FIVE: FREQUENCY-AWARE AND INTENT-GUIDED VISUAL EVIDENCE LEARNING FOR REMOTE SENSING COMPOSED IMAGE RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Remote Sensing Composed Image Retrieval (RSCIR) is an emerging retrieval paradigm that retrieves target images from complex scenes with multi-scale objects guided by diverse modification intents. However, existing methods still suffer from limitations in both visual representation granularity and modification intent modeling. On the one hand, small objects are prevalent in remote sensing images, yet their visual cues are often overwhelmed by large background regions in the resulting image representations, limiting fine-grained retrieval. On the other hand, most methods inject the query text into the retrieval representation as a unified condition, lacking explicit modeling of the visual evidence selection and semantic transfer mechanisms required by different editing operations. To address these issues, we propose **F**requency-aware and **I**ntent-guided **V**isual **E**vidence Learning (**FIVE**), a novel framework for RSCIR that improves the discriminability of retrieval representations from two perspectives: object-level visual representation enhancement and intent-guided semantic selection. Specifically, we design a Frequency-aware Object Enhancement module that introduces the Fast Fourier Transform during visual feature modeling to explicitly capture frequency-domain cues, thereby enhancing the representation of contours, textures, and local structures of small objects. Furthermore, we develop an Intent-guided Visual Evidence Selection module that leverages modification intents to guide the selection and reorganization of object regions, attribute cues, and spatial relations in the reference image, generating intent-aware query representations. Experimental results on two RSCIR datasets demonstrate that our method outperforms existing approaches, and ablation studies further verify the effectiveness of frequency-domain representation enhancement and modification intent modeling.

1 INTRODUCTION

With the rapid development of remote sensing observation technologies and the explosive growth of large-scale remote sensing imagery Wang et al. (2025), how to efficiently and accurately retrieve target images of user interest from massive remote sensing archives has become an important research problem in multimodal information retrieval. Traditional remote sensing image retrieval methods usually rely on single-modal queries, such as image-to-image retrieval or text-to-image retrieval Choudhury et al. (2025); Huang et al. (2026); Zheng et al. (2025); Chen et al. (2025a), which are insufficient to express complex retrieval needs involving local modifications to a reference image. To address this limitation, Remote Sensing Composed Image Retrieval (RSCIR) has recently attracted increasing attention.

RSCIR follows the basic formulation of Composed Image Retrieval (CIR). Given a reference image and a query text as a joint query, the task aims to retrieve a target image that is semantically related to the reference image and satisfies the specified modification. Early CIR methods typically encoded the image and text separately and constructed a joint query representation through modality fusion, gated composition, or attention-based mechanisms Vo et al. (2019); Chen & Bazzani (2020); Wen et al. (2021); Dodds et al. (2020). With the emergence of vision-language pretrained models, subsequent studies further exploited the cross-modal alignment capabilities of models such as CLIP Radford et al. (2021) and BLIP-2 Li et al. (2023), improving composed retrieval through feature com-

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

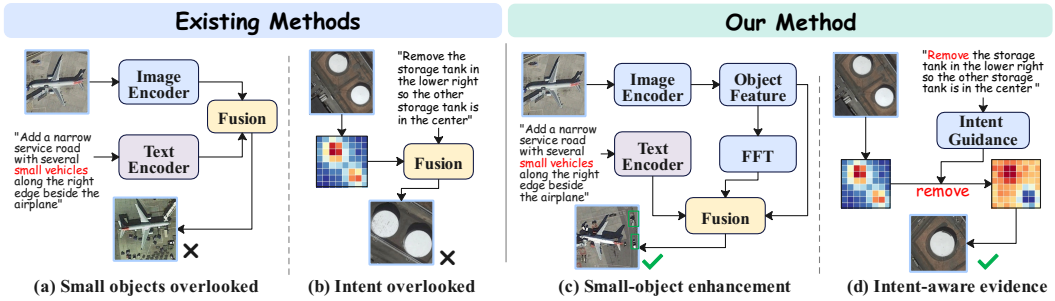


Figure 1: Limitations of existing RSCIR methods and our solutions. (a) Small objects are overlooked. (b) Modification intents are insufficiently modeled. (c) Our method enhances small-object features using frequency cues. (d) Our method reorganizes visual evidence according to modification intents.

position, prompt learning, pseudo-word mapping, and candidate re-ranking Bai et al. (2024); Chen et al. (2024); Baldrati et al. (2022); Goenka et al. (2022). These studies have established an important foundation for composed-query modeling in RSCIR. WEICOM Psomas et al. (2024) was the first to introduce CIR into the remote sensing domain, performing composed retrieval by weighted fusion of image- and text-based retrieval results. However, its queries are primarily composed of discrete attribute words, limiting its ability to support fine-grained modifications expressed in natural language. Subsequently, TisCIR Liang et al. (2025) addressed cross-modal conflicts by identifying and filtering inconsistent information between the reference image and the query text, while jointly exploiting effective representations from both modalities to improve the semantic alignment between composed queries and target images in remote sensing scenarios. Despite recent progress, existing RSCIR methods still face two key challenges: insufficient fine-grained visual representation and limited modeling of modification intents. On the one hand, as illustrated in Figure 1(a), remote sensing images are characterized by complex backgrounds and substantial scale variations, causing fine-grained cues of small objects, such as contours, textures, and local structures, to be easily overwhelmed by dominant background semantics during global feature aggregation. Although a few studies have attempted to emphasize target regions through spatial-domain strategies, such as region selection and segmentation guidance, spatial filtering alone remains insufficient to fully capture the discriminative characteristics of small objects. On the other hand, as shown in Figure 1(b), existing methods typically encode the modification text as a whole without explicitly parsing its underlying intent, thereby limiting their ability to select and reorganize visual evidence from the reference image in accordance with different modification intents.

To address the above issues, we propose **F**requency-aware and **I**ntent-guided **V**isual **E**vidence Learning (**FIVE**), a unified RSCIR framework that enhances retrieval representations from two complementary perspectives: fine-grained object modeling and intent-aware evidence reasoning. Specifically, the Frequency-aware Object Enhancement module introduces the Fast Fourier Transform to capture contour, texture, and local structural cues, thereby strengthening the representation of small objects, as shown in Figure 1(c). Meanwhile, the Intent-guided Visual Evidence Selection module exploits modification intents to selectively reorganize object regions, attribute cues, and spatial relations in the reference image, yielding an intent-aware query representation, as shown in Figure 1(d). By explicitly determining what should be preserved, what should be modified, and how the modification should be performed, FIVE more effectively models the semantic transition from the reference image to the target image.

To facilitate a comprehensive evaluation of FIVE, we construct PatternCIRRS, a reference-dependent RSCIR dataset with query texts that explicitly describe the semantic transformations from reference images to target images. We then conduct extensive experiments on both PatternCIR and PatternCIRRS. The contributions are summarized as follows:

- We propose FIVE, a unified RSCIR framework that jointly enhances fine-grained object representations and models intent-aware visual evidence transformations.
- We develop Frequency-aware Object Enhancement, which integrates spatial localization with frequency-domain modeling to strengthen the structural cues of small objects.

- We introduce Intent-guided Visual Evidence Selection, which uses parsed modification intents and selects and reorganizes visual evidence according to the corresponding editing operations.
- We construct PatternCIRRS, an RSCIR dataset whose queries explicitly describe modifications from reference images to target images.

2 Related Work

Composed Image Retrieval. Early CIR methods mainly follow a feature composition paradigm, extracting visual and textual representations separately and combining them through gated fusion, attention mechanisms, or cross-modal interaction modules to obtain a joint query representation Vo et al. (2019); Chen & Bazzani (2020); Dodds et al. (2020); Kim et al. (2021). However, these methods usually treat the modification text as a global condition and lack explicit modeling of object-level changes and fine-grained visual transformations. Recently, vision-language pretrained models have been introduced into CIR due to their strong cross-modal representation ability. Methods based on CLIP exploit large-scale image-text alignment for composed retrieval by fine-tuning vision-language features or introducing lightweight composition modules Baldrati et al. (2022); Goenka et al. (2022). Other works further explore prompt-based representations to bridge image features and textual instructions Bai et al. (2024); Li et al. (2025a). Despite their promising performance, existing methods struggle to generalize effectively to remote sensing scenarios due to the substantial domain gap between natural and remote sensing images.

Remote Sensing Composed Image Retrieval. Remote Sensing Composed Image Retrieval (RSCIR) extends the CIR task to remote sensing scenarios, aiming to retrieve target images according to a reference image and a semantic modification description. Early research in this direction mainly focused on constructing benchmarks for evaluating composed retrieval in remote sensing images. WEICOM first introduced composed image retrieval into the remote sensing domain and built an attribute-based benchmark by considering factors such as color, context, density, existence, quantity, and shape Psomas et al. (2024). However, its attribute-level query formulation lacks natural-language modifications and complete reference-target triplet annotations, limiting its application to existing CIR frameworks. To further advance RSCIR research, TisCIR proposed the PatternCIR dataset with paired images and textual queries Liang et al. (2025). Nevertheless, the generated queries mainly describe target image attributes and categories rather than the semantic transformation from the reference image to the target image, which may weaken the modeling of modification intents. Moreover, remote sensing images contain complex backgrounds and multi-scale objects, where small-scale targets and local structures are easily overwhelmed by global representations. These characteristics pose additional challenges for composed retrieval in remote sensing scenarios.

3 Method

3.1 Overview

Given a reference image I_r and a query text m, FIVE retrieves a target image I_t as illustrated in Figure 2. Image and text encoders produce F_r, F_t, and F_m, while an instruction parser converts m into a semantic graph $\mathcal{G}$. Frequency-aware Object Enhancement (FOE) extracts spatial and frequency features, and Intent-guided Visual Evidence Selection (IVES) models intent-specific evidence changes. These features enrich the Q-Former-composed query through cross-attention and gated fusion, yielding $\bar{f}_Q$ for matching against the pooled target representation $\bar{f}_t$.

3.2 Multimodal Input Representation

Given a training triplet (I_r, m, I_t) consisting of a reference image, query text, and target image, we extract their basic multimodal representations, denoted as $\mathbf{F}_r$, $\mathbf{F}_m$, and $\mathbf{F}_t$, respectively. The holistic text representation $\mathbf{F}_m$ preserves the overall context of the modification instruction, but does not explicitly separate the involved object, modification intent, and spatial or attribute constraints. We therefore adopt a LoRA-finetuned Hu et al. (2022) Qwen3-4B Yang et al. (2025) model as an instruction parser to transform the modification text into an object–intent–constraint semantic graph

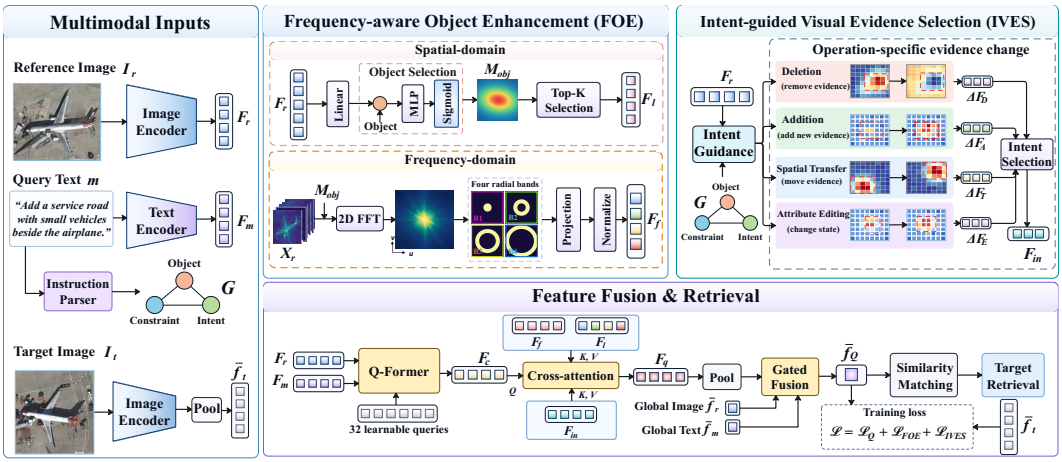


Figure 2: Overview of the proposed FIVE framework. FOE enhances small-object representations using local and frequency-domain features, while IVES reconstructs visual evidence according to modification intents. The Feature Fusion Module integrates these features to generate the final composed query representation.

$\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{o_i, a_i, l_i\}_{i=1}^{M}$ contains the object, intent, and location or attribute constraint of each semantic unit, and $\mathcal{E}$ encodes their object–intent and intent–constraint relations. Here, M denotes the number of semantic units. For each semantic unit, the object, intent, and constraint are further encoded and projected into semantic representations $\mathbf{h}_{o_i}$, $\mathbf{h}_{a_i}$, and $\mathbf{h}_{l_i}$, respectively.

3.3 Frequency-aware Object Enhancement

Remote sensing images often contain large-scale backgrounds, complex land-cover distributions, and objects with significant scale variations, where small-object cues can be overwhelmed by global semantics. To address this issue, we construct object-level visual evidence from spatial localization and frequency-domain enhancement.

Spatial-domain Features. Let the reference image features be denoted as $\mathbf{F}_r = \{\mathbf{f}_r^i\}_{i=1}^{N}$, where $\mathbf{f}_r^i \in \mathbb{R}^d$. The module first estimates the relevance between each visual patch and the object semantics parsed from the structured representation:

$$s_{i,j} = \phi_{\text{obj}}\left(\mathbf{f}_r^i, \mathbf{h}_{o_j}\right), \quad \mathbf{M}_{\text{obj}}^i = \sigma\left(\max_{j=1}^{M} s_{i,j}\right), \tag{1}$$

where $\phi_{obj}(\cdot)$ denotes the learnable patch-object relevance scorer and $\sigma(\cdot)$ is the sigmoid function. $\mathbf{M}_{\text{obj}}^i$ indicates the relevance between the i-th patch and the target object. We select the Top-K most relevant patch features with indices I_K. The spatial-domain feature representation is then constructed as:

$$\mathbf{F}_l = \{\text{L2Norm}(\mathbf{W}_l \mathbf{f}_r^i) \mid i \in I_K\} \in \mathbb{R}^{K \times d}. \tag{2}$$

where W_l denotes the projection matrix.

Frequency-domain Features. Contours, textures, and edges of small objects exhibit characteristic responses across different frequency bands, providing discriminative structural cues for distinguishing objects from complex backgrounds. Recent studies on tiny-object detection have demonstrated that frequency-domain modeling can strengthen the spectral signatures and fine-grained structural details of tiny objects while suppressing background interference Sun et al. (2025); Wen et al. (2025). Motivated by these findings, we restore the reference tokens and their object relevance weights to the original spatial patch layout, obtaining $\mathbf{X}_r \in \mathbb{R}^{H \times W \times d}$ and $\mathbf{M}_{\text{obj}} \in \mathbb{R}^{H \times W}$, respectively. The object-aware feature is constructed by spatially weighting each token, i.e., $\widetilde{\mathbf{X}}_r(u, v, :) = \mathbf{M}_{\text{obj}}(u, v)\mathbf{X}_r(u, v, :)$. We then compute its amplitude spectrum using the Fast

Fourier Transform (FFT), followed by spectrum centering and logarithmic compression:

$$\mathbf{A}_{\text{fft}} = \log\left(1 + \left|\text{Shift}\left(\text{FFT}(\widetilde{\mathbf{X}}_r)\right)\right|\right). \tag{3}$$

Here, $\text{Shift}(\cdot)$ moves the low-frequency component to the spectrum center, and $\mathbf{A}_{\text{fft}}$ denotes the amplitude spectrum of the object-aware feature. To convert the spectrum into token representations, we define a normalized radius $r(u, v) \in [0, 1]$ from the spectrum center and divide it into four concentric frequency bands using fixed radial boundaries $\boldsymbol{\rho} = [\rho_0, \rho_1, \rho_2, \rho_3, \rho_4]$. For each band $b \in \{1, \ldots, 4\}$, its frequency index set is defined as $\Omega_b = \{(u, v) \mid \rho_{b-1} \leq r(u, v) < \rho_b\}$. The corresponding frequency token is obtained by averaging the spectral responses within the band:

$$\mathbf{f}_f^b = \text{L2Norm}\left(\mathbf{W}_f \frac{1}{|\Omega_b|} \sum_{(u,v)\in\Omega_b} \mathbf{A}_{\text{fft}}(u, v, :)\right). \tag{4}$$

Here, W_f is the frequency-domain projection matrix. The resulting frequency tokens are stacked to form the frequency-domain object representation F_f. To align F_l and F_f with retrieval semantics, we introduce auxiliary contrastive supervision:

$$\mathcal{L}_{\text{FOE}} = \lambda_l \ell_{\text{NCE}}(\bar{\mathbf{f}}_l, \bar{\mathbf{f}}_t; \tau_l) + \lambda_f \ell_{\text{NCE}}(\bar{\mathbf{f}}_f, \bar{\mathbf{f}}_t; \tau_f), \tag{5}$$

where $\ell_{\text{NCE}}(\cdot)$ denotes the batch-wise contrastive retrieval loss. $\bar{\mathbf{f}}_l$, $\bar{\mathbf{f}}_f$, and $\bar{\mathbf{f}}_t$ are the retrieval vectors of spatial evidence, frequency evidence, and the target image, respectively. λ_l and λ_f are balancing weights.

3.4 Intent-guided Visual Evidence Selection

To explicitly model how different modification intents reshape the reference image, IVES formulates deletion, addition, spatial transfer, and attribute editing as operation-specific changes of visual evidence. The parsed intent directly activates the corresponding branch. We reuse the object relevance map $\mathbf{M}_{\text{obj}}$ to localize the object evidence involved in the modification, and summarize its visual representation through masked pooling:

$$\mathbf{r}_{\text{obj}} = \mathbf{W}_r \frac{\sum_{i=1}^{N} M_{\text{obj}}^i \mathbf{f}_r^i}{\epsilon + \sum_{i=1}^{N} M_{\text{obj}}^i}. \tag{6}$$

For addition, since the requested object is absent from the reference image, a lightweight MLP projects its object semantics and associated constraint into the visual space, yielding $\mathbf{e}_{\text{add}} = \phi_{\text{add}}([\mathbf{h}_o; \mathbf{h}_l])$. For addition and spatial transfer, the target-location semantics are further used to estimate a destination mask $\mathbf{M}_{\text{dst}}$, where $M_{\text{dst}}^i = \sigma\big(\phi_{\text{loc}}(\mathbf{f}_r^i, \mathbf{h}_l^{\text{dst}})\big)$, and ϕ_{loc} is a learnable location-relevance scorer.

Rather than reconstructing complete target features, each branch predicts the visual-evidence change induced by the corresponding operation. Deletion removes the object evidence from the reference representation:

$$\Delta\mathbf{F}_D = -\mathbf{M}_{\text{obj}} \odot \mathbf{F}_r. \tag{7}$$

Addition injects the requested object evidence into the destination region. Specifically, $\mathbf{M}_{\text{dst}}$ determines where the new evidence should appear, while $\mathbf{e}_{\text{add}}$ specifies what visual content should be introduced:

$$\Delta\mathbf{F}_A = \alpha_{\text{inj}} \mathbf{M}_{\text{dst}} \odot \mathbf{e}_{\text{add}}, \tag{8}$$

where α_{inj} is a bounded learnable coefficient. Spatial transfer removes the object evidence from its original region and places its aggregated visual representation at the destination:

$$\Delta\mathbf{F}_T = -\mathbf{M}_{\text{obj}} \odot \mathbf{F}_r + \mathbf{M}_{\text{dst}} \odot (\mathbf{W}_T \mathbf{r}_{\text{obj}}). \tag{9}$$

Attribute editing preserves the object identity and location while modifying its visual state. A lightweight modulation MLP predicts channel-wise scaling and shifting parameters as $[\boldsymbol{\gamma}, \boldsymbol{\beta}] = \phi_{\text{att}}([\mathbf{h}_a; \mathbf{h}_l])$, and produces the edited object representation $\mathbf{r}_{\text{edit}} = \boldsymbol{\gamma} \odot \mathbf{r}_{\text{obj}} + \boldsymbol{\beta}$. The corresponding evidence change is:

$$\Delta\mathbf{F}_E = \mathbf{M}_{\text{obj}} \odot (\mathbf{W}_E \mathbf{r}_{\text{edit}} - \mathbf{F}_r). \tag{10}$$

According to the parsed intent $c \in \{D, A, T, E\}$, the corresponding change feature is projected into a common feature space: $\mathbf{F}_{\text{in}} = \mathcal{P}_c(\Delta \mathbf{F}_c)$. The masks are broadcast along the feature dimension when applied to visual features. To align the learned change representation with the actual semantic transition from the reference image to the target image, we employ

$$\mathcal{L}_{\text{IVES}} = \lambda_{\text{in}} \left(1 - \cos\left(\bar{\mathbf{f}}_{\text{in}}, \bar{\mathbf{f}}_t - \bar{\mathbf{f}}_r\right)\right), \tag{11}$$

where $\bar{\mathbf{f}}_{\text{in}} = \text{Pool}(\mathbf{F}_{\text{in}})$ denotes the intent-aware transformation representation, $\bar{\mathbf{f}}_r$ is the retrieval representation of the reference image and λ_{in} is the loss weight.

3.5 Feature Fusion

We employ a Q-Former to compose the reference image features $\mathbf{F}_r$ and query text features $\mathbf{F}_m$ with 32 learnable queries, obtaining the initial composed representation $\mathbf{F}_c$. The frequency-domain, spatial-domain, and intent-aware representations are concatenated as $\mathbf{Z} = [\mathbf{F}_f; \mathbf{F}_l; \mathbf{F}_{\text{in}}]$. Taking $\mathbf{F}_c$ as the query and $\mathbf{Z}$ as the key and value, cross-attention incorporates the multi-granularity visual evidence and produces the enhanced representation $\mathbf{F}_q$. We further fuse it with the global image and text semantics to obtain the final query representation:

$$\bar{\mathbf{f}}_Q = \text{L2Norm}\left(\text{GateFusion}\left(\text{Pool}(\mathbf{F}_q), \bar{\mathbf{f}}_r, \bar{\mathbf{f}}_m\right)\right). \tag{12}$$

Here, $\bar{\mathbf{f}}_r$ and $\bar{\mathbf{f}}_m$ denote the global representations of the reference image and modification text, respectively. The final query representation is aligned with the target image representation $\bar{\mathbf{f}}_t$ using a bidirectional contrastive retrieval loss:

$$\mathcal{L}_Q = \frac{1}{2}\left[\ell_{\text{NCE}}\left(\bar{\mathbf{f}}_Q, \bar{\mathbf{f}}_t\right) + \ell_{\text{NCE}}\left(\bar{\mathbf{f}}_t, \bar{\mathbf{f}}_Q\right)\right]. \tag{13}$$

The overall objective is $\mathcal{L} = \mathcal{L}_Q + \mathcal{L}_{\text{FOE}} + \mathcal{L}_{\text{IVES}}$.

4 Experiments

In this section, we describe the construction of the PatternCIRRS dataset and present the corresponding experimental evaluation, including baseline performance comparisons, ablation studies, motivation experiments, and detailed analyses.

4.1 PatternCIRRS Dataset Construction

PatternCIR generates query texts using the fixed template *"A satellite image of, [attribute value], [class name], with, in the background of."*, where the attribute and category information are derived primarily from the target image. Consequently, these queries mainly describe the target image rather than its transformation from the reference image. To address this limitation, we construct PatternCIRRS.

Image subset construction. PatternCIRRS retains the official train/validation/test partitions of PatternCIR. Following CIRR, we independently construct six-image subsets within each split for each category, denoted as $S = \{I_1, \ldots, I_6\}$. We randomly select a seed image I_1 and extract ImageNet-pretrained ResNet152 He et al. (2016) features. Candidates with cosine similarity $k_i \geq 0.85$ to I_1 are removed as near duplicates. The remaining images are selected by similarity ranking, while candidates whose similarity differs by less than 0.005 from previously selected images are discarded to preserve subset diversity.

Triplet construction and verification. Based on the constructed subsets, we generate CIR triplets (I_r, m, I_t). Remote sensing images often contain small objects, dense categories, and subtle visual variations, making it difficult for a VLM alone to infer the transformation from I_r to I_t. To address this challenge, we propose an object-aware query generation strategy. Specifically, LAE-DINO Pan et al. (2025) is applied to both images for open-vocabulary object detection, and the detected objects are converted into structured representations of categories, quantities, and spatial locations. These representations, together with the image pair and candidate images, are fed into GPT-5.4 to generate the modification query:

$$m = \Phi(I_r, I_t, \mathcal{O}_r, \mathcal{O}_t, \mathcal{C}), \tag{14}$$

Table 1: Comparison of performance on the PatternCIRRS test set (%). Avg. denotes the mean of R@5 and R_{subset}@1. The best results are shown in bold and the second-best results underlined.

Method	R@K				R_{subset}@K			Avg.
	$K=1$	$K=5$	$K=10$	$K=50$	$K=1$	$K=2$	$K=3$	
C-Former (TMM'23)	22.08	53.11	68.76	95.27	42.47	66.45	82.07	47.79
CIR-TFMUR (ICLR'24)	20.19	48.50	63.09	92.30	53.03	74.16	86.98	50.77
SPRC (ICLR'24)	<u>27.96</u>	<u>58.96</u>	<u>72.41</u>	<u>96.02</u>	<u>65.46</u>	<u>84.36</u>	<u>93.08</u>	<u>62.21</u>
ENCODER (AAAI'25)	18.85	49.98	65.93	95.70	32.12	55.07	72.68	41.05
OFFSET* (ACM MM'25)	20.15	49.85	65.64	94.40	37.82	61.02	77.91	43.84
TME (CVPR'25)	27.43	57.71	71.09	95.30	64.81	83.09	92.32	61.26
TisCIR (IJCAI'25)	5.16	15.96	25.41	64.63	26.37	47.89	67.29	21.17
INTENT (AAAI'26)	17.49	40.50	53.74	86.38	43.32	64.54	79.23	41.91
HUG (AAAI'26)	8.35	26.78	40.65	84.75	27.81	50.61	68.94	27.30
HABIT (AAAI'26)	20.82	47.83	61.78	91.93	54.90	76.52	88.22	51.37
FIVE (Ours)	**29.74**	**61.62**	**75.10**	**96.29**	**66.23**	**85.47**	**93.66**	**63.93**

where $\Phi(\cdot)$ denotes the VLM-based query generation process, $\mathcal{O}_r$ and $\mathcal{O}_t$ are the structured object representations, and $\mathcal{C}$ denotes candidate images introduced to enhance retrieval discriminability. By explicitly grounding the query in object-level differences, our strategy encourages it to describe the required transformation rather than directly summarize the target image, yielding more reference-dependent and CIR-aligned queries. We further combine automatic and manual verification to examine edited objects, attributes, spatial relations, and query discriminability. The resulting PatternCIRRS dataset contains 17,469 triplets divided into training, validation, and test sets. Figure 7 demonstrates that PatternCIRRS can effectively evaluate models' composed image retrieval capabilities. More details about the dataset construction process are provided in the Appendix.

4.2 Experimental Setup

Datasets. To evaluate the effectiveness of the proposed model, we conduct experiments on PatternCIR and PatternCIRRS.

Baselines. To comprehensively evaluate the effectiveness of our method, we compare FIVE with ten representative baselines, including C-Former Xu et al. (2023), CIR-TFMUR Chen et al. (2024), SPRC Bai et al. (2024), ENCODER Li et al. (2025b), OFFSET Chen et al. (2025b), TME Li et al. (2025a), TisCIR Liang et al. (2025), INTENT Chen et al. (2026), HUG Tang et al. (2026), and HABIT Li et al. (2026). All baselines are reproduced on the same dataset splits, with their image encoders kept frozen. Since OFFSET originally jointly optimizes its image encoder, we denote its adaptation to the frozen-encoder setting as OFFSET*.

Implementation Details. The proposed framework is implemented based on BLIP-2. The dimensions of the visual patch features and textual features are $d_v = 1408$ and $d_t = 768$, respectively. Both modalities are projected into a shared retrieval embedding space with dimension $d = 256$. The number of selected patches in Top-K is set to $K = 4$, and the radial boundaries of the four frequency bands are set to $\rho = [0, 0.25, 0.5, 0.75, 1]$. The loss weights are set to $\lambda_l = 0.6$, $\lambda_f = 0.8$, and $\lambda_{in} = 0.5$. The model is optimized using AdamW with an initial learning rate of 1×10^{-5} and a weight decay of 0.05. The batch size is set to 64, and the model is trained for 35 epochs, with the first 3 epochs used for learning-rate warmup. All experiments are conducted on an NVIDIA A100 GPU. More details are in the Appendix.

Metrics. We use $R@K$ to measure the percentage of target images appearing in the top-K retrieved results. We also report $\mathrm{R}_{\text{subset}}@K$ to evaluate retrieval performance within each subset.

4.3 Performance Comparison

Tables 1 and 2 report the retrieval results of FIVE on PatternCIRRS and PatternCIR, respectively. As shown in Table 1, FIVE achieves the best performance across all metrics on PatternCIRRS.

Table 2: Comparison of performance on the PatternCIR test set (%). Avg. denotes the mean of R@5 and R_{subset}@1. The best results are shown in bold and the second-best results underlined.

Method	R@K				R_{subset}@K			Avg.
	$K=1$	$K=5$	$K=10$	$K=50$	$K=1$	$K=2$	$K=3$	
C-Former (TMM'23)	39.43	<u>72.96</u>	<u>84.04</u>	97.46	54.70	74.17	82.07	63.83
CIR-TFMUR (ICLR'24)	<u>42.83</u>	70.35	80.34	95.31	66.27	83.47	91.41	68.31
SPRC (ICLR'24)	42.49	72.51	81.81	96.48	<u>70.88</u>	<u>85.82</u>	<u>93.00</u>	<u>71.70</u>
ENCODER (AAAI'25)	33.23	68.09	81.66	<u>97.95</u>	47.18	69.26	82.91	57.64
OFFSET* (ACM MM'25)	36.93	71.34	83.40	**98.37**	56.40	75.95	88.01	63.87
TME (CVPR'25)	37.99	68.88	80.15	95.91	67.29	84.04	91.98	68.09
TisCIR (IJCAI'25)	7.03	20.11	31.04	67.59	42.15	65.10	80.87	31.13
INTENT (AAAI'26)	38.14	61.39	70.47	86.84	64.12	77.80	87.78	62.76
HUG (AAAI'26)	11.34	33.76	49.30	90.89	40.79	63.63	68.94	37.28
HABIT (AAAI'26)	22.45	48.99	64.23	92.89	49.86	70.92	85.29	49.43
FIVE (Ours)	**47.10**	**74.21**	**84.46**	97.24	**72.40**	**86.72**	**93.89**	**73.31**

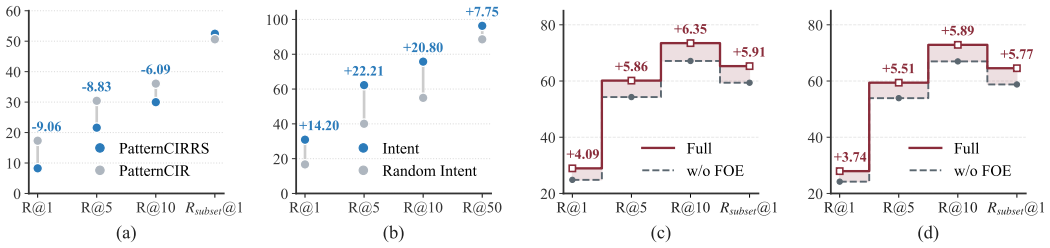


Figure 3: Motivation experiments. (a) Text-only retrieval on PatternCIRRS and PatternCIR. (b) Effect of original and random intents on retrieval performance. (c–d) Effect of removing FOE on small-object retrieval ($\sigma < 0.4$ and $\sigma < 0.3$).

Specifically, it obtains 29.74% R@1, 61.62% R@5, and 66.23% R_{subset}@1, with an overall score of 63.93%. Compared with the strongest baseline, SPRC, FIVE improves R@1, R@5, and the overall score by 1.78, 2.66, and 1.72 percentage points, respectively. These results demonstrate that FIVE can more accurately model modification intents and distinguish visually similar candidates.

As shown in Table 2, FIVE achieves 47.10% R@1, 72.40% R_{subset}@1, and an overall score of 73.31% on PatternCIR, outperforming the strongest baselines by 4.27, 1.52, and 1.61 percentage points, respectively. FIVE maintains competitive overall performance, with its advantages mainly reflected in Top-1 retrieval accuracy and fine-grained subset retrieval.

4.4 Motivation Experiments

To validate the dataset design and the retrieval mechanism of the model, we conduct the experiments shown in Figure 3. **#1: Assessing Reference Dependence of PatternCIRRS.** Figure 3(a) shows that PatternCIR achieves higher accuracy under the text-only retrieval setting, whereas PatternCIRRS yields an R@1 score that is 9.06 percentage points lower, indicating stronger dependence on the reference image. **#2: Evaluating Sensitivity to Modification Intent.** As shown in Figure 3(b), randomly shuffling the intents leads to substantial performance degradation, with R@1 and R@5 decreasing by 14.20 and 22.21 percentage points, respectively. This result indicates that the model relies on modification intents rather than merely matching object categories. **#3: Evaluating FOE for Small-Object Retrieval.** Figures 3(c) and (d) show that FOE improves retrieval for samples with $\sigma < 0.4$ and $\sigma < 0.3$, respectively, where σ denotes the percentile rank of the target-object size among all targets, with lower values indicating smaller objects.

Table 3: Ablation study on the validation sets of PatternCIRRS and PatternCIR. $\overline{R}$ and $\overline{R}_{\text{subset}}$ are averaged over $K = \{1, 5, 10, 50\}$ and $K = \{1, 2, 3\}$, respectively.

D#	Variant	PatternCIRRS		PatternCIR	
		$\overline{R}$	$\overline{R}_{\text{subset}}$	$\overline{R}$	$\overline{R}_{\text{subset}}$
Baseline and Core Modules					
1	Q-Former	60.19	61.56	66.13	71.19
2	w/o FOE	64.18	75.43	65.66	79.88
3	w/o IVES	70.70	80.11	73.45	83.83
Fine-grained Component Analysis					
4	w/o F_l	70.64	81.73	75.50	83.57
5	w/o F_f	64.02	76.00	71.39	82.02
6	w/o M_{obj}	71.58	81.08	75.46	84.25
Loss Functions					
7	w/o $\mathcal{L}_{\text{FOE}}$	70.25	80.89	74.92	84.09
8	w/o $\mathcal{L}_{\text{IVES}}$	71.74	81.05	75.52	83.97
FIVE (Ours)		**71.84**	**81.97**	**75.75**	**84.33**

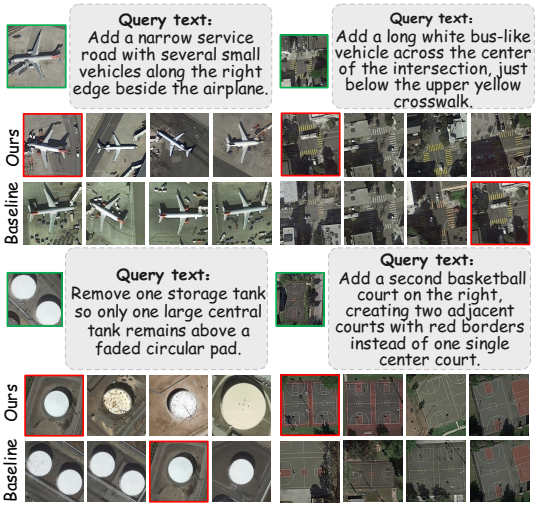


Figure 4: Case study. The reference and target images are highlighted with green and red outlines, respectively. Results are ranked from Top-1 to Top-4 from left to right.

4.5 Ablation Studies

To evaluate the effectiveness of different components, we conduct three groups of ablation experiments on the validation sets of PatternCIRRS and PatternCIR. The results are reported in Table 3. **G1: Baseline and Core Modules.** D#(1) adopts a BLIP-2-based baseline that retains only the Q-Former composition pathway. D#(2) removes the entire FOE module, while D#(3) removes the entire IVES module. The complete model substantially outperforms the baseline, while removing either FOE or IVES leads to performance degradation. These results demonstrate the complementary effectiveness of the two modules. **G2: Fine-grained Component Analysis.** D#(4) removes the spatial-domain object features F_l, and the resulting performance drop demonstrates the effectiveness of local spatial evidence for object-level representation. D#(5) removes the frequency-domain structural features F_f, causing a relatively pronounced degradation on both datasets and verifying the importance of frequency-domain structural information. D#(6) removes the object-aware mask M_{obj}, indicating that it supports both object feature enhancement and intent-guided visual evidence selection. **G3: Auxiliary Loss Functions.** D#(7) removes $\mathcal{L}_{\text{FOE}}$, while D#(8) removes $\mathcal{L}_{\text{IVES}}$. Both auxiliary losses provide supervision for their corresponding feature branches.

4.6 Case Study

Figure 4 presents the Top-4 retrieval results of FIVE and the baseline on PatternCIRRS. Our method achieves better and more consistent retrieval performance, ranking the target images at Top-1 across challenging queries involving small objects and diverse modification intents. In contrast, the baseline often retrieves globally similar images that fail to satisfy the specified query texts. More cases are provided in the Appendix.

5 Conclusion

In this paper, we propose FIVE, a visual evidence learning framework for remote sensing composed image retrieval. By enhancing object-level representations and modeling modification-aware evidence transformation, FIVE effectively improves retrieval performance. Extensive experiments on two datasets demonstrate its superiority over existing methods, highlighting the importance of fine-grained visual evidence and intent-aware modeling for RSCIR.

AI USE STATEMENT

We used generative AI tools to assist with language editing, relevant literature searches, the organization of research content, and the drafting and refinement of conceptual illustrations to improve the clarity and readability of the manuscript. During PatternCIRRS construction, GPT-5.4 generated modification queries from image pairs and object-level information, followed by automatic and manual verification of query quality. FIVE also uses a LoRA-finetuned Qwen3-4B model as its instruction parser, as described in Section 3.2. The authors reviewed and revised AI-assisted text and figures, checked the accuracy and relevance of references, and verified consistency with the actual method and experimental results. The authors take full responsibility for the accuracy, originality, and academic integrity of the final manuscript and associated research artifacts.

ETHICS STATEMENT

This study investigates remote sensing composed image retrieval using PatternCIR and PatternCIRRS, which is constructed from PatternCIR imagery. It does not involve human-subject experiments or additional collection of personally identifiable information or sensitive personal data. The data are used for academic research and model evaluation in remote sensing image understanding and retrieval. Automatically generated modification queries undergo automatic and manual verification to assess their consistency with image content and retrieval discriminability, reducing potential annotation errors. Use and redistribution of the underlying imagery remain subject to the applicable terms of the source dataset. We follow principles of research integrity throughout data processing, experimental design, result analysis, and manuscript preparation.

REPRODUCIBILITY STATEMENT

The main text describes the proposed method and its principal experimental settings to support reproducibility. Section 3 presents the FIVE framework, including multimodal input representations, instruction parsing, Frequency-aware Object Enhancement, Intent-guided Visual Evidence Selection, feature fusion, and training objectives. Section 4.1 describes the construction of PatternCIRRS, including image subset selection, object information extraction, modification query generation and verification, and data partitioning. Section 4.2 reports the datasets, evaluation metrics, baseline configurations, and main implementation details, including the pretrained backbone, feature dimensions, key hyperparameters, optimizer, training schedule, and hardware. Comparisons use consistent data splits and frozen image encoders. These descriptions support understanding and reproduction of the proposed method and verification of the experimental results.

REFERENCES

Yang Bai, Xinxing Xu, Yong Liu, Salman Khan, Fahad Khan, Wangmeng Zuo, Rick Siow Mong Mong, and Chun-Mei Feng. Sentence-level prompts benefit composed image retrieval. In *International Conference on Learning Representations*, volume 2024, pp. 30531–30548, 2024.

Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Conditioned and composed image retrieval combining and partially fine-tuning clip-based features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4959–4968, 2022.

Xiumei Chen, Xiangtao Zheng, and Xiaoqiang Lu. Context-aware local-global semantic alignment for remote sensing image-text retrieval. *IEEE Transactions on Geoscience and Remote Sensing*, 2025a.

Yanbei Chen and Loris Bazzani. Learning joint visual semantic matching embeddings for language-guided retrieval. In *European Conference on Computer Vision*, pp. 136–152. Springer, 2020.

Yiyang Chen, Zhedong Zheng, Wei Ji, Leigang Qu, and Tat-Seng Chua. Composed image retrieval with text feedback via multi-grained uncertainty regularization. In *International Conference on Learning Representations*, volume 2024, pp. 54915–54927, 2024.

Zhiwei Chen, Yupeng Hu, Zixu Li, Zhiheng Fu, Xuemeng Song, and Liqiang Nie. Offset: Segmentation-based focus shift revision for composed image retrieval. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 6113–6122, 2025b.

Zhiwei Chen, Yupeng Hu, Zhiheng Fu, Zixu Li, Jiale Huang, Qinlei Huang, and Yinwei Wei. Intent: Invariance and discrimination-aware noise mitigation for robust composed image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 20463–20471, 2026.

Shabnam Choudhury, Yash Salunkhe, Sarthak Mehrotra, and Biplab Banerjee. Rejepa: A novel joint-embedding predictive architecture for efficient remote sensing image retrieval. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2373–2382, 2025.

Eric Dodds, Jack Culpepper, Simao Herdade, Yang Zhang, and Kofi Boakye. Modality-agnostic attention fusion for visual search with text feedback. *arXiv preprint arXiv:2007.00145*, 2020.

Sonam Goenka, Zhaoheng Zheng, Ayush Jaiswal, Rakesh Chada, Yue Wu, Varsha Hedau, and Pradeep Natarajan. Fashionvlp: Vision language transformer for fashion retrieval with feedback. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14105–14115, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

Jirui Huang, Yaxiong Chen, Chuang Du, Shengwu Xiong, and Xiaoqiang Lu. Recotr: Reducing semantic cognitive shift via dual-consensus token compression for remote sensing image-text retrieval. *IEEE Transactions on Image Processing*, 2026.

Jongseok Kim, Youngjae Yu, Hoeseong Kim, and Gunhee Kim. Dual compositional learning in interactive image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 1771–1779, 2021.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

Shuxian Li, Changhao He, Xiting Liu, Joey Tianyi Zhou, Xi Peng, and Peng Hu. Learning with noisy triplet correspondence for composed image retrieval. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19628–19637, 2025a.

Zixu Li, Zhiwei Chen, Haokun Wen, Zhiheng Fu, Yupeng Hu, and Weili Guan. Encoder: Entity mining and modification relation binding for composed image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 5101–5109, 2025b.

Zixu Li, Yupeng Hu, Zhiwei Chen, Shiqi Zhang, Qinlei Huang, Zhiheng Fu, and Yinwei Wei. Habit: Chrono-synergia robust progressive learning framework for composed image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 6762–6770, 2026.

Zhechun Liang, Tao Huang, Fangfang Wu, Shiwen Xue, Zhenyu Wang, Weisheng Dong, Xin Li, and Guangming Shi. Patterncir benchmark and tiscir: Advancing zero-shot composed image retrieval in remote sensing. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pp. 1530–1538, 2025.

Jiancheng Pan, Yanxing Liu, Yuqian Fu, Muyuan Ma, Jiahao Li, Danda Pani Paudel, Luc Van Gool, and Xiaomeng Huang. Locate anything on earth: Advancing open-vocabulary object detection for remote sensing community. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 6281–6289, 2025.

Bill Psomas, Ioannis Kakogeorgiou, Nikos Efthymiadis, Giorgos Tolias, Ondřej Chum, Yannis Avrithis, and Konstantinos Karantzalos. Composed image retrieval for remote sensing. In *IGARSS 2024-2024 IEEE International Geoscience and Remote Sensing Symposium*, pp. 8526–8534. IEEE, 2024.

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

Huixin Sun, Runqi Wang, Yanjing Li, Linlin Yang, Shaohui Lin, Xianbin Cao, and Baochang Zhang. Set: Spectral enhancement for tiny object detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 4713–4723, 2025.

Haomiao Tang, Jinpeng Wang, Minyi Zhao, Guanghao Meng, Ruisheng Luo, Long Chen, and Shu-Tao Xia. Heterogeneous uncertainty-guided composed image retrieval with fine-grained probabilistic learning. *arXiv preprint arXiv:2601.11393*, 2026.

Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. Composing text and image for image retrieval-an empirical odyssey. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6439–6448, 2019.

Guoping Wang, Gang Wan, Zhijuan Su, Yang Wang, Yutong Jia, Gong Li, and Shi Liang. High-performance on-orbit intelligent computing and real-time services for remote sensing satellites based on large-scale computing power in space. *IEEE Access*, 2025.

Haokun Wen, Xuemeng Song, Xin Yang, Yibing Zhan, and Liqiang Nie. Comprehensive linguistic-visual composition network for image retrieval. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pp. 1369–1378, 2021.

Zixiao Wen, Peifeng Li, Yuhan Liu, Jingming Chen, Xiantai Xiang, Yuan Li, Huixian Wang, Yongchao Zhao, and Guangyao Zhou. Fanet: Frequency-aware attention-based tiny-object detection in remote sensing images. *Remote Sensing*, 17(24):4066, 2025.

Yahui Xu, Yi Bin, Jiwei Wei, Yang Yang, Guoqing Wang, and Heng Tao Shen. Multi-modal transformer with global-local alignment for composed query image retrieval. *IEEE Transactions on Multimedia*, 25:8346–8357, 2023.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Chengyu Zheng, Xiu Li, Xinyue Liang, Lei Huang, Shan Du, Jie Nie, and Junyu Dong. Cross-modal progressive perspective matching network for remote sensing image-text retrieval. *IEEE Transactions on Multimedia*, 27:3966–3978, 2025.

A Appendix Overview

This appendix provides supplementary materials, including the detailed construction of PatternCIRRS in Section B, hyperparameter sensitivity analysis in Section C, and additional qualitative analyses in Section D.

B Construction Details of PatternCIRRS

The main paper introduces the image subset partitioning and triplet construction process of PatternCIRRS. This section provides further details on query generation, automatic and manual verification, and dataset text statistics.

B.1 Query Generation

For each reference–target image pair, we first employ a remote sensing open-vocabulary object detector to extract object categories, quantities, and approximate spatial locations. We then compare the detection results of the two images to obtain structured information, including added objects, removed objects, and quantity changes. The detection results are not directly used as the final

```
user_prompt_text = (
    "Task: Generate one remote-sensing CIR query describing how to modify the reference image to match the target image. You will receive
two images: the reference image followed by the target image. Inspect the images directly and use them as the primary evidence. Use
captions and detector-derived object summaries only as auxiliary hints. Identify the most retrieval-discriminative visible change from
reference to target. "
    f"Reference caption: {['ref_caption']} "
    f"Target caption: {['tgt_caption']} "
    f"Reference objects: {json.dumps(signals['ref_summary'])} "
    f"Target objects: {json.dumps(signals['tgt_summary'])} "
    f"Added objects: {json.dumps(signals['added']} "
    f"Removed objects: {json.dumps(signals['removed'])} "
    f"Count changes: {json.dumps(signals['count_changed'])} "
    "Compare the reference and target images directly, select the single most important visible change, and describe it as an imperative edit
from reference to target. Focus on one dominant and retrieval-discriminative change. Prioritize object changes, attributes, count, density,
scale, orientation, position, spatial location, and background differences. Mention objects, locations, and changed properties when possible.
Use spatial descriptions for similar objects. Trust image evidence over captions or detector outputs when they conflict. Avoid vague
descriptions such as 'the scene changes'. Keep query_text within 25 English words. Set hard_query to true for fine-grained objects, attributes,
orientation changes, or precise spatial descriptions. "
    f"Allowed change types: {','.join(sorted(VALID_CHANGE_TYPES))}.
"
)
```

Figure 5: Prompt used for query generation in PatternCIRRS.

annotations. Instead, they serve as auxiliary visual localization cues and are jointly provided to the vision-language model together with the reference image, target image, image captions, and candidate images from the same subset.

The prompt used to generate the query text is shown in Figure 5. Its design follows three principles. First, the LLM is explicitly instructed to directly compare the reference and target images and use the image content as the primary evidence. The image captions and detected object summaries are used only for auxiliary localization. Second, the query text must describe an explicit modification from the reference image to the target image, rather than providing a static description of the target image. The LLM is required to select a single most salient change and express it as an imperative sentence. For scenes containing multiple similar objects, the prompt further requires spatial expressions such as "upper-left," "right of center," and "along the lower edge" for disambiguation. Finally, we provide the model with candidate images from the same subset, encouraging it not only to focus on the visual differences between the reference and target images, but also to prioritize modifications that distinguish the target image from visually similar candidates. For example, in parking lots, roads, or intersections, a simple description such as "add a vehicle" may apply to multiple candidate images. Therefore, the prompt encourages the model to describe the position, color, orientation, or local-region change of a specific vehicle, rather than providing only a generic quantity difference. The model outputs the generation results in a structured JSON format, where *query text* denotes the final query text, *change type* represents the corresponding modification type, and *key objects* records the main objects involved in the modification. To avoid generating queries that are overly long or contain multiple unrelated changes, the prompt requires the text to remain concise and, in principle, limits it to no more than 25 words.

B.2 Automatic and Manual Verification

Remote sensing images contain numerous objects that are small in scale, visually similar, or densely arranged. Therefore, we introduce an independent automatic verification stage after query generation, whose prompt is shown in Figure 6.

The verification model simultaneously receives the reference image, target image, generated query text, and candidate distractor image, and evaluates the sample quality from the following aspects: whether the query correctly describes the modification from the reference image to the target image; whether the edited object is correct; whether attributes such as color, quantity, scale, and orientation are consistent with the images; whether the location or spatial reference is accurate; whether the query invents a nonexistent global background change; and whether the target image is more compatible with the query than the candidate distractor image. As in the generation stage, the verification model is instructed to prioritize visual evidence rather than performing semantic matching solely based on the existing text or object detection results. The verification results are represented by multiple Boolean fields, including object correctness, attribute correctness, location correctness,

```
user_prompt_text = (
    "Task: Verify whether this remote-sensing CIR sample is valid and suitable for retrieval. "
    f"Reference image path: {ref_image} "
    f"Target image path: {tgt_image} "
    f"Candidate distractor image path: {candidate_note} "
    f"Query text: {query_text} "
    "Inspect the images directly and focus on the most important and discriminative changes. Verify object correctness, attribute correctness,
spatial correctness, background consistency, and retrieval discriminativeness. If objects have changed direction (e.g., vehicles or airplanes), pay
special attention to orientation or heading changes. In crosswalks or parking areas, do not focus only on vehicle count changes; instead,
consider specific vehicles or local region-level changes. Check whether the query correctly describes how to edit the reference image into the
target image. Check whether the edited object, changed attribute, and spatial reference are correct. Avoid accepting queries that invent
unsupported global background changes. Determine whether the target image is more compatible with the query than the candidate image. The
confidence score must be a float between 0 and 1 indicating verification accuracy. The explanation must be exactly one short sentence. "
    "Return JSON only with this schema: "
    '{"object_correct":true,"attribute_correct":true,"location_correct":true,'
    '"background_consistent":true,"retrieval_discriminative":true,'
    '"final_pass":true,"confidence":0.93,'
    '"explanation":"target matches the queried edit."}'
)
```

Figure 6: Prompt used for triplet verification in PatternCIRRS.

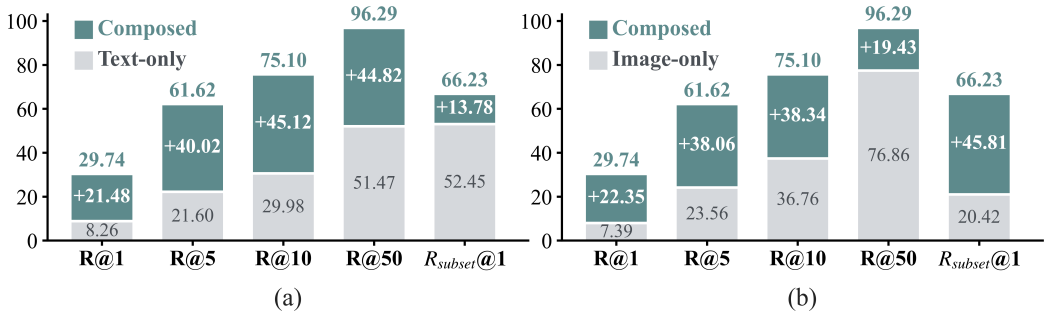


Figure 7: Comparison of image-only, text-only, and composed retrieval performance on PatternCIRRS.

background consistency, and retrieval discriminability. The model also outputs an overall decision, final pass, a confidence score, and a short explanatory sentence. We further conduct manual verification on the triplets that pass the automatic verification stage. The expert team consists of four members, and the manual verification process takes two months to complete. By combining automatic verification with manual review, we ultimately retain 17,469 CIR triplets with explicit reference dependence.

B.3 Query Text Statistics

Figure 8(a) presents the word-count distribution of the query texts in PatternCIRRS. Most queries contain approximately 16 to 24 English words, with higher frequencies around 19 to 21 words. This distribution indicates that the generated texts generally contain sufficient information to describe the modified object, operation type, and spatial constraints, while avoiding the semantic insufficiency of overly short texts and the potentially redundant changes contained in overly long texts.

Figure 8(b) presents the word cloud of the query texts. Operation words such as replace, add, remove, and move occur frequently, indicating that the texts in the dataset primarily describe explicit image modifications. Meanwhile, spatial words such as left, right, center, and diagonal also appear frequently, reflecting the need for precise positional and directional disambiguation in remote sensing scenes. The word cloud further contains attribute words such as white, dark, large, and small, as well as diverse scene objects such as building, parking, road, court, airplane, and boat, demonstrating the diversity of PatternCIRRS in terms of object categories, modification operations, and visual attributes.

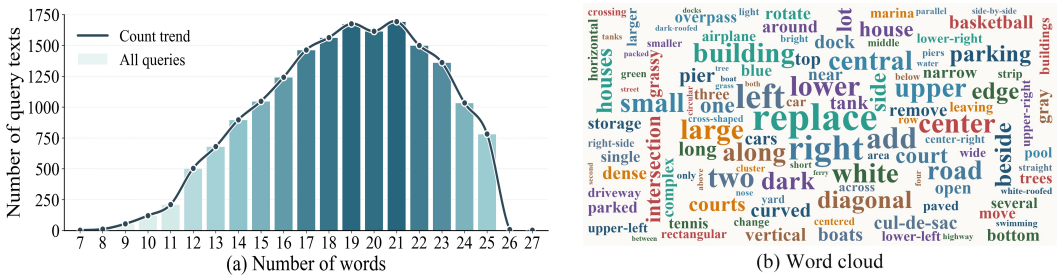


Figure 8: Statistics of the query texts in PatternCIRRS. (a) Word-count distribution. (b) Word cloud of frequently occurring words.

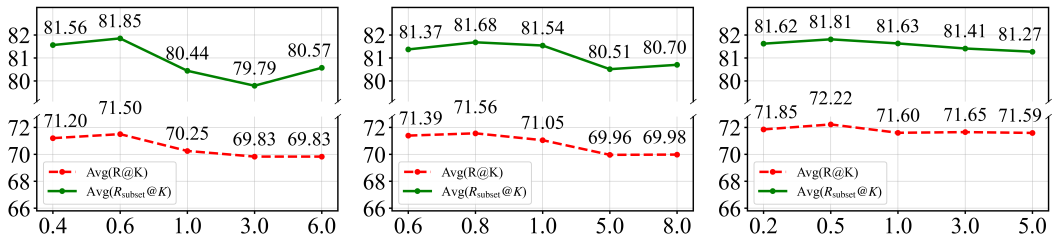


Figure 9: Hyperparameter sensitivity analysis of the loss weights (a) λ_l, (b) λ_f, and (c) λ_{in}. The averaged global recall and subset recall are reported.

C Hyperparameter Sensitivity Analysis

C.1 Sensitivity to Loss Weights

We analyze the sensitivity of FIVE to the loss weights λ_l, λ_f, and λ_{in}. As shown in Figure 9, the retrieval performance remains relatively stable within a reasonable range of each hyperparameter. The best overall results are obtained with $\lambda_l = 0.6$, $\lambda_f = 0.8$, and $\lambda_{in} = 0.5$, which are used as the default settings. Increasing λ_l or λ_f excessively causes a slight performance drop, while the model is comparatively less sensitive to λ_{in}. These results demonstrate that FIVE is generally robust to the selection of loss weights.

C.2 Sensitivity to the Number of Frequency Bands

We further analyze the sensitivity of FIVE to the number of frequency bands B. Specifically, we vary B from 2 to 10 with an interval of 2. As shown in Figure 10, $B = 4$ achieves the best performance across all evaluation metrics. A smaller B leads to a coarse partition of the frequency spectrum, which may be insufficient to distinguish complementary structural cues. In contrast, excessively increasing B fragments the spectral information and may introduce redundant or noisy frequency representations. Therefore, we set $B = 4$ as the default configuration, indicating that a moderate frequency partition provides a favorable balance between spectral granularity and representation robustness.

D Additional Qualitative Results

Figures 11, 12, 13 and 14 present additional qualitative comparisons. FIVE retrieves the correct targets for object replacement, addition, relocation, and attribute changes, while the baseline often favors globally similar but mismatched images. These results further demonstrate the effectiveness of our method in modeling fine-grained modification intents.

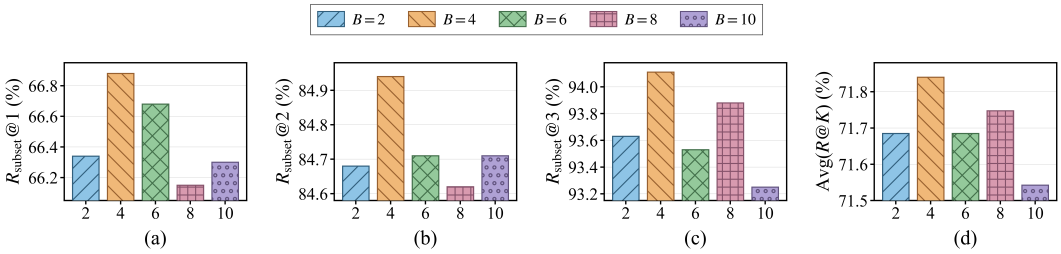


Figure 10: Hyperparameter sensitivity analysis of the number of frequency bands B. We report (a) R_{subset}@1, (b) R_{subset}@2, (c) R_{subset}@3, and (d) the global recall averaged over $K = \{1, 5, 10, 50\}$.

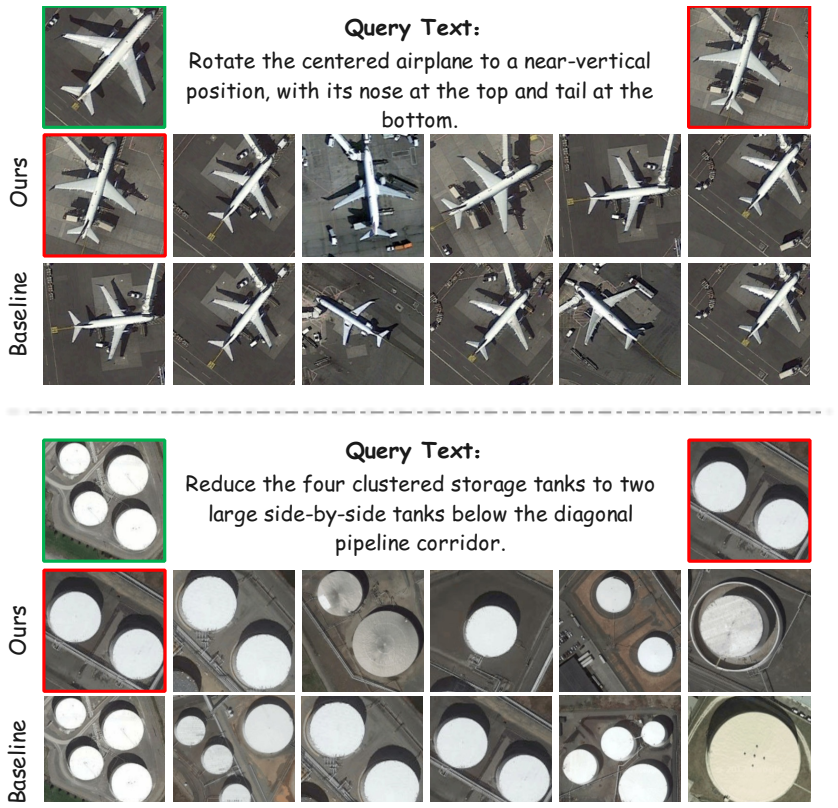


Figure 11: Additional qualitative comparisons. The reference and target images are outlined in green and red, respectively. From left to right, the retrieval results are ranked from Top-1 to Top-6.



Figure 12: Additional qualitative comparisons. The reference and target images are outlined in green and red, respectively. From left to right, the retrieval results are ranked from Top-1 to Top-6.

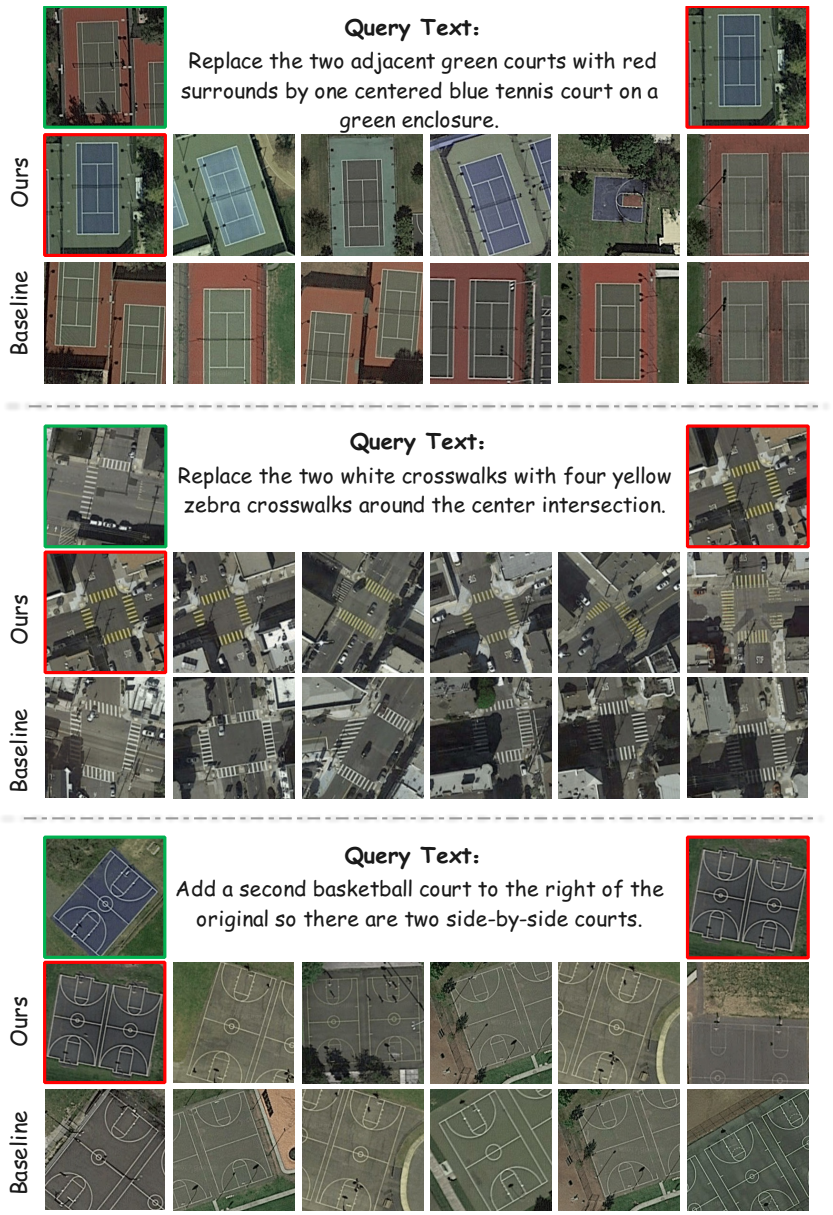


Figure 13: Additional qualitative comparisons. The reference and target images are outlined in green and red, respectively. From left to right, the retrieval results are ranked from Top-1 to Top-6.

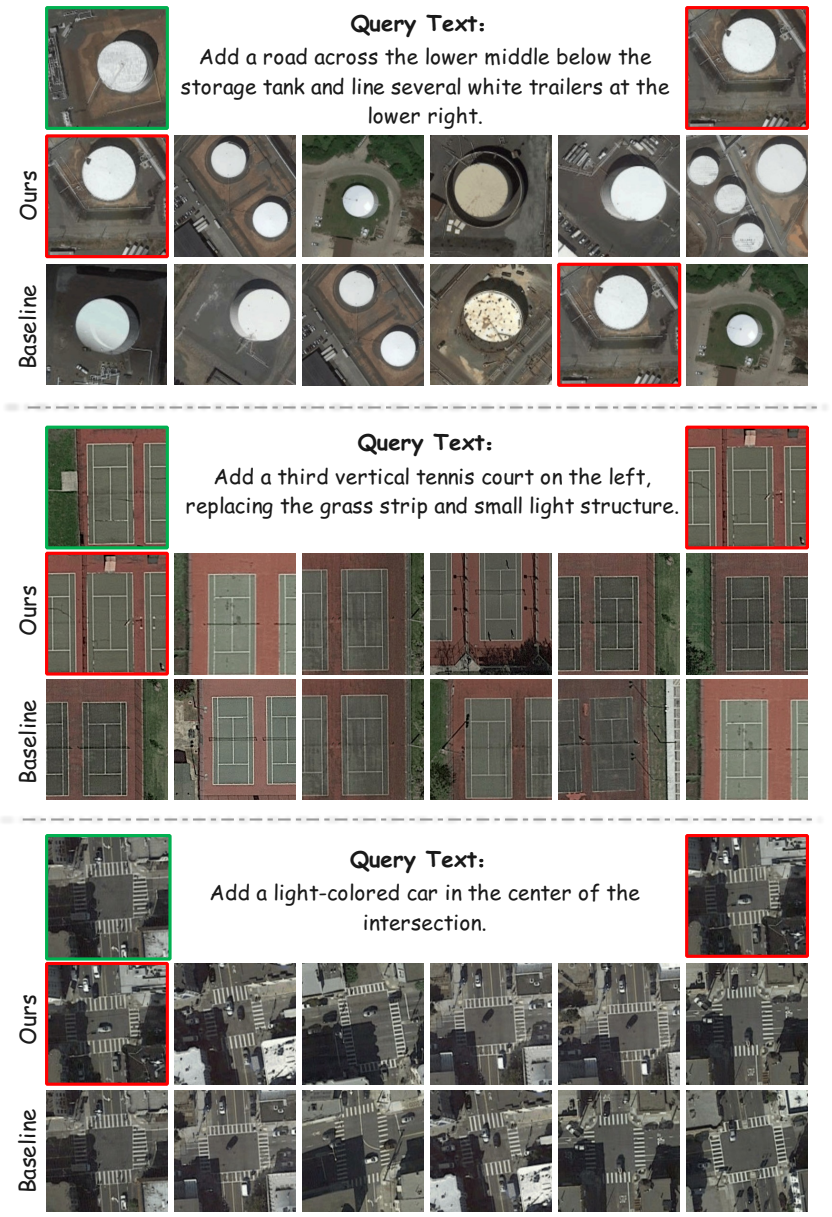


Figure 14: Additional qualitative comparisons. The reference and target images are outlined in green and red, respectively. From left to right, the retrieval results are ranked from Top-1 to Top-6.