Benchmarking Composed Image Retrieval for Applied Earth Observation

Bill Psomas[a], Dionysis Christopoulos[b], Thanasis Petropoulos[b], Nikos Efthymiadis[a], Ioannis Kakogeorgiou[c], Ondřej Chum[a], Yannis Avrithis[d], Giorgos Tolias[a] and Konstantinos Karantzalos[b]

[a]*Visual Recognition Group, Department of Cybernetics, Czech Technical University in Prague, Czechia*
[b]*Remote Sensing Laboratory, School of Rural, Surveying and Geoinformatics Engineering, National Technical University of Athens, Greece*
[c]*Institute of Informatics & Telecommunications, National Centre for Scientific Research "Demokritos", Greece*
[d]*Department of Informatics and Telecommunications, National and Kapodistrian University of Athens, Greece*

ARTICLE INFO

Keywords:
Remote Sensing Image Retrieval
Composed Image Retrieval
Multimodal Retrieval
Vision–Language Models
Earth Observation
Benchmarking

ABSTRACT

Remote sensing composed image retrieval (RSCIR) enables search in large satellite image archives using composed queries that combine a reference image with a textual modifier. Although RSCIR offers a flexible interface for expressing targeted retrieval intent, the transferability of modern composition methods to Earth observation (EO) imagery and their relevance to operational EO workflows remain underexplored. We address this gap through a unified benchmark and an application-oriented study. First, we systematically adapt and evaluate representative composed image retrieval methods with six vision-language backbones on PatternCom under a standardized protocol, analyzing their behavior across backbones, composition strategies, and query types. Second, we introduce xView2-CIR, a change-centric dataset for disaster and damage monitoring, where retrieval is conditioned on scene identity and a target post-event state. Our results show that training-free composition methods provide strong and scalable baselines for EO retrieval, while change-centric retrieval presents different challenges from attribute-based retrieval, particularly due to the need to preserve scene identity. Overall, this study establishes a practical benchmark for RSCIR and positions composed retrieval as a complementary tool for remote sensing image retrieval, archive exploration, and change analysis.

1. Introduction

The explosive growth of earth observation (EO) data has enabled large-scale monitoring of the planet, but has also made it increasingly difficult for users to navigate massive satellite image archives and retrieve content matching specific information needs. Remote sensing image retrieval (RSIR) [1] addresses this challenge by searching EO archives from a query, most commonly an image. Prior RSIR research spans unisource and cross-source [110] settings, as well as single-label [30, 80, 14] and multi-label [37, 34, 75] formulations. Despite this progress, a fundamental limitation persists: most RSIR systems are driven by a *single modality* query. A user must typically choose between an image query or a text query, which *restricts expressivity* when the retrieval intent includes *targeted modifications* beyond overall similarity.

Many operational remote sensing workflows are inherently *compositional*. For example, an analyst may wish to retrieve locations visually *similar* to a reference region but with a different attribute, such as a parking lot with lower occupancy or a facility with a larger number of storage tanks, as shown in Figure 1(a,b). In change-centric applications, the goal may instead be to retrieve imagery of the *same* location under a different state, such as after a wildfire or recent construction, as shown in Figure 1(c,d). Such intent is difficult to express with unimodal queries alone. Remote sensing composed image retrieval (RSCIR) [64] addresses this limitation by combining a reference image with a natural-language modifier, aiming to retrieve images that preserve the relevant visual context while satisfying the textual specification [71, 19, 65].

Despite its promise, RSCIR remains insufficiently studied from an applied EO perspective. Existing work [64] introduced the task, but it remains unclear how modern composed image retrieval methods behave under a unified evaluation protocol, how well they transfer across general-purpose and remote-sensing vision–language backbones, and whether composed retrieval remains effective when relevance depends on scene identity and a target post-event state rather than class-level attribute matching. These questions are important for assessing whether RSCIR can support practical EO workflows such as archive exploration, infrastructure monitoring, disaster response, and urban change analysis.

This work advances RSCIR through a unified benchmark and an application-oriented study. We first benchmark representative composition methods on PatternCom [64], an attribute-modification testbed derived from PatternNet [112]. We evaluate six vision–language backbones, including CLIP [67], SigLIP [105], RemoteCLIP [51], and SkyCLIP [94], and adapt representative composition methods such as Pic2Word [71], SEARLE [2], FreeDom [19], MagicLens [106], and BASIC [65] to EO imagery. Beyond aggregate retrieval metrics, we analyze method behavior across backbones, query types, and composition strategies to provide practical guidance for selecting RSCIR pipelines.

ORCID(s): 0000-0001-5381-0312 (Bill Psomas)

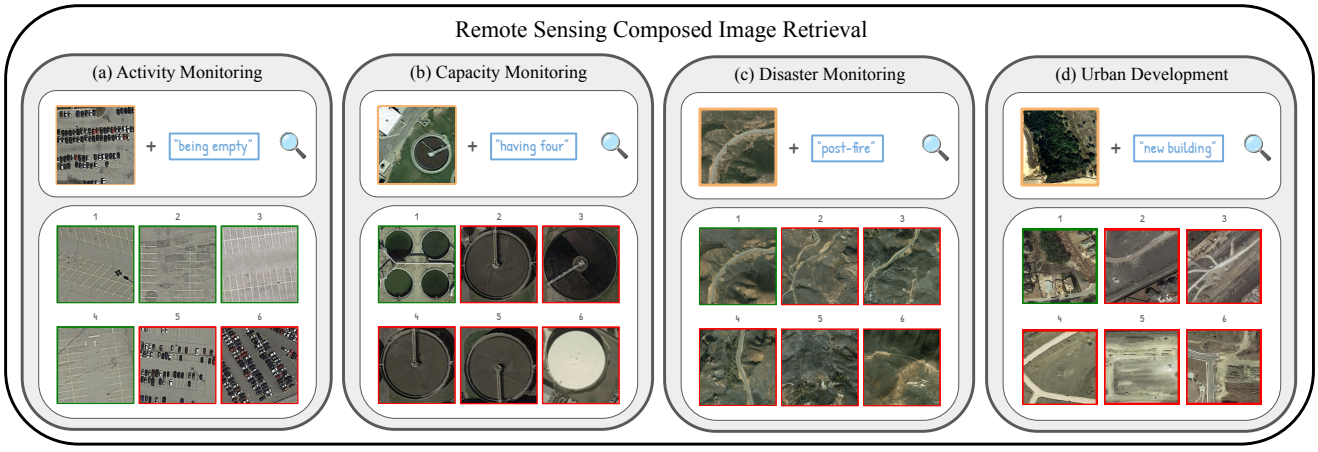

Figure 1: Composed Image Retrieval for applied Earth Observation. We illustrate how composed queries (query image + query text) enable controllable retrieval by specifying a targeted change. (a) Activity monitoring: a parking-lot query image composed with `being empty` retrieves visually *similar* parking areas with low vehicle occupancy. (b) Capacity monitoring: A facility query image composed with `having four` retrieves visually *similar* sites exhibiting a different quantity attribute. (c) Disaster monitoring: a pre-event query image composed with `post-fire` retrieves post-event imagery of the *same* scene consistent with wildfire impact. (d) Urban development: a pre-construction query image composed with `new buildings` retrieves the *same* scene exhibiting recent construction. For each query, retrieved candidates (ranked left-to-right, top-to-bottom) are produced by FreeDom [19] with OpenAI CLIP [67]; green borders indicate correct retrievals under the respective evaluation criterion (a,b: same class + target attribute value, c,d: same scene + target change), while red borders indicate mismatches. Images in (a,b) are from PatternNet [112], (c) from xView2 [41], and (d) from LEVIR-CC [50].

To connect benchmarking with operational EO needs, we further introduce xView2-CIR, a new dataset derived from xView2 [41] for change-centric composed retrieval in disaster and damage monitoring. In xView2-CIR, a query consists of a pre-event reference image and a textual modifier such as `post-fire`, and the goal is to retrieve the corresponding post-event image of the same location. This setting evaluates whether composed retrieval can support location-aware and semantically steerable search, complementing established change-centric workflows such as change detection, damage assessment, and rapid mapping.

In summary, we make the following contributions:

1. We establish a unified benchmark for RSCIR on PatternCom, with domain-grounded adaptations of representative composition methods and a standardized protocol spanning six vision-language backbones.
2. We introduce xView2-CIR, a new evaluation dataset for change-centric composed retrieval in disaster and damage monitoring, where relevance depends on both scene identity and target post-event state.
3. We provide empirical analysis and practical guidance showing when composed retrieval is useful for remote sensing applications, and how its behavior differs between attribute-based and change-centric settings.

2. Related Work

Remote Sensing Image Retrieval. With the aim to effectively *search* and *retrieve* information from extensive remote sensing (RS) image archives, remote sensing image retrieval (RSIR) can be categorized into *unisource* and *cross-source* [110]. Initially, RSIR methods focus on handcrafted and low-level visual features [58, 56, 63, 45, 5, 88, 74, 89, 9, 17]. With the advent of deep learning, neural networks are utilized for unisource *single-label* retrieval: (a) as feature extractors [47, 31, 6, 100, 59, 22, 83, 33, 70, 30], (b) trained from scratch [111, 107, 113, 54, 92, 80, 93, 10], (c) integrating attention modules [90, 91, 96, 11] and (d) using metric learning [109, 8, 14, 53, 20, 52]. Neural networks are also used for unisource *multi-label* [9, 37, 79, 81, 15, 34, 75], cross-source *cross-sensors* [46, 57, 98, 97], cross-source *cross-modal* [11, 99, 82, 78, 55, 103, 102] and cross-source *cross-view* retrieval [32, 104, 85, 48, 39, 76]. Our work fills a notable gap and enhances user intent expression in RSIR by combining image with text.

Composed Image Retrieval. Image-to-image [66, 23, 61] and text-to-image [72, 108, 21] retrieval provide ways to explore large image archives. However, the most accurate and flexible way to express the user intent is a query *composed* of both an image and a text. Composed image retrieval (CIR) [87, 13, 29, 3, 42, 71] aims to retrieve images not only visually similar to the query image, but also altered to align with the specifics of the query text. Traditionally, CIR methods are supervised by *triplets* of the form *query image, query text, target image* [87, 12, 101, 29, 68, 13, 42, 18]. The labor-intensive process of labeling such triplets limits early works to specific applications in fashion [28, 4, 95], physical states [35], object attributes and composition [87, 49, 60]. The emergence of vision-language models (VLMs) such as CLIP [67], ALIGN [36], or BLIP [44] has recently enabled

zero-shot composed image retrieval (ZS-CIR): one can compose a query by manipulating embeddings alone, without any task-specific training. Early ZS-CIR attempts include Pic2Word [71] and SEARLE [2], which invert the visual embedding back to text by either pre-training a small decoder or optimizing a pseudo token at test-time respectively. More recent work pushes this idea further with memory-based inversion (FreeDom [19]) or caption-and-LLM pipelines (CIReVL [38]). CompoDiff [25] casts the composition as a diffusion [69] problem, gradually blending visual and textual semantics. MagicLens [106] revisits triplet supervision at web scale, where a VLM with extra attention layers is fine-tuned to project the image-text input to a single embedding. Despite this rapid progress, the extent to which ZS-CIR methods transfer to EO imagery and how they should be benchmarked and used in real-world remote sensing workflows remains insufficiently understood. This paper addresses this gap through a unified benchmark, complemented by an application-oriented study.

Vision-Language Models. Foundational VLMs such as CLIP [67], ALIGN [36] and BLIP/BLIP-2 [44, 43] are pre-trained on hundreds of millions, or even billions, of web image–text pairs (*e.g.* LAION-2B [73]). Their joint embedding spaces enable strong zero-shot transfer to a wide range of tasks, including open-vocabulary classification [67], detection [26], segmentation [77], and captioning [84]. OpenCLIP [16] re-implements CLIP and releases checkpoints trained directly on LAION-2B. SigLIP [105] replaces the softmax contrastive loss with a sigmoid variant and is trained on the multilingual WebLI corpus. Several works adapt CLIP to the remote sensing domain: RemoteCLIP [51] fine-tunes OpenAI CLIP on caption-augmented RS datasets, CLIP LAION-RS uses a 726k RS subset of LAION-2B (LAION-RS), and SkyCLIP [94] leverages SkyScript, a million-scale image-text pair dataset, constructed by linking satellite images from Google Earth Engine [24] with OpenStreetMap [27] annotations. In this work, we benchmark various VLMs under identical composed image retrieval protocols, revealing how generic versus RS-specialised pre-training affects the performance on the proposed RSCIR task.

3. Task & Methodology

3.1. Problem formulation

In remote sensing composed image retrieval (RSCIR), the query consists of a reference image y and a textual modifier t, denoted as $q = (y, t)$. The goal is to rank images $x \in X$ from a database according to a composed similarity score $s(q, x) \in \mathbb{R}$. Depending on the application, the query image y is associated with either a semantic *class* label C_y or a *scene* label S_y identifying a geographic location. We denote by A_y the attribute or state visible in y, and by A_t the target attribute or state specified by the text.

We consider two relevance protocols:

- **Same class + target attribute.** A result x is relevant if it depicts the same class as the query and matches the target attribute, i.e., $C_x = C_y$ and $A_x = A_t$. This setting captures attribute-oriented search, such as retrieving similar facilities with a different color, shape, density, or quantity.
- **Same scene + target state.** A result x is relevant if it corresponds to the same scene/location as the query and matches the target state or change specified by the text, i.e., $S_x = S_y$ and $A_x = A_t$. This setting captures location-aware change-centric retrieval, such as retrieving the post-event image of the same site after a wildfire or flood.

The two protocols reflect different *operational goals*. The first evaluates class-conditioned attribute modification, while the second evaluates identity-preserving state retrieval. The latter is not a minor variant of the former: preserving scene identity while satisfying a target state may favor different composition mechanisms.

To define s, we make use of pre-trained VLMs that consist of a *visual encoder* $f : \mathcal{I} \to \mathbb{R}^d$ and a *text encoder* $g : \mathcal{T} \to \mathbb{R}^d$, which map input images from image space $\mathcal{I}$ and words from the text space $\mathcal{T}$ to the same embedding space with dimension d. We extract the visual embedding $\mathbf{v}_y = f(y) \in \mathbb{R}^d$ and the text embedding $\mathbf{v}_t = g(t) \in \mathbb{R}^d$ to use as queries. Finally, the embedding of a dataset image $x \in X$ is denoted as $\mathbf{v}_x = f(x) \in \mathbb{R}^d$. All embeddings are ℓ_2-normalized.

3.2. Baselines

Unimodal. These baselines score each database image using only one component of the composed query. Using the VLM encoders f and g, we define the *text-only* baseline as $s_g(q, x) = g(t)^\top f(x)$ and the *image-only* baseline as $s_f(q, x) = f(y)^\top f(x)$. These baselines isolate the contribution of the textual modifier and visual reference, respectively.

Multimodal. These baselines combine the two unimodal scores by fusing the image- and text-based similarities. A simple fusion is *averaging*: $s_a(q, x) = \frac{1}{2}[s_g(q, x) + s_f(q, x)]$, which yields the same ranking as *summing* the similarities. We denote this baseline as *text + image*. While straightforward, this fusion can be biased toward the image signal in practice, since same-modality similarities (image–image) often have a different scale than cross-modal similarities (text–image), even under ℓ_2 normalization. A second fusion computes similarity by *multiplying* the unimodal scores as $s_m(q, x) = s_f(q, x) \times s_g(q, x)$. This method emphasizes retrieval results with high agreement between modalities, inherently penalizing disagreement, and serving as a form of *soft normalization* that mitigates bias toward either modality. We denote this baseline as: *text* × *image*.

3.3. Methods

Beyond unimodal and score-fusion baselines, we adapt and evaluate representative composed image retrieval methods from computer vision. These methods cover textual

inversion, memory-based query construction, caption-and-LLM composition, diffusion-based embedding composition, supervised multimodal pooling, and training-free feature calibration.

WeiCom [64] is a training-free score-fusion method. It computes image-to-image and text-to-image similarities, standardizes each score distribution over the database, maps the standardized scores to $[0, 1]$ using the Gaussian CDF, and combines them as $s_{\mathrm{WC}}(q, x) = \lambda s'_g(q, x) + (1 - \lambda)s'_f(q, x)$, where s'_f and s'_g are the calibrated image and text scores, and $\lambda \in [0, 1]$ controls the modality weight.

Pic2Word [71] casts composition as *textual inversion*: it learns an approximate inverse of the text encoder, mapping an image embedding back into a *text-space token representation* that can be composed with the modifier. Concretely, given $\mathbf{v}_y = f(y)$, it predicts a pseudo-text representation $y^\star \approx g^{-1}(\mathbf{v}_y)$ with a small network trained on large-scale image–text pairs. The composed query is then formed in text space via concatenation and encoded as $g([y^\star; t])$, which is used for text-to-image retrieval.

SEARLE [2] also follows textual inversion, but avoids pretraining by performing *test-time optimization*. For each query image, it directly optimizes a pseudo token $y^\star$ such that $g(y^\star)$ matches $\mathbf{v}_y$ under a cosine objective, complemented by an LLM/GPT-based [7] regularizer that encourages linguistic plausibility. The optimized $y^\star$ is then concatenated with the modifier and encoded with $g(\cdot)$ for retrieval. This per-query optimization is computationally heavier, but can yield strong inversion quality.

FreeDom [19] replaces explicit inversion with a *memory* of textual anchors. It pre-encodes a vocabulary $W = \{w_i\}_{i=1}^N$ through the text encoder $g(\cdot)$ to obtain $V_W = \{g(w_i)\}_{i=1}^N$ and performs nearest-neighbor search from the image embedding $\mathbf{v}_y$ to retrieve a set of words whose embeddings best match $\mathbf{v}_y$. Since the mapping between words and embeddings is explicit, this provides an interpretable textual surrogate for the image query without optimization or pretraining. The retrieved words are concatenated with the modifier and encoded with $g(\cdot)$ for text-to-image retrieval. FreeDom further improves robustness through query expansion using visually similar images.

CIReVL [38] follows a *caption-and-LLM* pipeline. It converts the image query to natural language via captioning rather than inversion. A captioner (*e.g.*, BLIP/BLIP-2 [44, 43]) produces an image description, which is then merged with the modifier using an LLM to form a refined query sentence. Retrieval is performed as text-to-image matching with the VLM. This pipeline favors human-readable intermediate representations and can benefit from strong captioning and rewriting quality.

CompoDiff [25] recasts composition as a *diffusion* process in the joint VLM embedding space. Starting from the image embedding $\mathbf{v}_y$, a denoising diffusion model gradually injects the textual condition t (with image self-conditioning), producing a composed embedding $\tilde{\mathbf{v}}_{y,t}$ after T steps. Retrieval is then performed with cosine similarity $\tilde{\mathbf{v}}_{y,t}^\top \mathbf{v}_x$.

MagicLens [106] follows a *supervised multimodal pooling* pipeline. It learns a dedicated composition module via supervision. It is trained on web-mined triplets (query image, query text, target image), and fine-tunes a VLM augmented with additional attention/pooling layers to map the joint image–text input to a single embedding. Database images are mapped to the same space and matched by cosine similarity.

BASIC [65] is a *training-free compositional scoring* method operating on frozen VLM features. It starts from the standard image-to-image and text-to-image similarities, and aims to make their combination more reliable by (i) reducing modality- and domain-specific bias, (ii) suppressing nuisance directions in the embedding space, and (iii) fusing the two modalities with an "AND"-like criterion. Concretely, BASIC first *centers* image and text embeddings by subtracting mean features (optionally estimated from a large external corpus), which mitigates global bias in the representation. It then applies a *semantic projection* derived from two textual corpora: a positive/object corpus C_+ and a negative/style/context corpus C_-. A contrastive PCA-style construction emphasizes directions that are informative for object content while de-emphasizing directions associated with style, viewpoint, or acquisition context. Optionally, BASIC further strengthens the visual cue via *query expansion* using top retrieved neighbors. Finally, BASIC computes modality scores against the database and applies a lightweight *score calibration* to reduce scale mismatch between modalities, before combining them with a multiplicative fusion regularized by a Harris-like penalty.

4. Experiments

4.1. Experimental setup

Datasets. We conduct our study on three datasets that cover both *attribute-driven* and *change-centric* composed retrieval in EO imagery. PatternCom [64] is an existing RSCIR benchmark derived from PatternNet [112], a high resolution dataset with 38 classes (800 images per class, 256×256 pixels). PatternCom organizes PatternNet into composed queries of the form (query image, query text), where the text specifies a target *attribute value* for the class depicted in the query image (*e.g.*, shape: `oval` for swimming pool). It spans six attribute types (*color, context, density, existence, quantity, shape*) with class-specific value sets and highly variable numbers of positives per query. In PatternCom, relevance is defined by the *same class + target attribute* criterion. As illustrated in Figure 1 (a,b), this protocol naturally supports attribute-oriented

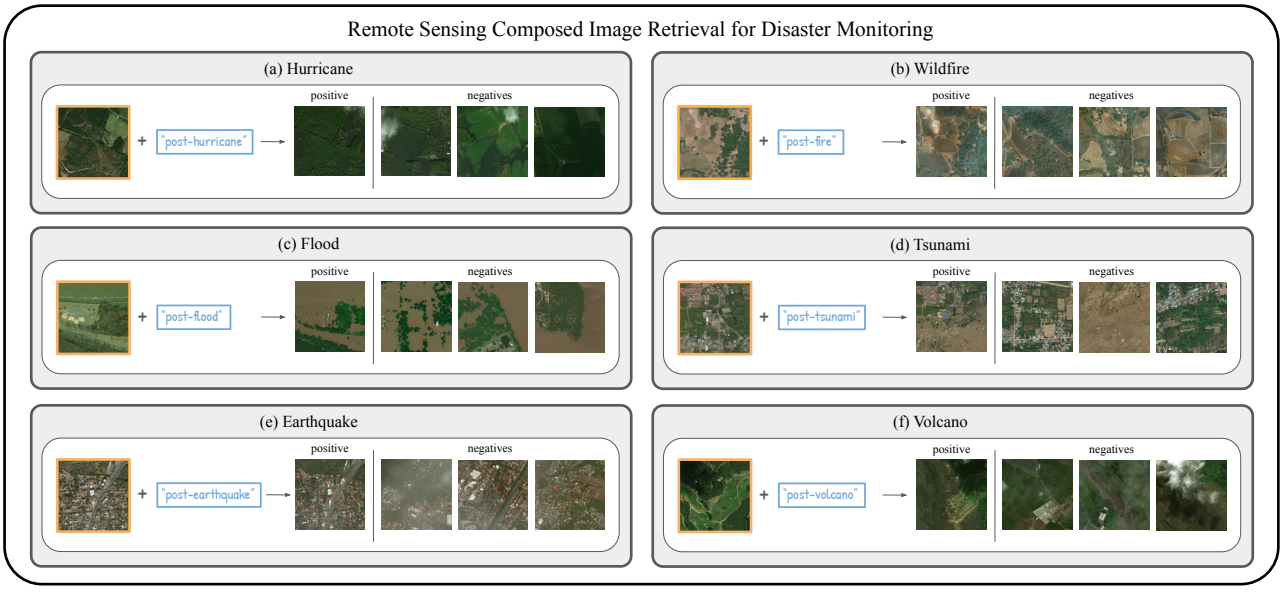

Figure 2: Remote sensing composed image retrieval for disaster monitoring. We restructure xView2 [41] into a composed retrieval setting where a query consists of a *pre-event* reference image of a specific location and a textual modifier describing the desired post-event state (*e.g.*, `post-hurricane`). For each disaster type (a–f), we illustrate the query and the corresponding relevance criterion: *positives* are images of the *same scene/location* depicting the specified post-event state, while *negatives* are non-matching candidates (*e.g.*, different locations with the target change or the same location without the target change).

RS use cases such as activity/occupancy monitoring and capacity/infrastructure monitoring.

PatternCom does not fully reflect operational scenarios where users seek *the same location* under a different state. To bridge this gap, we introduce xView2-CIR, a new dataset built by restructuring xView2 [41] into a composed retrieval setting for disaster and damage monitoring. In xView2-CIR, each query pairs a *pre-event* reference image of a geo-registered scene with a text modifier describing the desired *post-event* state (*e.g.*, `post-hurricane`, `post-fire`); relevance follows the *same scene/location + target state/change* criterion, as depicted in Figure 2. The unique positive of each query is the post-event image of the same geo-registered location. All remaining images are treated as negatives, including images from different locations that exhibit the same disaster category and images from the same location that do not match the target post-event state, when applicable. This construction isolates the challenge of identity-preserving change retrieval.

Finally, to further highlight the applicability of RSCIR to urban change analysis, we include qualitative examples from LEVIR-CC [50], which provides paired imagery for building change detection. We use a small, partially structured subset to form composed queries for urban development (*e.g.*, `new building`), following the *same scene/location + target change* criterion; these examples are used for qualitative analysis only (*e.g.*, Figure 1 (d)), and are not included in the quantitative benchmark due to the absence of a fully curated CIR protocol for this dataset in our study. A comprehensive overview of the dataset statistics is provided in A.1.

Networks. We evaluate six vision-language models, all using the ViT-L/14 vision backbone: CLIP [16] LAION-2B [73], RemoteCLIP [51], OpenAI CLIP [67], SigLIP [105], CLIP LAION-RS [94], and SkyCLIP-50 [94]. We provide further details on the networks in A.2.

Vocabularies. Recent CIR methods [2, 19, 65] leverage a vocabulary of possible categories or concepts. However, many of these, such as Open Images v7 [40] with 21k concepts, originate from general-purpose computer vision datasets. To better align these methods with remote sensing semantics, we create a family of synthetic vocabularies using GPT-4o [62], which we will distribute to avoid redistribution of any copyrighted content. These vocabularies, denoted as RSText, consist of increasing sizes: 150 to two thousand distinct concepts. The generation process (presented in subsection A.3) follows a structured prompt that emphasizes fine-grained, diverse, and remote-sensing-relevant terminology across a wide range of thematic areas. To combine both general-purpose and domain-specific semantics, we further merge RS-Text-2k with the Open Images v7, resulting in a hybrid vocabulary of 23k classes, referred to as HybridText-23k.

Evaluation protocols. For a query q, we compute Average Precision (AP) as the mean of the precision values at the ranks where relevant items are retrieved. We then aggregate AP across queries to obtain mAP, which reflects both relevance and ranking quality. For PatternCom, we follow an *attribute-balanced* protocol: AP is first averaged over queries within each attribute type (*e.g.*, *color*), and the final score is the unweighted average across attribute types.

This corresponds to a *macro-averaged mAP* (mAP_{macro}), ensuring that each attribute contributes equally. For xView2-CIR, the query distribution across disaster types is imbalanced. Therefore, we report both (i) the same *macro-averaged* mAP over disaster categories, mAP_{macro}, which weights each disaster type equally, and (ii) the standard *overall* mAP, $mAP_{overall}$, computed by averaging AP over *all* queries irrespective of category. Reporting both provides a fair view of performance: mAP_{macro} reflects robustness across rare disaster types, while $mAP_{overall}$ reflects expected performance under the natural query frequency.

Hyperparameters. Both PatternCom and xView2-CIR are used strictly as evaluation sets. To avoid tuning on test data, we keep method hyperparameters fixed to the default settings reported in the original papers (or their official implementations) whenever applicable. When a method exposes several knobs (*e.g.*, visual/textual expansion sizes, inversion iterations, diffusion steps), we do *not* optimize them on PatternCom or xView2-CIR. Instead, we report a sensitivity analysis in the ablation section to characterize how performance varies with key hyperparameters, and we use a single fixed configuration for all main comparisons. For instance, for WeiCom we report the λ sensitivity separately, while using a single fixed λ in the main tables for fair comparison.

4.2. Experimental analysis

Benchmark. Table 1, 2, and 3 report the main quantitative results on PatternCom and xView2-CIR, respectively, covering a broad range of composition methods and vision–language models. We include unimodal baselines (*text-only*, *image-only*) and two simple multimodal fusions: score averaging (*text* + *image*) and score multiplication (*text* × *image*). We further benchmark representative state-of-the-art CIR methods from the computer vision literature: CompoDiff [25], Pic2Word [71], SEARLE [2], CIReVL [38], MagicLens [106], WeiCom [64], BASIC [65], and FreeDom [19], after applying domain-grounded adaptations required for EO imagery. Our adaptation principle is intentionally conservative: for each method, we preserve the original inference mechanism and introduce only minimal domain-grounded modifications needed to make it meaningful for EO imagery, such as replacing generic vocabularies, adjusting prompt templates, or using remote-sensing-relevant textual resources. We generally avoid benchmark-specific tuning beyond what is required for compatibility. Detailed adaptation choices and implementation details are provided in A.4.

4.3. Experimental results

Quantitative results on PatternCom Table 1 and 2 summarize attribute-modification retrieval performance on PatternCom across six VLM backbones. Unimodal baselines are consistently weak: *image-only* usually outperforms *text-only*, confirming that visual similarity alone is useful but insufficient for satisfying the textual modifier. Simple multimodal fusion provides clear gains over unimodal retrieval. In particular, *text* × *image* is often stronger than *text* + *image*, although it is less stable on some backbones, such as SigLIP.

Among composition methods, FreeDom is the strongest performer, achieving the best average mAP across all six backbones, with particularly high absolute performance on SigLIP (46.61%) and strong results on RemoteCLIP, SkyCLIP-50, and CLIP LAION-RS. CIReVL and BASIC form the next tier. CIReVL performs particularly well on remote-sensing-adapted CLIP backbones like RemoteCLIP and SkyCLIP-50. BASIC is the second strongest with OpenAI CLIP and SigLIP. WeiCom remains a competitive lightweight baseline, while MagicLens shows that supervised multimodal heads trained on natural-image triplets

Table 1
Attribute modification performance on PatternCom (mAP, %). Results for three vision—language backbones and up to eight composition methods. We report *macro-averaged mAP* (mAP_{macro}): for each attribute type, AP is averaged over all queries targeting a given attribute value (*e.g.*, `rectangular` for Shape), then averaged over the remaining values of the same attribute type (*e.g.*, `oval`, `kidney-shaped`), and finally averaged across attribute types. Avg.: resulting mAP_{macro} over all attribute types and classes; **bold**: best; underline: second.

(a) CLIP LAION-2B

Method	Color	Context	Density	Existence	Quantity	Shape	Avg.
Text-only	13.47	4.83	3.58	2.00	3.31	6.22	5.57
Image-only	14.66	8.32	13.49	16.47	7.84	15.76	12.74
Text + Image	23.13	11.02	15.87	16.93	10.13	21.38	16.41
Text × Image	40.97	11.87	14.48	14.36	20.31	23.99	21.00
SEARLE	14.75	7.98	13.63	15.23	8.01	15.86	12.58
CIReVL	17.79	28.55	17.69	35.14	14.95	25.91	23.34
WeiCom	46.08	17.45	16.49	8.36	18.15	23.97	21.75
BASIC	29.26	6.38	15.29	18.54	12.18	17.18	16.47
FreeDom	**46.55**	**43.49**	**21.32**	**46.98**	**25.09**	**48.13**	**38.59**

(b) RemoteCLIP

Method	Color	Context	Density	Existence	Quantity	Shape	Avg.
Text-only	10.75	8.87	22.16	6.98	8.25	24.12	13.52
Image-only	14.40	6.62	15.11	13.10	6.99	15.18	11.90
Text + Image	23.67	10.01	18.45	13.98	7.97	19.63	15.62
Text × Image	47.20	19.65	27.09	12.97	14.59	40.60	27.02
CompoDiff	8.58	17.88	14.41	6.04	9.34	12.46	11.45
Pic2Word	40.88	**40.26**	17.41	16.92	9.18	27.98	25.44
SEARLE	14.44	6.00	13.49	12.95	7.29	14.86	11.51
CIReVL	42.80	36.79	21.34	**43.81**	19.35	35.55	33.27
WeiCom	43.68	31.45	**39.94**	14.92	20.51	29.78	30.05
BASIC	47.65	15.45	17.55	22.03	20.76	25.00	24.74
FreeDom	**49.80**	38.80	28.31	36.64	**26.77**	**39.49**	**36.64**

(c) OpenAI CLIP

Method	Color	Context	Density	Existence	Quantity	Shape	Avg.
Text-only	3.49	7.75	2.51	3.31	1.01	2.94	3.50
Image-only	13.61	7.86	18.19	13.30	8.30	15.08	12.72
Text + Image	16.59	11.13	19.88	15.13	8.96	17.55	14.87
Text × Image	27.59	14.65	15.63	17.97	9.08	17.77	17.12
CompoDiff	10.70	8.32	16.83	12.36	4.67	13.50	11.06
Pic2Word	27.09	20.93	19.04	9.81	8.05	19.24	17.36
SEARLE	15.24	7.25	16.20	13.30	9.00	15.78	12.80
CIReVL	29.32	31.14	16.02	37.23	11.70	22.88	24.72
MagicLens	34.63	18.98	14.54	21.46	13.21	17.70	20.09
WeiCom	27.42	25.15	13.63	21.98	7.01	15.33	18.42
BASIC	38.31	24.53	26.35	30.94	**18.03**	**35.73**	28.98
FreeDom	**39.08**	**45.87**	**36.72**	**37.25**	13.22	28.69	**33.47**

Table 2
Attribute modification performance on PatternCom (mAP, %). Results for three vision—language backbones and up to seven composition methods. We report *macro-averaged mAP* (mAP_{macro}): for each attribute type, AP is averaged over all queries targeting a given attribute value (*e.g.*, `rectangular` for Shape), then averaged over the remaining values of the same attribute type (*e.g.*, `oval`, `kidney-shaped`), and finally averaged across attribute types. Avg.: resulting mAP_{macro} over all attribute types and classes; **bold**: best; underline: second.

(a) SigLIP

Method	Color	Context	Density	Existence	Quantity	Shape	Avg.
Text-only	7.99	3.93	2.30	9.93	0.58	31.16	9.32
Image-only	15.03	10.67	21.46	21.13	8.94	17.37	15.77
Text + Image	25.11	11.46	20.47	21.31	10.66	27.92	19.49
Text × Image	18.80	5.96	6.57	18.77	1.28	43.66	15.84
WeiCom	24.77	15.32	9.06	22.37	4.59	43.64	19.96
BASIC	36.11	10.30	21.76	24.68	20.43	19.46	22.12
FreeDom	**47.99**	**49.88**	**22.84**	**54.70**	**37.05**	**67.17**	**46.61**

(b) CLIP LAION-RS

Method	Color	Context	Density	Existence	Quantity	Shape	Avg.
Text-only	3.85	6.13	2.60	5.56	0.99	4.57	3.95
Image-only	13.35	7.51	15.82	15.79	7.50	15.91	12.65
Text + Image	20.94	10.07	16.54	16.91	8.75	20.78	15.67
Text × Image	21.38	12.98	10.91	16.91	10.20	22.32	15.78
CompoDiff	19.70	11.32	17.98	17.97	9.67	20.36	16.17
Pic2Word	31.18	19.58	16.53	11.87	7.88	19.03	17.68
SEARLE	14.89	10.23	14.34	17.20	8.31	16.51	13.58
CIReVL	32.57	36.45	15.33	34.68	**13.78**	24.88	26.28
WeiCom	29.63	30.45	11.19	21.13	8.60	25.85	21.14
BASIC	35.51	14.06	16.32	25.56	12.18	21.81	20.91
FreeDom	**41.78**	**51.34**	**21.88**	**38.22**	13.29	**33.58**	**33.35**

(c) SkyCLIP-50

Method	Color	Context	Density	Existence	Quantity	Shape	Avg.
Text-only	4.60	8.44	3.44	5.61	1.02	7.34	5.08
Image-only	14.48	9.03	17.82	20.11	8.77	16.32	14.42
Text + Image	23.49	12.64	19.54	21.42	10.24	21.80	18.19
Text × Image	37.03	15.84	16.24	23.99	12.90	31.51	22.92
CompoDiff	20.59	12.67	17.57	21.98	10.23	22.31	17.56
Pic2Word	34.57	21.58	18.23	13.09	10.03	21.93	19.91
SEARLE	16.57	11.68	16.41	20.22	9.22	16.88	15.16
CIReVL	34.90	34.93	17.09	35.03	14.39	28.92	27.54
WeiCom	40.46	38.10	18.17	27.91	10.10	31.52	27.71
BASIC	34.89	12.20	16.64	28.98	12.32	18.30	20.56
FreeDom	**49.88**	**50.47**	**29.71**	**38.00**	**14.58**	**39.99**	**37.11**

can transfer to EO imagery to some extent. In contrast, continuous inversion and diffusion-based approaches are generally less competitive. Pic2Word and SEARLE yield modest gains at best, while CompoDiff is often unstable and can even underperform the *image-only* baseline. This suggests that methods relying on synthesized query embeddings or pseudo-token inversion are more sensitive to domain shift in EO imagery.

Overall, PatternCom favors methods that preserve class identity while injecting attribute-specific textual cues. Strong textual surrogates and domain-grounded vocabularies are therefore particularly beneficial, explaining the consistent advantage of FreeDom in this setting.

Table 3
Disaster monitoring performance on xView2-CIR (mAP, %). Results for two vision–language backbones and three composition methods. We report per-disaster mAP (%) for `post-*` modifiers (*e.g.*, Hurricane). Avg.: mAP_{macro} across disasters (equal weight per disaster); Total: $mAP_{overall}$ (weighted by the number of queries per disaster); **bold**: best; underline second.

(a) OpenAI CLIP

Method	Hurricane	Wildfire	Flood	Tsunami	Earthquake	Volcano	Avg.	Total
Text-only	2.50	1.61	3.25	18.08	4.19	15.77	7.57	2.97
Image-only	6.99	2.39	1.44	25.17	8.35	9.69	9.00	5.53
Text + Image	14.61	7.45	4.75	37.46	16.63	25.49	17.73	12.14
Text × Image	18.11	12.12	8.36	37.63	17.93	**61.46**	25.94	16.35
WeiCom	**25.94**	13.88	9.43	42.96	**36.48**	55.34	**30.67**	**21.40**
BASIC	22.71	**18.49**	**14.59**	**46.15**	1.69	59.20	27.14	21.38
FreeDom	18.01	17.46	3.77	35.03	3.21	25.88	17.23	16.84

(b) SigLIP

Method	Hurricane	Wildfire	Flood	Tsunami	Earthquake	Volcano	Avg.	Total
Text-only	2.32	1.81	1.09	7.10	0.88	13.77	4.49	2.31
Image-only	10.65	9.74	10.68	24.15	14.44	20.77	15.07	10.97
Text + Image	15.86	19.94	13.05	26.29	19.92	26.17	20.20	17.50
Text × Image	10.74	13.69	14.20	17.54	**23.17**	53.96	22.22	13.08
WeiCom	20.19	**30.61**	9.95	29.46	18.56	32.09	23.48	**23.14**
BASIC	**21.54**	12.30	13.98	**36.52**	13.86	**56.79**	**25.83**	18.52
FreeDom	13.07	15.37	**21.22**	19.33	1.60	48.61	19.99	15.37

Quantitative results on xView2-CIR Table 3 reports performance on xView2-CIR under the *same scene/location + target state* criterion. Unimodal baselines remain weak, while score fusion provides clear gains across both backbones. *Text + image* is consistently reliable, whereas *text × image* is more variable across disaster types. Among composition methods, WeiCom is the most robust overall, achieving the best TOTAL score for both OpenAI CLIP and SigLIP, suggesting that calibrated modality fusion is well suited to identity-preserving change retrieval. In contrast, FreeDom and BASIC are less effective than on PatternCom, partly because their query-expansion mechanisms can introduce visually similar but geographically different scenes, which is harmful when relevance depends on scene identity. These results show that change-centric RSCIR poses different challenges from attribute-based retrieval and should be evaluated as a distinct setting.

Qualitative results on PatternCom. In Figure 3, we visualize representative retrieval outputs on PatternCom with OpenAI CLIP, comparing unimodal baselines (image-only, text-only) against two multimodal methods (WeiCom and FreeDom). The examples clearly illustrate the limitations of unimodal retrieval. WeiCom generally improves over unimodal baselines by leveraging both modalities, but it can still be sensitive to modality dominance depending on the attribute type. In the *activity monitoring* example (parking lot + `empty`), the retrieved results remain biased toward the visual layout of the reference image, yielding parking-lot-like structures that do not clearly satisfy the "empty" intent. In the *capacity monitoring* example (storage tanks + `one`), the textual cue can dominate, pulling results with fewer tanks that are only loosely aligned with the specific query instance. In contrast, FreeDom is consistently the most reliable across

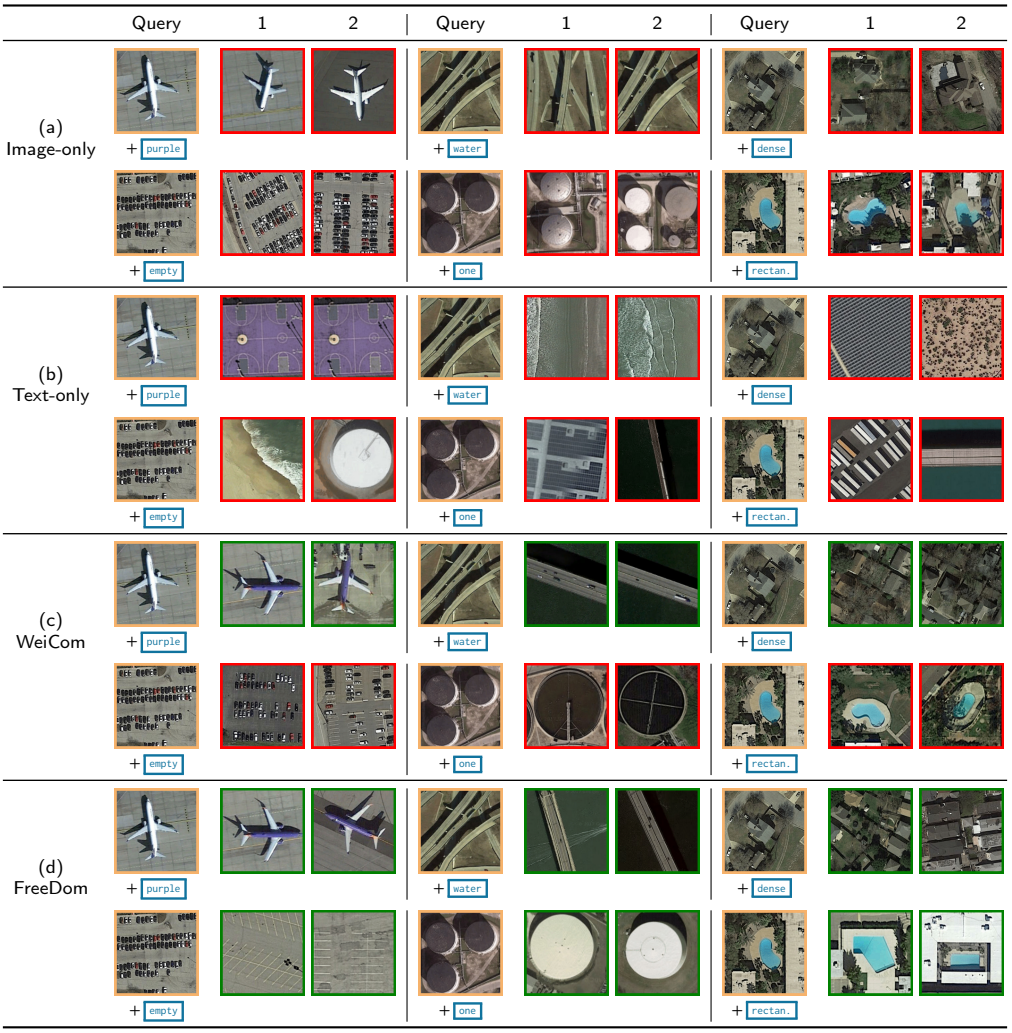

Figure 3: Qualitative composed retrieval results on PatternCom with OpenAI CLIP. Comparison between unimodal and multimodal methods. Each query is shown as a reference image combined with a boxed textual modifier. Columns report the top-2 retrieved results. Retrieval is evaluated under the *same class + target attribute value* relevance criterion.

all shown cases, producing results that simultaneously respect the query class and the target attribute value, aligning well with PatternCom's relevance criterion.

Qualitative results on xView2-CIR. In Figure 4, we visualize qualitative composed retrieval results on xView2-CIR using OpenAI CLIP with WeiCom. Unlike PatternCom, relevance in xView2-CIR requires *the same scene or location and* the *target state* (post-event), which exposes a key failure mode of generic composition: many retrieved images match the requested disaster state but drift to a different geographic area with similar visual cues. This is especially apparent for hurricanes and wildfires, where widespread debris/burn patterns and textured backgrounds can look plausible across locations, leading to several off-scene false positives.

Still, WeiCom succeeds in a subset of cases where the scene has distinctive geometry or landmarks (*e.g.*, road layouts, coastline structure, dense urban blocks), allowing the query image to anchor location while the text steers the retrieval toward the correct post-event state. Overall, the figure highlights that xView2-CIR is less forgiving than attribute-only benchmarks: effective composed retrieval must jointly preserve *instance identity* (same place) while applying a *state change* constraint, and methods that lean too heavily on "state-like" appearance cues tend to retrieve the right disaster but the wrong scene.

Qualitative results on LEVIR-CC. To illustrate more applied EO use cases, Figure 5 shows qualitative results on LEVIR-CC [50] using WeiCom with OpenAI CLIP. As shown in the examples for *building construction* and *road construction*, the method takes a reference image together with a change-oriented modifier (*e.g.*, `new building`) and retrieves candidate scenes that match the requested change, under the *same location/scene + target state* relevance criterion. In these cases, the top retrieved results frequently correspond to plausible construction changes, suggesting that composed retrieval can serve as a lightweight interface for change-focused search in real remote sensing workflows (*e.g.*, rapid infrastructure monitoring).

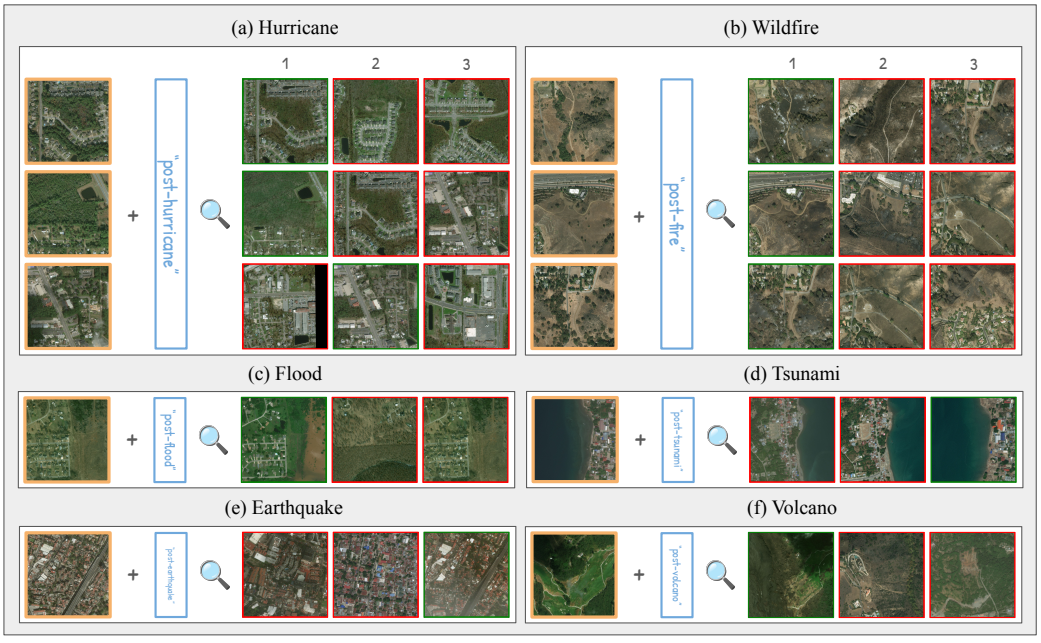

Figure 4: Qualitative composed retrieval results on xView2-CIR with OpenAI CLIP and WeiCom. Each query combines a *pre-disaster* reference image with a boxed textual modifier (`post-*`) indicating the target post-event state. We show the top retrieved *post-disaster* results per disaster type. Retrieval is evaluated under the *same scene/location + target state* relevance criterion, so visually plausible post-event matches from different locations are considered negatives.

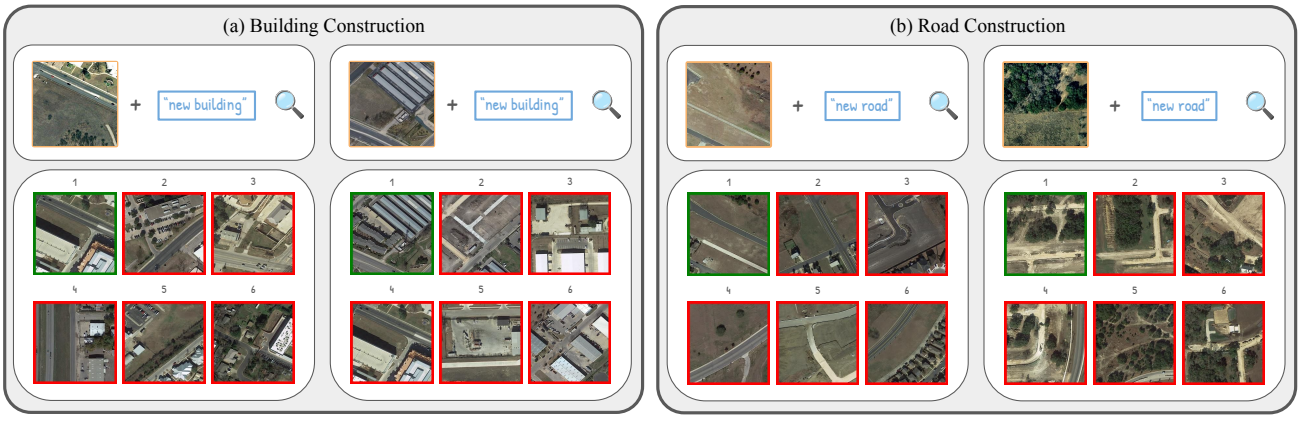

Figure 5: Qualitative composed retrieval on LEVIR-CC with OpenAI CLIP and WeiCom. Each query is shown as a reference image combined with a textual change modifier. We report the top retrieved candidates under the *same location + target state* relevance criterion.

4.4. Sensitivity and Ablation Analysis

Impact of λ in WeiCom. Figure 6 analyzes the modality-control parameter λ of WeiCom with RemoteCLIP, where λ=0 corresponds to *image-only* and λ=1 to *text-only* retrieval. Although all main results use a fixed λ=0.5 to avoid tuning on PatternCom, the sensitivity analysis shows that performance is maximized at intermediate values, with the best average mAP obtained at λ=0.6. Similar trends are observed across backbones, where the optimal λ consistently lies away from the unimodal extremes (e.g., λ=0.3 for CLIP LAION-2B). This broad mid-range optimum indicates that both the visual reference and textual modifier contribute meaningfully to attribute-based retrieval. The main exception is *color*, which peaks at smaller λ values, suggesting stronger dependence on the visual signal.

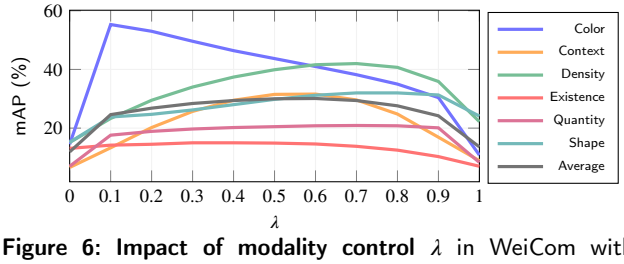

Figure 6: Impact of modality control λ in WeiCom with RemoteCLIP on PatternCom. Curves report attribute-wise mAP (%) as λ varies.

Impact of vocabulary on FreeDom. Figure 7 analyzes how the textual memory affects FreeDom on PatternCom with CLIP LAION-2B. Open Images v7 provides strong performance due to its large and diverse vocabulary of roughly 21k concepts. However, the EO-specific RSText vocabularies remain competitive despite being much smaller, with up to 200× fewer entries for RSText-150 and 20× fewer entries for RSText-1k. This suggests that compact, domain-grounded vocabularies can provide effective textual surrogates for EO imagery. The best performance is obtained by HybridText-23k, which combines Open Images with RSText-2k, indicating that broad semantic coverage and EO-specific grounding are complementary.

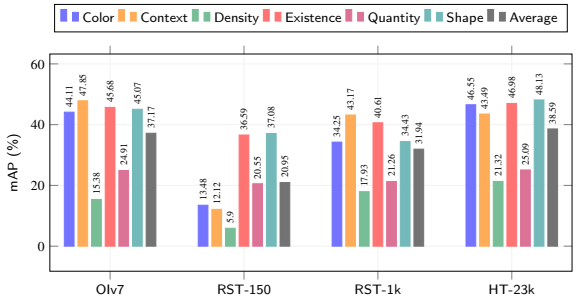

Figure 7: Impact of vocabulary on FreeDom with CLIP LAION-2B on PatternCom. We compare Open Images v7, EO-specific RSText vocabularies, and their hybrid combination HybridText-23k. Bars report mAP (%) per attribute and the average.

Impact of textual expansion on FreeDom. Figure 8 analyzes frequency-based textual expansion in FreeDom by disabling visual query expansion (k=1) and varying the number of aggregated textual concepts (m=n). On PatternCom, increasing m=n improves average mAP from 28.99% at m=n=1 to 35.89% at m=n=7, with the largest gains on shape and context. The same moderate-expansion trend appears on xView2-CIR, where performance improves from 9.59% to 16.84% at m=n=7. In both settings, larger expansions plateau or degrade, suggesting that overly long textual surrogates introduce noisy or off-target concepts. The corresponding xView2-CIR analysis is reported in A.5.

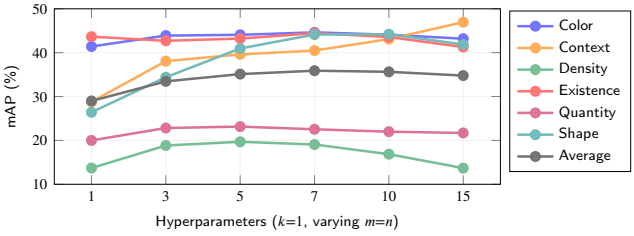

Figure 8: Impact of textual expansion on FreeDom with CLIP LAION-2B on PatternCom. Visual expansion is disabled (k=1), and the number of aggregated textual concepts is varied as m=n.

Impact of visual expansion on FreeDom. Figure 9 and 10 analyze visual query expansion in FreeDom by varying the number of proxy images k while disabling textual expansion (m=n=1). On PatternCom, visual expansion is *beneficial* up to a moderate range: average mAP increases from 28.99% at k=1 to 30.41% at k=100, before dropping to 26.71% at k=1000. This is consistent with the *same class + target attribute* criterion, where nearest neighbors often remain class-consistent and provide useful attribute evidence, until overly large proxy sets introduce noisy distractors. In contrast, visual expansion is *harmful* on xView2-CIR: Total mAP decreases from 9.59% at k=1 to 4.51% at k=500. Since xView2-CIR requires the *same scene/location + target state*, visually similar proxies usually correspond to different locations and inject irrelevant content. This explains why proxy-based expansion benefits attribute-oriented retrieval but can degrade identity-preserving change retrieval.

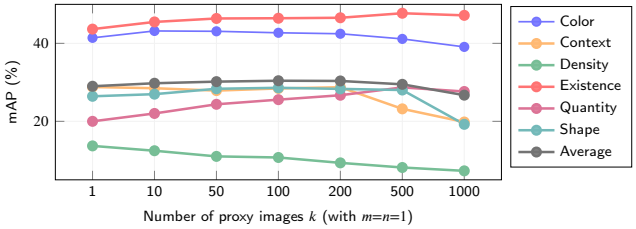

Figure 9: Impact of visual expansion on FreeDom with CLIP LAION-2B on PatternCom. The number of proxy images k is varied while textual expansion is disabled (m=n=1).

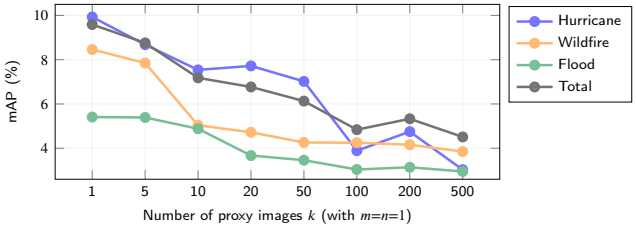

Figure 10: Impact of visual expansion on FreeDom with OpenAI CLIP on xView2-CIR. The number of proxy images k is varied while textual expansion is disabled (m=n=1).

Ablation of BASIC components. We further analyze the contribution of BASIC components, with full results reported in A.5. On PatternCom, the largest drops occur when removing *centering* or *semantic projection*, indicating that modality calibration and semantic subspace alignment are the main drivers of BASIC's performance. Harris weighting, min-based normalization, text contextualization, and query expansion provide smaller, secondary gains. A similar trend appears on xView2-CIR, where centering and semantic projection remain the most important components, while visual query expansion becomes harmful under the *same scene/location + target state* criterion. This supports the broader finding that proxy-based expansion can help attribute-oriented retrieval but may degrade identity-preserving change retrieval.

Sensitivity to text query phrasing. We also evaluate whether composed retrieval is sensitive to the phrasing of the textual modifier, with full results reported in A.5. On PatternCom, the original plain attribute values (*e.g.*, `blue`, `water`, `dense`) are generally the most reliable, especially for

stronger multimodal methods. More verbose templates such as `having blue color` or `with the main object modified to have blue color` often reduce performance, suggesting that additional function words can dilute the compact attribute signal in this benchmark.

On xView2-CIR, prompt sensitivity is more dataset- and method-dependent. Rephrasing `post-{disaster}` into impact-oriented or disaster-explicit descriptions (*e.g.*, `burned area`, `flooded area`, `fire-affected region`) can benefit methods that rely more strongly on semantic text structure, particularly BASIC. Overall, these results indicate that prompt phrasing is an important practical factor: compact modifiers are preferable for controlled attribute edits, whereas change-centric retrieval may benefit from more descriptive event semantics.

4.5. Discussion

Generalization across backbones and EO adaptation. Our benchmark suggests that both remote-sensing adaptation and general VLM quality matter for RSCIR. RS-adapted CLIP variants, such as RemoteCLIP, CLIP LAION-RS, and SkyCLIP-50, often improve over generic CLIP-style backbones under the same protocol. However, the gains from domain adaptation can be comparable to, or smaller than, the gains obtained by using stronger general-purpose backbones and training objectives, such as SigLIP. This indicates that advances in VLM architecture, pre-training data, and loss design transfer directly to composed retrieval in EO imagery. Consequently, future RSCIR systems should consider both EO-specific adaptation and improvements in general VLM representation quality.

Scalability and deployment trade-offs. The results show that training-free composition methods provide strong baselines for EO retrieval. FreeDom is the strongest method on PatternCom, while WeiCom and BASIC remain competitive lightweight alternatives. This is important for operational EO systems, where latency, scalability, and engineering complexity matter alongside accuracy. CIReVL can be effective, but requires image captioning and LLM-based rewriting, introducing additional computational cost and possible failure modes. In contrast, score-fusion or feature-calibration methods such as WeiCom and BASIC are simpler to deploy over large archives. Supervised multimodal heads such as MagicLens also transfer to EO imagery to some extent, but their performance may remain limited without training on remote-sensing compositional data.

Attribute-based and change-centric RSCIR require different mechanisms. The contrast between PatternCom and xView2-CIR shows that change-centric RSCIR should be treated as a distinct retrieval setting rather than a direct extension of attribute-based retrieval. PatternCom rewards methods that preserve class identity while injecting attribute-specific textual cues, which explains the strong performance of vocabulary-based methods such as FreeDom. In xView2-CIR, however, relevance requires the same scene/location and a target post-event state. This makes proxy-based visual expansion less reliable, because visually similar images often correspond to different locations and can dilute the scene-identity signal. These findings suggest that methods designed for attribute editing may not directly transfer to identity-preserving change retrieval.

Role in operational EO workflows. RSCIR is not a replacement for pixel-level change detection, damage assessment, or rapid mapping. Instead, it provides a complementary interface for controllable archive exploration and rapid evidence gathering. In practical workflows, composed retrieval can support location-aware search, retrieval of visually similar historical cases, analyst-in-the-loop exploration, quality control through hard-negative discovery, and weak supervision for downstream change models. This is especially useful when dense annotations, perfect co-registration, or complete temporal coverage are unavailable.

Interpretability and analyst interaction. Interpretability is also important for EO deployment. Methods based on explicit textual memories, such as FreeDom, provide human-readable surrogate concepts for the visual query, making it easier to inspect which semantic cues drive retrieval. Score-fusion methods such as WeiCom expose the relative contribution of image and text through the modality weight, while BASIC highlights the role of feature calibration and semantic projection. Such properties can help analysts diagnose whether failures arise from excessive visual dominance, weak text grounding, or scene-identity drift.

Sensitivity to text phrasing. Prompt phrasing is a practical factor in composed retrieval. On PatternCom, compact attribute modifiers such as `blue`, `water`, or `dense` are generally most effective, whereas verbose templates can dilute the attribute signal. In contrast, xView2-CIR can benefit from more descriptive, impact-oriented formulations, particularly for methods that rely more strongly on semantic text structure. This suggests that prompt design should be treated as part of the retrieval system: short modifiers are suitable for controlled attribute edits, while change-centric settings may require more explicit event or impact descriptions.

5. Limitations

This study has several limitations. First, although xView2-CIR extends RSCIR toward operationally relevant change-centric retrieval, it remains relatively small and imbalanced, especially for rare disaster categories; results on these categories should therefore be interpreted as indicative rather than definitive. Second, our evaluation focuses on image-level retrieval and does not address finer-grained localization or dense temporal search. Third, some composition methods are more naturally tied to specific backbones or representation spaces, which limits strict cross-backbone comparability. Finally, prompt design and auxiliary vocabularies influence performance, particularly in change-centric scenarios, and require more systematic

study. We therefore view this work as a benchmark-and-analysis foundation for RSCIR rather than a final solution to composed retrieval in Earth observation.

6. Conclusion

We studied remote sensing composed image retrieval (RSCIR), where a query combines a reference image with a textual modifier to express targeted retrieval intent. We established a unified benchmark on PatternCom with domain-grounded adaptations of representative composition methods and a standardized protocol spanning six vision–language backbones. To connect benchmarking with operational EO needs, we introduced xView2-CIR, a change-centric benchmark for disaster and damage monitoring, where retrieval requires both scene identity and a target post-event state.

Our experiments show that stronger vision–language backbones transfer directly to RSCIR, and that training-free composition strategies can be highly effective in EO settings. FreeDom achieves the strongest performance on PatternCom, while lightweight methods such as WeiCom and BASIC remain competitive and attractive for practical deployment. At the same time, xView2-CIR reveals that change-centric retrieval differs substantially from attribute-based editing: preserving scene identity changes the method ranking and exposes weaknesses in proxy-based expansion strategies.

Overall, our results position RSCIR as a practical complement to standard RSIR and change-analysis pipelines, especially when analysts need controllable, semantically guided access to large EO archives. Future work should expand change-centric benchmarks, study prompt and vocabulary design more systematically, and connect scene-level composed retrieval with finer-grained localization and temporal reasoning.

7. Data and Code Availability Statement

The code and data supporting the findings of this study are publicly available at https://github.com/billpsomas/rscir. PatternCom is based on publicly available source data. The proposed xView2-CIR benchmark and the scripts used for data preparation, evaluation, and reproduction of the reported results are also provided in the repository.

8. Acknowledgment

This work was supported by the Czech Technical University in Prague grant No. SGS23/173/OHK3/3T/13, the EU Horizon Europe programme MSCA PF RAVIOLI (No. 101205297), and the Junior Star GACR GM 21-28830M. We acknowledge VSB–Technical University of Ostrava, IT4Innovati ons National Supercomputing Center, Czech Republic, for awarding this project (OPEN-33-67) access to the LUMI supercomputer, owned by the EuroHPC Joint Undertaking, hosted by CSC (Finland) and the LUMI consortium, through the Ministry of Education, Youth and Sports of the Czech Republic via the e-INFRA CZ project (ID: 90254). The access to the computational infrastructure of the OP VVV funded project CZ.02.1.01/0.0/0.0/16_019/0000765 “Research Center for Informatics” is also gratefully acknowledged.

References

[1] Agouris, P., Carswell, J., Stefanidis, A., 1999. An environment for content-based image retrieval from large spatial databases. ISPRS J. Photogramm. Remote Sens. .

[2] Baldrati, A., Agnolucci, L., Bertini, M., Del Bimbo, A., 2023. Zero-shot composed image retrieval with textual inversion, in: Proc. IEEE/CVF Int. Conf. Comput. Vis., pp. 15338–15347.

[3] Baldrati, A., Bertini, M., Uricchio, T., Del Bimbo, A., 2022. Effective conditioned and composed image retrieval combining clip-based features, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.

[4] Berg, T.L., Berg, A.C., Shih, J., 2010. Automatic attribute discovery and characterization from noisy web data, in: Proc. Eur. Conf. Comput. Vis., Springer.

[5] Bhagavathy, S., Manjunath, B.S., 2006. Modeling and detection of geospatial objects using texture motifs. IEEE Trans. Geosci. Remote Sens. 44, 3706–3715.

[6] Boualleg, Y., Farah, M., 2018. Enhanced interactive remote sensing image retrieval with scene classification convolutional neural networks model, in: Proc. IEEE Int. Geosci. Remote Sens. Symp., pp. 4748–4751.

[7] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al., 2020. Language models are few-shot learners. Adv. Neural Inf. Process. Syst. 33, 1877–1901.

[8] Cao, R., Zhang, Q., Zhu, J., Li, Q., Li, Q., Liu, B., Qiu, G., 2020. Enhancing remote sensing image retrieval using a triplet deep metric learning network. Int. J. Remote Sens. 41, 740–751.

[9] Chaudhuri, B., Demir, B., Chaudhuri, S., Bruzzone, L., 2017. Multilabel remote sensing image retrieval using a semisupervised graph-theoretic method. IEEE Trans. Geosci. Remote Sens. 56, 1144–1158.

[10] Chaudhuri, U., Banerjee, B., Bhattacharya, A., 2019. Siamese graph convolutional network for content based remote sensing image retrieval. Comput. Vis. Image Underst. .

[11] Chaudhuri, U., Banerjee, B., Bhattacharya, A., Datcu, M., 2021. Attention-driven graph convolution network for remote sensing image retrieval. IEEE Geosci. Remote Sens. Lett. 19, 1–5.

[12] Chen, Y., Bazzani, L., 2020. Learning joint visual semantic matching embeddings for language-guided retrieval, in: Proc. Eur. Conf. Comput. Vis.

[13] Chen, Y., Gong, S., Bazzani, L., 2020. Image search with text feedback by visiolinguistic attention learning, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.

[14] Cheng, Q., Gan, D., Fu, P., Huang, H., Zhou, Y., 2021a. A novel ensemble architecture of residual attention-based deep metric learning for remote sensing image retrieval. Remote Sens. 13, 3445.

[15] Cheng, Q., Huang, H., Ye, L., Fu, P., Gan, D., Zhou, Y., 2021b. A semantic-preserving deep hashing model for multi-label remote sensing image retrieval. Remote Sens. 13, 4965.

[16] Cherti, M., Beaumont, R., Wightman, R., Zhai, X., Beyer, L., Kolesnikov, A., Dosovitskiy, A., Houlsby, N., Minderer, M., 2022. Openclip: An open source implementation of clip .

[17] Dai, O.E., Demir, B., Sankur, B., Bruzzone, L., 2017. A novel system for content based retrieval of multi-label remote sensing images, in: Proc. IEEE Int. Geosci. Remote Sens. Symp., pp. 1744–1747.

[18] Delmas, G., de Rezende, R.S., Csurka, G., Larlus, D., 2022. Artemis: Attention-based retrieval with text-explicit matching and implicit similarity.

[19] Efthymiadis, N., Psomas, B., Laskar, Z., Karantzalos, K., Avrithis, Y., Chum, O., Tolias, G., 2025. Composed image retrieval for

training-free domain conversion, in: Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.

[20] Fan, L., Zhao, H., Zhao, H., 2020. Global optimization: Combining local loss with result ranking loss in remote sensing image retrieval. IEEE Trans. Geosci. Remote Sens. 59, 7011–7026.

[21] Frome, A., Corrado, G.S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., Mikolov, T., 2013. Devise: A deep visual-semantic embedding model. Adv. Neural Inf. Process. Syst. 26.

[22] Ge, Y., Jiang, S., Xu, Q., Jiang, C., Ye, F., 2018. Exploiting representations from pre-trained convolutional neural networks for high-resolution remote sensing image retrieval. Multimed. Tools Appl. 77, 17489–17515.

[23] Gordo, A., Almazan, J., Revaud, J., Larlus, D., 2016. Deep image retrieval: Learning global representations for image search, in: Proc. Eur. Conf. Comput. Vis.

[24] Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R., 2017. Google earth engine: Planetary-scale geospatial analysis for everyone. Remote Sens. Environ. 202, 18–27.

[25] Gu, G., Chun, S., Kim, W., Jun, H., Kang, Y., Yun, S., a. Compodiff: Versatile composed image retrieval with latent diffusion. Transactions on Machine Learning Research .

[26] Gu, X., Lin, T.Y., Kuo, W., Cui, Y., b. Open-vocabulary object detection via vision and language knowledge distillation.

[27] Haklay, M., Weber, P., 2008. Openstreetmap: User-generated street maps. IEEE Pervasive Comput. 7, 12–18.

[28] Han, X., Wu, Z., Huang, P.X., Zhang, X., Zhu, M., Li, Y., Zhao, Y., Davis, L.S., 2017. Automatic spatially-aware fashion concept discovery, in: Proc. IEEE/CVF Int. Conf. Comput. Vis.

[29] Hosseinzadeh, M., Wang, Y., 2020. Composed query image retrieval using locally bounded features, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.

[30] Hou, D., Miao, Z., Xing, H., Wu, H., 2020. Exploiting low dimensional features from the mobilenets for remote sensing image retrieval. Earth Sci. Inform. 13, 1437–1443.

[31] Hu, F., Tong, X., Xia, G.S., Zhang, L., 2016. Delving into deep representations for remote sensing image retrieval, in: Proc. IEEE Int. Conf. Signal Process., pp. 198–203.

[32] Hu, S., Feng, M., Nguyen, R.M., Lee, G.H., 2018. Cvm-net: Cross-view matching network for image-based ground-to-aerial geo-localization, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 7258–7267.

[33] Imbriaco, R., Sebastian, C., Bondarev, E., de With, P.H., 2019. Aggregated deep local features for remote sensing image retrieval. Remote Sens. 11, 493.

[34] Imbriaco, R., Sebastian, C., Bondarev, E., de With, P.H., 2021. Toward multilabel image retrieval for remote sensing. IEEE Trans. Geosci. Remote Sens. 60, 1–14.

[35] Isola, P., Lim, J.J., Adelson, E.H., 2015. Discovering states and transformations in image collections, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., pp. 1383–1391.

[36] Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T., 2021. Scaling up visual and vision-language representation learning with noisy text supervision, in: Proc. Int. Conf. Mach. Learn.

[37] Kang, J., Fernandez-Beltran, R., Hong, D., Chanussot, J., Plaza, A., 2020. Graph relation network: Modeling relations between scenes for multilabel remote-sensing image classification and retrieval. IEEE Trans. Geosci. Remote Sens. 59, 4355–4369.

[38] Karthik, S., Roth, K., Mancini, M., Akata, Z., 2024. Vision-by-language for training-free compositional image retrieval, in: Int. Conf. Learn. Represent., pp. 16926–16941.

[39] Khurshid, N., Hanif, T., Tharani, M., Taj, M., 2019. Cross-view image retrieval-ground to aerial image retrieval through deep learning, in: Proc. Int. Conf. Neural Inf. Process., Springer. pp. 210–221.

[40] Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., et al., 2020. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. Int. J. Comput. Vis. .

[41] Lam, D., Kuzma, R., McGee, K., Dooley, S., Laielli, M., Klaric, M., Bulatov, Y., McCord, B., 2018. xview: Objects in context in overhead imagery. arXiv preprint arXiv:1802.07856 .

[42] Lee, S., Kim, D., Han, B., 2021. Cosmo: Content-style modulation for image retrieval with text feedback, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.

[43] Li, J., Li, D., Savarese, S., Hoi, S., 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, in: Proc. Int. Conf. Mach. Learn., pp. 19730–19742.

[44] Li, J., Li, D., Xiong, C., Hoi, S., 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, in: Proc. Int. Conf. Mach. Learn.

[45] Li, J., Narayanan, R.M., 2004. Integrated spectral and spatial information mining in remote sensing imagery. IEEE Trans. Geosci. Remote Sens. 42, 673–685.

[46] Li, Y., Zhang, Y., Huang, X., Ma, J., 2018. Learning source-invariant deep hashing convolutional neural networks for cross-source remote sensing image retrieval. IEEE Trans. Geosci. Remote Sens. 56, 6521–6536.

[47] Li, Y., Zhang, Y., Tao, C., Zhu, H., 2016. Content-based high-resolution remote sensing image retrieval via unsupervised feature learning and collaborative affinity metric fusion. Remote Sens. 8, 709.

[48] Lin, T.Y., Cui, Y., Belongie, S., Hays, J., 2015. Learning deep representations for ground-to-aerial geolocalization, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 5007–5015.

[49] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 2014. Microsoft coco: Common objects in context, in: Proc. Eur. Conf. Comput. Vis.

[50] Liu, C., Zhao, R., Chen, H., Zou, Z., Shi, Z., 2022. Remote sensing image change captioning with dual-branch transformers: A new method and a large scale dataset. IEEE Trans. Geosci. Remote Sens. 60, 1–20.

[51] Liu, F., Chen, D., Guan, Z., Zhou, X., Zhu, J., Ye, Q., Fu, L., Zhou, J., 2024. Remoteclip: A vision language foundation model for remote sensing. IEEE Trans. Geosci. Remote Sens. 62, 1–16.

[52] Liu, Y., Ding, L., Chen, C., Liu, Y., 2020a. Similarity-based unsupervised deep transfer learning for remote sensing image retrieval. IEEE Trans. Geosci. Remote Sens. 58, 7872–7889.

[53] Liu, Y., Han, Z., Chen, C., Ding, L., Liu, Y., 2020b. Eagle-eyed multitask cnns for aerial image retrieval and scene classification. IEEE Trans. Geosci. Remote Sens. 58, 6699–6721.

[54] Liu, Y., Liu, Y., Chen, C., Ding, L., 2020c. Remote-sensing image retrieval with tree-triplet-classification networks. Neurocomputing 405, 48–61.

[55] Lv, Y., Xiong, W., Zhang, X., Cui, Y., 2021. Fusion-based correlation learning model for cross-modal remote sensing image retrieval. IEEE Geosci. Remote Sens. Lett. 19, 1–5.

[56] Ma, C., Dai, Q., Liu, J., Liu, S., Yang, J., 2014. An improved svm model for relevance feedback in remote sensing image retrieval. Int. J. Digit. Earth 7, 725–745.

[57] Ma, J., Shi, D., Tang, X., Zhang, X., Han, X., Jiao, L., 2021. Cross-source image retrieval based on ensemble learning and knowledge distillation for remote sensing images, in: Proc. IEEE Int. Geosci. Remote Sens. Symp., IEEE. pp. 2803–2806.

[58] Mamatha, Y., Ananth, A., 2010. Content based image retrieval of satellite imageries using soft query based color composite techniques. Int. J. Comput. Appl. 7, 0975–8887.

[59] Napoletano, P., 2018. Visual descriptors for content-based retrieval of remote-sensing images. Int. J. Remote Sens. 39, 1343–1376.

[60] Neculai, A., Chen, Y., Akata, Z., 2022. Probabilistic compositional embeddings for multimodal image retrieval, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.

[61] Noh, H., Araujo, A., Sim, J., Weyand, T., Han, B., 2017. Large-scale image retrieval with attentive deep local features, in: Proc. IEEE/CVF Int. Conf. Comput. Vis.

[62] OpenAI, 2024. Gpt-4o system card.

[63] Piedra-Fernandez, J.A., Ortega, G., Wang, J.Z., Canton-Garbin, M., 2013. Fuzzy content-based image retrieval for oceanic remote sensing. IEEE Trans. Geosci. Remote Sens. 52, 5422–5431.

[64] Psomas, B., Kakogeorgiou, I., Efthymiadis, N., Tolias, G., Chum, O., Avrithis, Y., Karantzalos, K., 2024. Composed image retrieval for remote sensing, in: Proc. IEEE Int. Geosci. Remote Sens. Symp.

[65] Psomas, B., Retsinas, G., Efthymiadis, N., Filntisis, P., Avrithis, Y., Maragos, P., Chum, O., Tolias, G., 2025. Instance-level composed image retrieval, in: Adv. Neural Inf. Process. Syst.

[66] Radenović, F., Tolias, G., Chum, O., 2019. Fine-tuning cnn image retrieval with no human annotation. IEEE Trans. Pattern Anal. Mach. Intell. .

[67] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al., 2021. Learning transferable visual models from natural language supervision, in: Proc. Int. Conf. Mach. Learn.

[68] Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. Adv. Neural Inf. Process. Syst. .

[69] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B., 2022. High-resolution image synthesis with latent diffusion models, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.

[70] Sadeghi-Tehran, P., Angelov, P., Virlet, N., Hawkesford, M.J., 2019. Scalable database indexing and fast image retrieval based on deep learning and hierarchically nested structure applied to remote sensing and plant biology. J. Imaging 5, 33.

[71] Saito, K., Sohn, K., Zhang, X., Li, C.L., Lee, C.Y., Saenko, K., Pfister, T., 2023. Pic2word: Mapping pictures to words for zero-shot composed image retrieval, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.

[72] Sarafianos, N., Xu, X., Kakadiaris, I.A., 2019. Adversarial representation learning for text-to-image matching, in: Proc. IEEE/CVF Int. Conf. Comput. Vis., pp. 5814–5824.

[73] Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al., 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. Adv. Neural Inf. Process. Syst. .

[74] Shao, Z., Zhou, W., Cheng, Q., 2014. Remote sensing image retrieval with combined features of salient region. Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci. 40, 83–88.

[75] Shao, Z., Zhou, W., Deng, X., Zhang, M., Cheng, Q., 2020. Multilabel remote sensing image retrieval based on fully convolutional network. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 13, 318–328.

[76] Shi, Y., Li, H., 2022. Beyond cross-view image retrieval: Highly accurate vehicle localization using satellite image, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., pp. 17010–17020.

[77] Stojnić, V., Kalantidis, Y., Matas, J., Tolias, G., 2025. Lposs: Label propagation over patches and pixels for open-vocabulary semantic segmentation, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., pp. 9794–9803.

[78] Sumbul, G., De Wall, A., Kreuziger, T., Marcelino, F., Costa, H., Benevides, P., Caetano, M., Demir, B., Markl, V., 2021a. Bigearthnet-mm: A large-scale, multimodal, multilabel benchmark archive for remote sensing image classification and retrieval. IEEE Geosci. Remote Sens. Mag. 9, 174–180.

[79] Sumbul, G., Demir, B., 2021. A novel graph-theoretic deep representation learning method for multi-label remote sensing image retrieval, in: Proc. IEEE Int. Geosci. Remote Sens. Symp., IEEE. pp. 266–269.

[80] Sumbul, G., Demir, B., 2022. Plasticity-stability preserving multi-task learning for remote sensing image retrieval. IEEE Trans. Geosci. Remote Sens. 60, 1–16.

[81] Sumbul, G., Ravanbakhsh, M., Demir, B., 2021b. Informative and representative triplet selection for multilabel remote sensing image retrieval. IEEE Trans. Geosci. Remote Sens. 60, 1–11.

[82] Sun, Y., Feng, S., Ye, Y., Li, X., Kang, J., Huang, Z., Luo, C., 2021. Multisensor fusion and explicit semantic preserving-based deep hashing for cross-modal remote sensing image retrieval. IEEE Trans. Geosci. Remote Sens. 60, 1–14.

[83] Tang, X., Zhang, X., Liu, F., Jiao, L., 2018. Unsupervised deep feature learning for remote sensing image retrieval. Remote Sens. 10, 1243.

[84] Tewel, Y., Shalev, Y., Schwartz, I., Wolf, L., 2022. Zerocap: Zero-shot image-to-text generation for visual-semantic arithmetic, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., pp. 17918–17928.

[85] Tian, Y., Deng, X., Zhu, Y., Newsam, S., 2020. Cross-time and orientation-invariant overhead image geolocalization using deep local features, in: Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis., pp. 2512–2520.

[86] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al., 2023. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 .

[87] Vo, N., Jiang, L., Sun, C., Murphy, K., Li, L.J., Fei-Fei, L., Hays, J., 2019. Composing text and image for image retrieval-an empirical odyssey, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.

[88] Wang, M., Song, T., 2012. Remote sensing image retrieval by scene semantic matching. IEEE Trans. Geosci. Remote Sens. 51, 2874–2886.

[89] Wang, M., Wan, Q., Gu, L., Song, T., 2013. Remote-sensing image retrieval by combining image visual and semantic features. Int. J. Remote Sens. 34, 4200–4223.

[90] Wang, S., Hou, D., Xing, H., 2022. A novel multi-attention fusion network with dilated convolution and label smoothing for remote sensing image retrieval. Int. J. Remote Sens. 43, 1306–1322.

[91] Wang, Y., Ji, S., Lu, M., Zhang, Y., 2020. Attention boosted bilinear pooling for remote sensing image retrieval. Int. J. Remote Sens. 41, 2704–2724.

[92] Wang, Y., Ji, S., Zhang, Y., 2021. A learnable joint spatial and spectral transformation for high resolution remote sensing image retrieval. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 14, 8100–8112.

[93] Wang, Y., Zhang, L., Tong, X., Zhang, L., Zhang, Z., Liu, H., Xing, X., Mathiopoulos, P.T., 2016. A three-layered graph-based learning approach for remote sensing image retrieval. IEEE Trans. Geosci. Remote Sens. 54, 6020–6034.

[94] Wang, Z., Prabha, R., Huang, T., Wu, J., Rajagopal, R., 2024. Skyscript: A large and semantically diverse vision-language dataset for remote sensing, in: Proc. AAAI Conf. Artif. Intell., pp. 5805–5813.

[95] Wu, H., Gao, Y., Guo, X., Al-Halah, Z., Rennie, S., Grauman, K., Feris, R., 2021. Fashion iq: A new dataset towards retrieving images by natural language feedback, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.

[96] Xiong, W., Lv, Y., Cui, Y., Zhang, X., Gu, X., 2019. A discriminative feature learning approach for remote sensing image retrieval. Remote Sens. 11, 281.

[97] Xiong, W., Lv, Y., Zhang, X., Cui, Y., 2020a. Learning to translate for cross-source remote sensing image retrieval. IEEE Trans. Geosci. Remote Sens. 58, 4860–4874.

[98] Xiong, W., Xiong, Z., Cui, Y., Lv, Y., 2020b. A discriminative distillation network for cross-source remote sensing image retrieval. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 13, 1234–1247.

[99] Xu, F., Yang, W., Jiang, T., Lin, S., Luo, H., Xia, G.S., 2020. Mental retrieval of remote sensing images via adversarial sketch-image feature learning. IEEE Trans. Geosci. Remote Sens. 58, 7801–7814.

[100] Ye, F., Xiao, H., Zhao, X., Dong, M., Luo, W., Min, W., 2018. Remote sensing image retrieval using convolutional neural network features and weighted distance. IEEE Geosci. Remote Sens. Lett. 15, 1535–1539.

[101] Yin, M., Yao, Z., Cao, Y., Li, X., Zhang, Z., Lin, S., Hu, H., 2020. Disentangled non-local neural networks, in: Proc. Eur. Conf. Comput. Vis.

[102] Yuan, Z., Zhang, W., Fu, K., Li, X., Deng, C., Wang, H., Sun, X., 2021. Exploring a fine-grained multiscale method for cross-modal remote sensing image retrieval. IEEE Trans. Geosci. Remote Sens. 60, 1–19.

[103] Yuan, Z., Zhang, W., Tian, C., Rong, X., Zhang, Z., Wang, H., Fu, K., Sun, X., 2022. Remote sensing cross-modal text-image retrieval based on global and local information. IEEE Trans. Geosci. Remote Sens. 60, 1–16.

[104] Zeng, Z., Wang, Z., Yang, F., Satoh, S., 2022. Geo-localization via ground-to-satellite cross-view image retrieval. IEEE Trans. Multimedia .

[105] Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L., 2023. Sigmoid loss for language image pre-training, in: Proc. IEEE/CVF Int. Conf. Comput. Vis., pp. 11975–11986.

[106] Zhang, K., Luan, Y., Hu, H., Lee, K., Qiao, S., Chen, W., Su, Y., Chang, M.W., 2024. Magiclens: Self-supervised image retrieval with open-ended instructions, in: Proc. Int. Conf. Mach. Learn., pp. 59403–59420.

[107] Zhang, M., Cheng, Q., Luo, F., Ye, L., 2021. A triplet nonlocal neural network with dual-anchor triplet loss for high-resolution remote sensing image retrieval. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 14, 2711–2723.

[108] Zhang, Q., Lei, Z., Zhang, Z., Li, S.Z., 2020. Context-aware attention network for image-text retrieval, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., pp. 3536–3545.

[109] Zhao, H., Yuan, L., Zhao, H., Wang, Z., 2021. Global-aware ranking deep metric learning for remote sensing image retrieval. IEEE Geosci. Remote Sens. Lett. 19, 1–5.

[110] Zhou, W., Guan, H., Li, Z., Shao, Z., Delavar, M.R., 2023. Remote sensing image retrieval in the past decade: Achievements, challenges, and future directions. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. .

[111] Zhou, W., Newsam, S., Li, C., Shao, Z., 2017. Learning low dimensional convolutional neural networks for high-resolution remote sensing image retrieval. Remote Sens. 9, 489.

[112] Zhou, W., Newsam, S., Li, C., Shao, Z., 2018. Patternnet: A benchmark dataset for performance evaluation of remote sensing image retrieval. ISPRS J. Photogramm. Remote Sens. 145, 197–209.

[113] Zhuo, Z., Zhou, Z., 2021. Remote sensing image retrieval with gabor-ca-resnet and split-based deep feature transform network. Remote Sens. 13, 869.

Appendix

A. Additional Experiments

A.1. Dataset statistics

Table 4 reports the detailed statistics of PatternCom, including the number of composed queries and positives for each attribute type, class, and target value. These statistics highlight the variability in the number of positives across attribute values, motivating the use of attribute-balanced macro-averaged mAP in the main evaluation.

Table 4
PatternCom statistics. Breakdown by attribute type, class, and attribute value. #Queries denotes the number of composed queries where the query text specifies the corresponding target value, and #Positives denotes the number of database images that satisfy the relevance criterion (*same class* and *target attribute value*).

Attribute	Class	Value	#Queries	#Positives
color	airplane	white	53	672
		purple	672	53
	nursing home	white	383	85
		gray	85	383
	crosswalk	white	388	412
		yellow	412	388
	tennis court	blue	287	339
		brown	624	2
		gray	576	50
		green	415	211
		red	602	24
context	bridge	concrete	800	800
		water	800	800
density	residential	sparse	800	800
		dense	800	800
existence	parking	with cars	653	947
		without cars	947	653
	pier	with boats	255	1345
		without boats	1345	255
quantity	storage tank	one	261	356
		two	498	119
		three	552	65
		four	540	77
	wast. tr. plant	one	78	724
		two	758	44
		three	792	10
		four	778	24
	basketball court	one	383	340
		two	437	286
		three	702	21
		half	662	61
		two-halfs	708	15
shape	swimming pool	rectangular	299	261
		oval	508	52
		kidney-shaped	313	247
	river	curved	623	177
		straight	177	623
	road	cross	800	800
		round	800	800

Table 5 reports the statistics of xView2-CIR. This dataset should be viewed as a first evaluation benchmark for change-centric RSCIR rather than a complete operational disaster-monitoring dataset. Some disaster categories are small, which motivates reporting both macro-averaged and overall metrics and cautions against over-interpreting fine-grained differences on rare categories.

Table 5
xView2-CIR statistics. Number of composed queries per disaster type and associated textual modifier. #Queries denotes the number of pre-event composed queries. Each query has exactly one positive match (the post-event image of the *same location* under the *target disaster*), hence #Positives = 1 per query.

Disaster	Text query	#Queries	#Positives
hurricane	post-hurricane	147	1
wildfire	post-fire	98	1
flood	post-flood	28	1
tsunami	post-tsunami	9	1
earthquake	post-earthquake	5	1
volcano	post-volcano	4	1

A.2. Vision–language backbones

We evaluate six vision–language models, all using a ViT-L/14 visual backbone:

- **CLIP LAION-2B**: A CLIP model trained on LAION-2B [73], a dataset of 2.3 billion web-collected image–text pairs. We use the publicly released `laion2b_s32b_b82k` checkpoint from OpenCLIP [16].
- **RemoteCLIP** [51]: A remote-sensing-adapted CLIP model initialized from OpenAI CLIP [67] and fine-tuned on image–text pairs derived from annotated remote sensing datasets using synthetic captions.
- **OpenAI CLIP** [67]: The original CLIP model released by OpenAI, trained on 400M web image–text pairs.
- **SigLIP** [105]: A vision–language model trained with a sigmoid-based contrastive loss instead of the standard softmax contrastive loss. We use the SigLIP model trained on WebLI.
- **CLIP LAION-RS** [94]: A remote-sensing-adapted CLIP model initialized from OpenAI CLIP and fine-tuned on LAION-RS, a 726K-image remote sensing subset of LAION-2B.
- **SkyCLIP-50** [94]: A remote-sensing-adapted CLIP model initialized from OpenAI CLIP and fine-tuned on SkyScript-50. SkyScript-50 contains 2.6M image–text pairs constructed by linking geo-referenced satellite imagery from Google Earth Engine [24] with OpenStreetMap [27] annotations.

A.3. Vocabulary construction

Some evaluated methods rely on a vocabulary or textual memory. To better align these methods with EO semantics, we generate a family of remote-sensing-specific vocabularies, denoted as RSText, using the prompt shown below. We use vocabularies of different sizes in the analysis and construct HybridText-23k by merging RSText-2k with Open Images v7.

Vocabulary generation prompt

I need a fine-grained, diverse vocabulary list suitable for detailed land-use/land-cover (LULC), remote sensing, and object-detection datasets. The vocabulary should be explicitly descriptive, detailed, and balanced across multiple thematic categories, including but not limited to:

- Urban infrastructure
- Cultural and historical sites
- Recreational and tourism areas
- Transportation and logistics
- Construction and housing
- Education and healthcare
- Natural features and ecosystems
- Biodiversity and wildlife habitats
- Sustainability initiatives and eco-friendly technologies
- Agriculture, farming, and traditional land-use practices
- Marine and aquatic environments

Each vocabulary entry should ideally be short (1–4 words), clear, explicit, self-contained, and suitable for remote sensing imagery annotation. Provide vocabulary entries systematically in batches of 100, carefully ensuring thematic diversity and granularity until a total of at least 2000 entries is reached.

A.4. Additional details for adapted methods

This section provides additional implementation details for the adapted composed image retrieval methods evaluated in the main paper. Unless otherwise stated, we follow the official implementations or default settings from the original papers and introduce only the minimal adaptations needed for EO imagery.

CompoDiff is a compositional diffusion framework [69] that samples an image-conditioned text embedding via a denoising diffusion process in the joint CLIP embedding space. Sampling is conditioned on (i) the query image embedding, (ii) the composed text prompt (*e.g.*, `red {*}`), and (iii) a negative prompt (*e.g.*, `low quality`). We evaluate CompoDiff in its native configuration and, where applicable, under the hybrid protocol described above. To better preserve identity/class information, we blend the sampled embedding with the original image embedding using a source weight parameter. We report results with 10 diffusion steps and source weight 0.4.

Pic2Word learns a lightweight MLP that maps image embeddings to a textual embedding suitable for composition, enabling image-to-text inversion with a single forward pass. For retrieval, we use the template `{attribute_value} {*}`, where `*` denotes the inverted image token. We evaluate Pic2Word in its standard setting and, when relevant, under the hybrid protocol described above.

SEARLE performs test-time textual inversion by optimizing a pseudo token such that its text-encoder embedding matches the query image embedding, regularized by an LLM-guided loss. To ground SEARLE in RS, we replace the Open Images v7 [40] vocabulary with our domain-specific HybridText-23k vocabulary and use the RS-adapted prompt template `a satellite image of {concept} that`. For retrieval, we use the composed text template `{attribute_value} {$}`, where the attribute value (*e.g.*, `red`) modifies the learned placeholder token `$`. We report results with 200 optimization steps, learning rate 0.002, and GPT-loss weight λ_{GPT}=0.25.

CIReVL is a caption-guided composition pipeline that converts the query image into text and then edits it to reflect the modifier. It consists of: (i) captioning the query image, (ii) editing the caption with an instruction-tuned LLM given the textual modifier, and (iii) encoding the edited description with the retrieval backbone for text-to-image search. We use BLIP-2 [43] (blip2-opt-2.7b) as the captioner and LLaMA-2 7B [86] (Llama-2-7b-chat-hf) as the editor. To ground CIReVL to PatternCom, we design a domain-specific modifier prompt covering the six attribute types to enforce faithful attribute changes while preserving scene semantics.

CIReVL modifier prompt for PatternCom

You are given a high-resolution satellite image caption describing the "Image Content", along with a single-word "Instruction" that specifies a modification to apply to the scene. Generate a complete, natural-language "Edited Description" that integrates the modification while preserving all other aspects of the original content. The instruction will always belong to one of six attribute types: color, context, density, existence, quantity, shape. Example (shape):

- Image Content: a satellite image of a rectangular swimming pool in a resort.
- Instruction: `oval`
- Edited Description: a satellite image of an *oval* swimming pool in a resort.

MagicLens augments an OpenAI CLIP backbone with a transformer based compositional head trained for image–text composition under supervised triplets. We use the official pretrained model (ViT-L/14) and apply it directly to both queries and database images, without additional adaptation.

Retrieval is performed using the joint embeddings produced by the MagicLens head.

WeiCom is a training-free score-fusion approach that combines image-to-image and text-to-image similarities after calibrating them to a comparable scale. We report results with λ=0.5, corresponding to equal modality contribution.

BASIC is a training-free composition method that combines image-to-image and text-to-image similarities after centering and semantic projection, optionally applying image-side query expansion, and fusing modalities with a Harris-regularized multiplicative score. To ground BASIC in RS, we augment C_+ with RSText-150. We report results using 250 principal components for the semantic projection, contrastive scaling α=0.2, query expansion with 25 nearest neighbors, and Harris regularization weight λ_{Harris}=0.1.

FreeDom is a memory-based textual inversion method that constructs an interpretable text representation of the query image using a large vocabulary and proxy-image expansion. Given a query image and text, it first retrieves k proxy images via image similarity, then retrieves the top-n vocabulary labels per proxy, and retains the most frequent m labels overall. Each retained label w is composed with the modifier using `{attribute_value} {w}`, and the resulting text embeddings are fused by frequency-weighted averaging to form the final query representation. We ground the textual memory to remote sensing using HybridText-23k. We report results using k=20, n=7, and m=7.

The evaluated methods span a broad deployment spectrum: some are fully training-free and lightweight at inference (*e.g.*, WeiCom, BASIC), some require large auxiliary vocabularies or proxy retrieval (*e.g.*, FreeDom), some rely on captioning and LLM editing (*e.g.*, CIReVL), and others incur heavier per-query optimization or sampling cost (*e.g.*, SEARLE, CompoDiff). This diversity is important when assessing their suitability for operational EO archives.

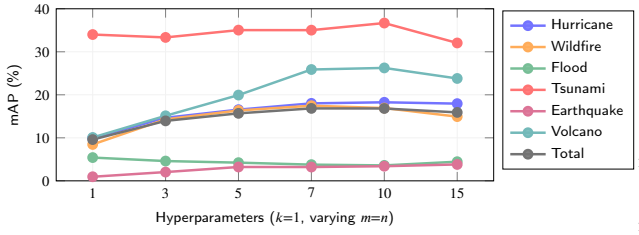

Figure 11: Impact of frequency-based textual expansion on FreeDom with OpenAI CLIP on xView2-CIR with k=1 and increasing m=n. Aggregating multiple textual labels improves mAP across most disaster types, especially hurricane and wildfire. Performance peaks at m=n=7, after which further expansion yields diminishing returns.

A.5. Additional sensitivity and ablation analyses

Textual expansion on xView2-CIR. Figure 11 complements the main-paper analysis of FreeDom textual expansion by reporting results on xView2-CIR. Visual expansion is disabled (k=1), while the number of aggregated textual concepts is varied as m=n. Moderate textual expansion improves performance, with the best Total score obtained at m=n=7. Larger expansions provide diminishing returns, suggesting that overly long textual surrogates can introduce noisy or off-target concepts.

CompoDiff hyperparameter sensitivity. Figure 12 analyzes CompoDiff sensitivity to the number of diffusion timesteps t and the source weight used to mix the reference image with the generated proxy. Performance improves when moving from very small source weights to the mid-range (0.3–0.5), indicating that both the original image and the generative signal are useful for composition. Moderate diffusion budgets (t=10–20) are generally sufficient, while larger budgets provide limited additional gains.

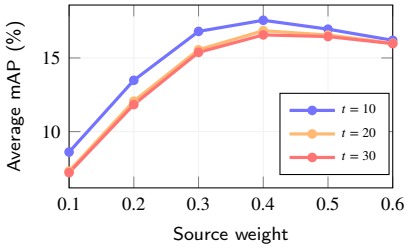

Figure 12: CompoDiff hyperparameter sweep with SkyCLIP-50 on PatternCom. We report average mAP as a function of source weight for different numbers of diffusion timesteps.

SEARLE hyperparameter sensitivity. Figure 13 evaluates SEARLE under different learning rates, inversion steps, and GPT mixing coefficients λ_{GPT}. Performance is relatively stable around λ_{GPT}=0.25 and moderate learning rates, while very small learning rates degrade results. Larger GPT mixing does not provide consistent gains. Overall, SEARLE is moderately sensitive to optimization settings but remains stable around prior-work defaults.

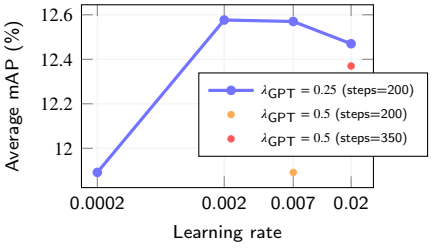

Figure 13: SEARLE hyperparameter sweep with CLIP LAION-2B on PatternCom. We report average mAP as a function of learning rate, number of inversion steps, and GPT mixing coefficient λ_{GPT}.

Ablation of BASIC components Table 6 and 7 analyze the contribution of BASIC components on PatternCom and xView2-CIR. Across both datasets, removing *centering* or *semantic projection* causes the largest drops, showing that modality calibration and semantic subspace alignment are the main drivers of BASIC's performance. Other components, including Harris weighting, min-based normalization, text contextualization, and query expansion, provide smaller

Table 6
BASIC component ablation on PatternCom with OpenAI CLIP. We report attribute-wise and average mAP (%) when removing components from the full pipeline: mean centering (**Cen.**), min-based normalization (**Norm.**), Harris weighting (**Har.**), text contextualization (**Con.**), semantic projection (**Proj.**), and query expansion (**Q. Ex.**).

Cen.	Norm.	Har.	Con.	Proj.	Q. Ex.	Color	Context	Density	Existence	Quantity	Shape	Avg.
✓	✓	✓	✓	✓	✓	38.31	24.53	26.35	30.94	18.03	35.73	**28.98**
✓	✓	✗	✓	✓	✓	36.69	22.87	24.34	28.40	16.14	32.74	26.86
✓	✗	✗	✓	✓	✓	26.57	30.51	24.63	11.93	26.08	23.27	23.83
✓	✓	✓	✗	✓	✓	34.60	24.56	19.17	12.81	26.44	17.03	22.44
✓	✓	✓	✓	✗	✓	23.22	24.41	18.97	3.11	24.40	20.19	19.05
✗	✓	✓	✓	✓	✓	28.90	18.95	22.77	8.02	12.57	16.33	17.92
✓	✓	✓	✓	✓	✗	37.16	24.51	24.93	28.62	16.57	34.22	27.67

Table 7
BASIC component ablation on xView2-CIR with OpenAI CLIP. We report disaster-wise mAP (%), macro-average mAP (Avg.), and overall mAP (Total) when removing components from the full pipeline. Component abbreviations follow Table 6.

Cen.	Norm.	Har.	Con.	Proj.	Q. Ex.	Hurricane	Fire	Flood	Tsunami	Earthquake	Volcano	Avg.	Total
✓	✓	✓	✓	✓	✓	9.36	7.74	7.54	19.52	1.18	13.96	9.88	8.88
✓	✓	✓	✓	✓	✗	22.71	18.49	14.59	46.15	1.69	59.20	**27.14**	**21.38**
✓	✓	✗	✓	✓	✗	21.16	17.45	14.60	47.63	1.79	51.04	25.61	20.18
✓	✗	✗	✓	✓	✗	16.39	17.08	11.11	29.13	1.14	64.17	23.17	16.90
✓	✓	✓	✗	✓	✗	18.48	16.69	14.38	32.37	2.62	57.25	23.63	18.17
✓	✓	✓	✓	✗	✗	16.03	14.74	13.68	24.25	1.26	59.01	21.49	15.96
✗	✓	✓	✓	✓	✗	0.68	0.71	1.24	17.02	1.39	6.29	4.56	1.34

and more dataset-dependent gains. Notably, query expansion is harmful on xView2-CIR, where relevance requires the same scene/location and visually similar proxies often introduce off-location evidence. Overall, BASIC's gains primarily come from representation calibration and semantic projection, while auxiliary heuristics play a secondary role.

A.6. Sensitivity to text query phrasing

We study how composed retrieval methods respond to alternative textual formulations of the modifier. For PatternCom, we test three templates applied to each attribute value v: R1 uses a *being* formulation (*e.g.*, `being blue`, `being rectangular-shaped`); R2 uses a *having* formulation (*e.g.*, `having blue color`, `having rectangular shape`); and R3 uses longer, relational descriptions such as `with the main object modified to have blue color` or `with the main object modified to have rectangular sha pe`. As shown in Table 8, the original plain attribute word (*e.g.*, `blue`, `water`, `dense`) is generally the most reliable choice, especially for stronger multimodal methods. More verbose templates tend to reduce performance, suggesting that additional function words can dilute the compact attribute signal.

For xView2-CIR, the plain query is `post-{disaster}`, while R1 uses impact-oriented rephrasing (*e.g.*, `burned area`, `flooded area`) and R2 uses disaster-explicit formulations (*e.g.*, `fire-affected region`, `seismic damag e`). As shown in Table 9, prompt sensitivity is more method-dependent in this change-centric setting. More descriptive formulations can benefit methods that rely more strongly on semantic text structure, particularly BASIC, while other methods are more sensitive to distribution shifts introduced by rewording. Overall, these results indicate that prompt design is dataset- and method-dependent: compact modifiers are effective for controlled attribute edits, whereas change-centric retrieval may benefit from more explicit event or impact descriptions.

Table 8
Text query rephrasing on PatternCom with CLIP LAION-2B. We report average mAP (%) for the original modifier (Plain) and three rephrased variants (R1–R3).

Method	Average			
	Plain	R1	R2	R3
Text-only	5.57	**9.16**	7.07	3.52
Text+Image	16.41	**17.70**	17.48	15.10
Text×Image	21.00	19.03	**21.58**	16.76
WeiCom	21.75	20.86	**23.31**	16.38
BASIC	16.47	**17.50**	17.73	17.34
FreeDom	**38.59**	33.06	31.48	29.53

Table 9
Text query rephrasing on xView2-CIR with OpenAI CLIP. We report macro-average mAP (Average) and overall mAP (Total) for the original query (Plain) and two rephrased variants (R1–R2).

Method	Average			Total		
	Plain	R1	R2	Plain	R1	R2
Text-only	**7.57**	4.53	4.30	**2.97**	1.76	2.06
Text+Image	**17.73**	15.27	15.63	**12.14**	9.14	10.22
Text×Image	**25.94**	14.97	12.13	**16.35**	10.38	9.65
WeiCom	**30.67**	24.39	19.22	**21.40**	12.58	13.13
BASIC	**27.14**	17.39	17.36	**21.38**	18.56	18.17
FreeDom	**17.23**	9.86	11.98	**16.84**	10.83	12.19