of our framework as follows:

Unlocking LVLMs' Full Potential. Our framework consistently outperforms all ten LVLM-based baselines. The gap is particularly striking for the LLaVA-OneVision families: localization through textual decoding almost collapses, whereas our method recovers robust performance on the same frozen LVLMs. This result shows that change signals are not missing from LVLMs; instead, they remain poorly accessible through a textual interface. Further, our training-free method applied directly to the original Qwen2-VL-7B substantially outperforms Migician-7B—which acquires multi-image localization capability using explicit fine-tuning on the MGrounding-630K dataset (Li et al., 2025a). This finding suggests that effectively recovering latent understanding can be more advantageous than labor-intensive fine-tuning. Moreover, our LLaVA-1.5-7B significantly outperforms previous single-image grounding-specialized LVLMs—including both fine-tuned methods (LISA and GLaMM) and training-free approaches (LocalizationHeads and EGG). Importantly, LocalizationHeads similarly identifies text-referring heads for localization; however, as demonstrated in Tables 1 and 2, these heads hardly capture changed regions. This result shows that heads dedicated to single-image grounding do not necessarily encode the relational evidence required to identify *changes* (see Appendix C.3 for further discussion). Our method explicitly captures such evidence by selecting heads that strongly activate to change and remain spatially consistent under input-order reversal.

Strong Localization Ability. Our method achieves performance comparable to previous localization-only approaches. As shown in Table 1, our InternVL3-14B performs on par with dedicated methods such as GeSCF and RSCD—which leverage vision foundation models' features (Kirillov et al., 2023; Oquab et al., 2023)—and even outperforms them on the SF-XL and St Lucia splits. Furthermore, our LVLM-based approach enables seamless adaptation across change scenarios through simple prompt-tuning (e.g., "*describe the change that satisfies* `[suburban scenes]`."⇒"*describe the change that satisfies* `[brain anomalies]`."). Despite the substantial domain shift from natural scenes to anomaly imagery—where specialized models are typically required—our method adapts to each scenario by

Table 1: **Performance on ChangeVPR.** Best is in red for localization-only and blue for LVLM-based methods; green denotes gains over previous LVLM baselines.

Methods	SF-XL	St Lucia	Nordland
Previous localization-only methods			
DR-TANet	24.5	19.5	16.8
C-3PO	25.3	10.9	13.6
RSCD	62.6	60.2	44.5
GeSCF	67.6	64.6	50.1
Previous LVLM-based methods			
LISA-7B	27.0	48.3	11.4
GLaMM-7B	28.7	13.8	10.9
LocalizationHeads-7B	7.8	2.9	4.5
EGG-7B	26.5	14.5	6.3
Migician-7B	34.2	28.4	26.0
LLaVA-1.5-7B	0.0	0.0	0.0
Qwen2-VL-7B	5.0	3.3	0.2
LLaVA-OneVision-7B	1.0	0.4	0.5
LLaVA-OneVision1.5-4B	0.0	0.6	0.0
LLaVA-OneVision1.5-8B	0.0	0.0	0.0
LLaVA-OneVision2-8B	23.9	39.2	23.5
Qwen3-VL-2B-Instruct	30.7	36.8	14.6
Qwen3-VL-8B-Instruct	53.4	60.7	33.6
InternVL3-8B	27.0	17.3	14.6
InternVL3-14B	24.3	23.5	1.4
Our LVLM-based methods			
LLaVA-1.5-7B	62.3 (+62.3)	57.7 (+57.7)	20.9 (+20.9)
Qwen2-VL-7B	57.2 (+52.2)	58.9 (+55.6)	27.1 (+26.9)
LLaVA-OneVision-7B	63.5 (+62.5)	65.9 (+65.5)	43.2 (+42.7)
LLaVA-OneVision1.5-4B	69.4 (+69.4)	61.8 (+61.2)	39.3 (+39.3)
LLaVA-OneVision1.5-8B	66.1 (+66.1)	61.2 (+61.2)	42.2 (+42.2)
LLaVA-OneVision2-8B	65.3 (+41.4)	68.0 (+28.8)	42.9 (+19.4)
Qwen3-VL-2B-Instruct	46.9 (+16.2)	50.7 (+13.9)	16.0 (+1.4)
Qwen3-VL-8B-Instruct	67.9 (+14.5)	67.5 (+6.8)	42.5 (+8.9)
InternVL3-8B	69.1 (+42.1)	67.8 (+50.5)	27.9 (+13.3)
InternVL3-14B	69.7 (+45.4)	71.7 (+48.2)	48.4 (+47.0)

Table 2: **Performance on MVTecAD and Brain AD.** Red, blue, and green are used as in Table 1.

Methods	MVTecAD	Brain AD
Previous localization-only methods		
GeSCF	29.8	47.9
MetaUAS	54.9	50.8
CLIP-AD	36.0	28.7
AdaCLIP	44.0	52.3
Previous LVLM-based methods		
LISA-7B	11.0	19.2
GLaMM-7B	7.8	7.5
LocalizationHeads-7B	1.1	10.1
EGG-7B	6.5	1.9
Migician-7B	24.8	53.3
LLaVA-1.5-7B	0.0	0.0
Qwen2-VL-7B	7.7	42.0
LLaVA-OneVision-7B	0.1	2.4
LLaVA-OneVision1.5-4B	0.0	2.3
LLaVA-OneVision1.5-8B	1.0	1.4
LLaVA-OneVision2-8B	27.5	44.6
Qwen3-VL-2B-Instruct	11.6	43.9
Qwen3-VL-8B-Instruct	34.5	49.3
InternVL3-8B	13.0	16.4
InternVL3-14B	13.5	20.7
Our LVLM-based methods		
LLaVA-1.5-7B	31.9 (+31.9)	26.2 (+26.2)
Qwen2-VL-7B	33.8 (+26.1)	68.2 (+26.2)
LLaVA-OneVision-7B	35.7 (+35.6)	68.0 (+65.6)
LLaVA-OneVision1.5-4B	33.0 (+33.0)	72.3 (+70.0)
LLaVA-OneVision1.5-8B	36.7 (+35.7)	75.0 (+73.6)
LLaVA-OneVision2-8B	35.8 (+8.3)	70.9 (+26.3)
Qwen3-VL-2B-Instruct	18.7 (+7.1)	67.6 (+23.7)
Qwen3-VL-8B-Instruct	35.3 (+0.8)	75.7 (+26.4)
InternVL3-8B	34.2 (+21.2)	73.1 (+56.7)
InternVL3-14B	40.8 (+27.3)	78.3 (+57.6)

using LVLMs' world knowledge—achieving strong results on Brain AD (see Table 2). In contrast, localization-only methods, such as GeSCF, remain suboptimal for anomaly detection despite strong results on their target benchmark. Our framework is the *first*, *training-free* LVLM-based approach to match pure localization models across two distinct domains. See Appendix D.4 for qualitative results.

Endowing LVLMs with the Ability to See. Different from previous localization-only approaches—which can *see* changes but cannot *say* what they denote—we take the opposite direction: starting from LVLMs that excel at change captioning (see results in Appendix C.4), we amplify their capabilities to achieve reliable change localization. This approach consolidates spatial localization and linguistic description in a concurrent process—marking a solid step towards human-like change understanding.

5.3 Ablation Studies

Effect of Criteria and Selection Strategies. We conducted comprehensive ablation studies across all datasets in Tables 1 and 2 with ten LVLMs (Tables 3 and 4)—No. 7 representing our default setting. First, each criterion alone substantially outperforms random head selection, while combining both yields the best performance—confirming their complementary roles (No. 1–3). Without subtracting the unchanged baseline (*No sub.*), attention artifacts remain and degrade results (No. 4–5). Moreover, we compare our method under two selection strategies: *fixed*, where heads are selected through the VL-CMU-CD training set and shared across all datasets, and *adapt*, where the top-k heads are selected per sample. Here, fixed selection performs better—suggesting that dataset-level statistics provide a more reliable estimate, whereas the sample-wise selection is more susceptible to noise (No. 6). Lastly, in line with findings in Section 4.2, using multiple heads surpasses a single head, largely saturating at k=3 (at most 0.7% of total heads).

Table 3: **Ablation study on head filtering criteria and selection strategies.**

No.	Criteria 1	Criteria 2	Selection Fixed	Selection Adapt	F1-Score
1	*Random*		✓		27.2
2	✓		✓		50.6
3		✓	✓		47.3
4	*No sub.*		✓		44.5
5	*No sub.*	✓	✓		46.5
6	✓	✓		✓	43.8
7	✓	✓	✓		52.6

Table 4: **Ablation study on the number of top-k heads for aggregation.**

k=1	k=2	k=3	k=4	k=5
50.3	51.4	52.6	52.3	52.2

Table 5: **Ablation study on SAD and SAM refinement.** Avg. Δ denotes the average performance difference from the raw seeds.

Methods	SF-XL	St Lucia	Nordland	MVTecAD	Brain AD	Avg. Δ
GeSCF (raw)	56.7	40.8	37.6	23.1	28.2	–
+ SAD	52.5	28.8	32.6	18.8	28.1	-5.1
+ SAM	67.6	64.6	50.1	29.8	47.9	+14.7
+ SAD & SAM	66.6	52.9	45.2	26.6	46.0	+10.2
InternVL3-14B (raw)	37.4	28.6	20.1	17.7	20.1	–
+ SAD	45.8	30.6	27.9	25.1	49.5	+11.0
+ SAM	56.1	62.2	41.1	28.0	57.0	+24.1
+ SAD & SAM	69.7	71.7	48.4	40.8	78.3	+37.0

Table 6: **Ablation study on prompts.** We report the mean and standard deviation over five prompt phrasings.

Prompt (InternVL3-14B)	SF-XL	St Lucia	Nordland	MVTecAD	Brain AD
"Describe the change."	31.8±1.6	43.2±1.3	4.8±2.2	18.8±0.7	23.8±2.7
"Describe the change that satisfies the `[context]`."	68.5±1.5 (+36.7)	71.6±1.0 (+28.4)	45.9±3.5 (+41.1)	40.7±0.3 (+21.9)	77.3±1.8 (+53.5)

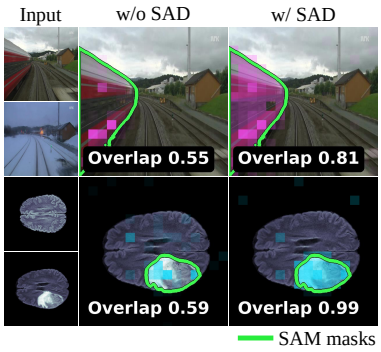


Figure 6: **Effect of SAD.** SAD stabilizes sparse attention seed maps.

Effect of SAD and SAM Refinement. As validated in GeSCF (Kim & Kim, 2025), SAM can refine coarse change seeds into instance-aligned masks. However, SAM does not identify changes by itself; its effectiveness depends on whether the input seed provides accurate signals for selecting valid mask proposals. Table 5 distinguishes raw seed quality from refinement compatibility. While GeSCF starts from stronger seed maps—as it directly extracts similarity maps from a vision foundation model—our framework surpasses it in the final change mask. Furthermore, SAD and SAM exhibit clear synergy: SAD stabilizes the spatial support of sparse LVLM change cues, allowing SAM to retain more valid proposals (see Figure 6), while it may offer little benefit for seeds that are already spatially stabilized.

Effect of Prompts. Our framework is robust to syntactic variation yet responsive to the semantic definition of changes (see Table 6). Generic instructions such as "*describe the change.*" produce poor results, whereas instructing the model to identify changes according to the given `[context]`—specifying meaningful changes to detect and irrelevant changes to ignore—substantially improves performance. This finding suggests that reliable change localization, similar to context-aware change captioning (Kim et al., 2026a), requires context-conditioned reasoning to accurately identify the task-relevant changes under a given scenario or dataset definition—highlighting the semantic complexity of the task and the significance of our prompt-controllable design over prior localization-only methods.

5.4 A Lens into How LVLMs See and Say Changes

Our method also provides a diagnostic lens into how LVLMs connect spatial evidence to linguistic descriptions. Intuitively, accurately describing a change should generally require first identifying where that change occurs. Figure 7 makes this relationship visible: localization based on bounding boxes followed by SAM remains poor even when the model produces correct change captions—obscuring the relationship and creating the misleading impression that LVLMs can *say* without *seeing*. In contrast, with our method, localization performance improves in tandem with captioning quality—revealing a clear coupling between *seeing* and *saying*. We leave further details and in-depth analysis in Appendix C.4.

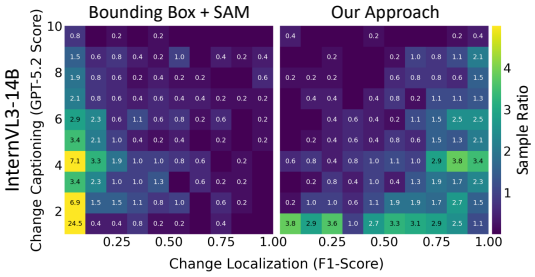


Figure 7: ***See–say* distribution on ChangeVPR.** *Seeing* tracks *saying* quality under our approach.

6 Conclusion

In this paper, we showed that frozen LVLMs encode rich signals for change localization that remain largely hidden from textual decoding, and proposed two principled criteria to identify them. Building on this insight, we introduced a simple training-free framework that combines text-to-image attention maps with semantic affinities in their frozen vision encoder for change localization. By exposing and exploiting this latent spatial knowledge, our method endows LVLM with the ability to accurately *see* changes—all without additional training data and sophisticated training strategies. We believe our work lays a solid foundation for comprehensive understanding of change in our *ever-evolving* world. Further discussion (e.g., origination of change-informative heads) is provided in Appendix F.

AI USE STATEMENT

In this work, we used generative AI tools solely for language polishing, including improving grammar, clarity, and readability. We did not use generative AI tools for research ideation, literature review, data collection or analysis, code generation, figure creation, experimental design, or interpretation of results, and the remaining disclosure categories are not applicable to this work. We carefully reviewed and revised all AI-assisted text to ensure its accuracy, consistency, and fidelity to our original scientific claims. We take full responsibility for the final content of this work, including any text produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide a detailed description of the proposed framework, including the head selection criteria and change mask generation procedure, in the main paper. The Appendix further explains the implementation details, model configurations, hyperparameters, prompts, dataset preprocessing, and evaluation protocols used in all experiments. We also provide additional analyses and qualitative results to clarify the behavior of each component. All source code, configuration files, and supporting materials required to reproduce our results have been prepared and will be publicly released upon publication.

REFERENCES

Pablo Fernández Alcantarilla, Simon Stent, Germán Ros, Roberto Arroyo, and Riccardo Gherardi. Street-view change detection with deconvolutional networks. *Autonomous Robots*, 42:1301 – 1322, 2016. 2, 3, 4, 29

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zhengxue Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Chun Yat Wu, Huajie Tan, Chunyuan Li, Jing Yang, Jiecao Yu, Xiyao Wang, Bin Qin, Yumeng Wang, Zizhen Yan, Ziyong Feng, Ziwei Liu, Bo Li, and Jiankang Deng. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *ArXiv*, abs/2509.23661, 2025. 7, 20

Xiang An, Yin Xie, Feilong Tang, Yunyao Yan, Huajie Tan, Didi Zhu, Changrui Chen, Xiuwei Zhao, Bin Qin, Kaicheng Yang, Yifei Shen, Yuanhan Zhang, Kaichen Zhang, Wenkang Zhang, Zhengxue Cheng, Nan Zhang, Chun Yat Wu, Chunjiang Ge, Zimin Ran, Dehua Song, Chunyuan Li, Shikun Feng, Mingxue Hu, Zhangquan Chen, Junbo Niu, Bo Li, Ziyong Feng, Ziwei Liu, Zongyuan Ge, and Jiankang Deng. Llava-onevision-2: Towards next-generation perceptual intelligence. *ArXiv*, abs/2605.25979, 2026. 1, 7, 20

Ujjwal Baid, Satyam Ghodasara, Michel Bilello, Suyash Mohan, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe Campos Kitamura, Sarthak Pati, Luciano M. Prevedello, Jeffrey D. Rudie, Chiharu Sako, Russell T. Shinohara, Timothy Bergquist, Rong Chai, James A. Eddy, Julia Elliott, Walter Caswell Reade, Thomas Schaffter, Thomas Yu, Jiaxin Zheng, BraTS Annotators, Christos Davatzikos, John T Mongan, Christopher Paul Hess, Soonmee Cha, Javier E. Villanueva-Meyer, John B. Freymann, Justin S. Kirby, Benedikt Wiestler, Priscila Sacilotto Crivellaro, Rivka R.Colen, Aikaterini Kotrotsou, Daniel Marcus, Mikhail Milchenko, Arash Nazeri, Hassan M. Fathallah-Shaykh, Roland Wiest, András Jakab, Marc-André Weber, Abhishek Mahajan, Bjoern H Menze, Adam E. Flanders, and Spyridon Bakas. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *ArXiv*, abs/2107.02314, 2021. 1, 7

W. G. C. Bandara and Vishal M. Patel. Revisiting consistency regularization for semi-supervised change detection in remote sensing images. *ArXiv*, abs/2204.08454, 2022. 34

Luca Bergamasco, Sudipan Saha, Francesca Bovolo, and Lorenzo Bruzzone. Unsupervised change detection using convolutional-autoencoder multiresolution features. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–19, 2022. 34

Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. Mvtec ad — a comprehensive real-world dataset for unsupervised anomaly detection. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9584–9592, 2019. 1, 7

Jorge Bernal, F. Javier Sánchez, Glòria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized medical imaging and graphics : the official journal of the Computerized Medical Imaging Society*, 43:99–111, 2015. 20

Maximilian Bernhard, Niklas Strauß, and Matthias Schubert. Mapformer: Boosting change detection by using pre-change information. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16791–16800, 2023. 34

Jing Bi, Junjia Guo, Yunlong Tang, Lianggong Bruce Wen, Zhang Liu, and Chenliang Xu. Unveiling visual perception in language models: An attention head analysis approach. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 4

Vivien Cabannes, Charles Arnal, Wassim Bouaziz, Alice Yang, François Charton, and Julia Kempe. Iteration head: A mechanistic study of chain-of-thought. *ArXiv*, abs/2406.02128, 2024. 4

Yunkang Cao, Jiangning Zhang, Luca Frittoli, Yuqi Cheng, Weiming Shen, and Giacomo Boracchi. Adaclip: Adapting clip with hybrid learnable prompts for zero-shot anomaly detection. *ArXiv*, abs/2407.15795, 2024. 7, 19, 20, 34

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9630–9640, 2021. 6

Chao Chen, Jun-Wei Hsieh, Ping-Yang Chen, Yi-Kuan Hsieh, and Bo Wang. Saras-net: Scale and relation aware siamese network for change detection. *ArXiv*, abs/2212.01287, 2022. 34

Shuo Chen, Kailun Yang, and Rainer Stiefelhagen. Dr-tanet: Dynamic receptive temporal attention network for street scene change detection. *2021 IEEE Intelligent Vehicles Symposium (IV)*, pp. 502–509, 2021. 7, 20, 34

Xuhai Chen, Jiangning Zhang, Guanzhong Tian, Haoyang He, Wuhao Zhang, Yabiao Wang, Chengjie Wang, Yunsheng Wu, and Yong Liu. Clip-ad: A language-guided staged dual-path model for zero-shot anomaly detection. *ArXiv*, abs/2311.00453, 2023. 7, 19, 20, 34

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does bert look at? an analysis of bert's attention. In *BlackboxNLP@ACL*, 2019. 4

Maxwell Forbes, Christine Kaeser-Chen, Piyush Sharma, and Serge J. Belongie. Neural naturalist: Generating fine-grained image comparisons. In *Conference on Empirical Methods in Natural Language Processing*, 2019. 2

Bin-Bin Gao. Metauas: Universal anomaly segmentation with one-prompt meta-learning. *ArXiv*, abs/2505.09265, 2025. 7, 19, 20, 34

Lila R. Gleitman, David January, Rebecca Nappa, and John C. Trueswell. On the give and take between event apprehension and utterance formulation. *Journal of memory and language*, 57 4: 544–569, 2007. 1

Rhys Gould, Euan Ong, George Ogden, and Arthur Conmy. Successor heads: Recurring, interpretable attention heads in the wild. *ArXiv*, abs/2312.09230, 2023. 4

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. *International Journal of Computer Vision*, 127:398 – 414, 2016. 40

Marcel Gröpl, Jaewoo Jung, Seungryong Kim, Marc Pollefeys, and Sung-Jin Hong. Entropy-gradient grounding: Training-free evidence retrieval in vision-language models. *ArXiv*, abs/2604.08456, 2026. 7, 20

Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. Anomalygpt: Detecting industrial anomalies using large vision-language models. In *AAAI Conference on Artificial Intelligence*, 2023. 33, 34

Zixin Guo, T. Wang, and Jorma T. Laaksonen. Clip4idc: Clip for image difference captioning. *ArXiv*, abs/2206.00629, 2022. 30

Fan Hao, Zongfang Ma, Hong peng Tian, Hao Wang, and Di Wu. Semi-supervised label propagation for multi-source remote sensing image change detection. *Comput. Geosci.*, 170:105249, 2023. 34

Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *ArXiv*, abs/1807.01697, 2019. 32

Andrew Hollingworth. Object-position binding in visual memory for natural scenes and object arrays. *Journal of experimental psychology. Human perception and performance*, 33 1:31–47, 2007. 1

Yanning Hou, Peiyuan Li, Zirui Liu, Yitong Wang, Yanran Ruan, Jianfeng Qiu, and Ke Xu. Visualad: Language-free zero-shot anomaly detection via vision transformer. *ArXiv*, abs/2603.07952, 2026. 40

Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6693–6702, 2019. 40

Jeremy Irvin, Emily Liu, Joyce Chen, Ines Dormoy, Jinyoung Kim, Samar Khanna, Zhuo Zheng, and Stefano Ermon. Teochat: A large vision-language assistant for temporal earth observation data. *ArXiv*, abs/2410.06234, 2024. 34

Harsh Jhamtani and Taylor Berg-Kirkpatrick. Learning to describe differences between pairs of similar images. *ArXiv*, abs/1808.10584, 2018. 33, 34

Hyun Woo Kang, Woohyung Lee, Jaewon Kim, and Hogun Park. Judo: A juxtaposed domain-oriented multimodal reasoner for industrial anomaly qa. *ArXiv*, abs/2605.20284, 2026. 33, 34

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. Your large vision-language model only needs a few attention heads for visual grounding. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9339–9350, 2025. 3, 4, 5, 7, 18, 19, 29, 30

Jae-Woo Kim, Hyeongbeom Kim, and Ue-Hwan Kim. C3-bench: A context-aware change captioning benchmark. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2026a. 2, 3, 5, 6, 9, 21, 30

Jaewoo Kim and Uehwan Kim. Towards generalizable scene change detection. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24463–24473, 2025. 1, 2, 3, 4, 6, 7, 9, 19, 20, 31, 34

Jinyeong Kim, Seil Kang, Jiwoo Park, Junhyeok Kim, and Seong Jae Hwang. Interpreting attention heads for image-to-text information flow in large vision-language models. *ArXiv*, abs/2509.17588, 2025. 3

Yirum Kim, Jaewoo Kim, and Uehwan Kim. Ma3dsg: Multi-agent 3d scene graph generation for large-scale indoor environments. *ArXiv*, abs/2602.04152, 2026b. 1

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloé Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. Segment anything. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3992–4003, 2023. 2, 7

Kartik Kuckreja, Muhammad Sohail Danish, Muzammal Naseer, Abhijit Das, Salman H. Khan, and Fahad Shahbaz Khan. Geochat:grounded large vision-language model for remote sensing. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27831–27840, 2023. 34

Sreejan Kumar, Theodore R. Sumers, Takateru Yamakoshi, Ariel Goldstein, Uri Hasson, Kenneth A. Norman, Thomas L. Griffiths, Robert D. Hawkins, and Samuel A. Nastase. Shared functional specialization in transformer-based language models and the human brain. *Nature Communications*, 15, 2023. 4

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9579–9589, 2023. 3, 7, 19

Jason J. Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. Descriptor : A dataset of clinically generated visual questions and answers about radiology images. 2018. 40

Seonhoon Lee and Jong-Hwan Kim. Semi-supervised scene change detection by distillation from feature-metric alignment. *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1215–1224, 2024. 34

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *ArXiv*, abs/2408.03326, 2024a. 7, 20

Chong Li, Shaonan Wang, Yunhao Zhang, Jiajun Zhang, and Chengqing Zong. Interpreting and exploiting functional specialization in multi-head attention under multi-task learning. In *Conference on Empirical Methods in Natural Language Processing*, 2023. 4

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *ArXiv*, abs/2407.07895, 2024b. 7

You Li, Heyuan Huang, Chi Chen, Kaiyu Huang, Chao Huang, Zonghao Guo, Zhiyuan Liu, Jinan Xu, Yuhua Li, Ruixuan Li, and Maosong Sun. Migician: Revealing the magic of free-form multi-image grounding in multimodal large language models. *ArXiv*, abs/2501.05767, 2025a. 3, 7, 20

Yuanze Li, Shihao Yuan, Haolin Wang, Qizhang Li, Ming Liu, Chen Xu, Guangming Shi, and Wangmeng Zuo. Triad: Empowering lmm-based anomaly detection with expert-guided region-of-interest tokenizer and manufacturing process. *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21917–21926, 2025b. 34

Yun-Cheng Li, Sen Lei, Yitao Zhao, Heng Li, Jun Li, and Antonio Plaza. Sam-based building change detection with distribution-aware fourier adaptation and edge-constrained warping. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–14, 2025c. 3

Chun-Jung Lin, Sourav Garg, Tat-Jun Chin, and Feras Dayoub. Robust scene change detection using visual foundation models and cross-attention mechanisms. *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 8337–8343, 2024. 2, 3, 7, 31, 32, 34

Bo Liu, Li-Ming Zhan, Li Xu, Lingxing Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pp. 1650–1654, 2021. 40

Chenyang Liu, Keyan Chen, Haotian Zhang, Zipeng Qi, Zhengxia Zou, and Zhen Xia Shi. Change-agent: Toward interactive comprehensive remote sensing change interpretation and analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–16, 2024. 1, 34

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *ArXiv*, abs/2304.08485, 2023. 3, 7, 20, 28

Jun Ma, Zongxin Yang, Sumin Kim, Bihui Chen, Mohammed Baharoon, Adibvafa Fallahpour, Reza Asakereh, Hongwei Lyu, and Bo Wang. Medsam2: Segment anything in 3d medical images and videos. *ArXiv*, abs/2504.03600, 2025. 6

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3190–3199, 2019. 40

Fanqing Meng, Jin Wang, Chuanhao Li, Quanfeng Lu, Hao Tian, Jiaqi Liao, Xizhou Zhu, Jifeng Dai, Yu Qiao, Ping Luo, Kaipeng Zhang, and Wenqi Shao. Mmiu: Multimodal multi-image understanding for evaluating large vision-language models. *ArXiv*, abs/2408.02718, 2024. 2

Hung Huy Nguyen, Pooyan Rahmanzadehgervi, Rosanne Liu, Long Mai, and Anh Totti Nguyen. Tab: Transformer attention bottlenecks enable user intervention and debugging in vision-language models. *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22551–22562, 2024a. 30, 34

Kiet A. Nguyen, Adheesh Sunil Juvekar, Tianjiao Yu, Muntasir Wahed, and Ismini Lourentzou. Calico: Part-focused semantic co-segmentation with large vision-language models. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4550–4561, 2024b. 7

Hyeoncheol Noh, Jingi Ju, Min seok Seo, Jong chan Park, and Dong geol Choi. Unsupervised change detection based on image reconstruction loss. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1351–1360, 2022. 34

OpenAI. GPT-5.2, 2025. URL https://developers.openai.com/api/docs/models/gpt-5.2. OpenAI API model documentation. 30

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russ Howes, Po-Yao (Bernie) Huang, Shang-Wen Li, Ishan Misra, Michael G. Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *ArXiv*, abs/2304.07193, 2023. 2, 7

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking : Bringing order to the web. In *The Web Conference*, 1999. 6

Dong Huk Park, Trevor Darrell, and Anna Rohrbach. Robust change captioning. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4623–4632, 2019. 1, 2, 30, 34

Woohyeon Park, Jaeik Kim, Sunghwan Cho, Pamela M. Hong, Wookyoung Jeong, Yoojin Nam, Namjoon Kim, Ginny Y. Wong, Ka Chun Cheung, and Jaeyoung Do. Medic-ad: Towards medical vision-language model's clinical intelligence. *ArXiv*, abs/2603.27176, 2026. 34

Zhen Qu, Xian Tao, Xiaoyi Bao, Dingrong Wang, Shichen Qu, Zhengtao Zhang, and Xingang Wang. Ag-vas: Anchor-guided zero-shot visual anomaly segmentation with large multimodal models. *ArXiv*, abs/2603.01305, 2026. 34

Maithra Raghu, Thomas Unterthiner, Simon Kornblith, Chiyuan Zhang, and Alexey Dosovitskiy. Do vision transformers see like convolutional neural networks? In *Neural Information Processing Systems*, 2021. 4

Vijaya Raghavan T. Ramkumar, Prashant Bhat, E. Arani, and Bahram Zonooz. Self-supervised pretraining for scene change detection. 2021. 4

Hanoona Abdul Rasheed, Muhammad Maaz, Sahal Shaji Mullappilly, Abdelrahman M. Shaker, Salman H. Khan, Hisham Cholakkal, Rao Muhammad Anwer, Eric P. Xing, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Glamm: Pixel grounding large multimodal model. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13009–13018, 2023. 3, 7, 19

Samyak Rawlekar, Amit kumar Swain, Yujun Cai, Yiwei Wang, Ming-Hsuan Yang, and Narendra Ahuja. Finding distributed object-centric properties in self-supervised transformers. *ArXiv*, abs/2603.26127, 2026. 4

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded sam: Assembling open-world models for diverse visual tasks. *ArXiv*, abs/2401.14159, 2024. 3, 7

Ken Sakurada, Weimin Wang, Nobuo Kawaguchi, and Ryosuke Nakamura. Dense optical flow based change detection network robust to difference of camera viewpoints. *ArXiv*, abs/1712.02941, 2017. 2

Ken Sakurada, Mikiya Shibuya, and Weimin Wang. Weakly supervised silhouette-based semantic scene change detection. *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6861–6867, 2018. 1, 29, 32

Xiangxi Shi, Xu Yang, Jiuxiang Gu, Shafiq R. Joty, and Jianfei Cai. Finding it at another side: A viewpoint-adapted matching encoder for change captioning. *ArXiv*, abs/2009.14352, 2020. 2

Yan Shu, Bin Ren, Zhitong Xiong, Xiaoxia Zhu, Begum Demir, Nicu Sebe, and Paolo Rota. Terrascope: Pixel-grounded visual reasoning for earth observation. *ArXiv*, abs/2603.19039, 2026. 34

Oriane Siméoni, Gilles Puy, Huy V. Vo, Simon Roburin, Spyros Gidaris, Andrei Bursuc, Patrick P'erez, Renaud Marlet, and Jean Ponce. Localizing objects with self-supervised transformers and no labels. *ArXiv*, abs/2109.14279, 2021. 4

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michael Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3. 2025. 19

Xinyu Tian, Shu Zou, Zhaoyuan Yang, and Jing Zhang. Identifying and mitigating position bias of multi-image vision-language models. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10599–10609, 2025. 5, 21

Yunbin Tu, L. Li, Li Su, Jun Du, Kelvin Lu, and Qin Huang. Viewpoint-adaptive representation disentanglement network for change captioning. *IEEE Transactions on Image Processing*, 32: 2620–2635, 2023a. 2, 30

Yunbin Tu, Liang Li, Li Su, Zheng-Jun Zha, Chenggang Clarence Yan, and Qin Huang. Self-supervised cross-view representation reconstruction for change captioning. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2793–2803, 2023b. 30

Yunbin Tu, Liang Li, Li Su, Chenggang Yan, and Qin Huang. Distractors-immune representation learning with cross-modal contrastive regularization for change captioning. In *European Conference on Computer Vision*, 2024. 1, 2, 30

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Neural Information Processing Systems*, 2017. 3

Elena Voita, David Talbot, Fédor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. *ArXiv*, abs/1905.09418, 2019. 4

Matthew Walmer, Saksham Suri, Kamal Gupta, and Abhinav Shrivastava. Teaching matters: Investigating the role of supervision in vision transformers. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7486–7496, 2022. 4

Fei Wang, Xingyu Fu, James Y. Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, Tianyi Yan, Wenjie Jacky Mo, Hsiang-Hui Liu, Pan Lu, Chunyuan Li, Chaowei Xiao, Kai-Wei Chang, Dan Roth, Sheng Zhang, Hoifung Poon, and Muhao Chen. Muirbench: A comprehensive benchmark for robust multi-image understanding. *ArXiv*, abs/2406.09411, 2024a. 2

Feng-Xiang Wang, Hong-Zhen Wang, Zonghao Guo, Di Wang, Yu-Lin Wang, Ming-Shuo Chen, Qiang Ma, Long Lan, Wenjing Yang, Jing Zhang, Zhiyuan Liu, and Maosong Sun. Xlrs-bench: Could your multimodal llms understand extremely large ultra-high-resolution remote sensing imagery? *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14325–14336, 2025. 34

G. Wang, Bin-Bin Gao, and Chengjie Wang. How to reduce change detection to semantic segmentation. *Pattern Recognit.*, 138:109384, 2022a. 1, 3, 7, 20, 34

Weiyun Wang, Zhe Chen, Wenhai Wang, Yue Cao, Yang-Zhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, and Jifeng Dai. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. *ArXiv*, abs/2411.10442, 2024b. 40

Yangtao Wang, XI Shen, Shell Xu Hu, Yuan Yuan, James L. Crowley, and Dominique Vaufreydaz. Self-supervised transformers for unsupervised object discovery using normalized cut. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14523–14533, 2022b. 4

Size Wu, Sheng Jin, Wenwei Zhang, Lumin Xu, Wentao Liu, Wei Li, and Chen Change Loy. F-lmm: Grounding frozen large multimodal models. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24710–24721, 2024a. 3

Tsung-Han Wu, Giscard Biamby, Jerome Quenum, Ritwik Gupta, Joseph Gonzalez, Trevor Darrell, and David M. Chan. Visual haystacks: Answering harder questions about sets of images. *ArXiv*, abs/2407.13766, 2024b. 5

Yunqiu Xu, Linchao Zhu, and Yi Yang. Mc-bench: A benchmark for multi-context visual grounding in the era of mllms. *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17675–17687, 2024. 2, 7

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Ke-Yang Chen, Kexin Yang, Mei Li, Min Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yunyang Wan, Yunfei Chu, Zeyu Cui, Zhenru Zhang, and Zhi-Wei Fan. Qwen2 technical report. *ArXiv*, abs/2407.10671, 2024. 7

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Jingren Zhou, Junyan Lin, Kai Dang, Keqin Bao, Ke-Pei Yang, Le Yu, Li-Chun Deng, Mei Li, Min Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yi-Chao Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. 2025. 1, 7, 20

Jia Yoon and Uehwan Kim. Vscd: Video-based scene change detection in unaligned scenes. *ArXiv*, abs/2605.20821, 2026. 1, 40

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Development of a large-scale medical visual question-answering dataset. *Communications Medicine*, 4, 2024. 40

Zhuo Zheng, Ailong Ma, Liangpei Zhang, and Yanfei Zhong. Change is everywhere: Single-temporal supervised object change detection in remote sensing imagery. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15173–15182, 2021. 2

Zifan Zheng, Yezhaohui Wang, Yuxin Huang, Shichao Song, Bo Tang, Feiyu Xiong, and Zhiyu Li. Attention heads of large language models. *Patterns*, 6, 2024. 4

Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection. *ArXiv*, abs/2310.18961, 2023. 34

Duowang Zhu, Xiaohu Huang, Haiyan Huang, Hao Zhou, and Zhenfeng Shao. Change3d: Revisiting change detection and captioning from a video modeling perspective. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24011–24022, 2025a. 34

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Yue Cao, Yangzhou Liu, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Han Lv, De-Hua Chen, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Jiapeng Luo, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Ying Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Lijun Wu, Kai Zhang, Hui Deng, Jiaye Ge, Kaiming Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *ArXiv*, abs/2504.10479, 2025b. 1, 5, 7, 20, 28, 32

Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *European Conference on Computer Vision*, 2022. 1, 20

Appendix

TABLE OF CONTENTS

A **Further Details of Experimental Setup** 19
A.1 Further Details of Benchmarks 19
A.2 Further Details of Baselines 19
B **Further Details of Our Framework** 20
B.1 Detailed Equation of Algorithm 20
B.2 Hyperparameters 20
B.3 Prompts 21
C **Further Analysis** 28
C.1 Generality of the Head Filtering Criteria 28
C.2 Generality of the Head Filtering Dataset 29
C.3 Further Comparison with LocalizationHeads (Kang et al., 2025) 29
C.4 Further Details of Lens into How LVLMs See and Say Changes 30
C.5 Latency Analysis 31
C.6 Temporal Consistency 31
D **Further Experimental Results** 32
D.1 Further Ablation Studies 32
D.2 Performance on PSCD and Robustness to Viewpoint Variation 32
D.3 Further Quantitative Results 33
D.4 Qualitative Results 33
E **Further Related Work** 34
F **Where Do Change-Informative Heads Originate From?** 40
G **Limitations** 40

A Further Details of Experimental Setup

All experiments in this paper were conducted using three NVIDIA H200 GPUs, each with 141 GB of memory, and 16 CPU cores from an Intel Xeon Platinum 8462Y processor. For consistency, all LVLMs employ a temperature of 0; we increased the temperature for unparsable outputs in bounding box prediciton. Images are standardized to a resolution of 512×512 pixels; when this resolution is incompatible with a model's patch-size requirement, we use the closest valid resolution to 512×512, while conducting all evaluations at 512×512. For every SAM-based mask refinement, we leverage SAM 2.1 with the Hiera-Large architecture in automatic mode. We use a point grid with `points_per_side`=64 and `points_per_batch`=128, and set both `pred_iou_thresh` and `stability_score_thresh` to 0.7. For MedSAM (in Brain AD), we use the `MedSAM2_latest` checkpoint based on the SAM 2.1 Hiera-Tiny architecture, also in automatic mask generation mode with `points_per_side`=64, and set `pred_iou_thresh` and `stability_score_thresh` to 0.6. Lastly, for head filtering, we randomly select 1,000 image pairs from the VL-CMU-CD training set and retain the heads consistently selected across five trials.

A.1 Further Details of Benchmarks

To evaluate the generalizability of our framework, we adopt two zero-shot change-centric benchmarks: *(i)* the *GeSCD benchmark*, which evaluates spatial robustness under a strict cross-domain evaluation protocol. Following this, we conduct head filtering solely on the VL-CMU-CD training set and directly transfer the selected heads to ChangeVPR. We further offer results on the Temporal Consistency (TC) score (Kim & Kim, 2025)—defined as the IoU between bidirectional change masks obtained via input-order reversal—in Appendix C.6; and *(ii)* the *zero-shot anomaly detection benchmark*, which evaluates anomalous-region localization using MVTecAD and Brain AD. For each test image, we retrieve its most similar normal reference from the corresponding training split using frozen DINOv3 (Siméoni et al., 2025) features—following the reference-selection paradigm in MetaUAS (Gao, 2025). Following conventions (Chen et al., 2023; Cao et al., 2024), we measure the F1-Score$_{\text{max}}$ for all anomaly baselines to report their best achievable performance.

A.2 Further Details of Baselines

Here, we provide detailed implementation of all baselines. For the prompts used with LVLM-based methods, please refer to Appendix B.3.

Multi-Image Adaptation of LLaVA-1.5-7B. To enable multi-image inference with LLaVA-1.5-7B and its derivatives—including LISA, GLaMM, LocalizationHeads, and EGG—we insert two consecutive `<image>` placeholders into the prompt and provide the corresponding image pair. Each image is independently processed by the original CLIP ViT-L/14 vision encoder and MLP projector. The resulting visual-token sequences are concatenated and placed before the textual instruction, enabling the language decoder to jointly attend to both images. This adaptation requires neither architectural modification nor parameter updates.

Single-Image Grounding LVLMs. We adapt all single-image grounding baselines to our multi-image setting using the formulation described above. For LISA (Lai et al., 2023), we use the official implementation and the publicly available `LISA_Plus_7b` checkpoint, which combines LLaVA-1.5-7B with a SAM ViT-H mask decoder. Each generated `[SEG]` embedding is independently decoded with respect to both frames, yielding one mask per `[SEG]` token and frame. We threshold the decoder logits at zero and take the union of all generated masks within each frame to obtain the final change mask. For GLaMM (Rasheed et al., 2023), we use the `GLaMM-FullScope` checkpoint and follow the same mask-prediction procedure as for LISA.

Following the original LocalizationHeads protocol (Kang et al., 2025), we randomly sample 1,000 image–text pairs from the RefCOCO training set and identify candidate heads based on their attention sum and selection frequency. We repeat the process over five independent trials and report the average results to assess the consistency of the selected heads. We then fix the top three heads and evaluate LLaVA-1.5-7B with the multi-image input formulation described above. The resulting aggregated attention maps are converted into bounding boxes—which are subsequently employed as prompts for SAM—following the original inference pipeline. We provide a more detailed analysis between our framework and LocalizationHeads in Appendix C.3.

Finally, for EGG (Gröpl et al., 2026), we adapt its initial Region-of-Interest (RoI) grounding stage—a training-free entropy-gradient method—to our multi-image setting. Following the official implementation, we compute the next-token prediction entropy at the final prompt position and backpropagate it to the visual embeddings of each input image. The resulting per-token gradient magnitudes constitute a visual relevance map, which is smoothed, binarized using the authors' thresholding strategy, and upsampled to the original image resolution. We then construct the final change masks using the same procedure as for LISA and GLaMM.

Multi-Image Capable LVLMs. For the ten multi-image-capable LVLM baselines (Liu et al., 2023; Li et al., 2024a; An et al., 2025; Zhu et al., 2025b; Yang et al., 2025; An et al., 2026) and Migician (Li et al., 2025a), we instruct each model to predict one or more bounding boxes for changed regions in both input images. The model outputs each prediction on a separate line—specifying the corresponding image index and the bounding box coordinates. We then parse the predicted bounding boxes for each image and use them as prompts to SAM to generate instance masks. Finally, we take the union of all SAM-generated masks within each image to obtain the final change mask.

Localization-Only Methods. For DR-TANet (Chen et al., 2021), we employ the default configuration described in the original paper. For C-3PO (Wang et al., 2022a), we adopt the (I+A+D+E) setting. Also, we evaluate GeSCF (Kim & Kim, 2025) using its default configuration. For a fair comparison, we standardize the refinement procedure of GeSCF to the shared SAM configurations described above. For AdaCLIP (Cao et al., 2024), we use the released checkpoint trained on VisA (Zou et al., 2022) and ClinicDB (Bernal et al., 2015); thus, both MVTecAD and Brain AD datasets remain unseen during training—preserving the zero-shot setting. For CLIP-AD (Chen et al., 2023), we adopt its training-free SDP configuration. For MetaUAS (Gao, 2025), we use the official implementation and released checkpoints. Its input formulation—a normal prompt image paired with a query image—directly corresponds to our paired-image protocol.

B Further Details of Our Framework

B.1 Detailed Equation of Algorithm

To localize changed instances in I_a, we extract two bidirectional deviation maps: $\big(\mathbf{a}_{1\text{st}}^{\ell,h} - \hat{\mathbf{a}}_{1\text{st}}^{\ell,h}\big)$ under (I_a, I_b) and $\big(\mathbf{a}_{2\text{nd}}^{\ell,h} - \hat{\mathbf{a}}_{2\text{nd}}^{\ell,h}\big)$ under (I_b, I_a), both defined over the spatial grid of I_a. Here, we apply Gaussian noise to each input image at σ=0.26 or 0.38—selected with equal probability to construct unchanged baselines. Then, to integrate the information from both attention maps and produce a spatially coherent seed mask for SAM refinement, we linearly interpolate their responses, diffuse the fused map with SAD, and then resize it to the input resolution via bilinear interpolation, as follows:

$$X_i = \text{resize}\Big(\text{SAD}\Big(\sum_{(\ell,h)\in\mathcal{H}_k}\Big[(1-p)\big(\mathbf{a}_{1\text{st}}^{\ell,h} - \hat{\mathbf{a}}_{1\text{st}}^{\ell,h}\big) + p\big(\mathbf{a}_{2\text{nd}}^{\ell,h} - \hat{\mathbf{a}}_{2\text{nd}}^{\ell,h}\big)\Big]\Big)\Big), \quad i \in \{a, b\}. \tag{3}$$

where $\mathcal{H}_k$ denotes the set of top-k selected attention heads. The interpolation weight assigned to each response is determined by the input slot occupied by the target frame: $(1-p)$ for the first slot and p for the second slot. We select p on the VL-CMU-CD training set and fix it across all datasets (see Appendix B.2). Finally, we threshold X_a by its mean value and take the union of SAM proposals with at least 0.65 overlap ratio with the resulting seed mask as the change mask M_a—following (Kim & Kim, 2025). Applying the same procedure to I_b yields M_b, and the final pairwise change mask is their union, $M{=}M_a \cup M_b$. For anomaly detection, where I_a serves as the normal reference and I_b as the anomalous image, we adopt M_b as the final change mask.

B.2 Hyperparameters

Table 7 summarizes the hyperparameters for criterion 2, interpolation, and SAD. To ensure that each LVLM retains a sufficient number of attention heads satisfying our criteria, we apply the threshold τ separately for each model. Also, we set the interpolation weight p independently for each model to balance attention responses obtained when the same instance occupies the first or second input position. Interestingly, the two directional responses are not always complementary (e.g., p=0.5);

Table 7: **Hyperparameters for criterion 2, interpolation, and SAD.**

Methods	Criterion 2	Interpolation	SAD	
	Threshold τ	Weight p	Top-n	Ratio α
LLaVA-1.5-7B	0.10	1.0	10	0.7
Qwen2-VL-7B	0.50	0.5	10	0.7
LLaVA-OneVision-7B	0.15	0.3	10	0.7
LLaVA-OneVision1.5-4B	0.30	0.9	10	0.7
LLaVA-OneVision1.5-8B	0.30	0.9	10	0.7
LLaVA-OneVision2-8B	0.50	0.5	10	0.7
Qwen3-VL-2B-Instruct	0.50	1.0	10	0.7
Qwen3-VL-8B-Instruct	0.50	1.0	10	0.7
InternVL3-8B	0.70	0.5	10	0.7
InternVL3-14B	0.50	1.0	10	0.7

instead, the optimal interpolation weights tend to be polarized toward one direction—indicating that many LVLMs favor detecting changes from a particular input direction—consistent with previous observations on multi-image LVLMs (Tian et al., 2025; Kim et al., 2026a). Lastly, we apply a single SAD configuration using the ViT features of the last layer across all LVLM baselines—keeping the neighborhood size Top-n and diffusion ratio α fixed without model- and data-specific tuning. Appendix D.1 further explores how the SAD hyperparameters affect change localization performance.

B.3 Prompts

We provide the prompts used for attention head filtering and LVLM inference. Their structure follows C3-Bench (Kim et al., 2026a)—explicitly specifying task-relevant changes, pseudo-changes to ignore, and the required output format. For LVLM-based baselines that produce textual bounding boxes, we replace the captioning instruction, "`Describe the visible change from Image 1 to Image 2 in one concise sentence while satisfying the following criteria:`", with "`For the changes that meet the following criteria, output multiple bounding boxes in both Image 1 and Image 2, one per line, as Image 1: [x1, y1, x2, y2] or Image 2: [x1, y1, x2, y2] (coordinates 0-1000). List ALL of them. If there is no change, output: none.`" Figure 8 shows the prompt used for attention head filtering on the VL-CMU-CD training set, whereas Figures 9–27 provide the inference prompts for ChangeVPR, MVTecAD, and Brain AD.

```
Describe the visible change from image 1 to image 2 in one concise
    sentence while satisfying the following criteria:

## Change to Detect
    - Objects that appear, disappear, move, or change on the street
    (e.g., vehicles, trash, signs, barriers, construction)

## No Change
    - Lighting or exposure variations
    - Viewpoint or camera shifts
    - Minor photometric or compression artifacts
    - Weather, seasonal, or shadow differences

## Tone
    - Neutral and objective

. If there is no visible change, output exactly: no change.
```

Figure 8: **Prompt for VL-CMU-CD.**

```
Describe the visible change from image 1 to image 2 in one concise
    sentence while satisfying the following criteria:
```

```
## Change to Detect
    - Movement, appearance, disappearance, or modification in the
    urban scene (e.g., vehicles, pedestrians, traffic lights, street
    signs, or roadside objects)

## No Change
    - Lighting or exposure variations
    - Viewpoint or camera shifts
    - Minor photometric or compression artifacts
    - Weather, seasonal, or shadow differences

## Tone
    - Neutral and objective

. If there is no visible change, output exactly: no change.
```

Figure 9: **Prompt for ChangeVPR (SF-XL).**

```
Describe the visible change from image 1 to image 2 in one concise
    sentence while satisfying the following criteria:

## Change to Detect
    - Objects that appear, disappear, move, or change in the suburban
    street scene (e.g., vehicles, bins, signs, fences, house features)

## No Change
    - Lighting or exposure variations
    - Viewpoint or camera shifts
    - Minor photometric or compression artifacts
    - Weather, seasonal, or shadow differences

## Tone
    - Neutral and objective

. If there is no visible change, output exactly: no change.
```

Figure 10: **Prompt for ChangeVPR (St Lucia).**

```
Describe the visible change from image 1 to image 2 in one concise
    sentence while satisfying the following criteria:

## Change to Detect
    - Appearance, disappearance, or modification of trains, vehicles,
    signs, buildings, or man-made structures along the railway

## No Change
    - Lighting or exposure variations
    - Viewpoint or camera shifts
    - Minor photometric or compression artifacts
    - Weather, seasonal, or shadow differences

## Tone
    - Neutral and objective

. If there is no visible change, output exactly: no change.
```

Figure 11: **Prompt for ChangeVPR (Nordland).**

```
Describe the visible change from image 1 to image 2 in one concise
    sentence while satisfying the following criteria:

## Change to Detect
```

```
    - Defects that appear on the bottle: e.g., large or small breaks,
    contamination
    - Any other physical damage, contamination, or abnormality on the
    object

## No Change
    - Lighting or exposure variations
    - Viewpoint or camera shifts
    - Natural variation between individual samples of the same product
    - Minor photometric or reflection differences

## Tone
    - Neutral and objective

. If there is no visible change, output exactly: no change.
```

Figure 12: **Prompt for MVTecAD (bottle).**

```
Describe the visible change from image 1 to image 2 in one concise
    sentence while satisfying the following criteria:

## Change to Detect
    - Defects that appear on the cable cross-section: e.g., bent
    wires, swapped or missing cables, cut or poked insulation
    - Any other physical damage, contamination, or abnormality on the
    object

## No Change
    - Lighting or exposure variations
    - Viewpoint or camera shifts
    - Natural variation between individual samples of the same product
    - Minor photometric or reflection differences

## Tone
    - Neutral and objective

. If there is no visible change, output exactly: no change.
```

Figure 13: **Prompt for MVTecAD (cable).**

```
Describe the visible change from image 1 to image 2 in one concise
    sentence while satisfying the following criteria:

## Change to Detect
    - Defects that appear on the capsule: e.g., cracks, faulty
    imprints, pokes, scratches, squeezed deformation
    - Any other physical damage, contamination, or abnormality on the
    object

## No Change
    - Lighting or exposure variations
    - Viewpoint or camera shifts
    - Natural variation between individual samples of the same product
    - Minor photometric or reflection differences

## Tone
    - Neutral and objective

. If there is no visible change, output exactly: no change.
```

Figure 14: **Prompt for MVTecAD (capsule).**

```
Describe the visible change from image 1 to image 2 in one concise
    sentence while satisfying the following criteria:

## Change to Detect
    - Defects that appear on the carpet: e.g., color patches, cuts,
    holes, metal contamination, thread damage
    - Any other physical damage, contamination, or abnormality on the
    object

## No Change
    - Lighting or exposure variations
    - Viewpoint or camera shifts
    - Natural variation between individual samples of the same product
    - Minor photometric or reflection differences

## Tone
    - Neutral and objective

. If there is no visible change, output exactly: no change.
```

Figure 15: **Prompt for MVTecAD (carpet).**

```
Describe the visible change from image 1 to image 2 in one concise
    sentence while satisfying the following criteria:

## Change to Detect
    - Defects that appear on the metal grid: e.g., bent or broken
    segments, glue marks, metal contamination, thread anomalies
    - Any other physical damage, contamination, or abnormality on the
    object

## No Change
    - Lighting or exposure variations
    - Viewpoint or camera shifts
    - Natural variation between individual samples of the same product
    - Minor photometric or reflection differences

## Tone
    - Neutral and objective

. If there is no visible change, output exactly: no change.
```

Figure 16: **Prompt for MVTecAD (grid).**

```
Describe the visible change from image 1 to image 2 in one concise
    sentence while satisfying the following criteria:

## Change to Detect
    - Defects that appear on the hazelnut: e.g., cracks, cuts, holes,
    print marks
    - Any other physical damage, contamination, or abnormality on the
    object

## No Change
    - Lighting or exposure variations
    - Viewpoint or camera shifts
    - Natural variation between individual samples of the same product
    - Minor photometric or reflection differences

## Tone
    - Neutral and objective

. If there is no visible change, output exactly: no change.
```

Figure 17: **Prompt for MVTecAD (hazelnut).**

```
Describe the visible change from image 1 to image 2 in one concise
    sentence while satisfying the following criteria:

## Change to Detect
    - Defects that appear on the leather surface: e.g., color patches,
    cuts, folds, glue marks, pokes
    - Any other physical damage, contamination, or abnormality on the
    object

## No Change
    - Lighting or exposure variations
    - Viewpoint or camera shifts
    - Natural variation between individual samples of the same product
    - Minor photometric or reflection differences

## Tone
    - Neutral and objective

. If there is no visible change, output exactly: no change.
```

Figure 18: **Prompt for MVTecAD (leather).**

```
Describe the visible change from image 1 to image 2 in one concise
    sentence while satisfying the following criteria:

## Change to Detect
    - Defects that appear on the metal nut: e.g., bent threads,
    discoloration, flipped orientation, scratches
    - Any other physical damage, contamination, or abnormality on the
    object

## No Change
    - Lighting or exposure variations
    - Viewpoint or camera shifts
    - Natural variation between individual samples of the same product
    - Minor photometric or reflection differences

## Tone
    - Neutral and objective

. If there is no visible change, output exactly: no change.
```

Figure 19: **Prompt for MVTecAD (metal_nut).**

```
Describe the visible change from image 1 to image 2 in one concise
    sentence while satisfying the following criteria:

## Change to Detect
    - Defects that appear on the pill: e.g., color anomalies,
    contamination, cracks, faulty imprints, wrong pill type, scratches
    - Any other physical damage, contamination, or abnormality on the
    object

## No Change
    - Lighting or exposure variations
    - Viewpoint or camera shifts
    - Natural variation between individual samples of the same product
    - Minor photometric or reflection differences
```

```
## Tone
    - Neutral and objective

. If there is no visible change, output exactly: no change.
```

Figure 20: **Prompt for MVTecAD (pill).**

```
Describe the visible change from image 1 to image 2 in one concise
    sentence while satisfying the following criteria:

## Change to Detect
    - Defects that appear on the screw: e.g., manipulated tip,
    scratched head or neck, damaged threads
    - Any other physical damage, contamination, or abnormality on the
    object

## No Change
    - Lighting or exposure variations
    - Viewpoint or camera shifts
    - Natural variation between individual samples of the same product
    - Minor photometric or reflection differences

## Tone
    - Neutral and objective

. If there is no visible change, output exactly: no change.
```

Figure 21: **Prompt for MVTecAD (screw).**

```
Describe the visible change from image 1 to image 2 in one concise
    sentence while satisfying the following criteria:

## Change to Detect
    - Defects that appear on the tile: e.g., cracks, glue strips, gray
    strokes, oil stains, rough areas
    - Any other physical damage, contamination, or abnormality on the
    object

## No Change
    - Lighting or exposure variations
    - Viewpoint or camera shifts
    - Natural variation between individual samples of the same product
    - Minor photometric or reflection differences

## Tone
    - Neutral and objective

. If there is no visible change, output exactly: no change.
```

Figure 22: **Prompt for MVTecAD (tile).**

```
Describe the visible change from image 1 to image 2 in one concise
    sentence while satisfying the following criteria:

## Change to Detect
    - Defects that appear on the toothbrush: e.g., broken, bent,
    missing, or contaminated bristles
    - Any other physical damage, contamination, or abnormality on the
    object

## No Change
    - Lighting or exposure variations
    - Viewpoint or camera shifts
```

```
    - Natural variation between individual samples of the same product
    - Minor photometric or reflection differences

## Tone
    - Neutral and objective

. If there is no visible change, output exactly: no change.
```

Figure 23: **Prompt for MVTecAD (toothbrush).**

```
Describe the visible change from image 1 to image 2 in one concise
    sentence while satisfying the following criteria:

## Change to Detect
    - Defects that appear on the transistor: e.g., bent or cut leads,
    damaged case, misplaced component
    - Any other physical damage, contamination, or abnormality on the
    object

## No Change
    - Lighting or exposure variations
    - Viewpoint or camera shifts
    - Natural variation between individual samples of the same product
    - Minor photometric or reflection differences

## Tone
    - Neutral and objective

. If there is no visible change, output exactly: no change.
```

Figure 24: **Prompt for MVTecAD (transistor).**

```
Describe the visible change from image 1 to image 2 in one concise
    sentence while satisfying the following criteria:

## Change to Detect
    - Defects that appear on the wood surface: e.g., color patches,
    holes, liquid stains, scratches
    - Any other physical damage, contamination, or abnormality on the
    object

## No Change
    - Lighting or exposure variations
    - Viewpoint or camera shifts
    - Natural variation between individual samples of the same product
    - Minor photometric or reflection differences

## Tone
    - Neutral and objective

. If there is no visible change, output exactly: no change.
```

Figure 25: **Prompt for MVTecAD (wood).**

```
Describe the visible change from image 1 to image 2 in one concise
    sentence while satisfying the following criteria:

## Change to Detect
    - Defects that appear on the zipper: e.g., broken, split, or
    squeezed teeth, frayed fabric, rough surface
    - Any other physical damage, contamination, or abnormality on the
    object
```

```
## No Change
    - Lighting or exposure variations
    - Viewpoint or camera shifts
    - Natural variation between individual samples of the same product
    - Minor photometric or reflection differences

## Tone
    - Neutral and objective

. If there is no visible change, output exactly: no change.
```

Figure 26: **Prompt for MVTecAD (zipper).**

```
Describe the visible change from image 1 to image 2 in one concise
    sentence while satisfying the following criteria:

## Change to Detect
    - Brain lesions or abnormal tissue that appear in the second
    image: e.g., tumors, edema, hemorrhage, or other focal
    abnormalities
    - Any other pathological change in brain tissue appearance or
    structure

## No Change
    - Intensity, contrast, or acquisition differences between scans
    - Slice position or head orientation shifts
    - Natural anatomical variation between individual subjects
    - Minor noise or reconstruction artifacts

## Tone
    - Neutral and objective

. If there is no visible change, output exactly: no change.
```

Figure 27: **Prompt for Brain AD.**

C Further Analysis

C.1 Generality of the Head Filtering Criteria

In Section 4.2 of the main paper, we demonstrated that our two principled criteria identify attention heads whose spatial responses align with ground-truth change masks in InternVL3-14B (Zhu et al., 2025b). We extend this analysis to other LVLMs (see Table 8). Across all evaluated multi-image-capable LVLMs, head rankings based on criterion 1 alone exhibit a modest Pearson correlation with attention mass within ground-truth change regions. This correlation becomes substantially stronger after filtering with criterion 2—reaching values close to or above 0.7 among the retained heads. This improvement underscores the role of attention spatial consistency in distinguishing change-informative heads. Moreover, only a small fraction of heads satisfy both criteria—allowing our selection process to isolate a sparse set with reliable spatial evidence of change. The sole exception is LLaVA-1.5-7B (Liu et al., 2023), which does not exhibit the same trend. We argue that this difference may stem from its single-image training regime, which may constrain the emergence of heads with consistent cross-image comparison signals.

Table 8: **Generality of attention head filtering criteria and the number of retained heads.** We measure the absolute Pearson correlation coefficient, $|r|$, between the criterion-induced rank and the GT attention mass.

Methods	Criterion 1	Criterion 2	# Heads
LLaVA-1.5-7B	0.47	0.16	13 (1.3%)
Qwen2-VL-7B	0.69	0.72	62 (7.9%)
LLaVA-OneVision-7B	0.56	0.69	62 (7.9%)
LLaVA-OneVision1.5-4B	0.65	0.77	12 (1.0%)
LLaVA-OneVision1.5-8B	0.66	0.93	8 (0.7%)
LLaVA-OneVision2-8B	0.76	0.88	10 (0.9%)
Qwen3-VL-2B-Instruct	0.70	0.76	25 (5.6%)
Qwen3-VL-8B-Instruct	0.64	0.94	7 (0.6%)
InternVL3-8B	0.61	0.78	21 (2.7%)
InternVL3-14B	0.63	0.78	33 (1.7%)

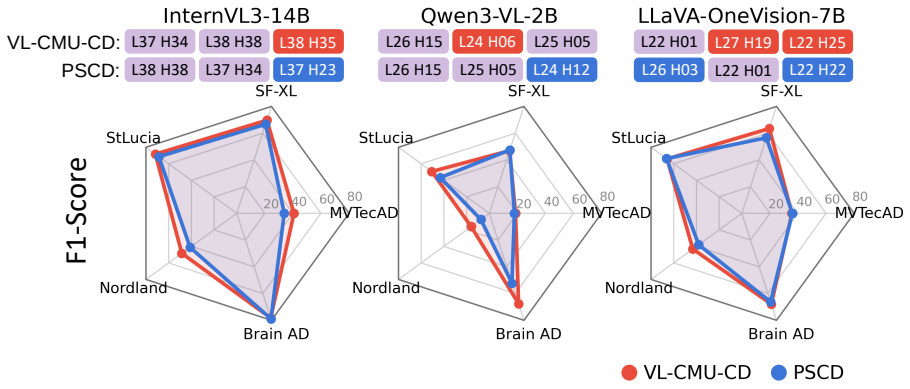


Figure 28: **Generality of the head filtering dataset.** We illustrate the top-3 attention heads in each layer when the filtering criteria are computed on VL-CMU-CD (Alcantarilla et al., 2016) and PSCD (Sakurada et al., 2018) datasets. Heads shared across datasets are marked with purple boxes.

C.2 Generality of the Head Filtering Dataset

For all results in the main paper, we select attention heads for each LVLM using the VL-CMU-CD training set and keep them fixed across all evaluation scenarios. We leverage VL-CMU-CD because changed instances appear in only one image per pair, reducing potential effects arising from changes present in both images. Furthermore, its substantial seasonal variations introduce challenging pseudo-changes—encouraging the selection of heads that respond to genuine changes rather than superficial appearance variations. Interestingly, the selected heads generalize well even to anomaly detection. To further assess robustness to the head filtering dataset, we repeat the same procedure on PSCD (Sakurada et al., 2018). As illustrated in Figure 28, PSCD-based selection yields comparable performance, although VL-CMU-CD-based selection exhibits a slight advantage. Importantly, both datasets display common high-ranking heads within their top-3 filtering results. This agreement in head selection—together with their consistent localization performance—supports the generality of our head filtering strategy to the choice of datasets.

C.3 Further Comparison with LocalizationHeads

Figure 29 compares the attention heads filtered by our criteria with those selected by LocalizationHeads (Kang et al., 2025). In particular, heads selected by the LocalizationHeads are responsive to explicit referring expressions—such as "*the railway track.*"—but struggle to localize the relevant change when prompted to compare the two images. In contrast, our selected change-informative heads expose spatial evidence associated with the underlying change. Their responses are further shaped by prompt semantics: under the context-aware prompt, they attend to changes relevant to the task definition—whereas under the generic prompt "*describe the change.*", they tend to highlight snow on the railway tracks—a visible difference that is excluded as a pseudo-change by the contextual prompt. These results demonstrate complementary roles for head selection and prompt semantics: selecting change-informative heads provides reliable paths for relational evidence, and contextual prompts offer trustworthy signals toward task-relevant differences.

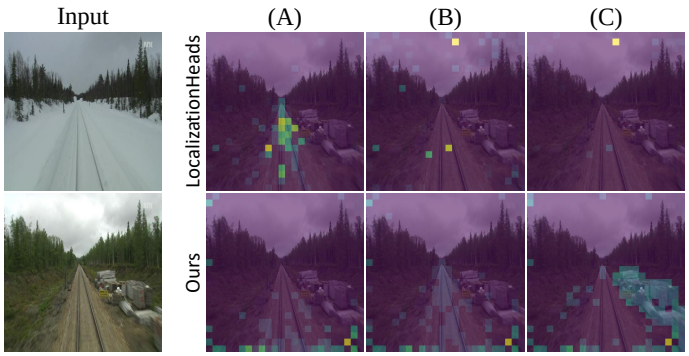
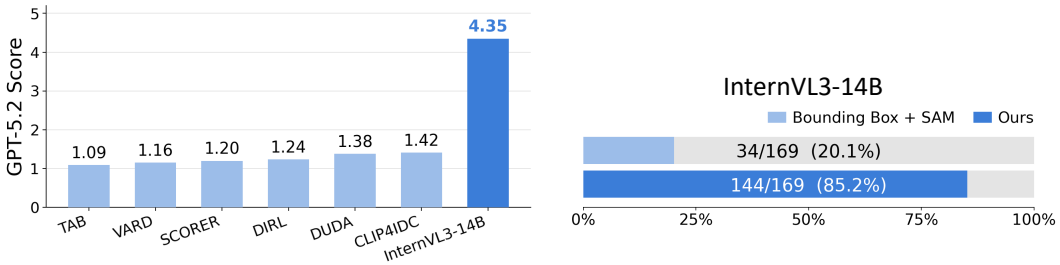


Figure 29: **Comparison between LocalizationHeads and change-informative heads.** We aggregate the top-3 attention heads of each protocol in LLaVA-1.5-7B: L14 H24, L14 H26, and L20 H5, L23 H31, and L22 H19 for our method. Labels (A)–(C) denote the prompts "*the railway track.*", "*describe the change.*", and "*describe the change that satisfies the* `[context]`.", respectively.



(a) Change captioning performance

(b) See—say performance gap

Figure 30: **(a) Change captioning performance.** Recent LVLM substantially outperforms conventional change captioning models (Kim et al., 2026a). **(b) The *see–say* performance gap.** Proportion of samples with localization F1-Scores ≥ 50 among the 169 samples with captioning scores ≥ 5, as evaluated by GPT-5.2 Aggregation score (OpenAI, 2025).

We further evaluate the effectiveness of our head-filtering criteria by directly replacing attention sum deviation with attention sum and attention spatial consistency with selection frequency (Kang et al., 2025), while keeping the rest of the framework unchanged. As shown in Table 9, performance degrades across all settings. Consistent with the preceding analysis, these results suggest that criteria designed for single-image text-referred localization do not generalize to change localization across two images with a contextual prompt, underscoring the importance of our criteria design for identifying change-informative heads.

Table 9: **Performance comparison depending on head filtering criteria on InternVL3-14B.**

Criteria	SF-XL	St Lucia	Nordland	MVTecAD	Brain AD
LocalizationHeads	29.7	23.3	12.2	17.1	32.8
Ours	69.7	71.7	48.4	40.8	78.3

C.4 Further Details of Lens into How LVLMs See and Say Changes

Here, we describe how Figure 7 in the main paper was created and provide a more detailed discussion.

To construct Figure 7, we quantify *see* and *say* performance on the identical image pairs. Given the limited availability of diverse natural image datasets that provide both pixel-wise change masks and change captions, we follow the data curation pipeline of C3-Bench (Kim et al., 2026a), with 3 data annotators, to annotate change captions for 523 image pairs from ChangeVPR—complementing its existing pixel-wise change masks.[1] For each pair, we measure *see* performance using the localization predictions of the Bounding Box + SAM baseline and our method against the ground-truth masks. We also measure *say* performance using the C3-Bench Aggregation score, evaluated by GPT-5.2. Finally, we group samples by their localization and captioning scores and visualize the proportion falling into each region of the resulting joint score distribution.

Using ChangeVPR with the additional caption annotations described above, we first compare recent LVLMs with conventional change captioning models (Park et al., 2019; Guo et al., 2022; Tu et al., 2023b;a; Nguyen et al., 2024a; Tu et al., 2024). Figure 30(a) shows that recent LVLMs substantially outperform conventional models, consistent with the findings of C3-Bench (Kim et al., 2026a)—highlighting substantial progress in describing visual changes. Yet, this strength does not translate reliably into explicit spatial localization through textual coordinate generation, as shown in Figure 1 and Tables 1 and 2 of the main paper. The coexistence of accurate descriptions and poor localization can therefore create the misleading impression that LVLMs can *say* without *seeing*. Figure 30(b) further examines this discrepancy among the 169 samples with captioning scores ≥ 5. Bounding Box prompting followed by SAM achieves successful localization (F1-Score ≥ 50) in only 20.1% of these cases, whereas our method raises this proportion to 85.2%—a 65.1 percentage-point gain—recovering accurate localization for the vast majority of samples with high-scoring captions. These results suggest that latent spatial evidence of change underpins accurate change captioning in LVLMs. Our method directly harnesses this evidence to produce explicit change localization, substantially closing the gap between what LVLMs can successfully *say* and what they can accurately *see*.

[1]We will publicly release these annotations to support reproducibility and future research.

C.5 Latency Analysis

Tables 10 and 11 compare inference latency with GeSCF (Kim & Kim, 2025) and detail the cost of seed generation on ChangeVPR. With the same SAM refinement configuration, our method reduces seed generation latency from 0.91 s to 0.30 s and total latency from 2.85 s to 2.24 s. Our seed generation requires only 0.22 s for extracting bidirectional deviation maps and 0.08 s for the SAD algorithm (for both input images).

Table 10: **Latency analysis (s)**. We report the averaged per-sample latency measured on ChangeVPR.

Methods	**Component**		**Total**
	Seed Generation	SAM	
GeSCF	0.91	1.94	2.85
Ours (InternVL3-14B)	0.30	1.94	2.24

Table 11: **Seed generation latency (s).**

Bidirectional Deviation maps	**SAD**
0.22	0.08

C.6 Temporal Consistency

Our final change mask is commutative by construction under deterministic inference. For each image pair, our method evaluates both input orders, (I_a, I_b) and (I_b, I_a). Therefore, reversing the input pair does not change the set of image orders evaluated by the framework. In Equation 3, each image is associated with the same attention deviations when it appears in the first and second input slots, weighted by $(1-p)$ and p, respectively. Reversing the input order thus only swaps the labels of the two images. With a fixed head set $\mathcal{H}_k$, identical photometric perturbation, and a fixed interpolation weight p, the resulting maps satisfy:

$$f(X_a, X_b) = f(X_b, X_a). \tag{4}$$

The subsequent mean-thresholding and SAM refinement are also equivariant to this exchange, yielding:

$$g(M_a, M_b) = g(M_b, M_a). \tag{5}$$

Consequently, as in GeSCF, the final change mask is invariant to the input order—demonstrating complete Temporal Consistency (Kim & Kim, 2025):

$$M = M_a \cup M_b = M_b \cup M_a. \tag{6}$$

Note that Temporal Consistency (Kim & Kim, 2025) assumes approximately aligned image pairs. When viewpoints differ, ground-truth masks are defined separately for each input image—as in RSCD (Lin et al., 2024).

D Further Experimental Results

D.1 Further Ablation Study

Effect of SAD Hyperparameters. Figure 31 reports localization performance under different neighborhood sizes n and diffusion ratios α. Overall, performance remains stable across moderate settings but degrades when either hyperparameter becomes too large. While an appropriate size of n can help propagate the change signal across different patches of the changed instance, an overly large n may spread the signal to other semantically related instances (but unchanged) in the scene—introducing false-positive responses. Similarly, while a moderate α expands the seed's spatial support through feature affinities while preserving its localization focus, an excessively large α can create long-range propagation paths and weaken the direct potential of the change-informative seed.

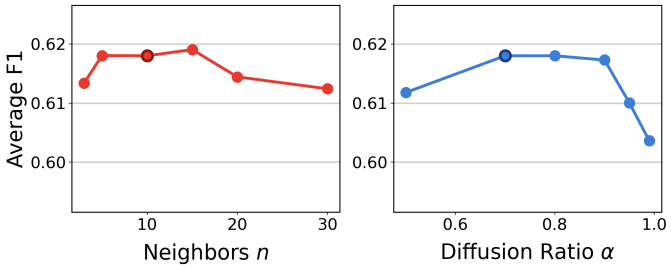


Figure 31: **Ablation study on SAD hyperparameters**. We measure the average results of InternVL3-14B across ChangeVPR, MVTecAD, and Brain AD. The default configuration is highlighted with a border.

Effect of Photometric Perturbation. To further investigate the role of the unchanged self-augmented baseline in Section 4.1 applied for criterion 1, we replace the self-augmented pairs with identical image pairs independently during head filtering and inference[2]. Table 12 demonstrates that photometric perturbations improve performance when applied during either head filtering or the inference step, with best results obtained when used at both stages. These findings indicate that exposing the unchanged baseline to pseudo-changes provides a more informative reference than simply duplicating the image—helping identify heads responsive to genuine changes and isolate change-relevant attention cues for prediction.

Table 12: **Ablation study on photometric perturbation**. We report the average F1-Score across all datasets on InternVL3-14B. No. 4 represents the default configuration.

No.	Perturbation		F1-Score
	Head Filtering	Inference	
1			60.9
2	✓		61.6
3		✓	61.4
4	✓	✓	61.8

D.2 Performance on PSCD and Robustness to Viewpoint Variation

To further evaluate the generalizability of our framework, we conduct experiments on the PSCD test split (Sakurada et al., 2018), following the evaluation protocol of RSCD (Lin et al., 2024) (see Table 13). In addition to the standard setting, this protocol introduces controlled spatial offsets to the query image—yielding the Diff-1 and Diff-2 variants with progressively larger viewpoint discrepancies. Across all unaligned settings, our InternVL3-14B (Zhu et al., 2025b) consistently outperforms the competing methods. Notably, it maintains strong performance under both Diff-1 and the more challenging Diff-2 conditions, despite using neither an explicit viewpoint-alignment module nor a task-specific correspondence mechanism. This result shows that LVLMs can use their innate spatial understanding to distinguish meaningful changes from appearance discrepancies induced by viewpoint variation—enabling robust change localization under imperfect cross-view alignment.

Table 13: **Performance on PSCD**. We use the same experimental settings as those in Tables 1 and 2.

Methods	Aligned	Diff-1	Diff-2	Avg.
GeSCF	36.3	11.1	5.7	17.7
RSCD	45.9	34.1	23.5	34.5
Ours (InternVL3-14B)	45.2	36.7	26.9	36.3

[2]We apply Gaussian noise following the ImageNet-C configurations (Hendrycks & Dietterich, 2019), selecting severity level 4 or 5 with equal probability, corresponding to standard deviations of 0.26 and 0.38, respectively.

Table 14: **Further quantitative results of prior captioning-model-based methods and LVLM-based methods.**

Methods	SF-XL	St Lucia	Nordland	MVTecAD	Brain AD
TAB	17.1	21.7	16.2	14.0	19.8
AnomalyGPT	52.4	55.1	25.8	57.7	53.0
JUDO	11.1	20.1	6.6	32.0	35.4
Ours (InternVL3-14B)	69.7	71.7	48.4	40.8	78.3

D.3 Further Quantitative Results

We illustrate additional quantitative results using alternative localization methods. Specifically, we compare our approach with a localization model based on conventional change captioning methods, TAB—which is trained on the Spot-the-Diff dataset (Jhamtani & Berg-Kirkpatrick, 2018)—as well as LVLM-based anomaly localization models—including AnomalyGPT and JUDO (Gu et al., 2023; Kang et al., 2026). To ensure a fair comparison, we evaluate AnomalyGPT and JUDO in the 1-shot setting (paired-image). Because JUDO predicts anomalous grid cells rather than pixel-level masks, we convert its predicted cells to a bounding box and leverage SAM to obtain a final change mask. As described in Table 14, the TAB model exhibits limited localization performance, suggesting that training on a small-scale change captioning dataset may not generalize to diverse change-centric domains. Moreover, anomaly-specialized LVLM (AnomalyGPT)—similar to GeSCF—transfers less effectively to other domains despite strong performance on industrial anomaly imagery. These results demonstrate the generality of our framework; by explicitly extracting spatial understanding from a frozen, general-purpose LVLM, our method achieves robust localization performance across both natural scenes and anomaly images without domain-specific training.

D.4 Qualitative Results

In Figures 32–36, we present qualitative results that demonstrate the effectiveness of our framework relative to prior localization-only methods and LVLM-based baselines. Unlike previous localization-only approaches, our LVLM-based method *concurrently* localizes changes and generates descriptive captions that characterize their semantic content. Moreover, whereas prior methods typically require domain-specific models or task-specific retraining, our framework can be adapted to diverse change-centric tasks through prompt design alone. Finally, in contrast to existing LVLM-based baselines that can *say* change without reliably grounding it, our framework couples change reasoning with reliable localization—enabling it not only to *say* what has changed, but also to *see* where that change occurs.

E Further Related Work

Visual change understanding research broadly spans three domains with distinct data characteristics: natural scenes, anomaly imagery, and satellite imagery. These domains have been studied independently due to their heterogeneous visual properties, acquisition conditions, and change patterns (Lee & Kim, 2024). In this study, we focus on natural scenes and anomaly domains, and formulate them within a unified framework by using LVLMs' world knowledge and implicit spatial understanding.

Change understanding in natural scenes and anomaly imagery can both be formulated as identifying meaningful differences between a reference and an observed state. In natural scenes, these differences reflect temporal changes at the same location, including object motion, appearance or disappearance, and structural alterations (Chen et al., 2021; Wang et al., 2022a; Lin et al., 2024; Kim & Kim, 2025). As discussed in the main paper, localization and captioning have largely evolved as separate research directions, likely in part because they require fundamentally different forms of supervision: dense pixel-level masks for localization and paired textual descriptions for captioning. Recent studies have sought to bridge this gap by adding a learnable layer on traditional change captioning models to aggregate localization cues (Nguyen et al., 2024a). However, these approaches rely on small-scale, narrowly scoped benchmarks (Jhamtani & Berg-Kirkpatrick, 2018; Park et al., 2019)—leaving their applicability to diverse real-world changes unclear. In contrast, we investigate the attention representations of emerging LVLMs and demonstrate that their latent spatial understanding can be directly amplified into reliable change localization.

In anomaly imagery, the objective is to identify deviations from normal appearance, such as industrial defects or medical abnormalities (Chen et al., 2023; Zhou et al., 2023; Cao et al., 2024; Gao, 2025). Recent studies have increasingly connected LVLM-based anomaly understanding with spatial localization. AnomalyGPT learns a visual decoder and prompt embeddings from synthetic defects to localize industrial anomalies (Gu et al., 2023). Triad combines expert-proposed regions of interest with manufacturing-aware instruction tuning for industrial defect recognition and explanation (Li et al., 2025b). AG-VAS learns semantic anchor tokens and a mask decoder for anomaly segmentation across industrial and medical imagery (Qu et al., 2026), while JUDO compares inspected images with normal references and uses supervised fine-tuning and reinforcement learning for industrial anomaly reasoning (Kang et al., 2026). Lastly, MEDIC-AD learns anomaly-aware and inter-image difference tokens for medical symptom tracking, then trains a mask-supervised heatmap decoder for lesion localization (Park et al., 2026). Although their architectures and target settings differ, these methods train anomaly-specific components or update model parameters using anomaly imagery data. In contrast, we draw on the latent understanding of a frozen LVLM to localize changes without task-specific training—applying the unified approach to both natural scenes and anomaly imagery.

Change understanding in remote sensing identifies surface changes over time using data captured by satellite or aerial platforms—offering a high-altitude perspective of phenomena such as urbanization, deforestation, and disaster damage (Chen et al., 2022; Bandara & Patel, 2022; Bergamasco et al., 2022; Noh et al., 2022; Bernhard et al., 2023; Hao et al., 2023). Although remote sensing change localization and captioning have likewise largely evolved as independent tasks, recent studies increasingly aim to unify both modalities. For example, Change3D formulates both tasks through video modeling with task-specific decoders (Zhu et al., 2025a), while Change-Agent leverages a shared visual backbone with separate detection and captioning branches (Liu et al., 2024). Moreover, several works support diverse change-analysis tasks through multimodal instruction tuning (Kuckreja et al., 2023; Irvin et al., 2024; Wang et al., 2025; Shu et al., 2026). In contrast, our framework offers a fundamentally different form of integration: it extracts and spatially amplifies change-informative latent attention signals within a frozen LVLM—enabling human-like change understanding through the concurrent process of *seeing* and *saying*. Importantly, this is achieved without task-specific branches or parameter updates—preserving the model's original general-purpose capabilities.

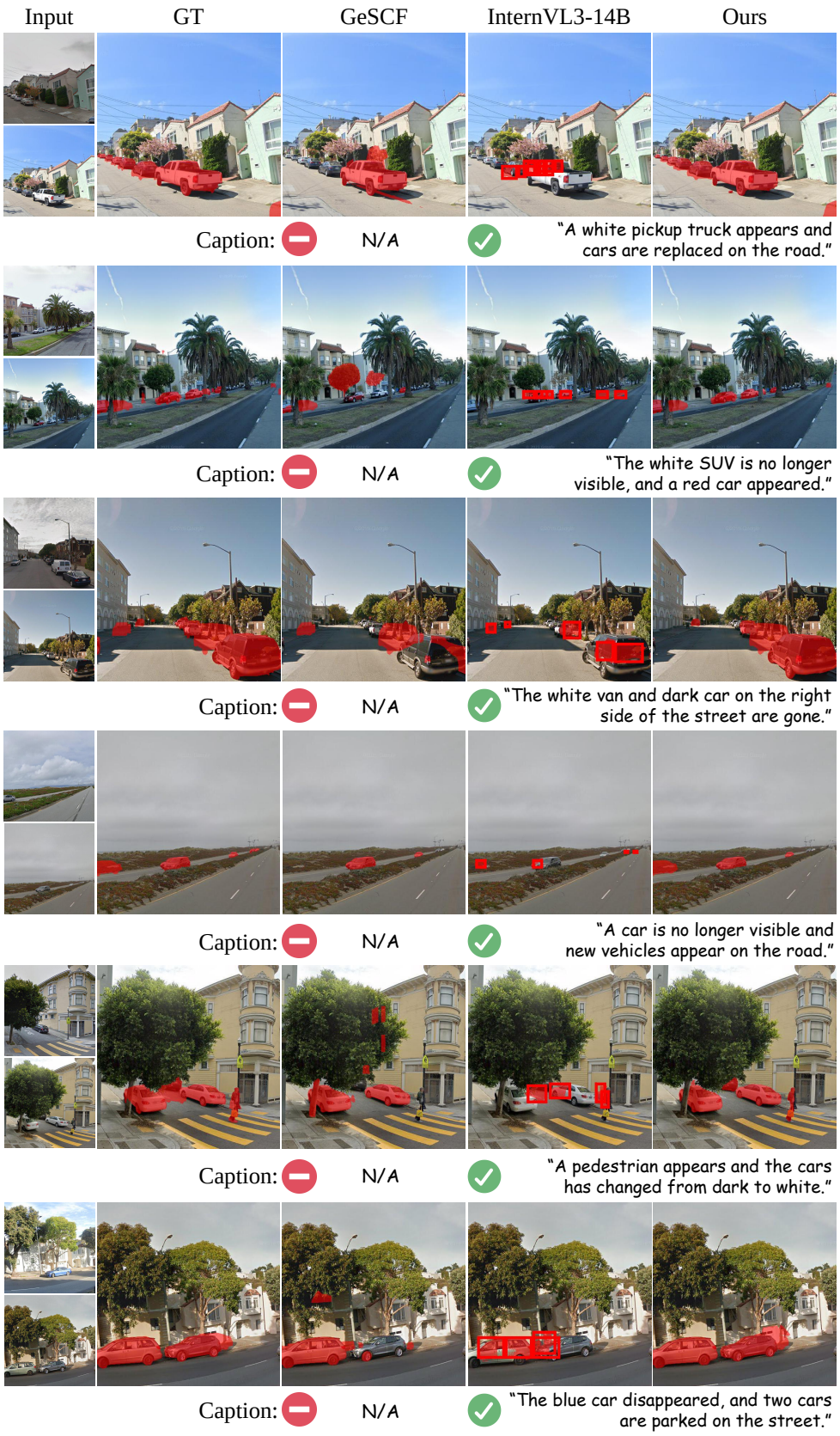


Figure 32: **Qualitative results on ChangeVPR SF-XL split**. Our method (InternVL3-14B-based) produces accurate change masks while concurrently describing the change in natural language.



Figure 33: **Qualitative results on ChangeVPR St Lucia split**. Our method (InternVL3-14B-based) produces accurate change masks while concurrently describing the change in natural language.

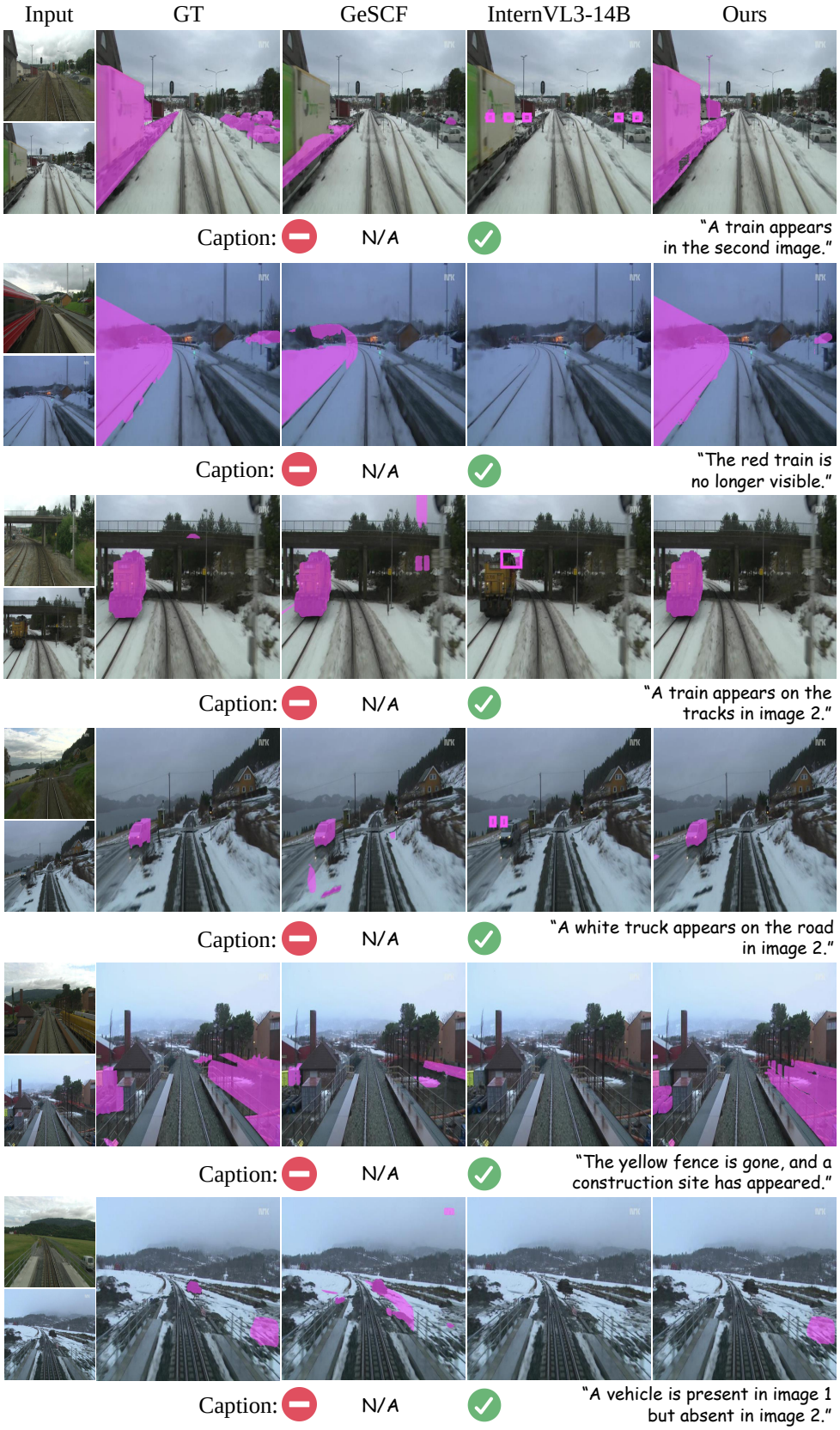


Figure 34: **Qualitative results on ChangeVPR Nordland split**. Our method (InternVL3-14B-based) produces accurate change masks while concurrently describing the change in natural language.

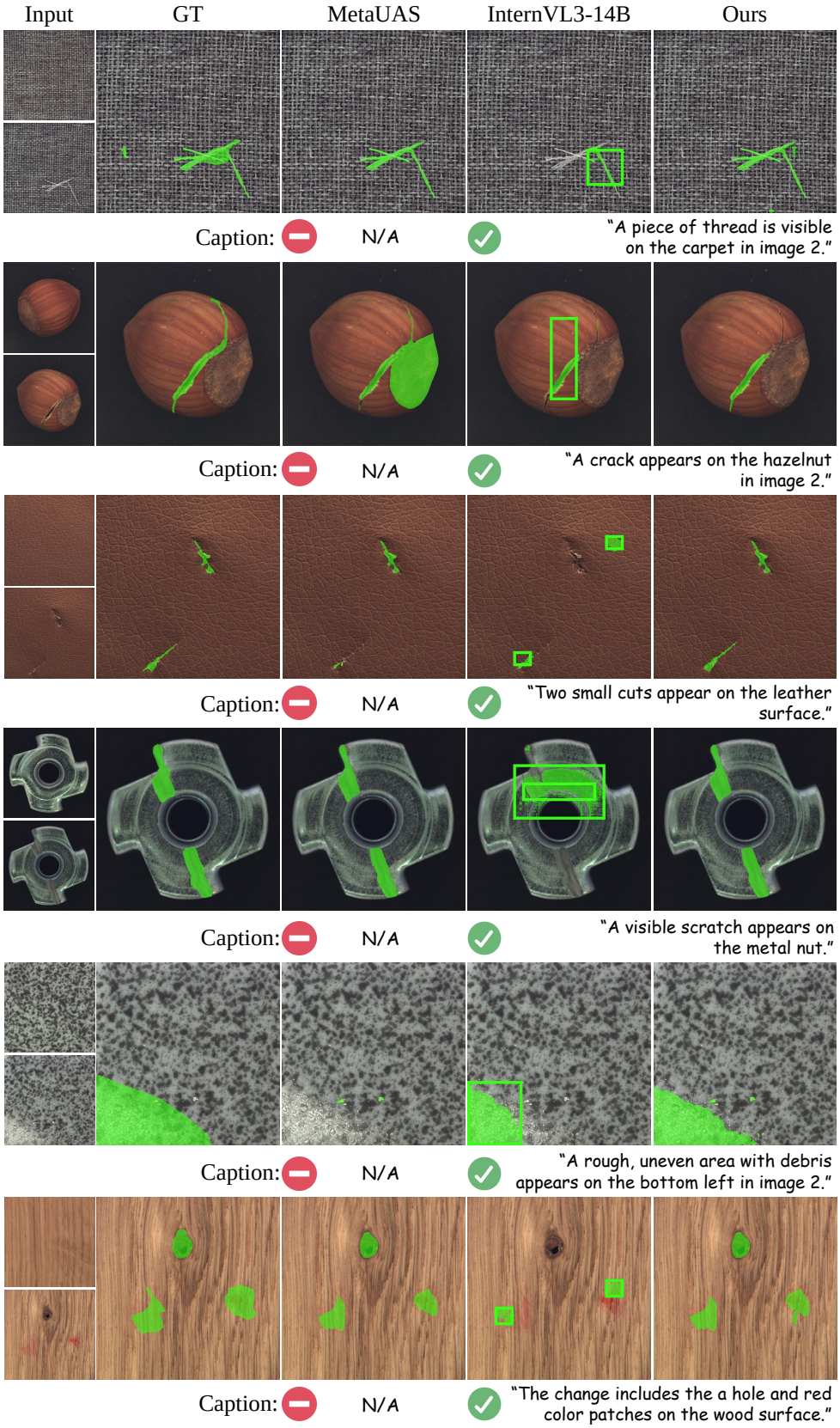


Figure 35: **Qualitative results on MVTecAD**. Our method (InternVL3-14B-based) produces accurate change masks while concurrently describing the change in natural language.

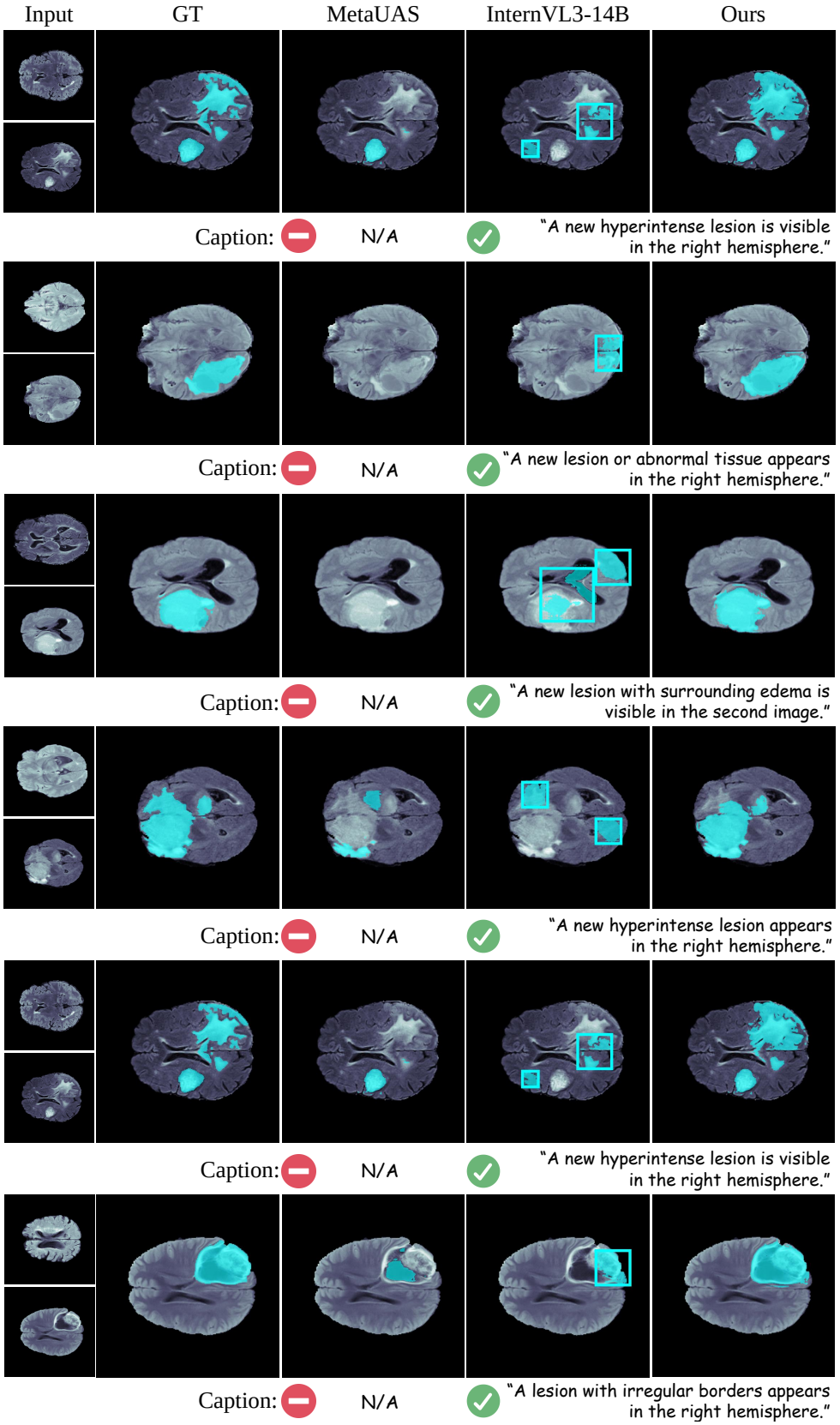


Figure 36: **Qualitative results on Brain AD**. Our method (InternVL3-14B-based) produces accurate change masks while concurrently describing the change in natural language.

Table 15: **Rank of final top-3 heads.**

Methods	L37 H34	L38 H38	L38 H35
PT	2	10	12
+ SFT	1	3	2
+ SFT & MPO	1	2	3

Table 16: **Performance across training pipelines.**

Methods	SF-XL	St Lucia	Nordland	MVTecAD	Brain AD
PT	54.8	68.1	30.2	33.9	58.0
+ SFT	69.4	69.8	47.5	38.3	75.8
+ SFT & MPO	69.7	71.7	48.4	40.8	78.3

F WHERE DO CHANGE-INFORMATIVE HEADS ORIGINATE FROM?

To investigate the question—"*where do change-informative heads originate from?*"—we apply our approach to InternVL3-14B checkpoints obtained at successive stages of its training pipeline, where training proceeds sequentially from native multimodal pretraining (PT), to supervised fine-tuning (SFT), and finally to reinforcement learning (Wang et al., 2024b) (MPO). As illustrated in Table 15, the head ranked first in the final model is already ranked second after PT—indicating that part of the eventual change-sensitive circuitry emerges during pretraining. In contrast, the other two final top-ranked heads initially appear at modest ranks of tenth and twelfth. Following SFT, however, all three heads rise to the top three positions, and their membership is maintained after MPO, with only a slight change in ordering. This observation suggests that PT sets an initial, partially change-relevant representation, whereas SFT amplifies a small set of heads into the dominant change-informative circuitry. This finding is consistent with the performance trends in Table 16: SFT yields significant localization improvements across all five benchmarks, while MPO delivers smaller but consistent additional gains. Together, these results suggest that supervised multimodal alignment is the primary stage at which change-relevant attention becomes reliably localized, whereas reinforcement learning further calibrates this capability.

G LIMITATIONS

First, while the proposed criteria effectively excavate change-informative heads, applying criteria 1 and 2 generally requires paired images. Extending the framework to video-based settings (Yoon & Kim, 2026) or single-image anomaly detection (Hou et al., 2026)—where change must be inferred without an explicit reference image—is an important direction for future work. Second, although our method is the first to demonstrate strong performance across both natural scenes and anomaly imagery, its performance is comparatively less pronounced on remote sensing imagery. The possible reason is—in line with Section F—likely limited exposure to remote sensing data during pretraining and supervised fine-tuning stages—as the training data largely consist of general-purpose visual question answering (VQA) data in the natural image domain (Goyal et al., 2016; Hudson & Manning, 2019; Marino et al., 2019) or medical domain (Lau et al., 2018; Liu et al., 2021; Zhang et al., 2024). This limitation reflects a broader challenge in current LVLMs rather than a constraint specific to our framework. As future foundation models are trained on more balanced and visually diverse data, we expect our training-free framework to extend naturally to a broader range of visual domains.