P^{2}CAL: A PRIOR-POSTERIOR COLLABORATIVE CALIBRATION FRAMEWORK FOR ROBUST REMOTE SENSING IMAGE-TEXT RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Remote sensing image-text retrieval (RSITR) aims to match remote sensing images with their corresponding natural language descriptions. Existing methods primarily focus on learning a discriminative cross-modal space, while paying insufficient attention to the collaborative calibration of image-text correspondences at the levels of prior supervision and posterior scores. Such calibration can facilitate both the learning of this space and the subsequent refinement of retrieval scores. To address this limitation, we propose P^{2}Cal, a prior-posterior collaborative calibration framework for robust RSITR. Specifically, we first propose a Prior Semantic Calibration Module (PSCM), which purifies supervision through prior semantic calibration to obtain highly reliable supervisory signals. These signals are then used to train our proposed Manifold Residual Adapter (MRA) and Posterior Score Calibration Flow (PSCF). MRA incorporates lightweight manifold residual adapters into the representation pathways to construct a stable discriminative score space. PSCF takes the resulting matching scores as its initial state and learns a continuous residual velocity field to progressively optimize retrieval scores. Extensive experiments on multiple RSITR datasets demonstrate that P^{2}Cal consistently outperforms existing state-of-the-art methods.

1 INTRODUCTION

With the rapid advancement of satellite platforms, airborne sensors, and remote sensing (RS) imaging technologies, vast amounts of Earth observation data are continually being generated and widely used in disaster response, urban planning, land-cover analysis, and ecological and environmental monitoring (Chi et al., 2016; Li et al., 2021; Joyce et al., 2009). Remote sensing image-text retrieval (RSITR) provides semantic access to these massive datasets by matching remote sensing images with natural language descriptions in both Image-to-Text (I2T) and Text-to-Image (T2I) directions. In recent years, numerous RSITR methods have been proposed. Existing methods mainly focus on discriminant space learning, typically through multi-scale and global-local embedding fusion (Yuan et al., 2022a;b), parameter-efficient adaptation of pretrained models (Yuan et al., 2023; Wang et al., 2025), and other strategies. In addition, some methods exploit additional prior knowledge to enhance cross-modal representations, such as PIR (Pan et al., 2023a) and KAMCL (Ji et al., 2023), while others refine retrieval results from a posterior perspective through similarity calibration or re-ranking, such as GaLR (Yuan et al., 2022b) and CFITR (Zhou et al., 2025). However, existing methods that exploit prior information mainly introduce additional information into the discriminative space to facilitate representation learning, whereas posterior methods typically perform inference-only calibration without jointly leveraging prior knowledge or the supervisory information used during discriminative space learning.

As illustrated in Figure 1, unlike existing methods, P^{2}Cal neither introduces additional information to facilitate discriminative space learning nor independently post-processes retrieval scores at the inference stage. Instead, it pays explicit attention to the collaborative calibration of image-text correspondences at the levels of prior supervision and posterior scores. Specifically, in the prior stage, P^{2}Cal leverages prior semantics to purify the existing supervision by masking incorrect image-text pairs, thereby providing more reliable supervisory signals for learning a robust and highly discriminative embedding space, rather than injecting additional knowledge into the discriminative space. In

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

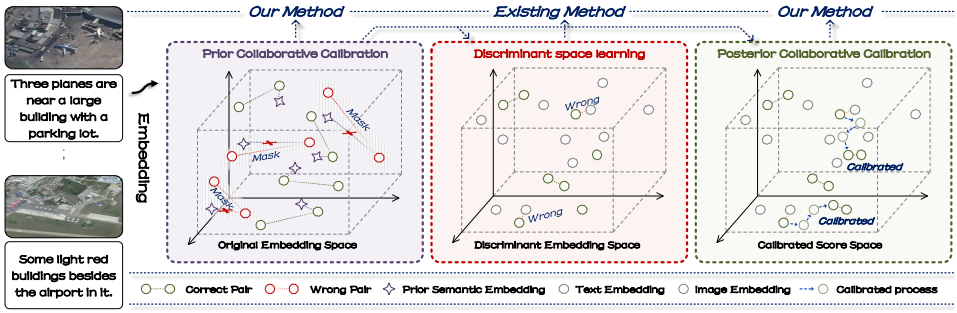

Figure 1: Comparison between existing methods and P^{2}Cal. Unlike existing methods that either introduce additional prior information or optimize retrieval scores only at inference, P^{2}Cal collaboratively calibrates image-text matching scores from both prior and posterior perspectives.

the posterior stage, P^{2}Cal further jointly exploits the purified supervision, the initial matching score matrix learned from the discriminative space, and the image-text embeddings to learn a continuous residual velocity field that progressively calibrates and refines the retrieval scores. In this way, P^{2}Cal no longer treats prior supervision optimization and posterior score calibration as two isolated processes, but instead establishes explicit connections among supervisory information, discriminative representations, and matching scores, enabling collaborative optimization from supervision purification to score refinement.

Based on the above analysis, we propose P^{2}Cal, a **P**rior-**P**osterior Collaborative **Cal**ibration Framework for robust RSITR. Specifically, we first propose a Prior Semantic Calibration Module (PSCM), which leverages a vision-language model (VLM) to extract semantic priors from images and calibrates supervision by comparing these priors with the corresponding textual descriptions, thereby obtaining highly reliable supervisory signals. These signals are then used to train our proposed Manifold Residual Adapter (MRA) and Posterior Score Calibration Flow (PSCF). MRA incorporates lightweight manifold residual adapters into the representation pathways. Through sample-conditioned orthogonal modulation in the bottleneck space, it provides a structurally constrained form of adaptation to transfer pretrained representations to the remote sensing domain and improve cross-modal discriminability, thereby supporting discriminative score-space learning. PSCF takes the resulting matching scores as its initial state and learns a continuous residual velocity field to progressively optimize retrieval scores. It increases the scores of underestimated reliable positive pairs while suppressing those of overestimated hard negative pairs, thereby improving the relative ranking of positive and negative pairs. Extensive experiments on multiple RSITR datasets demonstrate that P^{2}Cal consistently outperforms existing state-of-the-art methods.

Our main contributions are summarized as follows:

- We propose P^{2}Cal, a prior-posterior collaborative calibration framework for robust RSITR, which integrates image-text correspondence learning guided by reliable supervision with subsequent retrieval score calibration.
- We design three complementary modules: PSCM purifies supervision to provide highly reliable supervisory signals for training MRA and PSCF; MRA constructs a stable and domain-adaptive discriminative score space; and PSCF progressively refines retrieval scores through a continuous residual velocity field.
- Extensive experiments on three widely used RSITR datasets demonstrate that P^{2}Cal consistently outperforms existing state-of-the-art methods, validating the effectiveness of prior-posterior collaborative calibration.

2 Related Work

In recent years, numerous RSITR methods have been proposed. Existing methods mainly focus on discriminant space learning. For example, AMFMN (Yuan et al., 2022a) and GaLR (Yuan et al., 2022b) enhance cross-modal representations through multi-scale embedding aggregation and global-local evidence modeling, respectively. FAAMI (Zheng et al., 2023),

DOVE (Ma et al., 2024), and IEFT (Tang et al., 2023) improve representation quality by enforcing multi-scale consistency, suppressing regional redundancy, and modeling cross-modal interactions, respectively. LAPS (Fu et al., 2024), CHAN (Pan et al., 2023b), and FGVLA (Li et al., 2025a) further explore language-guided patch-level alignment, hard-fragment alignment, and patch-noun alignment, respectively. HVSA (Zhang et al., 2023) enhances embedding structures through hyperspherical regularization, while McRSA (Xiao et al., 2026) models coarse-to-fine cross-modal relationships via posterior probability estimation, centroid updating, and embedding-scale adaptation. DSDMR (Shi et al., 2024) improves robustness against noisy correspondences by modeling clean and noisy similarity distributions with a dynamic-margin ranking objective. RRSITR (Song et al., 2026) and CFITR (Zhou et al., 2025) further improve robustness and efficiency through reliability-aware margins, hash-based candidate generation, and Transformer-based refinement. Overall, these methods mainly aim to learn more discriminative cross-modal representations and enlarge the separation between matched and mismatched image-text pairs.

In addition, some studies exploit additional prior knowledge to improve cross-modal alignment. PIR (Pan et al., 2023a) introduces scene-level instructions, while KAMCL (Ji et al., 2023) leverages knowledge-guided concept filtering and alignment. Meanwhile, other methods focus on refining initial similarity estimation during retrieval inference. For example, GaLR (Yuan et al., 2022b) performs multivariate reranking by exploiting neighborhood and reverse-retrieval relationships, and CFITR (Zhou et al., 2025) adopts post-processing similarity reranking strategies. However, existing methods that exploit prior information mainly introduce additional information into the discriminative space to facilitate representation learning, whereas posterior methods typically perform inference-only calibration without jointly leveraging prior knowledge or the supervisory information used during discriminative space learning. Therefore, existing methods pay insufficient attention to the collaborative calibration of image-text correspondences at the levels of prior supervision and posterior scores. A unified framework that tightly integrates prior semantic guidance with posterior score calibration is essential for achieving more robust and accurate RSITR.

3 METHOD

The overall architecture of P^{2}Cal is illustrated in Figure 2. Built upon a frozen BLIP (Li et al., 2022) backbone, P^{2}Cal first employs the Prior Semantic Calibration Module (PSCM) to purify supervision from the remote sensing image-text dataset $\mathcal{D} = \{(I_i, T_{ij})\}_{i,j}$ and obtain reliable training signals for MRA and PSCF. MRA constructs a stable domain-adaptive discriminative score space, while PSCF progressively refines retrieval scores by learning a residual velocity field from the initial matching scores.

3.1 PRIOR SEMANTIC CALIBRATION MODULE

To reduce the impact of supervision noise introduced by unreliable textual descriptions on image-text correspondence learning, we propose a PSCM to purify supervision before training MRA and PSCF. Specifically, for each training image I_i, Qwen2.5-VL (Bai et al., 2025) generates an auxiliary description G_i offline, independently of the original annotations. PSCM then constructs a correspondence embedding vector $\boldsymbol{\phi}_{ij}$ by exploiting the semantic consistency between the generated description and the annotated text, as well as the consistency among texts associated with the same image. The semantic agreement between the generated description and the text is defined as $q_{ij} = \mathrm{sim}_{\mathrm{IDF}}(G_i, T_{ij})$, where $\mathrm{sim}_{\mathrm{IDF}}(\cdot,\cdot)$ denotes the IDF-weighted token-level cosine similarity. The intra-group semantic consistency between T_{ij} and other texts associated with the same image is defined as $c_{ij\ell} = \mathrm{sim}_{\mathrm{IDF}}(T_{ij}, T_{i\ell}), \ell \neq j$. This semantic information is aggregated into $\boldsymbol{\phi}_{ij}$.

To estimate correspondence noise, PSCM trains a noise probability estimator on synthetic noisy validation data. Noisy image-text pairs are generated by randomly perturbing clean correspondences, and each pair is assigned a binary noise label. A logistic regression model is fitted on the normalized correspondence embeddings $\widetilde{\boldsymbol{\phi}}_{ij}$ to predict the posterior noise probability:

$$p_{ij}^{\mathrm{noise}} = \sigma_{\mathrm{sig}}\left(\boldsymbol{w}^{\top}\widetilde{\boldsymbol{\phi}}_{ij} + b\right), \tag{1}$$

where $\boldsymbol{w}$ and b are the learned regression parameters and $\sigma_{\mathrm{sig}}(\cdot)$ denotes the sigmoid function. The predicted value $p_{ij}^{\mathrm{noise}} \in [0,1]$ represents the probability that the image-text pair (I_i, T_{ij}) contains noisy supervision, where a larger value indicates a higher likelihood of unreliable correspondence.

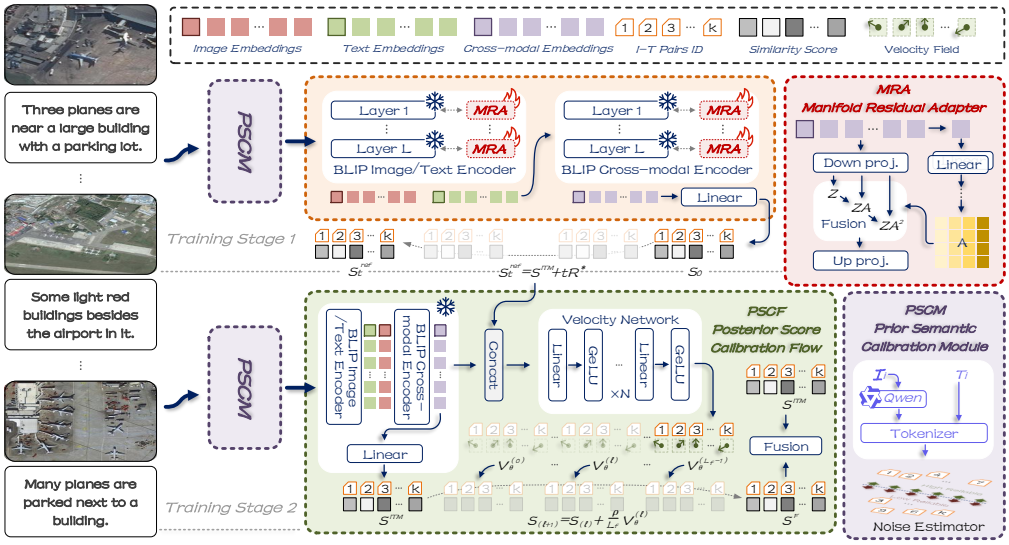

Figure 2: Overview of P^{2}Cal. PSCM provides reliable supervision for MRA and PSCF. MRA adapts the visual, textual, and cross-modal pathways to construct a stable discriminative score space, while PSCF progressively refines the resulting matching scores through a continuous residual velocity field.

Before training MRA and PSCF, PSCM applies the trained estimator to all training pairs. The average predicted noise probability is calculated as $\bar{p}^{\text{noise}} = \frac{1}{N_{\text{pair}}} \sum_{i,j} p_{ij}^{\text{noise}}$, where N_{pair} denotes the total number of image-text pairs. Then, the PSCM constructs an empirical calibration mapping using synthetic validation data. Specifically, the mean predicted probabilities under different known corruption ratios r are collected, and the mapping $\mathcal{C}_{\text{val}} : \bar{p}_r^{\text{noise}} \mapsto r$ is obtained through monotonic interpolation. The estimated noise proportion of the training set is then calculated as

$$\zeta = \mathcal{C}_{\text{val}}(\bar{p}^{\text{noise}}), \tag{2}$$

where $\mathcal{C}_{\text{val}}(\cdot)$ denotes the empirical calibration mapping. Since the training and validation sets are sampled from the same RSITR dataset, their semantic characteristics and annotation noise patterns are assumed to follow similar distributions. Based on the estimated noisy proportion ζ, PSCM removes the most suspicious correspondences. Specifically, the number of filtered pairs is determined as $N_{\text{drop}} = \text{round}(\zeta N_{\text{pair}})$, and the pairs with the largest predicted noise probabilities are discarded. The binary retention mask is defined as $k_{ij}^{\text{keep}} = \mathbf{1}\{(I_i, T_{ij}) \text{ is retained}\}$. The remaining pairs form the purified supervision set $\mathcal{D}' = \{(I_i, T_{ij}) \mid k_{ij}^{\text{keep}} = 1\}$. The purified set $\mathcal{D}'$ is subsequently used to train both MRA and PSCF. See Appendix A.1 for details of noise-estimator construction and empirical calibration.

3.2 Manifold Residual Adapter

To bridge the domain gap and improve cross-modal discriminability, we propose MRA, which introduces lightweight manifold residual adapters into BLIP's visual, textual, and cross-modal pathways. MRA applies sample-conditioned orthogonal modulation in the bottleneck space, providing a structurally constrained adaptation mechanism for learning a discriminative score space. For layer l and pathway $m \in \{\text{vis}, \text{text}, \text{cross}\}$, the update at the feed-forward stage is written as

$$H^l = F^{l,m}(X) + \text{MRA}^{l,m}(X), \qquad l \geq 1, \tag{3}$$

with $H^0 = E$, where E contains the initial token embeddings and X is the hidden state after the corresponding attention operations. Here, $F^{l,m}$ denotes the original feed-forward update, including its residual connection and pathway-specific normalization.

For an N-token input, MRA computes the low-dimensional bottleneck embeddings $Z = XW_{\text{down}}^{l,m}$ and the global context $h = \text{Pool}(X) = \frac{1}{N}\sum_{n=1}^{N} X_n$. Learnable mappings produce $u = \phi_u^{l,m}(h)$ and $v = \phi_v^{l,m}(h)$, defining the low-rank skew-symmetric operator $A = uv^\top - vu^\top$, where u and

v are column vectors in the bottleneck space. This sample-specific operator consistently modulates all tokens through

$$\widetilde{Z} = Z + \alpha_1 ZA + \alpha_2 ZA^2, \tag{4}$$

where $\vartheta = \sqrt{\|u\|_2^2\|v\|_2^2 - (u^\top v)^2}$, $\alpha_1 = \sin\vartheta/\vartheta$, and $\alpha_2 = (1-\cos\vartheta)/\vartheta^2$, with continuous limits $\alpha_1 = 1$ and $\alpha_2 = 1/2$ at $\vartheta = 0$. The first- and second-order terms model direct embedding coupling and its propagation and redistribution, respectively; ϑ measures the area spanned by u and v and parameterizes the bottleneck rotation (Lezcano-Casado & Martínez-Rubio, 2019). For the skew-symmetric operator A of rank at most two, this expression equals $\widetilde{Z} = Z\exp(A)$. Since $\exp(A)$ is orthogonal, the rotation preserves pairwise inner products among bottleneck tokens within each sample, satisfying $\widetilde{Z}\widetilde{Z}^\top = ZZ^\top$. This preservation property is specific to the within-sample bottleneck rotation. The residual output is

$$\mathrm{MRA}^{l,m}(X) = s\,\sigma_{\mathrm{ReLU}}(\widetilde{Z})W_{\mathrm{up}}^{l,m}, \tag{5}$$

where $W_{\mathrm{down}}^{l,m}$ and $W_{\mathrm{up}}^{l,m}$ are down- and up-projections (Houlsby et al., 2019), σ_{ReLU} is ReLU function, and s scales the residual. The final global image and text representations, denoted by f_i^I and f_j^T, are projected into a shared embedding space, and their cosine similarities form the initial similarity matrix S^{ITC}. The top-K_{cand} candidate texts are selected according to S^{ITC}, and the corresponding image-text pairs are processed by the cross-modal encoder through all Transformer blocks to obtain the final cross-modal embedding h_{ij}^{IT} from the `[CLS]` embedding. The embedding h_{ij}^{IT} is then fed into a linear head to predict the discriminative score matrix S^{ITM}.

3.3 Posterior Score Calibration Flow

To progressively transport the initial discriminative scores toward a calibrated score state while retaining their initial discriminative structure, we propose Posterior Score Calibration Flow (PSCF), which initializes from the resulting matching scores S^{ITM} and learns a continuous residual velocity field to increase the scores of underestimated reliable positive pairs and suppress those of overestimated hard negative pairs, thereby improving their relative ranking.

For the i-th query, we collect all correctly matched sample positions j into $\mathcal{P}_i$ and let $S_{i,+}^{\mathrm{ITM}} = \max_{j\in\mathcal{P}_i} S_{ij}^{\mathrm{ITM}}$. We select the highest-scoring negatives as $\mathcal{H}_i = \mathrm{TopK}_{\mathrm{hard}}(\{j \mid j \notin \mathcal{P}_i\}, S_{ij}^{\mathrm{ITM}})$, with at most K_{hard} entries per query. Their correction magnitudes are $e_{ij} = [\mu + S_{ij}^{\mathrm{ITM}} - S_{i,+}^{\mathrm{ITM}}]_+ + \eta$, where μ is the ranking margin, η is a baseline correction for selected hard negatives, and $[\cdot]_+$ denotes ReLU. For nonempty $\mathcal{P}_i$ and $\mathcal{H}_i$, the target correction is

$$R_{ij}^* = \begin{cases} \dfrac{1}{|\mathcal{P}_i|\,|\mathcal{H}_i|}\displaystyle\sum_{k\in\mathcal{H}_i} e_{ik}, & j \in \mathcal{P}_i, \\ -\dfrac{e_{ij}}{|\mathcal{H}_i|}, & j \in \mathcal{H}_i, \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

This target correction increases positive scores, decreases hard-negative scores with larger corrections for margin violations, and assigns zero correction to other positions while controlling the correction scale. Then, PSCF learns along the reference path $S_t^{\mathrm{ref}} = S^{\mathrm{ITM}} + tR^*$, which connects the original state at $t = 0$ to the target-calibrated state at $t = 1$, with $t \sim \mathcal{U}(0,1)$ sampled during training. Its velocity network is

$$V_\theta = f_\theta\big(S, S^{\mathrm{ITC}}, \{f_i^I\}_i, \{f_j^T\}_j, \{h_{ij}^{\mathrm{IT}}\}_{i,j}, t\big), \tag{7}$$

where θ denotes the trainable network parameters and we set $S = S_t^{\mathrm{ref}}$ for flow-matching supervision. The trained flow model is then used to integrate from the initial score matrix S^{ITM} along the learned trajectory, yielding the calibrated score matrix S^{F}.

3.4 Training and Inference

Training proceeds in two stages on the filtered set produced by PSCM. We first optimize the MRA-enhanced discriminative retriever to construct the discriminative score space, then freeze it and train

PSCF for posterior retrieval-score calibration. Stage one minimizes $\mathcal{L}_{\mathrm{S1}} = \mathcal{L}_{\mathrm{ITC}} + \mathcal{L}_{\mathrm{ITM}}$, where $\mathcal{L}_{\mathrm{ITC}}$ and $\mathcal{L}_{\mathrm{ITM}}$ denote the image-text contrastive loss and image-text matching loss, respectively. For a batch of B pairs, the ITC objective aligns global embeddings through a bidirectional contrastive objective (Radford et al., 2021):

$$\mathcal{L}_{\mathrm{ITC}} = -\frac{1}{2B}\sum_{i=1}^{B}\left[\log\frac{\exp(S_{ii}^{\mathrm{ITC}}/\tau_c)}{\sum_{j=1}^{B}\exp(S_{ij}^{\mathrm{ITC}}/\tau_c)} + \log\frac{\exp(S_{ii}^{\mathrm{ITC}}/\tau_c)}{\sum_{k=1}^{B}\exp(S_{ki}^{\mathrm{ITC}}/\tau_c)}\right], \tag{8}$$

where $i = 1, \ldots, B$ and τ_c is the contrastive temperature. ITM predicts matching probabilities p_r^{ITM} from cross-modal embeddings and is optimized using

$$\mathcal{L}_{\mathrm{ITM}} = -\frac{1}{M}\sum_{r=1}^{M}\left[y_r^{\mathrm{ITM}}\log p_r^{\mathrm{ITM}} + (1 - y_r^{\mathrm{ITM}})\log(1 - p_r^{\mathrm{ITM}})\right], \tag{9}$$

where M denotes the number of candidate pairs and y_r^{ITM} is 1 for matches and 0 otherwise.

Stage two freezes the MRA-enhanced retriever to preserve its discriminative score space and optimizes only PSCF using $\mathcal{L}_{\mathrm{S2}} = \lambda_{\mathrm{FR}}\mathcal{L}_{\mathrm{FR}} + \lambda_{\mathrm{FM}}\mathcal{L}_{\mathrm{FM}}$, where $\mathcal{L}_{\mathrm{FR}}$ and $\mathcal{L}_{\mathrm{FM}}$ are the flow ranking and flow matching losses, weighted by λ_{FR} and λ_{FM}, respectively. For a batch of B pairs, the FM objective (Lipman et al., 2023) trains the velocity field to match the target corrections:

$$\mathcal{L}_{\mathrm{FM}} = \frac{1}{B}\sum_{i=1}^{B}\frac{\sum_j a_{ij}\left\|V_{\theta,ij} - R_{ij}^{*}\right\|_2^2}{\sum_j a_{ij}}, \tag{10}$$

where $V_{\theta,ij}$ denotes the correction predicted by the velocity field for the matching score between the i-th query and the j-th candidate at the reference state S_t^{ref}. We set $a_{ij} = 1$ for $j \in \mathcal{P}_i \cup \mathcal{H}_i$ and $a_{ij} = \beta_{\mathrm{bg}}$ for $j \in \mathcal{B}_i$, emphasizing corrections on positive and hard-negative pairs while reducing the influence of the remaining valid positions. Here, $\mathcal{B}_i$ is the set of valid candidates outside $\mathcal{P}_i \cup \mathcal{H}_i$, β_{bg} is their loss weight, and the sums over j run over valid candidates.

The FR objective supervises the ranking produced by the velocity field after integration, encouraging calibrated positive scores to exceed hard-negative scores:

$$\mathcal{L}_{\mathrm{FR}} = \frac{1}{B}\sum_{i=1}^{B}\frac{1}{|\mathcal{H}_i|}\sum_{j\in\mathcal{H}_i}\tau_{\mathrm{FR}}\log\left(1 + \exp\left(\frac{\mu + \widehat{S}_{ij} - \widehat{S}_{i,+}}{\tau_{\mathrm{FR}}}\right)\right), \tag{11}$$

where $\widehat{S}_{ij}$ denotes the (i,j)-th entry of the calibrated score matrix $\widehat{S} = S_{L_f}$, and $\widehat{S}_{i,+} = \max_{j\in\mathcal{P}_i}\widehat{S}_{ij}$ is the highest calibrated positive score for query i. Here, μ is the ranking margin, and τ_{FR} controls the smoothness of the margin loss. The matrix $\widehat{S}$ is obtained by integrating the learned velocity field from $S_0 = S^{\mathrm{ITM}}$ through the following Euler updates:

$$S_{\ell+1} = S_\ell + \frac{\rho}{L_f}V_\theta^{(\ell)}, \qquad \ell = 0, \ldots, L_f - 1, \tag{12}$$

where $V_\theta^{(\ell)}$ denotes the velocity predicted by f_θ with state input $S = S_\ell$ and time input $t = \ell/L_f$, L_f is the number of integration steps, and ρ controls the integration scale.

At inference, ITC first computes the global similarity matrix S^{ITC}. For each query, the top-K_{cand} candidates are retained, defining the valid-candidate mask M_{valid}, where $M_{\mathrm{valid},ij} = 1$ for selected pairs and 0 otherwise. ITM then evaluates these selected pairs (Li et al., 2022) and places their positive-class logits in S^{ITM}, with all other positions masked.

Starting from $S_0 = S^{\mathrm{ITM}}$, PSCF performs L_f Euler steps with step size $1/L_f$, updating only valid candidate entries. The resulting calibrated score matrix is

$$S^{\mathrm{F}} = S_{L_f} = S^{\mathrm{ITM}} + \frac{\rho}{L_f}\sum_{\ell=0}^{L_f-1}M_{\mathrm{valid}} \odot V_\theta^{(\ell)}, \tag{13}$$

where $V_\theta^{(\ell)}$ denotes the velocity predicted by f_θ at the current state S_ℓ and time $t_\ell = \ell/L_f$, and $\odot$ denotes elementwise multiplication. Invalid positions remain masked throughout the updates.

For stability, the final retrieval score retains a small contribution from the original ITM scores while assigning a larger weight to the PSCF-calibrated scores:

$$S_{\text{final}} = (1-\gamma)S^{\text{ITM}} + \gamma S^{\text{F}}, \tag{14}$$

where $1-\gamma$ denotes the smaller weight assigned to the ITM scores, while γ denotes the dominant weight assigned to the PSCF-calibrated scores. Retaining a small contribution from the ITM scores helps anchor the final scores to the discriminative matching scores produced by the frozen retriever. This design mitigates overly aggressive score adjustments while assigning greater weight to the calibrated scores. See Appendix A.2 for the training and inference algorithm.

4 Experiments

4.1 Datasets and Evaluation Metrics

We evaluate the efficacy of our proposed method using three datasets: RSICD (Lu et al., 2018), RSITMD (Yuan et al., 2022a), and NWPU (Cheng et al., 2022). Following standard protocols, we employ Recall at K ($\text{Recall@}K$, with $K = 1, 5, 10$) as our primary metric and report $\text{Recall@}K$ for both I2T and T2I retrieval. We further report the directional mean Recall, denoted as mR_{I2T} and mR_{T2I}, by averaging the three Recall values within each retrieval direction, and use the overall mean Recall (mR), averaged over all six Recall values, as the overall performance indicator. Best and second-best results are shown in bold and underlined, respectively.

4.2 Implementation Details

We initialize P^{2}Cal from pretrained BLIP with a ViT-B encoder at 384×384 resolution. We generate descriptions offline using Qwen2.5-VL-7B-Instruct. We insert MRA into the visual, textual, and cross-modal pathways with a residual scale of $s = 0.1$. We train the adapters, projection layers, and ITM head for 30 epochs in stage one, and then freeze the retriever and train PSCF for 10 epochs in stage two. We set $K_{\text{hard}} = 3$, $\mu = 0.2$, $\eta = 0.05$, $\lambda_{\text{FR}} = 1.0$, $\lambda_{\text{FM}} = 0.1$, $\beta_{\text{bg}} = 0.05$, $L_f = 3$, and $\rho = 1.0$. We use AdamW (Loshchilov & Hutter, 2019) with cosine learning-rate decay in both stages. We retain the top-128 candidates per retrieval direction using ITC at inference and compute the final score as $S_{\text{final}} = (1-\gamma)S^{\text{ITM}} + \gamma S^{\text{F}}$, with $\gamma = 0.9$. PSCM is used solely to facilitate training and is excluded from the inference process.

4.3 Compared Methods

We compare P^{2}Cal with representative RSITR and image-text retrieval methods in Table 1, including AMFMN (Yuan et al., 2022a), GaLR (Yuan et al., 2022b), PIR (Pan et al., 2023a), PE (Yuan et al., 2023), LAPS (Fu et al., 2024), FGVLA (Li et al., 2025a), EAPA (Li et al., 2025b), CUP (Wang et al., 2025), McRSA (Xiao et al., 2026), and RRSITR (Song et al., 2026). Most existing methods improve image-text discriminability through embedding enhancement and cross-modal alignment. Some approaches exploit additional prior knowledge, such as PIR (Pan et al., 2023a) and KAMCL (Ji et al., 2023), while others refine retrieval results from the posterior perspective through similarity calibration or reranking, such as GaLR (Yuan et al., 2022b).

4.4 Main Results

As shown in Table 1, P^{2}Cal achieves the best performance across all metrics on all three datasets. Compared with methods mainly focusing on embedding enhancement and cross-modal alignment, such as AMFMN (Yuan et al., 2022a), LAPS (Fu et al., 2024), and FGVLA (Li et al., 2025a), P^{2}Cal considers both the reliability of supervision for learning the retrieval space and the accuracy of the resulting ranking scores. Specifically, MRA facilitates pretrained knowledge transfer and discriminative embedding learning, while PSCM and PSCF address misleading image-text correspondences and residual score miscalibration, respectively. These complementary components enable more reliable supervision to support better embeddings, whose initial matching scores can be further optimized. P^{2}Cal also outperforms PIR (Pan et al., 2023a), which introduces prior scene instructions, and GaLR (Yuan et al., 2022b), which incorporates posterior re-ranking. These comparisons further demonstrate the benefit of connecting prior guidance with posterior optimization. In P^{2}Cal, the

Table 1: Quantitative comparison with existing methods on RSICD, RSITMD, and NWPU for image-to-text and text-to-image retrieval. Best results are shown in bold, and second-best results are underlined.

	RSICD Dataset				*RSITMD Dataset*				*NWPU Dataset*			
Method	$\text{Recall}_{\text{I2T}}$			mR_{I2T}	$\text{Recall}_{\text{I2T}}$			mR_{I2T}	$\text{Recall}_{\text{I2T}}$			mR_{I2T}
	@1	@5	@10		@1	@5	@10		@1	@5	@10	
AMFMN TGRS'22	5.95	16.83	27.72	16.83	12.17	27.43	39.38	26.33	14.03	41.75	60.06	38.61
GaLR TGRS'22	7.23	16.65	26.35	16.74	11.06	30.09	42.26	27.80	12.76	41.56	60.89	38.40
PIR ACMMM'23	9.06	24.89	36.14	23.36	18.14	38.72	51.11	35.99	20.89	55.65	72.51	49.68
PE TGRS'23	13.08	30.56	43.46	29.03	20.58	44.03	57.96	40.86	13.71	39.30	56.10	36.37
LAPS CVPR'24	10.16	23.97	36.60	23.58	15.71	34.96	48.67	33.11	15.37	45.56	64.76	41.90
FGVLA TGRS'25	16.01	33.21	47.03	32.08	26.77	47.12	61.28	45.06	14.92	44.13	62.48	40.51
EAPA ACMMM'25	13.36	30.92	43.92	29.40	22.79	45.80	57.96	42.18	25.81	62.19	78.54	55.51
CUP TNNLS'25	7.96	24.70	37.51	23.39	18.58	36.06	51.55	35.40	9.30	30.03	45.94	28.42
McRSA TMM'26	16.74	34.77	48.40	33.30	23.89	46.46	58.41	42.92	24.95	58.70	75.21	52.95
RRSITR CVPR'26	15.55	32.85	46.94	31.78	24.12	44.25	55.53	41.30	24.41	59.37	76.10	53.29
P^2Cal	**25.71**	**45.11**	**57.82**	**42.88**	**30.09**	**52.65**	**64.38**	**49.04**	**27.37**	**62.25**	**79.05**	**56.22**
Method	$\text{Recall}_{\text{T2I}}$			mR_{T2I}	$\text{Recall}_{\text{T2I}}$			mR_{T2I}	$\text{Recall}_{\text{T2I}}$			mR_{T2I}
	@1	@5	@10		@1	@5	@10		@1	@5	@10	
AMFMN TGRS'22	4.56	17.33	27.98	16.62	10.04	32.61	45.31	29.32	9.59	32.13	48.23	29.98
GaLR TGRS'22	4.43	15.59	26.09	15.37	10.58	35.18	51.99	32.58	9.54	32.43	49.07	30.35
PIR ACMMM'23	6.28	22.16	36.21	21.55	13.58	41.28	64.38	39.75	11.61	36.97	53.08	33.89
PE TGRS'23	10.14	33.16	49.50	30.93	21.90	50.97	67.70	46.86	9.18	29.54	44.47	27.73
LAPS CVPR'24	6.88	23.81	38.79	23.16	14.25	41.06	58.67	37.99	10.26	33.24	49.59	31.03
FGVLA TGRS'25	11.64	33.74	51.47	32.28	21.90	54.65	72.83	49.79	9.74	31.59	47.33	29.55
EAPA ACMMM'25	11.00	32.11	47.30	30.14	20.66	51.90	67.74	46.77	14.25	40.39	56.00	36.88
CUP TNNLS'25	7.32	26.77	41.74	25.28	14.34	43.58	63.10	40.34	6.90	24.14	37.73	22.92
McRSA TMM'26	13.89	37.38	53.63	34.97	22.30	52.35	67.92	47.52	14.62	40.77	56.65	37.35
RRSITR CVPR'26	10.85	31.56	45.78	29.40	20.53	49.47	66.90	45.63	14.53	40.32	56.03	36.96
P^2Cal	**18.46**	**44.17**	**59.19**	**40.61**	**27.35**	**59.03**	**75.97**	**54.12**	**14.98**	**41.34**	**56.68**	**37.67**

purified supervision obtained by PSCM is shared by both MRA and PSCF, thereby linking discriminative space learning with retrieval-score calibration and providing a coherent basis for improving image-text correspondence reliability across the framework.

4.5 Ablation Analysis

To validate the effectiveness of P^2Cal, we conduct ablation experiments on RSICD under clean and severely noisy training correspondences (Tables 2 and 3). Under clean supervision, MRA-based adaptation provides the largest gain over the unadapted backbone, consistent with its role in transferring pretrained BLIP knowledge to the remote sensing domain through lightweight residual adaptation of a frozen backbone. Removing any visual, textual, or cross-modal adapter reduces both directional means, with the largest reduction occurring when cross-modal adaptation is removed. These results support MRA's role in establishing the discriminative score space on which P^2Cal builds its collaborative calibration.

The remaining comparisons show how prior and posterior calibration further strengthen this foundation. PSCM substantially improves retrieval under severe correspondence noise by supplying purified supervision to both MRA and PSCF, establishing the shared supervisory basis of the two stages. PSCF further improves both retrieval directions under clean and noisy conditions, with larger overall mR gains when combined with PSCM. This supports the motivation of P^2Cal: reliable prior supervision facilitates discriminative embedding learning and also guides subsequent score refinement. The complete model achieves the strongest overall and directional mean Recall in both settings, supporting the effectiveness of P^2Cal through the complementary contributions of PSCM, MRA, and PSCF. See Appendix A.4 for the noisy-training protocol.

Table 2: Ablation study on RSICD with 0% correspondence noise.

D#	*PSCM*	*MRA*			*PSCF*	*RSICD*		
		Vis	*Text*	*Cross*		mR_{I2T}	mR_{T2I}	mR
1	×	×	×	×	×	18.23	21.91	20.07
2	×	✓	✓	✓	×	42.00	39.26	40.63
3	✓	✓	✓	✓	×	42.36	39.09	40.73
4	×	✓	✓	✓	✓	42.57	39.73	41.15
5	✓	×	✓	✓	✓	38.46	37.30	37.88
6	✓	✓	×	✓	✓	40.43	38.07	39.25
7	✓	✓	✓	×	✓	37.15	36.65	36.90
8	✓	✓	✓	✓	✓	42.88	40.61	41.75

Table 3: Ablation study on RSICD with 90% correspondence noise.

D#	*PSCM*	*MRA*			*PSCF*	*RSICD*		
		Vis	*Text*	*Cross*		mR_{I2T}	mR_{T2I}	mR
1	×	×	×	×	×	18.23	21.91	20.07
2	×	✓	✓	✓	×	7.72	5.82	6.77
3	✓	✓	✓	✓	×	28.15	29.92	29.04
4	×	✓	✓	✓	✓	8.11	6.78	7.45
5	✓	×	✓	✓	✓	27.88	28.55	28.21
6	✓	✓	×	✓	✓	28.58	29.38	28.98
7	✓	✓	✓	×	✓	26.96	25.69	26.33
8	✓	✓	✓	✓	✓	29.55	31.01	30.28

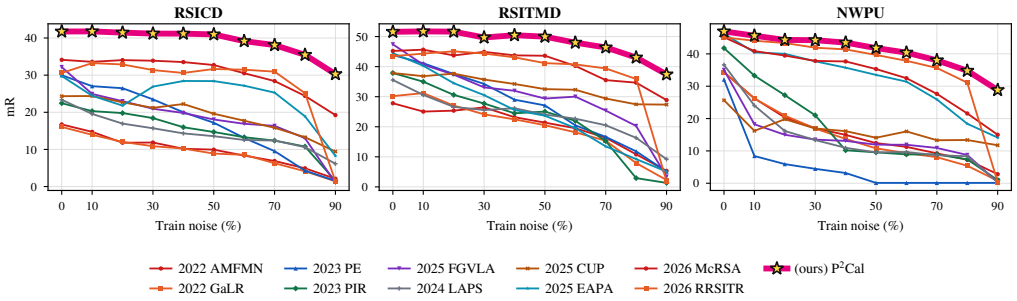

Figure 3: Robustness comparison under correspondence noise from 0% to 90% on RSICD, RSITMD, and NWPU. mR is reported for P^{2}Cal and competing methods.

Table 4: Cross-dataset generalization: trained on NWPU and evaluated on RSICD and RSITMD. Recall results are reported for I2T and T2I retrieval, with $\mathrm{mR}_{\mathrm{I2T}}$ and $\mathrm{mR}_{\mathrm{T2I}}$ denoting directional mean Recall.

Method	*NWPU→RSICD*								*NWPU→RSITMD*							
	$\mathrm{Recall}_{\mathrm{I2T}}$			$\mathrm{mR}_{\mathrm{I2T}}$	$\mathrm{Recall}_{\mathrm{T2I}}$			$\mathrm{mR}_{\mathrm{T2I}}$	$\mathrm{Recall}_{\mathrm{I2T}}$			$\mathrm{mR}_{\mathrm{I2T}}$	$\mathrm{Recall}_{\mathrm{T2I}}$			$\mathrm{mR}_{\mathrm{T2I}}$
	@1	@5	@10		@1	@5	@10		@1	@5	@10		@1	@5	@10	
AMFMN TGRS'22	5.67	13.36	23.06	14.03	3.84	16.10	25.56	15.17	8.85	22.57	30.53	20.65	6.06	21.50	34.38	20.65
GaLR TGRS'22	6.13	15.28	23.06	14.82	3.31	12.88	21.94	12.71	7.74	19.25	30.31	19.10	5.35	18.72	28.67	17.58
PE TGRS'23	9.42	22.42	33.67	21.84	8.38	25.53	37.69	23.87	15.49	32.30	45.35	31.05	12.43	35.62	51.90	33.32
PIR ACMMM'23	7.69	19.03	28.82	18.51	4.90	17.11	27.15	16.39	12.39	28.54	38.50	26.48	8.76	26.06	38.23	24.35
FGVLA TGRS'25	9.06	22.51	32.02	21.20	6.92	21.45	34.00	20.79	14.16	31.64	42.70	29.50	11.81	33.32	49.34	31.49
LAPS CVPR'24	8.42	23.70	33.12	21.75	6.46	23.88	36.12	22.15	13.27	30.75	43.36	29.13	10.18	33.45	48.45	30.69
CUP TNNLS'25	10.52	25.43	36.78	24.24	8.64	27.58	42.01	26.08	13.94	37.39	48.67	33.33	14.73	42.30	58.58	38.54
EAPA ACMMM'25	12.81	30.38	42.09	28.43	10.67	28.84	41.68	27.06	19.03	39.16	51.55	36.58	16.64	39.65	54.07	36.79
McRSA TMM'26	11.71	27.63	40.62	26.65	9.26	28.62	41.98	26.62	18.58	36.28	51.11	35.32	15.71	38.89	53.81	36.14
RRSITR CVPR'26	12.72	30.10	40.81	27.88	10.80	29.75	41.32	27.29	19.03	35.84	47.79	34.22	15.88	39.47	54.47	36.61
P^{2}Cal	**21.23**	**39.52**	**51.51**	**37.42**	**15.72**	**36.74**	**51.05**	**34.50**	**25.88**	**44.91**	**57.08**	**42.62**	**20.71**	**47.35**	**62.43**	**43.50**

4.6 Generalization and Robustness

To evaluate the generalization capability and robustness of P^{2}Cal, we conduct cross-dataset transfer and correspondence noise robustness experiments. We train on NWPU and directly evaluate on RSICD and RSITMD without fine-tuning. As shown in Table 4, P^{2}Cal achieves superior transfer performance, indicating that it learns generalizable image-text correspondence rather than dataset-specific patterns. This is attributed to the joint design of prior supervision calibration, discriminative score-space construction, and posterior retrieval-score refinement, enabling robust cross-modal alignment. Furthermore, we evaluate robustness under correspondence noise. As shown in Figure 3, noise ratios range from 0.0% to 90.0% with a step of 10.0%. P^{2}Cal achieves the best performance across all noise levels on RSICD, RSITMD, and NWPU, consistently outperforming competing methods. While existing approaches degrade significantly under noisy supervision, P^{2}Cal mitigates noise effects through calibration, discriminative score learning, and posterior score calibration. Even at 90.0% correspondence noise, P^{2}Cal maintains a clear advantage, demonstrating strong robustness for remote sensing image-text retrieval. See Appendix A.4 for the correspondence-noise protocol and detailed results.

5 Conclusion

In this paper, we propose P^{2}Cal, a prior-posterior collaborative calibration framework for robust remote sensing image-text retrieval. The PSCM leverages visual semantic priors to purify supervision and provide reliable signals for training MRA and PSCF. MRA incorporates lightweight manifold residual adapters into representation pathways to construct a stable, domain-adaptive discriminative score space. Based on matching scores, PSCF learns a continuous residual velocity field to optimize retrieval scores, promoting reliable positives and suppressing hard negatives. Together, these components couple reliable correspondence learning with retrieval-score calibration. Extensive experiments on multiple RSITR datasets demonstrate strong retrieval performance and validate the effectiveness of the method we proposed.

AI USE STATEMENT

We used generative AI tools to assist with manuscript preparation, including checking language and notation consistency, supporting searches for relevant literature on remote sensing image-text retrieval, learning with noisy correspondences, and retrieval-score calibration, and improving the organization and presentation of the appendix. Within the proposed method, Qwen2.5-VL-7B-Instruct generates auxiliary descriptions offline for PSCM, as described in Sections 3.1 and 4.2. The authors are responsible for reviewing and verifying AI-assisted content, including the accuracy and relevance of references and the consistency of the text and figures with the actual method. The authors take full responsibility for the accuracy, originality, and academic integrity of the final manuscript.

ETHICS STATEMENT

This study focuses on remote sensing image-text retrieval using the publicly available RSICD, RSITMD, and NWPU datasets. The expanded-gallery evaluation additionally uses RS5M images and texts as distractors, as described in Appendix A.8. No human-subject experiments or additional collection of personally identifiable or sensitive personal data are involved. The data are used for academic research and model evaluation, including bidirectional retrieval, robustness to noisy training correspondences, and cross-dataset generalization. We follow research-integrity principles throughout data use, experimental design, result analysis, and manuscript preparation.

REPRODUCIBILITY STATEMENT

The main manuscript and appendix document the proposed P^2Cal framework and its experimental evaluation. Section 3 presents the Prior Semantic Calibration Module (PSCM), Manifold Residual Adapter (MRA), and Posterior Score Calibration Flow (PSCF), together with the training objectives and inference procedure. Sections 4.1 and 4.2 describe the datasets, evaluation metrics, pretrained models, training settings, and inference configuration. Appendix A.1 details noise-estimator construction and empirical calibration, and Appendix A.2 provides the training and inference algorithm, while Appendices A.3-A.5 supplement the dataset statistics, metric definitions, training and evaluation protocols, and scene-level diagnostics. Appendices A.6-A.8 provide qualitative examples and protocols and results for visual-degradation and expanded-gallery evaluations. These descriptions and reported comparisons are intended to support independent examination of the method and reproduction of the experimental results.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Yi Chang, Luxin Yan, Houzhang Fang, and Chunan Luo. Anisotropic spectral-spatial total variation model for multispectral remote sensing image destriping. *IEEE Transactions on Image Processing*, 2015.

Qimin Cheng, Haiyan Huang, Yuan Xu, Yuzhuo Zhou, Huanying Li, and Zhongyuan Wang. NWPU-Captions dataset and MLCA-Net for remote sensing image captioning. *IEEE Transactions on Geoscience and Remote Sensing*, 2022.

Mingmin Chi, Antonio Plaza, Jón Atli Benediktsson, Zhongyi Sun, Jinsheng Shen, and Yangyong Zhu. Big data for remote sensing: Challenges and opportunities. *Proceedings of the IEEE*, 2016.

Zheren Fu, Lei Zhang, Hou Xia, and Zhendong Mao. Linguistic-aware patch slimming framework for fine-grained cross-modal alignment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Pascal Gwosdek, Sven Grewenig, Andrés Bruhn, and Joachim Weickert. Theoretical foundations of Gaussian convolution by extended box filtering. In *International Conference on Scale Space and Variational Methods in Computer Vision*, 2012.

Kaiming He, Jian Sun, and Xiaoou Tang. Single image haze removal using dark channel prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning*, 2019.

Zhong Ji, Changxu Meng, Yan Zhang, Yanwei Pang, and Xuelong Li. Knowledge-aided momentum contrastive learning for remote-sensing image text retrieval. *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

Karen E. Joyce, Stella E. Belliss, Sergey V. Samsonov, Stephen J. McNeill, and Phil J. Glassey. A review of the status of satellite remote sensing and image processing techniques for mapping natural hazards and disasters. *Progress in Physical Geography*, 2009.

Robert G. Keys. Cubic convolution interpolation for digital image processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1981.

Charis Lanaras, José Bioucas-Dias, Silvano Galliani, Emmanuel Baltsavias, and Konrad Schindler. Super-resolution of Sentinel-2 images: Learning a globally applicable deep neural network. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2018.

Mario Lezcano-Casado and David Martínez-Rubio. Cheap orthogonal constraints in neural networks: A simple parametrization of the orthogonal and unitary group. In *International Conference on Machine Learning*, 2019.

Jin Li, Zilong Liu, and Si Liu. Suppressing the image smear of the vibration modulation transfer function for remote-sensing optical cameras. *Applied Optics*, 2017.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, 2022.

Shuo Li, Haoyang Ji, Fang Liu, Licheng Jiao, Xutong Min, Xinyan Huang, Jiahao Wang, Long Sun, Lingling Li, and Xu Liu. Fine-grained visual-language alignment for remote sensing image–text retrieval. *IEEE Transactions on Geoscience and Remote Sensing*, 2025a.

Shuoshuo Li, Shuli Cheng, and Liejun Wang. Entity-level alignment with prompt-guided adapter for remote sensing image-text retrieval. In *ACM International Conference on Multimedia*, 2025b.

Yansheng Li, Jiayi Ma, and Yongjun Zhang. Image retrieval from remote sensing big data: A survey. *Information Fusion*, 2021.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

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

Xiaoqiang Lu, Binqiang Wang, Xiangtao Zheng, and Xuelong Li. Exploring models and data for remote sensing image caption generation. *IEEE Transactions on Geoscience and Remote Sensing*, 2018.

Qing Ma, Jiancheng Pan, and Cong Bai. Direction-oriented visual-semantic embedding model for remote sensing image-text retrieval. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

Jiancheng Pan, Qing Ma, and Cong Bai. A prior instruction representation framework for remote sensing image-text retrieval. In *ACM International Conference on Multimedia*, 2023a.

Zhengxin Pan, Fangyu Wu, and Bailing Zhang. Fine-grained image-text matching by cross-modal hard aligning network. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023b.

Manjun Qin, Fengying Xie, Wei Li, Zhenwei Shi, and Haopeng Zhang. Dehazing for multispectral remote sensing images based on a convolutional neural network with the residual architecture. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2018.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

Haitao Shi, Meng Liu, Xiaoxuan Mu, Xuemeng Song, Yupeng Hu, and Liqiang Nie. Breaking through the noisy correspondence: A robust model for image-text matching. *ACM Transactions on Information Systems*, 2024.

Qiya Song, Yiqiang Xie, Yuan Sun, Renwei Dian, and Xudong Kang. Robust remote sensing image-text retrieval with noisy correspondence. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Xu Tang, Yijing Wang, Jingjing Ma, Xiangrong Zhang, Fang Liu, and Licheng Jiao. Interacting-enhancing feature Transformer for cross-modal remote-sensing image and text retrieval. *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

Gregory K. Wallace. The JPEG still picture compression standard. *Communications of the ACM*, 1991.

Yijing Wang, Xu Tang, Jingjing Ma, Xiangrong Zhang, Fang Liu, and Licheng Jiao. Cross-modal remote sensing image–text retrieval via context and uncertainty-aware prompt. *IEEE Transactions on Neural Networks and Learning Systems*, 2025.

Yi Xiao, Jun Zhang, and Aiguo Song. Multi-constraint relational semantic alignment toward image-text retrieval. *IEEE Transactions on Multimedia*, 2026.

Yuan Yuan, Yang Zhan, and Zhitong Xiong. Parameter-efficient transfer learning for remote sensing image-text retrieval. *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

Zhiqiang Yuan, Wenkai Zhang, Kun Fu, Xuan Li, Chubo Deng, Hongqi Wang, and Xian Sun. Exploring a fine-grained multiscale method for cross-modal remote sensing image retrieval. *IEEE Transactions on Geoscience and Remote Sensing*, 2022a.

Zhiqiang Yuan, Wenkai Zhang, Changyuan Tian, Xuee Rong, Zhengyuan Zhang, Hongqi Wang, Kun Fu, and Xian Sun. Remote sensing cross-modal text-image retrieval based on global and local information. *IEEE Transactions on Geoscience and Remote Sensing*, 2022b.

A. Zabala and X. Pons. Effects of lossy compression on remote sensing image classification of forest areas. *International Journal of Applied Earth Observation and Geoinformation*, 2011.

Weihang Zhang, Jihao Li, Shuoke Li, Jialiang Chen, Wenkai Zhang, Xin Gao, and Xian Sun. Hypersphere-based remote sensing cross-modal text–image retrieval via curriculum learning. *IEEE Transactions on Geoscience and Remote Sensing*, 2023.

Zilun Zhang, Tiancheng Zhao, Yulong Guo, and Jianwei Yin. RS5M and GeoRSCLIP: A large-scale vision-language dataset and a large vision-language model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

Fuzhong Zheng, Xu Wang, Luyao Wang, Xiong Zhang, Hongze Zhu, Long Wang, and Haisu Zhang. A fine-grained semantic alignment method specific to aggregate multi-scale information for cross-modal remote sensing image retrieval. *Sensors*, 2023.

Wenqian Zhou, Hanlin Wu, and Pei Deng. Toward efficient and accurate remote sensing image–text retrieval with a coarse-to-fine approach. *IEEE Geoscience and Remote Sensing Letters*, 2025.

A Appendix

A.1 Details of Prior Semantic Calibration

This section supplements Section 3.1 with noise-estimator construction and empirical calibration. We retain the standard dataset splits. Synthetic copies of the clean validation subset are used to fit the estimator and calibration mapping, while the original clean subset supports checkpoint selection. This reuses existing validation annotations without requiring a separate calibration set. Synthetic validation pairs are not added to the training set for MRA and PSCF, and the test split is reserved for final evaluation.

Following Appendix A.4.1, we construct five validation copies with text-shuffling fractions $\rho_{\text{noise}} \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. Each copy contains N_{val} image-text pairs. Let $\widetilde{T}_{ij}$ denote the reassigned text and $\mathcal{T}_i$ the set of original clean text strings associated with image I_i. The noise label and corresponding corruption ratio are

$$y_{ij} = \mathbf{1}\left[\widetilde{T}_{ij} \notin \mathcal{T}_i\right], \qquad r = \frac{1}{N_{\text{val}}} \sum_{i,j} y_{ij}. \tag{15}$$

Here, $y_{ij} = 1$ indicates a noisy correspondence, and r is the resulting mismatch fraction used for calibration.

For each copy, we recompute q_{ij} and $c_{ij\ell}$ using its assigned texts and the fixed auxiliary descriptions G_i, obtaining $\boldsymbol{\phi}_{ij}$ as in Section 3.1. We pool the five copies and standardize each embedding coordinate using its pooled mean and standard deviation. The normalized embeddings enter the same estimator, $p_{ij}^{\text{noise}} = \sigma_{\text{sig}}(\boldsymbol{w}^\top \widetilde{\boldsymbol{\phi}}_{ij} + b)$. Indexing the $N_{\text{syn}} = 5N_{\text{val}}$ pooled records by n, we denote their labels and predictions by y_n and p_n^{noise}. The estimator is trained with class-balanced binary cross-entropy and ℓ_2 regularization:

$$\mathcal{L}_{\text{PSCM}} = -\frac{1}{N_{\text{syn}}} \sum_{n=1}^{N_{\text{syn}}} \omega_{y_n} \left[y_n \log p_n^{\text{noise}} + (1 - y_n) \log(1 - p_n^{\text{noise}}) \right] + \frac{\lambda_{\text{PSCM}}}{2} \|\boldsymbol{w}\|_2^2. \tag{16}$$

The sample weight ω_{y_n} is $\omega_0 = N_{\text{syn}}/(2N_0)$ for a clean record ($y_n = 0$) and $\omega_1 = N_{\text{syn}}/(2N_1)$ for a noisy record ($y_n = 1$), where N_0 and N_1 count clean and noisy records, respectively. This gives the two classes equal total weight. We set $\lambda_{\text{PSCM}} = 10^{-4}$ and keep the normalization statistics and fitted estimator fixed after fitting.

For calibration, we evaluate the same five synthetic copies and the original clean validation subset. Each provides a mean predicted probability $\bar{p}_r^{\text{noise}}$ paired with its known corruption ratio r, with $r = 0$ for the clean subset. Sorting these six points by mean predicted probability gives $(\bar{p}_k^{\text{noise}}, r_k)$, $k = 0, \ldots, 5$. We set $\widehat{r}_k = \max\{r_0, \ldots, r_k\}$ to make the calibration values nondecreasing, then interpolate between adjacent points:

$$\mathcal{C}_{\text{val}}\left(\bar{p}^{\text{noise}}\right) = \widehat{r}_k + \frac{\bar{p}^{\text{noise}} - \bar{p}_k^{\text{noise}}}{\bar{p}_{k+1}^{\text{noise}} - \bar{p}_k^{\text{noise}}} \left(\widehat{r}_{k+1} - \widehat{r}_k\right). \tag{17}$$

Here, $\bar{p}^{\text{noise}}$ is the input mean predicted noise probability, with $\bar{p}_k^{\text{noise}} \le \bar{p}^{\text{noise}} \le \bar{p}_{k+1}^{\text{noise}}$, $\bar{p}_k^{\text{noise}} < \bar{p}_{k+1}^{\text{noise}}$, and $k = 0, \ldots, 4$. Inputs outside the observed range take the corresponding endpoint value.

On the training set, $\zeta = \mathcal{C}_{\text{val}}(\bar{p}^{\text{noise}})$ determines the filtering budget, while p_{ij}^{noise} determines which pairs are removed, following Section 3.1. Across the three datasets, calibration uses existing validation annotations. For a new dataset, a small, manually verified subset can instead be drawn from development data after reserving an independent test set. This subset should represent the training data's scene content, text characteristics, and correspondence noise. Future work will explore smaller calibration budgets and more direct noise estimation from cross-modal consistency and model uncertainty.

A.2 Training and Inference Algorithm

To complement Sections 3.1-3.4, Algorithm 1 summarizes the operations and parameter updates from PSCM filtering through MRA training to PSCF calibration and inference. Detailed mathematical definitions are given in the corresponding Method subsections.

Algorithm 1 Training and Inference of P^{2}Cal

Require: Image-text training set $\mathcal{D} = \{(I_i, T_{ij})\}_{i,j}$; pretrained BLIP; candidate count K_{cand}; hard-negative count K_{hard}; flow steps L_f.
Ensure: Calibrated retrieval score matrix S_{final}.
Phase 0: Prior semantic calibration with PSCM
Fit the noise estimator and calibration mapping $\mathcal{C}_{\text{val}}$ on synthetic noisy validation data.
For each training image I_i, generate an auxiliary description G_i with Qwen2.5-VL.
Compare each text T_{ij} with G_i and with the other texts of I_i to obtain q_{ij} and $c_{ij\ell}$; aggregate these agreements into $\boldsymbol{\phi}_{ij}$ and predict p_{ij}^{noise} from its normalized embeddings.
Estimate $\zeta = \mathcal{C}_{\text{val}}(\bar{p}^{\text{noise}})$ and discard the $\text{round}(\zeta N_{\text{pair}})$ highest-noise pairs to obtain $\mathcal{D}'$.
Phase 1: Discriminative score-space construction with MRA
Freeze BLIP's backbone and insert MRA into its visual, textual, and cross-modal pathways. During each forward pass, apply the following operations at each adapted feed-forward layer (Section 3.2):
(a) Down-project the post-attention tokens X to bottleneck embeddings Z; mean-pool X into context h and map h to sample-specific vectors u, v.
(b) Form the skew-symmetric operator $A = uv^\top - vu^\top$ and rotate the tokens by $\widetilde{Z} = Z\exp(A)$, evaluated using the closed-form expression in Section 3.2.
(c) Activate and up-project $\widetilde{Z}$, then add the s-scaled residual to the original feed-forward output.
for each training mini-batch from $\mathcal{D}'$ **do**
  Obtain projected global embeddings f_i^I, f_j^T; form cosine similarities S^{ITC} and compute $\mathcal{L}_{\text{ITC}}$.
  Select top-K_{cand} ITC pairs; obtain cross-modal `[CLS]` embeddings h_{ij}^{IT}, ITM matching probabilities, and positive-class logits S^{ITM}.
  Compute $\mathcal{L}_{\text{ITM}}$; update MRA parameters, projection layers, and the ITM head by minimizing $\mathcal{L}_{\text{S1}} = \mathcal{L}_{\text{ITC}} + \mathcal{L}_{\text{ITM}}$.
end for
Phase 2: Posterior score calibration with PSCF
Freeze the MRA-enhanced discriminative retriever.
for each training mini-batch from $\mathcal{D}'$ **do**
  Obtain frozen S^{ITM}, S^{ITC}, global embeddings f_i^I, f_j^T, and cross-modal embeddings h_{ij}^{IT}.
  for each query i **do**
    Identify positive positions $\mathcal{P}_i$ and select at most K_{hard} highest-scoring negatives as $\mathcal{H}_i$.
    Compare hard-negative scores with the best positive score to obtain correction magnitudes e_{ij}, using margin μ and offset η (Section 3.3).
    For nonempty $\mathcal{P}_i, \mathcal{H}_i$, construct R^* using Section 3.3: decrease each hard-negative score by $e_{ij}/|\mathcal{H}_i|$, distribute an equal total increase evenly over the positives, and assign zero elsewhere.
  end for
  Sample $t \sim \mathcal{U}(0,1)$; form a reference state between the initial and target scores: $S_t^{\text{ref}} = S^{\text{ITM}} + tR^*$.
  Condition the velocity network on S_t^{ref}, S^{ITC}, the frozen global and cross-modal embeddings, and t; use $\mathcal{L}_{\text{FM}}$ to match its predicted velocity V_θ to R^*.
  Integrate from $S_0 = S^{\text{ITM}}$ to $\widehat{S} = S_{L_f}$ (Section 3.4); use $\mathcal{L}_{\text{FR}}$ to encourage the highest positive score to exceed hard-negative scores.
  Update only PSCF parameters θ by minimizing $\mathcal{L}_{\text{S2}} = \lambda_{\text{FR}}\mathcal{L}_{\text{FR}} + \lambda_{\text{FM}}\mathcal{L}_{\text{FM}}$.
end for
Inference: calibrated retrieval decision
Obtain top-K_{cand} ITC candidates, evaluate their ITM logits, and build S^{ITM} with valid mask M_{valid}.
Initialize $S_0 = S^{\text{ITM}}$.
for $\ell = 0$ to $L_f - 1$ **do**
  Predict $V_\theta^{(\ell)}$ at state S_ℓ and time ℓ/L_f; update only valid candidate scores: $S_{\ell+1} = S_\ell + \frac{\rho}{L_f} M_{\text{valid}} \odot V_\theta^{(\ell)}$.
end for
Set $S^{\text{F}} = S_{L_f}$; blend the initial and calibrated scores for retrieval ranking: $S_{\text{final}} = (1-\gamma)S^{\text{ITM}} + \gamma S^{\text{F}}$.
return S_{final}.

A.3 Details of Datasets

To explain the complementary evaluation conditions provided by the three datasets, this section describes the scale, scene coverage, text characteristics, and visual diversity of RSICD, RSITMD, and NWPU.

Tables 5, 6, and 7 list scene-wise image counts to clarify the category coverage and imbalance behind the scene-level experiments in Appendix A.5. The corresponding category-wise Recall@5 results

are reported in Tables 11-16. Experimental settings and metrics are introduced with the experiments below.

RSICD. RSICD is a remote sensing image-text dataset collected for image description generation and later widely reused for image-text retrieval evaluation (Lu et al., 2018). It contains 10,921 images from 30 scene categories, and each image is associated with five natural language texts. Its diverse scene distribution makes it suitable for evaluating general remote sensing image-text retrieval performance.

Table 5: Scene-wise image statistics of the RSICD dataset.

RSICD Dataset Categories and Counts			
Category	*#*	*Category*	*#*
airport	420	farmland	370
bare land	310	forest	250
baseball field	276	industrial	390
beach	400	meadow	280
bridge	459	medium residential	290
center	260	mountain	340
church	240	commercial	350
dense residential	410	desert	300
park	350	parking	390
playground	1031	pond	420
port	389	railway station	260
resort	290	river	410
school	300	sparse residential	300
square	330	stadium	290
storage tanks	396	viaduct	420
		Total	**10,921**

RSITMD. RSITMD is a fine-grained remote sensing image-text matching dataset introduced with AMFMN (Yuan et al., 2022a). It contains 4,743 images from 32 scene categories, with five texts for each image. Compared with RSICD, RSITMD contains more fine-grained textual descriptions involving objects, attributes, and spatial relations, making it useful for testing cross-modal matching under detailed text semantics.

Table 6: Scene-wise image statistics of the RSITMD dataset.

RSITMD Dataset Categories and Counts			
Category	*#*	*Category*	*#*
industrial area	207	port	178
stadium	201	farmland	177
storage tank	201	resort	166
square	196	school	165
amusement park	191	park	160
river	188	dense residential	157
commercial bridge	188	sparse residential	157
pond	184	bridge	155
seashore	154	desert	126
commercial area	148	forest	116
center	146	railway station	112
parking lot	143	mountain	109
airport	140	baseball field	105
church	136	crossroad	67
medium residential	130	bare land	58
meadow	127	ship	55
		Total	**4,743**

NWPU. NWPU is constructed from NWPU remote sensing scene imagery and contains 31,500 images from 45 scene categories (Cheng et al., 2022). Each category contains 700 images, and each image is paired with five texts. Its larger scale and broader scene coverage make it useful for evaluating retrieval robustness across diverse remote sensing visual concepts.

Table 7: Scene-wise image statistics of the NWPU dataset.

NWPU Dataset Categories and Counts			
Category	*#*	*Category*	*#*
airplane	700	airport	700
baseball diamond	700	basketball court	700
beach	700	bridge	700
chaparral	700	church	700
circular farmland	700	cloud	700
commercial area	700	dense residential	700
desert	700	forest	700
freeway	700	golf course	700
ground track field	700	harbor	700
industrial area	700	intersection	700
island	700	lake	700
meadow	700	medium residential	700
mobile home park	700	mountain	700
overpass	700	palace	700
parking lot	700	railway	700
railway station	700	rectangular farmland	700
river	700	roundabout	700
runway	700	sea ice	700
ship	700	snowberg	700
sparse residential	700	stadium	700
storage tank	700	tennis court	700
terrace	700	thermal power station	700
wetland	700		
		Total	**31,500**

A.4 Details of Correspondence-Noise Robustness Experiments

To supplement the correspondence-noise robustness experiments in Section 4.6, we detail the experimental settings and provide the retrieval results underlying Figure 3.

A.4.1 Details of Experimental Settings

To detail the experimental setup, we follow RRSITR (Song et al., 2026) in introducing correspondence noise through text shuffling within the training split. The dataset partitions remain the same as in the main experiments, and the validation and test splits remain clean and fixed across noise levels. Let $\mathcal{D} = \{(I_i, T_{ij})\}_{i,j}$ denote the original training set, as in the Method section, and let $N_{\text{pair}} = |\mathcal{D}|$ be its number of image-text pairs. For $\rho_{\text{noise}} \in \{10\%, 20\%, \ldots, 90\%\}$, we randomly select $\lfloor \rho_{\text{noise}} N_{\text{pair}} \rfloor$ distinct pair positions and shuffle the texts among these positions, with their images held fixed. Writing the reassigned text as $\widetilde{T}_{ij}$, the selected pair becomes $(I_i, \widetilde{T}_{ij})$ in the noisy training set $\mathcal{D}_{\text{noise}}$. Selection operates on individual pairs, so texts associated with the same image need not be reassigned together. Image pixels and unselected pairs remain unchanged.

For each noise ratio, we train P^{2}Cal on four NVIDIA A100 GPUs using the training settings in Section 4.2. We select the best checkpoint according to performance on the clean validation split and report its results on the original clean test split; all six Recall values come from the same selected checkpoint. All methods are evaluated on the same test split for each dataset. Thus, changes in performance reflect the ability to learn from unreliable training correspondences under a fixed evaluation protocol. Following Section 4.1, Recall@K measures the proportion of queries with at least one ground-truth match among the top K results, with $K \in \{1, 5, 10\}$, and is reported as a percentage. The directional means mR_{I2T} and mR_{T2I} average the three Recall values within each direction; overall mR is the mean of all six values.

Algorithm 2 Text-Level Training Correspondence Corruption

Require: Training set $\mathcal{D} = \{(I_i, T_{ij})\}_{i,j}$; requested noise ratio ρ_{noise}
Ensure: Noisy training set $\mathcal{D}_{\text{noise}}$ with unchanged image pixels
1: Initialize $\mathcal{D}_{\text{noise}} \leftarrow \mathcal{D}$
2: Set $N_{\text{pair}} \leftarrow |\mathcal{D}|$
3: Uniformly select $\lfloor \rho_{\text{noise}} N_{\text{pair}} \rfloor$ pair positions without replacement
4: Copy the texts at the selected positions and randomly permute this copy
5: **for** each selected pair position (i, j) **do**
6: Let $\widetilde{T}_{ij}$ be the text assigned to this position by the permutation
7: Replace (I_i, T_{ij}) with $(I_i, \widetilde{T}_{ij})$ in $\mathcal{D}_{\text{noise}}$
8: **end for**
9: Keep every unselected correspondence unchanged
10: Keep the validation and test splits clean
11: **return** $\mathcal{D}_{\text{noise}}$

A.4.2 Detailed Results and Analysis

To complement the mR curves in Figure 3, Tables 8, 9, and 10 report the complete Recall@1/5/10 results for both retrieval directions, together with the aggregate metrics. These tables show how the overall robustness trend is supported at individual retrieval ranks on each dataset.

Table 8: Full noisy-correspondence robustness comparison on RSICD. The noise ratio specifies the requested text-level correspondence corruption in the training split; validation and test splits remain clean.

RSICD Dataset									
Method	$\text{Recall}_{\text{I2T}}$			$\text{Recall}_{\text{T2I}}$			mR_{I2T}	mR_{T2I}	mR
	@1	@5	@10	@1	@5	@10			
Noise Ratio = 10%									
AMFMN TGRS'22	5.12	16.01	24.43	3.82	14.84	24.17	15.19	14.28	14.73
GaLR TGRS'22	5.95	14.64	22.32	3.15	13.91	23.68	14.30	13.58	13.94
PIR ACMMM'23	7.69	21.23	33.58	5.69	19.45	34.25	20.83	19.80	20.31
PE TGRS'23	11.99	28.18	41.17	8.58	28.76	43.51	27.11	26.95	27.03
LAPS CVPR'24	6.95	20.40	31.75	5.11	19.60	33.54	19.70	19.42	19.56
FGVLA TGRS'25	9.06	24.43	37.05	7.80	27.25	44.06	23.51	26.37	24.94
EAPA ACMMM'25	10.16	24.52	35.96	7.78	26.70	41.57	23.55	25.35	24.45
CUP TNNLS'25	10.61	24.34	34.95	8.38	26.55	41.46	23.30	25.46	24.38
McRSA TMM'26	15.74	32.85	47.58	13.43	37.40	54.57	32.06	35.13	33.59
RRSITR CVPR'26	17.20	35.13	48.67	12.04	35.15	51.29	33.67	32.83	33.25
P^2Cal	**25.71**	**46.20**	**56.91**	**18.63**	**44.14**	**59.12**	**42.94**	**40.63**	**41.78**
Noise Ratio = 20%									
AMFMN TGRS'22	3.93	10.43	18.76	3.20	12.94	21.57	11.04	12.57	11.80
GaLR TGRS'22	3.29	11.25	19.67	2.74	12.94	22.27	11.40	12.65	12.03
PIR ACMMM'23	7.04	20.49	33.21	4.90	19.21	33.87	20.25	19.33	19.79
PE TGRS'23	10.43	27.81	41.54	8.49	27.37	43.22	26.59	26.36	26.48
LAPS CVPR'24	5.49	17.84	27.45	4.06	16.83	30.01	16.93	16.97	16.95
FGVLA TGRS'25	8.87	21.50	33.85	7.41	25.47	40.90	21.41	24.59	23.00
EAPA ACMMM'25	8.42	22.51	34.95	6.77	22.40	35.72	21.96	21.63	21.80
CUP TNNLS'25	8.60	22.14	34.13	7.12	24.43	39.69	21.62	23.75	22.69
McRSA TMM'26	16.47	35.68	48.86	13.38	36.74	53.16	33.67	34.43	34.05
RRSITR CVPR'26	17.93	35.59	47.48	11.58	34.38	50.28	33.67	32.08	32.88
P^2Cal	**24.06**	**45.20**	**57.82**	**18.35**	**44.52**	**58.72**	**42.36**	**40.53**	**41.45**

Continued on the next page.

Table 8: Full noisy-correspondence robustness comparison on RSICD (continued).

Method	$\text{Recall}_{\text{I2T}}$ @1	$\text{Recall}_{\text{I2T}}$ @5	$\text{Recall}_{\text{I2T}}$ @10	$\text{Recall}_{\text{T2I}}$ @1	$\text{Recall}_{\text{T2I}}$ @5	$\text{Recall}_{\text{T2I}}$ @10	mR_{I2T}	mR_{T2I}	mR
RSICD Dataset									
Noise Ratio = 30%									
AMFMN TGRS'22	3.57	13.45	20.49	2.58	11.25	19.84	12.50	11.22	11.86
GaLR TGRS'22	3.48	11.53	18.39	2.93	10.21	18.68	11.13	10.61	10.87
PIR ACMMM'23	6.59	18.39	31.29	4.28	18.30	31.78	18.76	18.12	18.44
PE TGRS'23	9.42	23.97	37.51	7.10	24.48	38.41	23.63	23.33	23.48
LAPS CVPR'24	4.94	15.37	25.89	3.75	16.32	27.90	15.40	15.99	15.70
FGVLA TGRS'25	7.14	19.12	31.11	6.09	22.89	39.36	19.12	22.78	20.95
EAPA ACMMM'25	10.89	26.72	39.25	9.35	30.03	45.22	25.62	28.20	26.91
CUP TNNLS'25	8.51	21.96	33.85	6.09	21.61	35.22	21.44	20.97	21.20
McRSA TMM'26	16.10	34.77	48.76	13.43	37.16	53.16	33.21	34.58	33.89
RRSITR CVPR'26	15.83	33.21	47.58	11.38	32.08	48.11	32.21	30.52	31.37
P^2Cal	**24.43**	**44.10**	**57.27**	**18.55**	**43.75**	**59.01**	**41.93**	**40.44**	**41.19**
Noise Ratio = 40%									
AMFMN TGRS'22	3.75	9.88	17.02	2.40	10.34	17.68	10.22	10.14	10.18
GaLR TGRS'22	3.11	11.16	16.83	2.56	10.17	17.58	10.37	10.10	10.23
PIR ACMMM'23	5.03	15.55	25.16	3.42	16.38	30.32	15.25	16.71	15.98
PE TGRS'23	8.78	20.68	31.20	5.27	20.16	33.92	20.22	19.78	20.00
LAPS CVPR'24	5.12	13.82	22.42	3.71	14.89	25.97	13.79	14.86	14.32
FGVLA TGRS'25	7.41	20.22	29.00	5.40	21.26	35.65	18.88	20.77	19.82
EAPA ACMMM'25	13.27	29.46	42.45	9.70	30.63	45.07	28.39	28.47	28.43
CUP TNNLS'25	9.52	22.69	35.41	6.02	22.85	36.85	22.54	21.91	22.23
McRSA TMM'26	16.38	33.39	49.31	12.72	36.43	52.88	33.03	34.01	33.52
RRSITR CVPR'26	15.74	33.39	46.29	10.12	31.33	46.68	31.81	29.38	30.59
P^2Cal	**22.32**	**44.65**	**57.91**	**18.57**	**44.17**	**59.36**	**41.63**	**40.70**	**41.17**
Noise Ratio = 50%									
AMFMN TGRS'22	2.84	10.34	17.20	2.47	10.12	16.63	10.13	9.74	9.94
GaLR TGRS'22	2.74	8.69	16.01	2.09	8.45	15.32	9.15	8.62	8.88
PIR ACMMM'23	4.85	14.00	23.79	3.09	14.60	27.80	14.21	15.16	14.69
PE TGRS'23	6.40	16.74	26.17	4.61	18.52	30.67	16.44	17.93	17.19
LAPS CVPR'24	4.21	11.89	21.77	3.42	14.35	25.78	12.62	14.52	13.57
FGVLA TGRS'25	5.40	17.11	27.26	5.22	19.38	34.03	16.59	19.54	18.06
EAPA ACMMM'25	14.18	29.28	42.27	9.72	30.61	44.32	28.58	28.22	28.40
CUP TNNLS'25	6.77	19.58	31.38	5.38	20.93	33.65	19.24	19.99	19.61
McRSA TMM'26	16.29	33.94	47.76	12.68	34.97	50.65	32.66	32.77	32.72
RRSITR CVPR'26	14.73	33.58	46.48	12.02	33.76	49.35	31.60	31.71	31.66
P^2Cal	**24.79**	**44.37**	**57.18**	**18.33**	**42.96**	**58.21**	**42.11**	**39.83**	**40.97**
Noise Ratio = 60%									
AMFMN TGRS'22	2.38	9.24	14.55	1.88	7.85	14.13	8.72	7.95	8.34
GaLR TGRS'22	2.74	9.70	15.55	1.85	7.96	13.60	9.33	7.80	8.56
PIR ACMMM'23	3.75	12.08	22.14	2.91	13.32	25.56	12.66	13.93	13.29
PE TGRS'23	3.66	13.08	20.68	3.48	14.27	23.44	12.47	13.73	13.10
LAPS CVPR'24	2.93	11.44	20.86	3.35	12.79	23.95	11.74	13.36	12.55
FGVLA TGRS'25	5.40	15.92	25.62	5.05	17.73	31.66	15.65	18.15	16.90
EAPA ACMMM'25	12.08	29.46	42.09	9.88	28.23	41.23	27.88	26.45	27.16
CUP TNNLS'25	6.31	18.85	28.09	4.46	17.86	30.83	17.75	17.72	17.73
McRSA TMM'26	14.36	32.94	44.83	11.02	32.13	47.81	30.71	30.32	30.52
RRSITR CVPR'26	14.64	34.77	48.40	10.65	32.92	47.43	32.60	30.33	31.46
P^2Cal	**22.42**	**42.36**	**53.80**	**17.53**	**41.88**	**57.00**	**39.53**	**38.80**	**39.16**
Noise Ratio = 70%									
AMFMN TGRS'22	2.56	7.04	12.35	1.72	6.59	11.00	7.32	6.44	6.88
GaLR TGRS'22	2.01	6.86	11.44	1.52	5.78	10.17	6.77	5.82	6.29
PIR ACMMM'23	4.12	14.18	22.87	2.20	10.30	20.59	13.72	11.03	12.38
PE TGRS'23	2.84	8.87	14.09	2.36	10.58	18.52	8.60	10.49	9.54
LAPS CVPR'24	3.57	10.80	19.85	3.09	13.03	24.01	11.41	13.38	12.39
FGVLA TGRS'25	6.59	16.74	25.25	4.08	16.93	28.42	16.19	16.48	16.34
EAPA ACMMM'25	11.89	26.99	38.15	8.45	25.98	40.33	25.68	24.92	25.30
CUP TNNLS'25	5.12	15.83	23.88	5.25	16.76	28.14	14.94	16.72	15.83
McRSA TMM'26	12.81	31.38	43.18	9.99	28.80	44.32	29.12	27.70	28.41
RRSITR CVPR'26	15.92	32.48	45.01	10.76	32.55	48.97	31.14	30.76	30.95
P^2Cal	**22.32**	**40.99**	**52.52**	**16.18**	**40.73**	**55.86**	**38.61**	**37.59**	**38.10**

Continued on the next page.

Table 8: Full noisy-correspondence robustness comparison on RSICD (continued).

RSICD Dataset									
Method	$\text{Recall}_{\text{I2T}}$			$\text{Recall}_{\text{T2I}}$			mR_{I2T}	mR_{T2I}	mR
	@1	@5	@10	@1	@5	@10			
Noise Ratio = 80%									
AMFMN TGRS'22	1.19	5.12	9.33	1.10	4.85	7.96	5.21	4.64	4.92
GaLR TGRS'22	1.56	4.67	7.32	0.99	3.59	6.55	4.52	3.71	4.12
PIR ACMMM'23	3.57	10.43	18.21	1.88	10.38	20.16	10.74	10.81	10.78
PE TGRS'23	1.28	3.75	7.23	0.93	4.65	7.87	4.09	4.48	4.29
LAPS CVPR'24	2.84	8.97	16.65	2.67	11.33	21.15	9.49	11.72	10.61
FGVLA TGRS'25	3.75	12.26	19.49	3.82	14.36	24.50	11.83	14.23	13.03
EAPA ACMMM'25	7.78	19.12	30.10	5.34	19.65	31.51	19.00	18.83	18.91
CUP TNNLS'25	4.03	13.54	20.77	3.73	13.87	23.82	12.78	13.81	13.29
McRSA TMM'26	9.88	26.90	37.15	8.25	24.90	39.40	24.64	24.18	24.41
RRSITR CVPR'26	11.25	26.17	38.52	8.62	26.22	40.09	25.31	24.98	25.14
P^2Cal	**18.21**	**36.41**	**49.31**	**14.57**	**39.23**	**55.06**	**34.64**	**36.29**	**35.47**
Noise Ratio = 90%									
AMFMN TGRS'22	0.55	2.01	4.30	0.31	1.77	3.66	2.29	1.91	2.10
GaLR TGRS'22	0.46	1.46	2.56	0.24	1.28	2.65	1.49	1.39	1.44
PIR ACMMM'23	0.09	1.83	3.48	0.22	1.15	2.60	1.80	1.32	1.56
PE TGRS'23	0.27	1.01	2.84	0.42	1.76	2.84	1.37	1.67	1.52
LAPS CVPR'24	2.29	4.57	8.78	1.45	7.04	12.79	5.21	7.09	6.15
FGVLA TGRS'25	0.27	1.01	2.01	0.22	0.84	1.59	1.10	0.88	0.99
EAPA ACMMM'25	3.11	8.78	14.18	2.21	8.14	13.16	8.69	7.84	8.27
CUP TNNLS'25	2.38	8.69	13.45	2.16	10.94	18.96	8.17	10.69	9.43
McRSA TMM'26	6.95	20.31	30.65	6.08	19.52	31.95	19.30	19.18	19.24
RRSITR CVPR'26	0.37	1.56	2.47	0.31	1.24	2.62	1.47	1.39	1.43
P^2Cal	**14.36**	**32.20**	**42.09**	**12.20**	**32.90**	**47.94**	**29.55**	**31.01**	**30.28**

Table 9: Full noisy-correspondence robustness comparison on RSITMD. The noise ratio specifies the requested text-level correspondence corruption in the training split; validation and test splits remain clean.

RSITMD Dataset									
Method	$\text{Recall}_{\text{I2T}}$			$\text{Recall}_{\text{T2I}}$			mR_{I2T}	mR_{T2I}	mR
	@1	@5	@10	@1	@5	@10			
Noise Ratio = 10%									
AMFMN TGRS'22	11.28	27.21	37.61	7.52	26.55	40.13	25.37	24.73	25.05
GaLR TGRS'22	13.50	30.31	44.91	9.91	35.71	52.83	29.57	32.82	31.20
PIR ACMMM'23	16.15	35.62	48.45	11.06	39.16	59.51	33.41	36.58	34.99
PE TGRS'23	20.13	42.04	54.87	18.45	45.62	65.71	39.01	43.26	41.13
LAPS CVPR'24	12.61	30.53	42.48	8.85	34.65	54.38	28.54	32.63	30.59
FGVLA TGRS'25	17.92	41.59	55.75	15.04	47.08	66.59	38.42	42.90	40.66
EAPA ACMMM'25	22.12	43.36	56.64	16.50	44.60	58.50	40.71	39.87	40.29
CUP TNNLS'25	17.92	35.84	51.99	12.96	41.86	60.40	35.25	38.41	36.83
McRSA TMM'26	25.00	46.02	59.51	22.88	52.65	67.79	43.51	47.77	45.64
RRSITR CVPR'26	23.01	43.81	57.52	19.87	52.26	69.56	41.45	47.23	44.34
P^2Cal	**31.42**	**51.55**	**65.71**	**27.26**	**58.72**	**75.62**	**49.56**	**53.87**	**51.72**
Noise Ratio = 20%									
AMFMN TGRS'22	9.29	24.78	34.73	7.65	29.78	45.93	22.93	27.79	25.36
GaLR TGRS'22	8.85	25.88	40.27	6.59	30.88	49.87	25.00	29.11	27.05
PIR ACMMM'23	13.05	30.09	44.03	8.81	33.10	54.56	29.06	32.16	30.61
PE TGRS'23	17.26	36.95	51.55	14.73	43.19	61.64	35.25	39.85	37.55
LAPS CVPR'24	8.63	23.23	38.05	8.36	31.02	50.40	23.30	29.93	26.62
FGVLA TGRS'25	17.92	37.61	49.56	13.45	42.61	62.88	35.03	39.65	37.34
EAPA ACMMM'25	14.38	37.39	53.54	12.52	37.57	51.64	35.10	33.91	34.50
CUP TNNLS'25	17.04	38.72	51.55	14.96	43.10	60.62	35.77	39.56	37.67
McRSA TMM'26	21.68	42.92	58.19	20.31	50.75	68.45	40.93	46.50	43.72
RRSITR CVPR'26	26.11	44.91	57.96	20.53	51.68	69.38	42.99	47.20	45.09
P^2Cal	**29.65**	**51.11**	**64.82**	**27.74**	**60.35**	**76.19**	**48.53**	**54.76**	**51.64**

Continued on the next page.

Table 9: Full noisy-correspondence robustness comparison on RSITMD (continued).

Method	$\text{Recall}_{\text{I2T}}$ @1	$\text{Recall}_{\text{I2T}}$ @5	$\text{Recall}_{\text{I2T}}$ @10	$\text{Recall}_{\text{T2I}}$ @1	$\text{Recall}_{\text{T2I}}$ @5	$\text{Recall}_{\text{T2I}}$ @10	mR_{I2T}	mR_{T2I}	mR
RSITMD Dataset									
Noise Ratio = 30%									
AMFMN TGRS'22	10.18	25.88	38.72	6.37	30.49	46.99	24.93	27.95	26.44
GaLR TGRS'22	6.64	24.12	35.62	6.24	26.73	45.49	22.13	26.15	24.14
PIR ACMMM'23	9.73	28.32	44.47	7.35	29.60	46.90	27.51	27.95	27.73
PE TGRS'23	16.81	34.29	44.91	12.92	39.73	56.81	32.00	36.49	34.25
LAPS CVPR'24	7.08	22.35	35.18	7.70	30.09	51.02	21.54	29.60	25.57
FGVLA TGRS'25	11.73	31.19	46.68	10.71	39.29	58.98	29.87	36.33	33.10
EAPA ACMMM'25	13.72	32.96	46.46	11.46	32.12	46.37	31.05	29.98	30.52
CUP TNNLS'25	15.27	34.96	49.56	13.85	40.53	60.09	33.26	38.16	35.71
McRSA TMM'26	24.34	44.69	57.08	20.53	54.07	68.81	42.04	47.80	44.92
RRSITR CVPR'26	21.02	45.58	59.51	20.04	50.49	69.47	42.04	46.67	44.36
P^2Cal	**28.10**	**48.01**	**63.94**	**25.71**	**57.92**	**74.47**	**46.68**	**52.70**	**49.69**
Noise Ratio = 40%									
AMFMN TGRS'22	7.52	21.46	34.51	6.37	25.09	44.20	21.16	25.22	23.19
GaLR TGRS'22	5.97	22.35	34.29	5.80	24.69	41.73	20.87	24.07	22.47
PIR ACMMM'23	7.96	23.89	40.49	6.99	25.49	42.08	24.11	24.85	24.48
PE TGRS'23	11.28	27.88	40.71	10.71	32.88	50.58	26.62	31.39	29.01
LAPS CVPR'24	9.51	26.11	37.61	7.04	29.20	47.43	24.41	27.89	26.15
FGVLA TGRS'25	11.73	30.75	47.57	11.37	35.53	54.82	30.02	33.91	31.96
EAPA ACMMM'25	10.84	28.98	39.82	8.58	26.64	38.67	26.55	24.63	25.59
CUP TNNLS'25	15.49	34.73	47.79	11.81	39.47	56.06	32.67	35.78	34.23
McRSA TMM'26	23.67	42.70	59.51	19.16	49.91	67.57	41.96	45.55	43.75
RRSITR CVPR'26	23.01	41.59	55.75	19.73	51.28	67.17	40.12	46.06	43.09
P^2Cal	**29.42**	**50.66**	**64.38**	**26.33**	**58.36**	**73.58**	**48.15**	**52.76**	**50.45**
Noise Ratio = 50%									
AMFMN TGRS'22	5.53	20.35	30.09	5.88	25.27	41.15	18.66	24.10	21.38
GaLR TGRS'22	5.31	19.03	33.63	4.69	21.99	38.50	19.32	21.73	20.52
PIR ACMMM'23	7.08	24.12	38.94	5.88	26.50	47.74	23.38	26.71	25.05
PE TGRS'23	11.95	28.76	38.50	9.42	30.40	43.58	26.40	27.80	27.10
LAPS CVPR'24	7.96	22.12	32.96	7.35	28.01	45.84	21.01	27.07	24.04
FGVLA TGRS'25	9.51	28.32	45.13	9.65	33.19	50.49	27.65	31.11	29.38
EAPA ACMMM'25	9.96	26.77	36.73	7.65	24.34	36.59	24.49	22.86	23.67
CUP TNNLS'25	12.17	33.41	46.68	11.02	36.50	55.35	30.75	34.29	32.52
McRSA TMM'26	24.34	43.36	58.41	19.16	50.09	66.37	42.04	45.21	43.62
RRSITR CVPR'26	19.47	43.36	55.31	18.85	46.59	63.45	39.38	42.96	41.17
P^2Cal	**30.53**	**50.66**	**61.95**	**26.68**	**57.30**	**72.70**	**47.71**	**52.23**	**49.97**
Noise Ratio = 60%									
AMFMN TGRS'22	5.75	20.13	29.87	4.78	21.55	35.09	18.58	20.47	19.52
GaLR TGRS'22	5.53	19.03	29.20	3.81	18.85	32.48	17.92	18.38	18.15
PIR ACMMM'23	8.19	20.13	33.19	5.62	22.79	41.99	20.50	23.47	21.98
PE TGRS'23	9.07	18.81	29.20	6.99	23.67	35.13	19.03	21.93	20.48
LAPS CVPR'24	7.74	20.58	31.64	6.02	26.11	44.03	19.99	25.39	22.69
FGVLA TGRS'25	13.94	30.97	45.13	9.60	31.86	48.67	30.01	30.04	30.02
EAPA ACMMM'25	7.74	21.24	32.30	5.75	19.47	30.88	20.43	18.70	19.56
CUP TNNLS'25	15.27	33.63	45.58	9.91	35.53	53.94	31.49	33.13	32.31
McRSA TMM'26	19.91	39.16	52.21	16.77	48.19	65.22	37.09	43.39	40.24
RRSITR CVPR'26	21.24	42.04	55.53	16.50	45.75	63.41	39.60	41.89	40.75
P^2Cal	**25.88**	**47.79**	**59.96**	**25.40**	**56.55**	**72.21**	**44.54**	**51.39**	**47.97**
Noise Ratio = 70%									
AMFMN TGRS'22	6.19	17.70	26.11	4.34	15.97	27.74	16.67	16.02	16.34
GaLR TGRS'22	5.09	16.81	26.11	3.41	14.82	25.93	16.00	14.72	15.36
PIR ACMMM'23	4.20	15.27	24.56	3.85	15.93	27.61	14.68	15.80	15.24
PE TGRS'23	6.64	15.71	23.23	6.11	19.47	28.67	15.19	18.08	16.63
LAPS CVPR'24	7.52	17.92	27.43	5.88	23.89	40.40	17.62	23.39	20.51
FGVLA TGRS'25	8.19	25.00	40.04	7.83	26.99	44.20	24.41	26.34	25.38
EAPA ACMMM'25	4.87	14.16	23.45	4.12	13.10	21.11	14.16	12.78	13.47
CUP TNNLS'25	13.27	26.55	42.26	9.47	32.74	52.08	27.36	31.43	29.39
McRSA TMM'26	19.69	35.18	48.89	14.07	39.07	56.19	34.59	36.44	35.52
RRSITR CVPR'26	18.58	38.27	52.43	17.30	45.88	63.94	36.43	42.37	39.40
P^2Cal	**26.33**	**46.90**	**58.63**	**23.67**	**53.23**	**69.56**	**43.95**	**48.82**	**46.39**

Continued on the next page.

Table 9: Full noisy-correspondence robustness comparison on RSITMD (continued).

RSITMD Dataset									
Method	$\text{Recall}_{\text{I2T}}$			$\text{Recall}_{\text{T2I}}$			mR_{I2T}	mR_{T2I}	mR
	@1	@5	@10	@1	@5	@10			
Noise Ratio = 80%									
AMFMN TGRS'22	5.09	12.39	17.48	2.12	9.87	18.14	11.65	10.04	10.84
GaLR TGRS'22	2.65	8.85	14.38	1.86	6.42	13.41	8.63	7.23	7.93
PIR ACMMM'23	0.44	2.43	5.09	0.66	3.10	5.80	2.65	3.19	2.92
PE TGRS'23	4.87	11.50	18.36	3.36	13.01	19.87	11.58	12.08	11.83
LAPS CVPR'24	3.54	15.04	24.12	3.54	18.72	32.96	14.23	18.41	16.32
FGVLA TGRS'25	7.74	21.02	30.97	6.02	21.86	33.98	19.91	20.62	20.27
EAPA ACMMM'25	2.43	9.51	14.60	2.92	9.78	16.19	8.85	9.63	9.24
CUP TNNLS'25	12.61	26.33	38.50	10.09	30.88	46.42	25.81	29.13	27.47
McRSA TMM'26	17.70	34.96	44.03	12.39	41.02	57.96	32.23	37.12	34.67
RRSITR CVPR'26	17.26	34.29	46.68	14.69	42.12	60.27	32.74	39.03	35.89
P^2Cal	**22.79**	**44.25**	**54.87**	**20.88**	**49.07**	**66.37**	**40.64**	**45.44**	**43.04**
Noise Ratio = 90%									
AMFMN TGRS'22	1.55	5.31	9.96	1.06	5.00	9.07	5.61	5.04	5.33
GaLR TGRS'22	0.88	1.99	4.65	0.31	2.08	3.81	2.51	2.07	2.29
PIR ACMMM'23	0.22	1.33	2.88	0.22	1.11	2.21	1.48	1.18	1.33
PE TGRS'23	1.33	5.09	7.96	1.64	5.71	8.89	4.79	5.41	5.10
LAPS CVPR'24	3.32	8.63	13.27	1.55	10.66	17.88	8.41	10.03	9.22
FGVLA TGRS'25	0.66	2.88	3.98	1.02	3.76	8.23	2.51	4.34	3.42
EAPA ACMMM'25	1.33	5.31	8.19	1.24	5.31	9.20	4.94	5.25	5.10
CUP TNNLS'25	12.61	25.66	38.72	7.92	31.06	48.27	25.66	29.08	27.37
McRSA TMM'26	11.73	28.76	40.49	9.03	33.32	50.00	26.99	30.78	28.88
RRSITR CVPR'26	0.44	1.99	3.76	0.31	1.50	3.58	2.06	1.80	1.93
P^2Cal	**18.81**	**36.73**	**49.12**	**17.52**	**42.61**	**59.65**	**34.89**	**39.93**	**37.41**

Table 10: Full noisy-correspondence robustness comparison on NWPU. The noise ratio specifies the requested text-level correspondence corruption in the training split; validation and test splits remain clean.

NWPU Dataset									
Method	$\text{Recall}_{\text{I2T}}$			$\text{Recall}_{\text{T2I}}$			mR_{I2T}	mR_{T2I}	mR
	@1	@5	@10	@1	@5	@10			
Noise Ratio = 10%									
AMFMN TGRS'22	8.00	30.57	47.43	6.36	24.86	40.26	28.67	23.83	26.25
GaLR TGRS'22	8.89	29.43	46.63	6.53	25.02	40.55	28.32	24.03	26.18
PIR ACMMM'23	13.11	42.10	60.63	8.48	29.64	45.67	38.61	27.93	33.27
PE TGRS'23	2.16	9.24	16.73	1.83	7.59	12.96	9.38	7.46	8.42
LAPS CVPR'24	7.75	27.59	44.25	5.54	22.21	36.82	26.53	21.52	24.02
FGVLA TGRS'25	4.83	20.06	33.52	4.58	16.90	29.48	19.47	16.99	18.23
EAPA ACMMM'25	18.44	52.48	70.44	11.95	36.55	52.99	47.12	33.83	40.47
CUP TNNLS'25	5.81	20.19	32.29	3.16	13.04	22.74	19.43	12.98	16.20
McRSA TMM'26	21.02	52.57	70.13	12.42	36.48	52.48	47.91	33.79	40.85
RRSITR CVPR'26	23.87	58.73	73.94	13.90	39.21	54.70	52.18	35.94	44.06
P^2Cal	**26.57**	**60.48**	**76.92**	**14.69**	**39.87**	**55.20**	**54.66**	**36.59**	**45.62**
Noise Ratio = 20%									
AMFMN TGRS'22	5.21	22.22	37.71	4.58	19.53	32.83	21.71	18.98	20.34
GaLR TGRS'22	5.78	24.03	37.24	5.05	20.03	33.42	22.35	19.50	20.93
PIR ACMMM'23	8.79	32.16	50.95	6.61	24.78	40.11	30.63	23.83	27.23
PE TGRS'23	1.24	5.78	10.98	1.50	5.92	10.14	6.00	5.85	5.92
LAPS CVPR'24	4.41	17.21	28.98	3.56	15.47	27.31	16.87	15.45	16.16
FGVLA TGRS'25	3.43	15.94	27.65	3.39	14.40	25.13	15.67	14.31	14.99
EAPA ACMMM'25	19.43	52.00	69.56	11.89	35.46	51.76	47.00	33.04	40.02
CUP TNNLS'25	6.67	23.65	37.56	4.77	17.41	28.77	22.63	16.98	19.80
McRSA TMM'26	19.02	51.02	68.03	11.66	35.62	51.53	46.02	32.94	39.48
RRSITR CVPR'26	23.56	58.16	73.65	13.20	38.01	53.59	51.79	34.93	43.36
P^2Cal	**25.21**	**58.25**	**75.90**	**14.04**	**38.62**	**53.86**	**53.12**	**35.51**	**44.31**

Continued on the next page.

Table 10: Full noisy-correspondence robustness comparison on NWPU (continued).

Method	$\text{Recall}_{\text{I2T}}$ @1	$\text{Recall}_{\text{I2T}}$ @5	$\text{Recall}_{\text{I2T}}$ @10	$\text{Recall}_{\text{T2I}}$ @1	$\text{Recall}_{\text{T2I}}$ @5	$\text{Recall}_{\text{T2I}}$ @10	mR_{I2T}	mR_{T2I}	mR
NWPU Dataset									
Noise Ratio = 30%									
AMFMN TGRS'22	4.00	17.90	31.37	3.87	16.41	28.47	17.76	16.25	17.01
GaLR TGRS'22	4.25	17.43	30.44	3.95	16.67	28.75	17.37	16.46	16.91
PIR ACMMM'23	5.30	24.32	40.63	4.51	18.66	32.61	23.42	18.59	21.01
PE TGRS'23	0.89	4.67	8.16	1.18	4.41	7.42	4.57	4.34	4.46
LAPS CVPR'24	3.71	14.25	24.22	2.76	12.46	22.77	14.06	12.66	13.36
FGVLA TGRS'25	3.52	13.46	23.37	3.19	13.54	23.86	13.45	13.53	13.49
EAPA ACMMM'25	16.67	48.32	66.22	10.89	34.18	49.90	43.74	31.66	37.70
CUP TNNLS'25	5.75	20.13	32.83	3.66	14.46	24.75	19.57	14.29	16.93
McRSA TMM'26	18.16	48.00	65.62	11.43	33.66	50.04	43.93	31.71	37.82
RRSITR CVPR'26	21.40	55.21	71.52	12.70	37.50	53.11	49.38	34.44	41.91
P^2Cal	**24.83**	**58.51**	**75.84**	**13.86**	**38.45**	**53.96**	**53.06**	**35.42**	**44.24**
Noise Ratio = 40%									
AMFMN TGRS'22	3.24	15.62	27.11	3.24	14.78	25.72	15.32	14.58	14.95
GaLR TGRS'22	3.33	14.16	25.21	2.94	13.10	23.56	14.23	13.20	13.71
PIR ACMMM'23	2.98	11.62	21.11	1.73	8.17	15.59	11.90	8.50	10.20
PE TGRS'23	0.70	2.98	6.03	0.66	3.16	5.52	3.24	3.11	3.17
LAPS CVPR'24	2.54	11.56	20.44	2.39	10.14	18.65	11.51	10.39	10.95
FGVLA TGRS'25	3.11	13.56	22.83	3.05	13.07	22.96	13.17	13.03	13.10
EAPA ACMMM'25	15.62	44.76	62.89	10.41	32.62	48.06	41.09	30.36	35.73
CUP TNNLS'25	6.32	20.38	31.87	3.18	12.67	22.23	19.52	12.69	16.11
McRSA TMM'26	18.89	48.48	65.62	10.56	33.27	49.10	44.33	30.98	37.66
RRSITR CVPR'26	21.33	53.71	71.17	12.48	36.97	52.66	48.74	34.04	41.39
P^2Cal	**24.73**	**57.27**	**74.54**	**13.63**	**37.86**	**52.81**	**52.18**	**34.77**	**43.48**
Noise Ratio = 50%									
AMFMN TGRS'22	2.73	12.29	22.79	2.57	12.21	21.94	12.60	12.24	12.42
GaLR TGRS'22	2.29	10.63	19.94	2.20	10.32	19.30	10.95	10.61	10.78
PIR ACMMM'23	2.41	11.87	20.25	1.56	7.10	14.11	11.51	7.59	9.55
PE TGRS'23	0.03	0.03	0.06	0.03	0.16	0.32	0.04	0.17	0.11
LAPS CVPR'24	2.22	9.49	18.54	1.83	8.77	16.53	10.08	9.04	9.56
FGVLA TGRS'25	2.76	12.03	21.75	2.57	11.52	20.98	12.18	11.69	11.93
EAPA ACMMM'25	13.33	41.05	59.71	9.42	30.98	46.52	38.03	28.97	33.50
CUP TNNLS'25	5.24	16.89	27.33	2.63	11.45	20.88	16.49	11.65	14.07
McRSA TMM'26	17.05	45.27	61.97	9.98	31.14	46.45	41.43	29.19	35.31
RRSITR CVPR'26	19.90	50.83	68.13	11.79	35.87	**51.50**	46.29	33.05	39.67
P^2Cal	**23.14**	**55.62**	**71.65**	**12.95**	**35.96**	51.49	**50.14**	**33.47**	**41.80**
Noise Ratio = 60%									
AMFMN TGRS'22	2.41	11.84	21.08	2.39	10.60	19.13	11.78	10.71	11.25
GaLR TGRS'22	2.13	9.17	17.37	2.03	8.79	16.41	9.56	9.08	9.32
PIR ACMMM'23	2.16	9.46	18.06	1.58	7.45	14.75	9.89	7.93	8.91
PE TGRS'23	0.03	0.03	0.06	0.03	0.16	0.32	0.04	0.17	0.11
LAPS CVPR'24	2.13	10.03	18.48	2.01	8.53	16.67	10.21	9.07	9.64
FGVLA TGRS'25	3.08	12.13	22.10	2.51	11.37	20.23	12.44	11.37	11.90
EAPA ACMMM'25	11.56	38.16	56.67	8.64	29.05	44.46	35.46	27.38	31.42
CUP TNNLS'25	6.16	19.84	32.00	3.33	12.64	22.17	19.33	12.71	16.02
McRSA TMM'26	14.48	41.78	58.29	8.75	28.38	43.24	38.18	26.79	32.48
RRSITR CVPR'26	19.08	48.79	65.11	11.25	33.80	49.16	44.33	31.40	37.86
P^2Cal	**21.84**	**52.73**	**69.49**	**12.36**	**35.30**	**50.62**	**48.02**	**32.76**	**40.39**

Continued on the next page.

Table 10: Full noisy-correspondence robustness comparison on NWPU (continued).

Method	$\text{Recall}_{\text{I2T}}$ @1	$\text{Recall}_{\text{I2T}}$ @5	$\text{Recall}_{\text{I2T}}$ @10	$\text{Recall}_{\text{T2I}}$ @1	$\text{Recall}_{\text{T2I}}$ @5	$\text{Recall}_{\text{T2I}}$ @10	mR_{I2T}	mR_{T2I}	mR
NWPU Dataset									
Noise Ratio = 70%									
AMFMN TGRS'22	2.00	9.71	16.95	1.99	8.98	15.81	9.55	8.93	9.24
GaLR TGRS'22	1.37	7.71	15.33	1.83	7.68	14.44	8.14	7.98	8.06
PIR ACMMM'23	2.32	9.49	18.63	1.46	7.30	14.23	10.15	7.66	8.91
PE TGRS'23	0.03	0.03	0.06	0.03	0.16	0.32	0.04	0.17	0.11
LAPS CVPR'24	1.84	8.57	16.54	1.73	7.99	15.43	8.98	8.38	8.68
FGVLA TGRS'25	2.25	10.63	19.81	2.38	10.69	19.67	10.90	10.91	10.91
EAPA ACMMM'25	10.48	31.43	46.54	6.81	23.30	36.88	29.48	22.33	25.91
CUP TNNLS'25	3.71	14.03	26.63	2.78	11.83	20.83	14.79	11.81	13.30
McRSA TMM'26	12.19	34.70	50.60	6.78	23.97	37.62	32.50	22.79	27.64
RRSITR CVPR'26	15.97	45.46	63.37	10.18	32.13	46.79	41.60	29.70	35.65
P²Cal	**19.90**	**48.10**	**66.51**	**11.68**	**33.45**	**48.41**	**44.84**	**31.18**	**38.01**
Noise Ratio = 80%									
AMFMN TGRS'22	1.33	7.78	14.25	1.49	6.67	11.96	7.79	6.71	7.25
GaLR TGRS'22	1.14	4.98	10.57	1.08	5.14	9.58	5.56	5.27	5.41
PIR ACMMM'23	1.30	7.17	13.94	1.33	6.89	13.61	7.47	7.28	7.38
PE TGRS'23	0.03	0.03	0.06	0.03	0.16	0.32	0.04	0.17	0.11
LAPS CVPR'24	1.65	8.35	15.78	1.59	7.79	14.86	8.59	8.08	8.34
FGVLA TGRS'25	2.13	7.87	15.24	1.98	8.69	16.58	8.41	9.08	8.75
EAPA ACMMM'25	6.10	19.97	32.41	5.05	17.74	29.10	19.49	17.30	18.39
CUP TNNLS'25	4.22	16.06	26.44	2.70	10.93	19.87	15.57	11.17	13.37
McRSA TMM'26	7.71	26.54	41.24	5.09	18.60	30.30	25.16	18.00	21.58
RRSITR CVPR'26	12.32	39.17	56.79	8.53	27.53	42.28	36.09	26.11	31.10
P²Cal	**18.06**	**45.46**	**60.92**	**10.32**	**30.27**	**43.72**	**41.48**	**28.10**	**34.79**
Noise Ratio = 90%									
AMFMN TGRS'22	0.57	2.54	5.27	0.53	2.77	5.20	2.79	2.83	2.81
GaLR TGRS'22	0.16	0.51	1.08	0.17	0.87	1.76	0.58	0.93	0.76
PIR ACMMM'23	0.16	1.17	2.22	0.14	0.76	1.49	1.18	0.80	0.99
PE TGRS'23	0.03	0.03	0.06	0.03	0.16	0.32	0.04	0.17	0.11
LAPS CVPR'24	0.06	0.29	0.41	0.06	0.41	0.88	0.25	0.45	0.35
FGVLA TGRS'25	0.06	0.22	0.35	0.06	0.32	0.59	0.21	0.32	0.27
EAPA ACMMM'25	4.35	14.73	25.08	3.53	13.51	23.50	14.72	13.51	14.12
CUP TNNLS'25	3.30	13.43	23.02	2.28	10.11	18.37	13.25	10.25	11.75
McRSA TMM'26	5.14	17.78	30.10	3.09	12.18	21.68	17.67	12.32	15.00
RRSITR CVPR'26	0.06	0.44	0.95	0.06	0.27	0.50	0.48	0.28	0.38
P²Cal	**13.37**	**36.48**	**51.08**	**8.55**	**25.40**	**37.98**	**33.64**	**23.98**	**28.81**

Across the three datasets, P^{2}Cal achieves the highest overall and directional mR in all 27 reported dataset-noise settings and leads 161 of the 162 individual Recall entries. The advantage therefore extends across both retrieval directions and Recall@1/5/10, covering accurate first-ranked matches as well as recovery within larger retrieved sets. At 90% noise, P^{2}Cal leads all six Recall metrics on every dataset, with gains of 5.02-20.98 percentage points over the strongest competing value for each entry. Thus, even when most training pairs are selected for shuffling, the model retains useful cross-modal retrieval ability across individual ranking thresholds. This broad advantage is consistent with the complementary roles of PSCM in filtering supervision, MRA in adapting the visual, textual, and cross-modal pathways, and PSCF in refining candidate scores.

As correspondence noise increases, P^{2}Cal remains relatively stable under low-to-moderate corruption and degrades more noticeably when supervision becomes severely unreliable. Nevertheless, it maintains a clear advantage in both retrieval directions. This distinction is particularly evident on RSICD and NWPU, where several competing methods lose much of their retrieval ability under severe noise, while P^{2}Cal continues to recover relevant image–text pairs. The trend is consistent with the RSICD ablations: purifying supervision helps preserve the benefits of pretrained embedding adaptation and enables subsequent score calibration to remain effective. Thus, the advantage of P^{2}Cal persists even when reliable training correspondences become scarce.

The sensitivity to noise also varies across datasets. P^{2}Cal exhibits comparable relative degradation on RSICD and RSITMD, whereas NWPU shows a larger overall decline. On RSITMD, some base-

lines, including CUP and McRSA, also retain useful retrieval ability under severe corruption. Its finer-grained texts may provide more informative object, attribute, and spatial-relation cues in the remaining reliable pairs, helping sustain learning despite unreliable supervision. P^{2}Cal nevertheless maintains the strongest overall and directional mean Recall across all tested noise levels and leads every reported Recall metric at the highest noise level on all three datasets. These trends support the robustness of prior-posterior collaborative calibration across different dataset conditions.

A.5 Scene-Level Retrieval Analysis

To identify scene-specific retrieval strengths and remaining semantic ambiguities, we conduct an additional category-wise analysis on RSICD, RSITMD, and NWPU.

A.5.1 Details of Experimental Settings

All methods retrieve from the complete test gallery: all test texts for image-to-text (I2T) and all test images for text-to-image (T2I). Scene labels, obtained from image annotations or filenames, only group queries after retrieval and do not restrict candidate selection. Categories are listed in Appendix A.3.

For I2T, an image query is a hit if any of its paired ground-truth texts appears in the top five retrieved texts. For scene s,

$$\mathrm{Recall}^{s}_{\mathrm{I2T}}@5 = \frac{1}{|\mathcal{Q}^{s}_{\mathrm{I2T}}|} \sum_{q_i \in \mathcal{Q}^{s}_{\mathrm{I2T}}} \mathbf{1}\left[\min_{g \in \mathcal{G}_i} \mathrm{rank}_i(g) \le 5\right], \tag{18}$$

where $\mathcal{Q}^{s}_{\mathrm{I2T}}$ contains scene-s test image queries, $\mathcal{G}_i$ contains query q_i's paired texts, and $\mathrm{rank}_i(g)$ is target g's one-based rank in the complete test-text gallery. For T2I, each text query inherits its paired image's scene label and is a hit if its paired image appears in the top five retrieved images. The corresponding Recall@5 is

$$\mathrm{Recall}^{s}_{\mathrm{T2I}}@5 = \frac{1}{|\mathcal{Q}^{s}_{\mathrm{T2I}}|} \sum_{q_i \in \mathcal{Q}^{s}_{\mathrm{T2I}}} \mathbf{1}\left[\min_{g \in \mathcal{G}_i} \mathrm{rank}_i(g) \le 5\right], \tag{19}$$

where $\mathcal{Q}^{s}_{\mathrm{T2I}}$ is the test text-query set for scene s, $\mathcal{G}_i$ denotes the query's ground-truth image set, and ranks are computed in the complete test-image gallery. In both directions, correctness follows the original image-text correspondences: an unpaired item from the same scene is not a match. Each query contributes at most one hit, even if multiple paired texts enter the top five in I2T. The denominator is the number of test queries in that scene and direction, not the gallery size or the full-dataset image count.

A.5.2 Results and Analysis

Tables 11-16 compare scene-level Recall@5 in both directions.

Table 11: Scene-level Recall@5 (%) on RSICD for image-to-text (I2T) retrieval. Higher is better.

Scene	*AMFMN* 2022	*GaLR* 2022	*PE* 2023	*PIR* 2023	*FGVLA* 2025	*LAPS* 2024	*CUP* 2025	*EAPA* 2025	*McRSA* 2026	*RRSITR* 2026	*P^{2}Cal*
airport	9.52	19.05	16.67	23.81	7.14	11.90	**26.19**	11.90	21.43	**30.95**	**33.33**
bare land	**16.13**	**16.13**	12.90	12.90	**19.35**	**19.35**	**16.13**	**19.35**	12.90	**19.35**	**25.81**
baseball field	23.33	16.67	**43.33**	26.67	**43.33**	26.67	**36.67**	30.00	**43.33**	**46.67**	**43.33**
beach	20.00	**27.50**	**27.50**	15.00	**32.50**	20.00	17.50	15.00	22.50	20.00	**37.50**
bridge	17.02	14.89	10.64	**29.79**	19.15	14.89	19.15	**23.40**	19.15	**29.79**	**38.30**
center	19.23	19.23	**53.85**	50.00	**61.54**	34.62	23.08	46.15	**65.38**	**65.38**	**65.38**
church	12.50	25.00	37.50	20.83	**58.33**	12.50	**45.83**	41.67	37.50	37.50	**62.50**
commercial	11.43	11.43	**22.86**	**34.29**	20.00	**22.86**	**22.86**	17.14	**22.86**	17.14	**31.43**
dense residential	9.76	9.76	**24.39**	**21.95**	14.63	12.20	19.51	**21.95**	**21.95**	14.63	**31.71**
desert	26.67	23.33	**36.67**	20.00	**36.67**	23.33	**43.33**	30.00	30.00	33.33	**63.33**
farmland	13.51	5.41	**18.92**	13.51	16.22	8.11	**18.92**	**32.43**	16.22	**18.92**	**37.84**
forest	24.00	28.00	20.00	**36.00**	28.00	**32.00**	24.00	28.00	24.00	**32.00**	**44.00**

Continued on the next page.

Table 11 (continued)

Scene	*AMFMN* 2022	*GaLR* 2022	*PE* 2023	*PIR* 2023	*FGVLA* 2025	*LAPS* 2024	*CUP* 2025	*EAPA* 2025	*McRSA* 2026	*RRSITR* 2026	*P²Cal*
industrial	17.95	7.69	25.64	23.08	**38.46**	25.64	23.08	**38.46**	23.08	**35.90**	**43.59**
meadow	21.43	21.43	**42.86**	21.43	**42.86**	28.57	35.71	32.14	**46.43**	35.71	**50.00**
medium residential	3.45	6.90	**20.69**	13.79	13.79	**17.24**	**17.24**	**20.69**	**31.03**	**17.24**	**17.24**
mountain	**23.53**	**20.59**	**35.29**	14.71	**23.53**	17.65	11.76	**20.59**	**23.53**	**23.53**	**20.59**
park	14.29	11.43	42.86	22.86	**45.71**	40.00	17.14	34.29	**48.57**	42.86	**57.14**
parking	7.69	20.51	20.51	10.26	25.64	12.82	12.82	23.08	**35.90**	**28.21**	**38.46**
playground	22.00	14.00	38.00	30.00	**42.00**	28.00	27.00	41.00	**45.00**	40.00	**55.00**
pond	19.05	19.05	35.71	21.43	**40.48**	35.71	33.33	38.10	**42.86**	35.71	**54.76**
port	7.69	10.26	**33.33**	**28.21**	**33.33**	23.08	20.51	**33.33**	**28.21**	25.64	**35.90**
railway station	7.69	23.08	**38.46**	26.92	30.77	**34.62**	26.92	**38.46**	**34.62**	**34.62**	**53.85**
resort	27.59	17.24	31.03	24.14	**55.17**	13.79	27.59	**48.28**	34.48	44.83	**51.72**
river	17.07	21.95	29.27	19.51	**34.15**	**34.15**	26.83	19.51	**41.46**	29.27	**36.59**
school	10.00	13.33	**36.67**	23.33	33.33	26.67	23.33	23.33	**40.00**	20.00	**60.00**
sparse residential	10.00	10.00	**33.33**	30.00	26.67	30.00	30.00	30.00	**43.33**	30.00	**56.67**
square	24.24	15.15	39.39	39.39	42.42	30.30	33.33	39.39	**60.61**	**51.52**	**54.55**
stadium	17.24	10.34	24.14	31.03	41.38	27.59	24.14	**48.28**	**51.72**	**58.62**	**58.62**
storage tanks	25.00	22.50	**40.00**	32.50	**45.00**	32.50	22.50	**47.50**	**40.00**	**45.00**	**40.00**
viaduct	19.05	26.19	30.95	28.57	**35.71**	23.81	26.19	33.33	**38.10**	28.57	**59.52**

Notes. The **highest**, **second-highest**, and **third-highest** distinct values in each scene row are marked in red, blue, and green, respectively. Ties share the same color; ranking uses the reported two-decimal values.

Table 12: Scene-level Recall@5 (%) on RSICD for text-to-image (T2I) retrieval. Higher is better.

Scene	*AMFMN* 2022	*GaLR* 2022	*PE* 2023	*PIR* 2023	*FGVLA* 2025	*LAPS* 2024	*CUP* 2025	*EAPA* 2025	*McRSA* 2026	*RRSITR* 2026	*P²Cal*
airport	11.90	11.43	**31.90**	18.10	19.05	19.05	16.19	16.19	**33.33**	29.52	**38.10**
bare land	12.26	12.26	**32.26**	16.77	**31.61**	18.06	23.23	25.81	25.81	22.58	**27.10**
baseball field	27.33	14.67	39.33	36.67	46.67	32.67	41.33	48.00	**53.33**	**56.00**	**62.67**
beach	17.50	19.00	24.00	22.50	19.00	17.00	**26.50**	22.00	**31.00**	22.00	**35.50**
bridge	15.74	16.60	17.87	10.64	21.70	16.17	15.32	**24.26**	**26.38**	20.43	**28.94**
center	20.00	21.54	53.85	25.38	**55.38**	28.46	40.00	53.85	**61.54**	51.54	**62.31**
church	14.17	10.83	41.67	26.67	**49.17**	25.83	35.00	42.50	**48.33**	39.17	**67.50**
commercial	9.14	7.43	**27.43**	18.29	23.43	18.86	14.86	24.57	**31.43**	20.00	**28.00**
dense residential	8.29	5.37	24.39	16.10	**25.37**	21.95	14.63	23.41	**37.07**	22.44	**32.68**
desert	22.67	23.33	**39.33**	20.00	35.33	24.00	**38.67**	38.00	32.67	32.67	**51.33**
farmland	14.05	5.95	13.51	14.59	**22.16**	7.03	20.00	**20.54**	17.84	17.84	**31.89**
forest	18.40	20.80	**34.40**	18.40	**36.80**	20.00	28.80	26.40	**40.80**	32.80	**40.80**
industrial	13.85	10.26	**29.74**	15.90	**29.74**	27.69	22.56	**32.82**	**32.82**	27.69	**36.41**
meadow	20.00	18.57	**44.29**	24.29	**40.71**	23.57	27.14	35.71	**47.86**	30.00	**47.86**
medium residential	5.52	8.28	17.93	8.97	**20.69**	13.10	20.00	17.24	**20.69**	**21.38**	**30.34**
mountain	18.24	23.53	29.41	18.82	**34.12**	20.00	22.35	25.29	**30.00**	24.71	**35.88**
park	18.29	16.57	31.43	24.57	**34.86**	30.29	33.14	32.00	**38.29**	28.57	**45.14**
parking	16.41	11.28	20.00	14.87	26.67	22.05	18.46	20.51	**27.18**	**27.69**	**38.97**
playground	22.40	17.60	38.60	30.20	**43.80**	28.80	30.60	42.00	**47.60**	42.60	**51.40**
pond	21.43	21.90	40.00	30.48	**44.29**	31.43	36.19	37.14	**42.86**	29.05	**56.19**
port	15.38	16.41	**34.87**	21.03	25.64	26.67	22.56	31.28	32.31	**36.92**	**45.64**
railway station	15.38	16.92	**45.38**	20.00	**43.85**	30.00	25.38	35.38	**45.38**	35.38	**46.15**
resort	18.62	20.00	35.86	28.28	37.93	33.10	37.93	**44.83**	**41.38**	32.41	**64.83**
river	17.07	12.20	**34.15**	19.02	30.24	31.71	28.29	26.83	**33.17**	26.83	**41.95**
school	14.00	14.00	33.33	18.00	**35.33**	16.00	19.33	31.33	**38.67**	24.67	**40.00**
sparse residential	17.33	20.00	**54.00**	25.33	**40.00**	27.33	30.00	34.67	38.67	33.33	**52.67**
square	18.79	16.97	**42.42**	29.09	35.15	20.00	37.58	40.61	**49.70**	**42.42**	**56.97**
stadium	22.76	18.62	37.24	37.24	38.62	35.17	37.93	**49.66**	**50.34**	**51.72**	**51.72**
storage tanks	24.50	18.00	41.00	28.00	**42.00**	24.50	35.00	**44.00**	41.00	39.50	**50.00**
viaduct	20.95	19.05	22.86	21.43	**32.86**	20.95	18.10	23.33	**30.48**	26.67	**40.00**

Notes. The **highest**, **second-highest**, and **third-highest** distinct values in each scene row are marked in red, blue, and green, respectively. Ties share the same color; ranking uses the reported two-decimal values.

Table 13: Scene-level Recall@5 (%) on RSITMD for image-to-text (I2T) retrieval. Higher is better.

Scene	*AMFMN* 2022	*GaLR* 2022	*PE* 2023	*PIR* 2023	*FGVLA* 2025	*LAPS* 2024	*CUP* 2025	*EAPA* 2025	*McRSA* 2026	*RRSITR* 2026	*P²Cal*
airport	33.33	41.67	**58.33**	50.00	50.00	50.00	**58.33**	**66.67**	**66.67**	**66.67**	**83.33**
bare land	**60.00**	**40.00**	**60.00**	**40.00**	**80.00**	**60.00**	**40.00**	**40.00**	**40.00**	**40.00**	**40.00**
baseball field	44.44	55.56	66.67	**77.78**	**100.00**	55.56	44.44	**88.89**	**77.78**	**88.89**	**100.00**
seashore	**26.67**	**53.33**	**26.67**	**40.00**	**26.67**	**26.67**	**40.00**	**26.67**	**40.00**	**40.00**	**53.33**
ship	0.00	**25.00**	**25.00**	**50.00**	**50.00**	**25.00**	**25.00**	**50.00**	**75.00**	**75.00**	**75.00**
bridge	26.67	26.67	26.67	40.00	53.33	40.00	33.33	**60.00**	**73.33**	40.00	**66.67**
center	21.43	35.71	**85.71**	28.57	**92.86**	28.57	57.14	**85.71**	**71.43**	**71.43**	**85.71**
church	46.15	53.85	**61.54**	38.46	**92.31**	46.15	53.85	**61.54**	38.46	**76.92**	**76.92**
commercial area	**35.71**	21.43	**35.71**	**50.00**	**50.00**	**50.00**	28.57	14.29	**35.71**	**50.00**	**57.14**
dense residential	**20.00**	**26.67**	**13.33**	**20.00**	**26.67**	**13.33**	6.67	**26.67**	**26.67**	**13.33**	**26.67**
desert	**41.67**	**41.67**	**41.67**	**41.67**	**58.33**	**41.67**	25.00	**66.67**	**66.67**	**41.67**	**66.67**
farmland	41.18	17.65	**58.82**	41.18	**58.82**	35.29	**58.82**	**70.59**	35.29	**52.94**	**70.59**
forest	**63.64**	**54.55**	**72.73**	**63.64**	**54.55**	36.36	27.27	**54.55**	**72.73**	45.45	45.45
industrial area	40.00	30.00	**55.00**	**60.00**	40.00	40.00	35.00	40.00	**50.00**	35.00	**55.00**
crossroad	**0.00**	**0.00**	**100.00**	**100.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
meadow	25.00	**41.67**	**41.67**	25.00	**50.00**	**33.33**	**41.67**	**33.33**	**41.67**	**41.67**	**33.33**
medium residential	30.77	30.77	**53.85**	**38.46**	30.77	15.38	30.77	23.08	30.77	23.08	**46.15**
mountain	**40.00**	30.00	**40.00**	**40.00**	**40.00**	**50.00**	**40.00**	30.00	30.00	**50.00**	**60.00**
park	25.00	25.00	**43.75**	**43.75**	**50.00**	25.00	**56.25**	37.50	31.25	**50.00**	**43.75**
parking lot	21.43	21.43	**42.86**	28.57	**35.71**	**42.86**	21.43	**50.00**	**35.71**	28.57	**42.86**
amusement park	22.22	11.11	**44.44**	33.33	**38.89**	27.78	22.22	**38.89**	**44.44**	**38.89**	**50.00**
pond	15.00	45.00	**50.00**	35.00	**55.00**	40.00	**50.00**	45.00	**50.00**	**50.00**	**75.00**
port	23.53	**35.29**	23.53	**35.29**	**41.18**	**47.06**	23.53	**47.06**	**35.29**	**41.18**	29.41
railway station	36.36	9.09	36.36	**45.45**	**54.55**	**54.55**	27.27	**45.45**	**72.73**	**54.55**	**54.55**
resort	31.25	18.75	**68.75**	50.00	**68.75**	31.25	**56.25**	50.00	**56.25**	**62.50**	43.75
river	16.67	22.22	**44.44**	**33.33**	27.78	16.67	**33.33**	**44.44**	**38.89**	**44.44**	**38.89**
school	18.75	**50.00**	**43.75**	**43.75**	**43.75**	**56.25**	37.50	**50.00**	**43.75**	**43.75**	**43.75**
sparse residential	13.33	**26.67**	**33.33**	**26.67**	**33.33**	**33.33**	**40.00**	**40.00**	**40.00**	**33.33**	**33.33**
square	10.53	10.53	**36.84**	21.05	**36.84**	**31.58**	26.32	**36.84**	**36.84**	**31.58**	**42.11**
stadium	21.05	10.53	**52.63**	42.11	42.11	42.11	31.58	**52.63**	**68.42**	36.84	**57.89**
storage tank	**26.09**	21.74	**26.09**	**26.09**	**26.09**	17.39	**26.09**	**34.78**	**43.48**	**34.78**	**43.48**
commercial bridge	16.67	**38.89**	16.67	**27.78**	**33.33**	16.67	**27.78**	**38.89**	22.22	**33.33**	**38.89**

Notes. The **highest**, **second-highest**, and **third-highest** distinct values in each scene row are marked in red, blue, and green, respectively. Ties share the same color; ranking uses the reported two-decimal values.

Table 14: Scene-level Recall@5 (%) on RSITMD for text-to-image (T2I) retrieval. Higher is better.

Scene	*AMFMN* 2022	*GaLR* 2022	*PE* 2023	*PIR* 2023	*FGVLA* 2025	*LAPS* 2024	*CUP* 2025	*EAPA* 2025	*McRSA* 2026	*RRSITR* 2026	*P²Cal*
airport	48.33	**68.33**	56.67	61.67	**66.67**	51.67	46.67	65.00	**68.33**	65.00	**76.67**
bare land	28.00	28.00	44.00	48.00	**72.00**	28.00	**52.00**	32.00	**56.00**	48.00	44.00
baseball field	68.89	66.67	**80.00**	68.89	**86.67**	68.89	71.11	**93.33**	**80.00**	**86.67**	**86.67**
seashore	36.00	48.00	**57.33**	52.00	**57.33**	37.33	37.33	**64.00**	**64.00**	52.00	**65.33**
ship	15.00	40.00	**65.00**	50.00	**65.00**	45.00	40.00	50.00	**65.00**	**60.00**	**70.00**
bridge	34.67	41.33	34.67	40.00	**68.00**	60.00	49.33	**61.33**	**89.33**	41.33	**68.00**
center	44.29	38.57	**71.43**	38.57	**78.57**	38.57	54.29	**71.43**	70.00	**75.71**	**78.57**
church	50.77	43.08	**84.62**	61.54	**80.00**	58.46	69.23	**84.62**	63.08	**92.31**	**80.00**
commercial area	21.43	27.14	31.43	32.86	**47.14**	34.29	25.71	42.86	38.57	**45.71**	**48.57**
dense residential	20.00	14.67	4.00	25.33	**30.67**	16.00	4.00	**26.67**	**38.67**	24.00	25.33
desert	38.33	35.00	**61.67**	35.00	**65.00**	53.33	53.33	**65.00**	**70.00**	58.33	60.00
farmland	30.59	29.41	**54.12**	32.94	**61.18**	45.88	51.76	51.76	36.47	47.06	**65.88**
forest	49.09	58.18	**67.27**	65.45	54.55	61.82	49.09	60.00	**83.64**	60.00	**76.36**
industrial area	48.00	35.00	**58.00**	52.00	50.00	45.00	51.00	**55.00**	52.00	43.00	**57.00**
crossroad	0.00	0.00	**60.00**	**20.00**	0.00	0.00	0.00	**20.00**	**20.00**	0.00	**40.00**
meadow	25.00	36.67	**68.33**	41.67	**48.33**	45.00	**48.33**	46.67	46.67	**56.67**	45.00
medium residential	15.38	23.08	23.08	24.62	**41.54**	26.15	15.38	**35.38**	29.23	18.46	**38.46**
mountain	50.00	60.00	**70.00**	58.00	60.00	**70.00**	62.00	**70.00**	**66.00**	64.00	**74.00**
park	33.75	30.00	**45.00**	36.25	**52.50**	35.00	36.25	38.75	40.00	42.50	**51.25**

Continued on the next page.

Table 14 (continued)

Scene	*AMFMN* 2022	*GaLR* 2022	*PE* 2023	*PIR* 2023	*FGVLA* 2025	*LAPS* 2024	*CUP* 2025	*EAPA* 2025	*McRSA* 2026	*RRSITR* 2026	*P²Cal*
parking lot	24.29	40.00	**54.29**	34.29	**54.29**	**45.71**	37.14	**51.43**	**45.71**	44.29	**54.29**
amusement park	28.89	31.11	**53.33**	33.33	**51.11**	35.56	36.67	50.00	48.89	44.44	**58.89**
pond	21.00	40.00	55.00	45.00	**66.00**	43.00	**63.00**	58.00	53.00	54.00	**70.00**
port	38.82	32.94	43.53	43.53	**45.88**	37.65	34.12	**44.71**	29.41	**47.06**	32.94
railway station	27.27	40.00	58.18	58.18	**72.73**	52.73	65.45	**80.00**	**76.36**	70.91	**80.00**
resort	33.75	38.75	51.25	41.25	**60.00**	38.75	55.00	43.75	53.75	**56.25**	**71.25**
river	28.89	32.22	54.44	54.44	**62.22**	36.67	**62.22**	46.67	**60.00**	**62.22**	**70.00**
school	37.50	38.75	**51.25**	41.25	41.25	41.25	35.00	**50.00**	**51.25**	40.00	**45.00**
sparse residential	21.33	30.67	**34.67**	20.00	28.00	32.00	26.67	30.67	**48.00**	32.00	**42.67**
square	10.53	12.63	**41.05**	24.21	34.74	26.32	33.68	34.74	**38.95**	24.21	**55.79**
stadium	41.05	27.37	**58.95**	42.11	**58.95**	35.79	53.68	**61.05**	57.89	53.68	**65.26**
storage tank	31.30	26.96	40.87	37.39	**44.35**	43.48	30.43	**45.22**	40.00	42.61	**52.17**
commercial bridge	25.56	26.67	**46.67**	26.67	**46.67**	23.33	32.22	35.56	28.89	**40.00**	**50.00**

Notes. The **highest**, **second-highest**, and **third-highest** distinct values in each scene row are marked in red, blue, and green, respectively. Ties share the same color; ranking uses the reported two-decimal values.

Table 15: Scene-level Recall@5 (%) on NWPU for image-to-text (I2T) retrieval. Higher is better.

Scene	*AMFMN* 2022	*GaLR* 2022	*PE* 2023	*PIR* 2023	*FGVLA* 2025	*LAPS* 2024	*CUP* 2025	*EAPA* 2025	*McRSA* 2026	*RRSITR* 2026	*P²Cal*
airplane	62.86	45.71	57.14	**71.43**	52.86	57.14	45.71	**74.29**	**70.00**	**70.00**	**71.43**
airport	41.43	40.00	44.29	**64.29**	48.57	48.57	27.14	**62.86**	57.14	57.14	**60.00**
baseball diamond	58.57	45.71	55.71	64.29	68.57	54.29	35.71	**72.86**	68.57	**70.00**	**84.29**
basketball court	51.43	58.57	54.29	67.14	60.00	61.43	41.43	**78.57**	**70.00**	68.57	**75.71**
beach	31.43	41.43	37.14	55.71	45.71	44.29	32.86	**67.14**	60.00	**61.43**	**64.29**
bridge	35.71	38.57	37.14	**65.71**	40.00	41.43	34.29	**74.29**	64.29	**65.71**	**78.57**
chaparral	34.29	41.43	31.43	**50.00**	40.00	38.57	15.71	**47.14**	40.00	**44.29**	**44.29**
church	42.86	44.29	47.14	52.86	57.14	**71.43**	47.14	67.14	68.57	**72.86**	**87.14**
circular farmland	51.43	45.71	44.29	**65.71**	51.43	60.00	21.43	**71.43**	**61.43**	**61.43**	**65.71**
cloud	54.29	62.86	44.29	51.43	58.57	48.57	41.43	**72.86**	64.29	**71.43**	**65.71**
commercial area	24.29	32.86	34.29	**50.00**	28.57	32.86	11.43	**51.43**	**51.43**	**54.29**	**51.43**
dense residential	41.43	38.57	35.71	**60.00**	38.57	42.86	18.57	**65.71**	**62.86**	**60.00**	**62.86**
desert	44.29	40.00	34.29	57.14	31.43	32.86	31.43	62.86	**67.14**	**74.29**	**64.29**
forest	52.86	42.86	34.29	50.00	40.00	37.14	35.71	**54.29**	52.86	**62.86**	**58.57**
freeway	60.00	55.71	62.86	65.71	64.29	61.43	34.29	**78.57**	65.71	**71.43**	**81.43**
golf course	44.29	45.71	35.71	60.00	51.43	50.00	17.14	**71.43**	**71.43**	**65.71**	**70.00**
ground track field	47.14	44.29	45.71	55.71	44.29	44.29	21.43	**62.86**	**62.86**	**60.00**	**65.71**
harbor	37.14	27.14	31.43	**48.57**	35.71	37.14	27.14	44.29	**58.57**	**51.43**	**51.43**
industrial area	37.14	48.57	35.71	58.57	51.43	51.43	37.14	**61.43**	**64.29**	57.14	**68.57**
intersection	41.43	50.00	34.29	62.86	52.86	45.71	34.29	**81.43**	64.29	**71.43**	**67.14**
island	44.29	31.43	42.86	47.14	42.86	35.71	40.00	**68.57**	**57.14**	54.29	**64.29**
lake	48.57	38.57	31.43	60.00	41.43	52.86	22.86	**61.43**	**61.43**	**70.00**	**64.29**
meadow	37.14	**42.86**	28.57	**52.86**	41.43	37.14	25.71	**54.29**	**52.86**	**54.29**	**54.29**
medium residential	27.14	32.86	28.57	**45.71**	31.43	35.71	12.86	**51.43**	**51.43**	**51.43**	**52.86**
mobile home park	51.43	45.71	44.29	**65.71**	50.00	54.29	37.14	**71.43**	58.57	60.00	**70.00**
mountain	30.00	45.71	31.43	45.71	38.57	38.57	18.57	**60.00**	**51.43**	**50.00**	38.57
overpass	47.14	54.29	47.14	**58.57**	52.86	47.14	40.00	57.14	51.43	**60.00**	**67.14**
palace	44.29	30.00	48.57	60.00	40.00	54.29	40.00	**67.14**	**64.29**	55.71	**77.14**
parking lot	44.29	45.71	35.71	**55.71**	41.43	48.57	12.86	**52.86**	47.14	**62.86**	51.43
railway	44.29	32.86	47.14	**64.29**	48.57	47.14	31.43	61.43	**64.29**	**67.14**	**62.86**
railway station	42.86	32.86	32.86	45.71	45.71	42.86	45.71	**65.71**	55.71	**61.43**	**68.57**
rectangular farmland	45.71	51.43	40.00	**61.43**	52.86	52.86	31.43	**67.14**	**67.14**	60.00	**68.57**
river	47.14	65.71	50.00	62.86	58.57	60.00	61.43	**77.14**	**70.00**	**74.29**	**74.29**
roundabout	52.86	50.00	47.14	61.43	50.00	58.57	31.43	**68.57**	**62.86**	**68.57**	**67.14**
runway	42.86	34.29	35.71	**50.00**	31.43	47.14	32.86	47.14	**65.71**	**57.14**	**57.14**
sea ice	37.14	37.14	20.00	38.57	31.43	34.29	14.29	**62.86**	**42.86**	40.00	**41.43**
ship	35.71	47.14	52.86	57.14	50.00	52.86	55.71	60.00	**62.86**	**64.29**	**77.14**
snowberg	31.43	28.57	30.00	**38.57**	32.86	31.43	17.14	30.00	**42.86**	**40.00**	31.43
sparse residential	37.14	38.57	34.29	54.29	51.43	40.00	31.43	52.86	**67.14**	**55.71**	**68.57**

Continued on the next page.

Table 15 (continued)

Scene	*AMFMN* 2022	*GaLR* 2022	*PE* 2023	*PIR* 2023	*FGVLA* 2025	*LAPS* 2024	*CUP* 2025	*EAPA* 2025	*McRSA* 2026	*RRSITR* 2026	*P²Cal*
stadium	48.57	42.86	51.43	**75.71**	50.00	52.86	34.29	70.00	**71.43**	70.00	**72.86**
storage tank	30.00	47.14	41.43	45.71	32.86	40.00	24.29	**58.57**	**55.71**	**52.86**	50.00
tennis court	41.43	37.14	51.43	**58.57**	37.14	44.29	21.43	**62.86**	45.71	51.43	**54.29**
terrace	18.57	24.29	21.43	**42.86**	18.57	30.00	17.14	**45.71**	**40.00**	**42.86**	38.57
thermal power station	31.43	30.00	15.71	47.14	28.57	27.14	24.29	**61.43**	**50.00**	**50.00**	**64.29**
wetland	22.86	12.86	21.43	**31.43**	24.29	22.86	15.71	**40.00**	**35.71**	25.71	25.71

Notes. The **highest**, **second-highest**, and **third-highest** distinct values in each scene row are marked in red, blue, and green, respectively. Ties share the same color; ranking uses the reported two-decimal values.

Table 16: Scene-level Recall@5 (%) on NWPU for text-to-image (T2I) retrieval. Higher is better.

Scene	*AMFMN* 2022	*GaLR* 2022	*PE* 2023	*PIR* 2023	*FGVLA* 2025	*LAPS* 2024	*CUP* 2025	*EAPA* 2025	*McRSA* 2026	*RRSITR* 2026	*P²Cal*
airplane	53.14	51.14	53.14	60.86	46.86	51.14	44.57	66.57	**67.43**	**68.86**	**67.71**
airport	30.57	30.00	28.57	35.43	30.29	32.00	27.14	**39.43**	**39.43**	**38.86**	**40.00**
baseball diamond	42.86	45.71	41.71	49.43	43.43	46.00	36.57	54.00	**55.71**	**58.57**	**60.00**
basketball court	41.14	45.43	44.00	46.57	46.57	46.00	36.57	**56.86**	53.14	**53.43**	**59.71**
beach	26.00	32.29	30.57	35.71	33.14	35.43	29.14	41.71	**42.00**	**43.71**	**43.14**
bridge	36.57	33.14	31.71	42.00	32.00	35.14	28.00	**43.14**	**48.00**	41.71	**50.57**
chaparral	28.57	28.57	20.00	**29.43**	24.29	28.29	14.29	**34.00**	**33.71**	**33.71**	28.86
church	36.86	33.71	38.86	39.71	40.86	41.14	31.43	**51.14**	50.29	**55.71**	**57.71**
circular farmland	37.43	37.14	31.43	40.57	34.29	34.29	20.00	**46.57**	**47.14**	44.86	**45.71**
cloud	41.14	44.29	38.00	46.57	43.71	41.14	31.43	**48.86**	**50.57**	**51.14**	48.00
commercial area	26.29	27.71	25.14	34.29	26.57	30.00	18.57	33.43	**36.29**	**37.71**	**34.86**
dense residential	30.57	28.00	26.29	32.29	26.86	28.57	18.57	**34.86**	**38.29**	**34.86**	**36.29**
desert	31.71	30.57	28.00	36.29	31.14	32.86	21.71	**42.29**	**41.71**	**42.57**	39.43
forest	36.86	35.14	29.43	37.71	28.00	32.00	21.43	**42.00**	**41.43**	**42.00**	**39.43**
freeway	47.43	46.29	47.43	48.86	48.00	52.00	37.71	**56.86**	**54.57**	50.86	**58.00**
golf course	31.43	33.43	28.29	35.71	31.14	31.14	20.29	**38.57**	**40.57**	**38.57**	**41.43**
ground track field	32.29	32.29	31.43	34.86	30.00	30.29	25.71	**38.29**	**37.43**	**41.71**	**38.29**
harbor	25.43	21.14	21.43	30.86	22.00	24.86	20.00	30.57	**32.00**	**32.86**	**32.29**
industrial area	30.00	29.71	24.57	32.57	30.29	30.00	20.86	**36.29**	**36.00**	**38.00**	**36.29**
intersection	30.00	31.71	27.14	36.57	29.43	32.29	20.86	**38.57**	**38.29**	**38.29**	**38.86**
island	32.86	30.57	25.14	38.57	28.57	31.43	24.86	**42.86**	**42.00**	38.00	**43.71**
lake	36.00	30.57	29.71	38.00	33.14	34.29	26.00	**40.86**	**43.43**	40.00	**41.71**
meadow	24.00	27.14	26.00	**31.71**	26.00	28.29	22.00	**31.43**	30.29	**32.29**	**31.71**
medium residential	26.29	24.29	20.57	28.00	25.14	25.14	18.86	30.57	**31.43**	**32.00**	**33.14**
mobile home park	30.86	32.29	25.43	**34.86**	29.71	33.14	20.57	**36.86**	**37.71**	**36.86**	**36.86**
mountain	23.14	25.71	20.86	28.86	26.00	27.43	17.43	**32.86**	**31.43**	**31.14**	29.14
overpass	34.00	35.14	34.00	37.43	35.71	36.29	26.57	**42.86**	**43.14**	42.00	**44.57**
palace	28.00	31.14	28.86	36.57	34.00	35.14	24.86	**42.86**	**41.43**	39.71	**44.86**
parking lot	30.29	28.86	28.86	32.86	26.86	30.00	20.86	**36.00**	**36.00**	**39.14**	**35.43**
railway	32.29	29.43	32.00	36.57	30.00	32.57	25.43	**40.29**	**42.57**	39.71	**40.86**
railway station	28.57	30.57	22.86	**34.00**	30.00	32.00	20.00	33.14	**34.00**	**34.57**	**37.71**
rectangular farmland	39.14	39.71	31.43	**44.29**	37.43	34.86	23.14	**45.43**	**46.29**	41.43	**45.43**
river	36.86	41.71	36.86	50.00	40.57	40.29	32.86	**52.00**	50.57	**52.86**	**54.00**
roundabout	42.86	42.86	37.43	46.57	40.86	46.00	28.86	**47.43**	**49.14**	47.14	**51.43**
runway	33.14	29.71	26.57	37.43	30.57	29.71	22.29	**38.86**	**45.14**	**39.43**	38.57
sea ice	23.71	25.71	22.57	28.86	20.00	26.00	13.14	**30.86**	**30.86**	**30.29**	**30.00**
ship	44.00	44.57	41.14	53.14	40.29	46.57	37.71	**58.57**	**55.14**	54.57	**62.00**
snowberg	27.43	25.14	22.29	**30.00**	23.71	25.43	17.43	27.43	**30.86**	**28.57**	28.29
sparse residential	28.29	28.57	28.86	32.00	30.29	29.71	23.71	**37.43**	**35.43**	34.00	**38.86**
stadium	29.43	32.29	34.29	39.71	33.71	35.43	26.57	40.86	**41.71**	**42.86**	**42.57**
storage tank	28.86	32.86	28.57	34.86	32.00	32.29	22.57	**36.57**	**38.00**	34.57	**38.57**
tennis court	31.43	29.71	32.57	32.86	30.00	30.57	24.00	**38.29**	**34.86**	**38.29**	**38.86**
terrace	19.14	19.14	14.57	22.29	18.00	20.86	14.86	**24.86**	**24.00**	**23.43**	22.57
thermal power station	21.71	25.43	16.57	26.29	22.00	21.43	14.57	**32.86**	**32.00**	30.57	**31.14**

wetland	17.14	18.57	14.00	21.43	18.29	16.57	12.29	**21.71**	**23.43**	**22.86**	**21.71**

Notes. The **highest**, **second-highest**, and **third-highest** distinct values in each scene row are marked in red, blue, and green, respectively. Ties share the same color; ranking uses the reported two-decimal values.

P^{2}Cal ranks among the top three in all 30 RSICD categories in both retrieval directions. It achieves top-three rankings in 29 and 27 of the 32 RSITMD categories, and in 39 and 38 of the 45 NWPU categories, for I2T and T2I, respectively. These counts include ties under the distinct-value ranking used in the tables. The results demonstrate consistently strong rankings across diverse scene types and both retrieval directions.

Bridges and railway stations repeatedly emerge as favorable scenes, while churches also show consistent advantages on RSICD and NWPU. Their distinctive objects and spatial layouts provide informative cues for cross-modal matching. MRA helps adapt pretrained visual, textual, and cross-modal representations to these cues; PSCM reduces inconsistent supervision that could weaken their associations; and PSCF refines the ordering of competing candidates. This collaboration provides a plausible explanation for the recurring strengths in scenes with recognizable structures. Clear gains also appear in challenging categories such as RSICD viaduct and school, while the strong results on RSITMD pond show that the benefits extend beyond built environments.

Repeated layouts and insufficiently distinctive descriptions remain challenging, particularly for T2I retrieval in dense residential scenes. Natural scenes show more varied behavior: P^{2}Cal leads both directions on RSICD desert and RSITMD mountain, whereas NWPU wetland remains difficult for all methods. These patterns suggest that the availability of distinguishing visual and textual cues matters more than a simple division between built and natural scenes. Embedding adaptation and supervision purification can improve matching, but cannot fully resolve ambiguity when such cues are absent. The qualitative examples in Appendix A.6 further illustrate these strengths and remaining limitations.

A.6 Qualitative Retrieval Analysis

To connect the scene-level findings in Appendix A.5.2 with individual retrieval decisions, we examine twelve examples across RSICD, RSITMD, and NWPU in Figures 4-15. Success and failure here refer to the first-ranked result. Green and red retain the original correspondence labels.

A.6.1 RSICD

In the bare-land and baseball-field examples (Figures 4 and 6), P^{2}Cal places the paired result first, while competing methods prioritize unpaired texts or images. The bare-land example shows an advantage in distinguishing paired texts from alternatives sharing the same broad scene category. In the baseball-field example, the playground and nearby blue building provide contextual cues beyond the field itself. These observations are consistent with the earlier strengths in scenes with recognizable structures and spatial context. MRA can help align these visual and textual cues, while reliable supervision from PSCM and score refinement through PSCF support the ordering of competing matches.

The desert and adjacent-court examples show the opposite outcome (Figures 5 and 7): several competing methods retrieve a paired result first, whereas P^{2}Cal places it below an unpaired candidate. In the desert case, descriptions of irregular patterns and rocky hills precede the paired text; in the court case, another sports complex is preferred despite the query's specific combination of neighboring facilities. These errors suggest that fine terrain patterns and local arrangements are not always sufficiently distinguished by the final scores. Strong scene-level performance can therefore coexist with individual ordering errors among closely related candidates.



Figure 4: RSICD I2T, bare-land query: P^{2}Cal returns correctly paired descriptions at ranks 1 and 2 despite many similar scene descriptions.



Figure 5: RSICD I2T, desert query: P^{2}Cal reaches the paired text at rank 5, after descriptions of similar dry terrain.

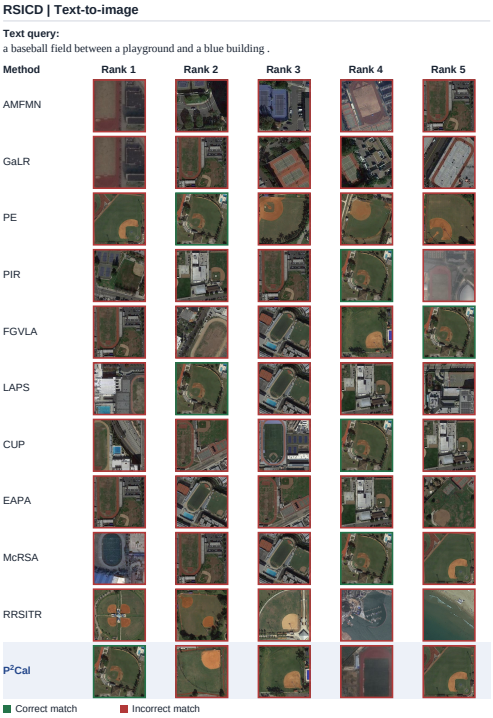

Figure 6: RSICD T2I, baseball-field query: P^{2}Cal ranks the scene with the annotated surrounding facilities first.

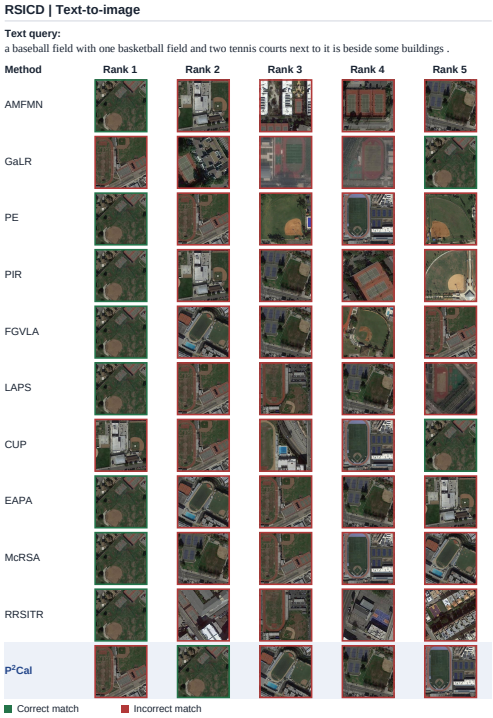

Figure 7: RSICD T2I, adjacent-court query: the paired image appears at rank 2 for P^{2}Cal, making this a top-1 error rather than a top-5 miss.

A.6.2 RSITMD

For the beach and short bridge queries (Figures 8 and 10), P^{2}Cal retrieves the paired result first, while competing methods select unpaired alternatives. The beach query contains a distinctive combination of shoreline, green water, and a white wave boundary, and the bridge query provides a recognizable crossing over water. These examples agree with the preceding scene analysis: informative matching cues can support retrieval in both natural and built environments. MRA's adaptation across the visual, textual, and cross-modal pathways, together with PSCM-filtered supervision and PSCF score refinement, provides a plausible basis for exploiting these cues more effectively.

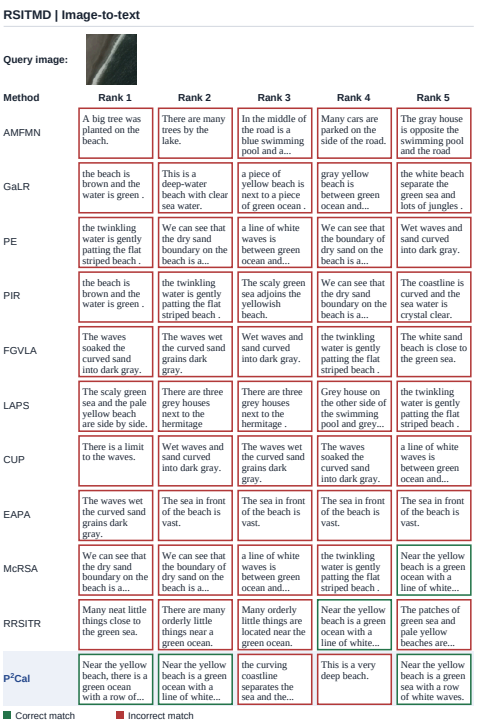

Figure 8: RSITMD I2T, beach query: P^{2}Cal retrieves correctly paired descriptions at ranks 1, 2, and 5.



Figure 9: RSITMD I2T, mountain query: all five P^{2}Cal results are incorrectly paired despite their mountain-related wording.

For the mountain and longer bridge queries (Figures 9 and 11), several competing methods recover the paired result first, whereas P^{2}Cal does not. Its mountain results emphasize snow-covered ridges or generic mountain content without recovering a paired text among the displayed candidates. In the longer bridge query, it prefers a different crossing despite the additional red- and blue-roof details. The contrast with the successful examples suggests that recognizing the dominant scene is easier than resolving all instance-specific attributes and spatial relations. These cases illustrate remaining fine-grained matching errors within scene types where the framework otherwise shows strong results.

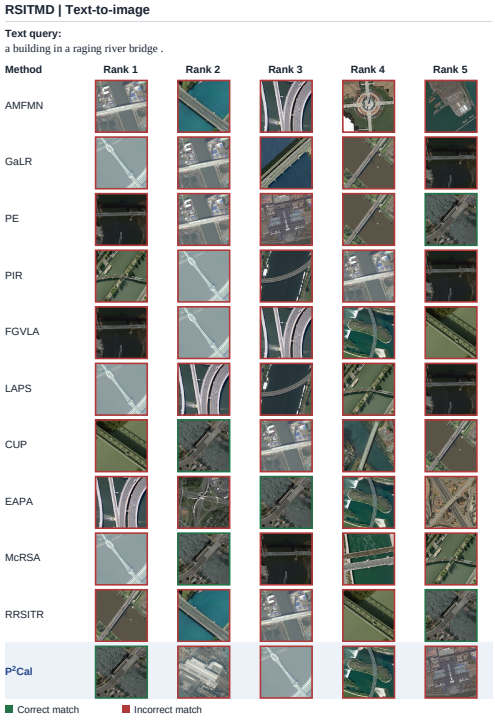

Figure 10: RSITMD T2I, short bridge query: P^{2}Cal ranks the paired bridge-over-water scene first.

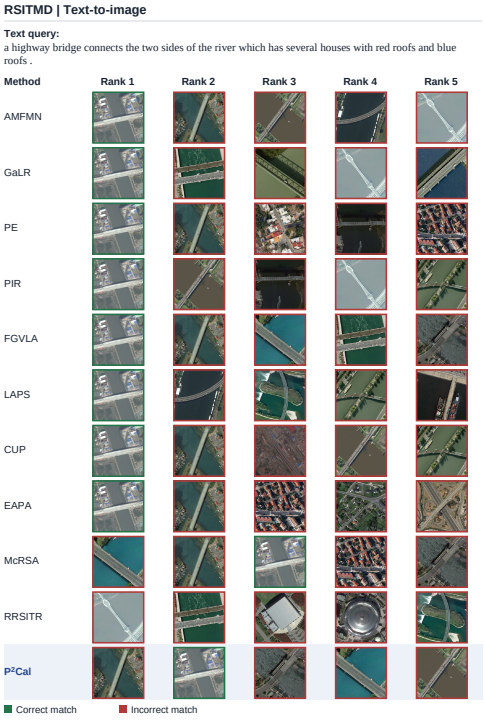

Figure 11: RSITMD T2I, riverside-building query: P^{2}Cal ranks a different crossing first and the paired scene second.

A.6.3 NWPU

In the aircraft and sports-facility examples (Figures 12 and 14), P^{2}Cal ranks a paired result first while competing methods begin with unpaired alternatives. Aircraft arrangement and the court's relation to parking, a baseball diamond, and housing provide concrete cues beyond generic object presence. This comparison reinforces the earlier advantage in scenes with distinctive objects and layouts. MRA can adapt the embeddings to these relationships, while PSCF refines candidate ordering using the reliable supervision shared with the discriminative retriever.

The cloud and meadow examples reverse this comparison (Figures 13 and 15): competing methods can retrieve a paired result first, but P^{2}Cal initially favors an unpaired one. Correctly labelled results remain farther down its lists, indicating an ordering failure in these cases. Repetitive cloud descriptions and the short grass description offer fewer distinguishing cues than the explicit object configurations above, which may make score refinement more sensitive to similarities shared by many candidates. This agrees with the earlier observation that natural scenes vary according to the specificity of their visual and textual cues, rather than being uniformly difficult.



Figure 12: NWPU I2T, aircraft query: P^{2}Cal retrieves a correct description at rank 1; competing matches first occur at ranks 2 or 3.



Figure 13: NWPU I2T, cloud query: the paired text appears at rank 4 for P^{2}Cal, following three similar but incorrectly matched descriptions.

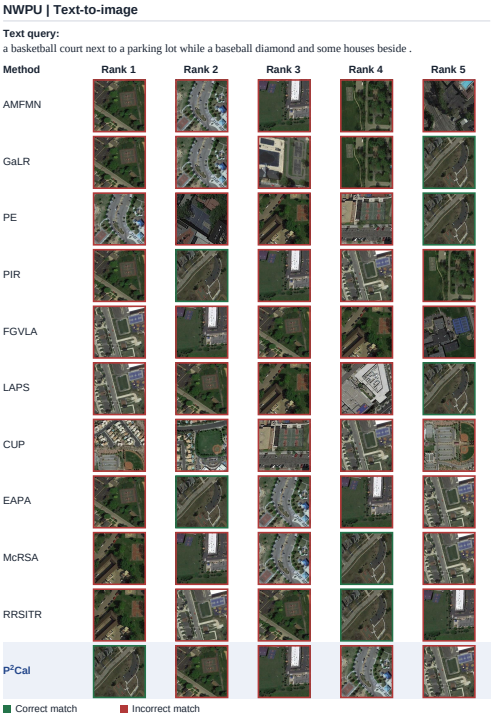

Figure 14: NWPU T2I, sports-facility query: P^{2}Cal ranks the annotated court and surrounding scene first.

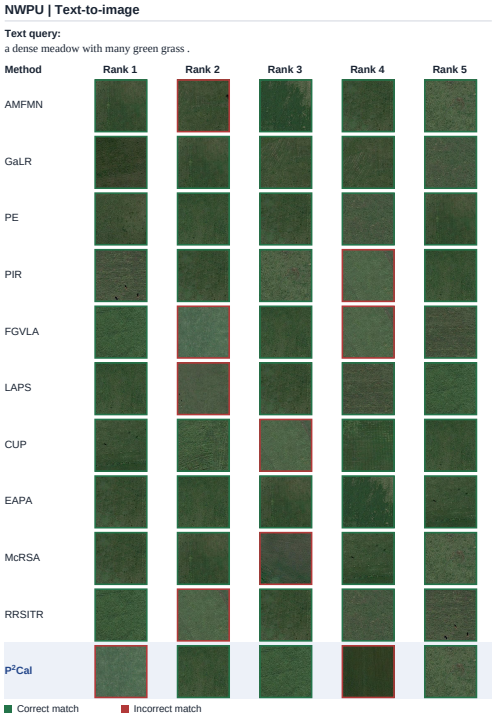

Figure 15: NWPU T2I, meadow query: P^{2}Cal fails at rank 1 but retains correctly labelled images at ranks 2, 3, and 5.

Limitations. Across these examples, precise counts, spatial relations, and insufficiently distinctive descriptions remain challenging. PSCM cannot supply missing query detail, and PSCF cannot recover a target outside its ITC candidates. The mountain panel alone does not distinguish candidate exclusion from later ranking errors. Semantically plausible unpaired results also show that correspondence labels need not capture all relevance.

A.7 RSICD Visual Degradation Robustness

Motivated by studies of haze, stripe noise, imaging blur, limited spatial resolution, and lossy compression in remote sensing (Qin et al., 2018; Chang et al., 2015; Li et al., 2017; Lanaras et al., 2018; Zabala & Pons, 2011), we construct synthetic visual degradations to evaluate retrieval under these imaging conditions.

A.7.1 Details of Experimental Settings

We use the original RSICD test split and apply five degradation types—haze, stripe noise, blur, low resolution, and JPEG compression—at severity $s \in \{1, 3, 5\}$. Each condition is applied independently to the original image I_i, producing an RGB image $I_i^{c,s}$ for degradation type c. The test pairs are $\{(I_i^{c,s}, T_{ij})\}_{i,j}$: texts and their original image associations remain unchanged. Degraded images serve as I2T queries and the T2I image gallery. All methods use the same RSICD checkpoints and retrieval settings as in the main experiments, without additional training. Evaluation follows the Recall and mean Recall definitions in Appendix A.4.1.

Algorithms 3-7 summarize the five constructions and their severity parameters. For an image of size $H \times W$, y and x index rows and columns, and a indexes RGB channels. The operator $Q_{\text{vis}}(z) = \lfloor \min\{255, \max\{0, z\}\} \rfloor$ keeps intensities within the valid integer range.

To evaluate retrieval under haze, we blend each image with a spatially varying white veil, following the mixture form of the atmospheric-scattering model (He et al., 2011). This weakens contrast without changing image-text associations; Algorithm 3 controls the veil strength across severity levels.

Algorithm 3 Haze Degradation

Require: Image I_i; severity $s \in \{1, 3, 5\}$
1: Select $(\lambda_1^{\text{haze}}, \lambda_3^{\text{haze}}, \lambda_5^{\text{haze}}) = (0.18, 0.38, 0.58)$.
2: Define the vertical veil profile $g(y) = 0.75 + 0.40y/(H-1)$.
3: Blend the image with the veil at each pixel and channel:
$I_i^{\text{haze},s}(y, x, a) = Q_{\text{vis}}\big((1 - \lambda_s^{\text{haze}} g(y)) I_i(y, x, a) + 255\lambda_s^{\text{haze}} g(y)\big)$.

To evaluate sensitivity to stripe noise, we introduce column-wise intensity offsets motivated by nonuniform detector responses (Chang et al., 2015). Sharing each offset across rows and channels produces coherent vertical bands while retaining the underlying scene layout. Algorithm 4 combines periodic, random, and sparse stronger offsets.

Algorithm 4 Stripe-Noise Degradation

Require: Image I_i; severity $s \in \{1, 3, 5\}$
1: Select amplitude A_s, period P_s, and sparse-offset probability p_s:
$(A_s, P_s, p_s) = (18, 18, 0.015), (38, 12, 0.035), (65, 8, 0.060)$ for $s = 1, 3, 5$.
2: For each column x, sample $\epsilon_x \sim \mathcal{N}(0, (0.35A_s)^2)$ and $b_x \sim \text{Bernoulli}(p_s)$.
3: Set $\xi_x = 0$ when $b_x = 0$; otherwise sample $\xi_x \in \{-1, +1\}$ with equal probability.
4: Form $\nu_x = A_s \sin(2\pi x/P_s) + \epsilon_x + 1.2A_s b_x \xi_x$.
5: Apply $I_i^{\text{stripe},s}(y, x, a) = Q_{\text{vis}}(I_i(y, x, a) + \nu_x)$ across rows and channels.

To evaluate retrieval when fine structures become blurred, we apply Gaussian smoothing followed by box smoothing (Gwosdek et al., 2012). Increasing the box radius progressively suppresses lo-

cal detail while preserving image dimensions, as specified in Algorithm 5. The table label "Blur" denotes this Gaussian-box construction rather than a directional motion kernel.

Algorithm 5 Blur Degradation

Require: Image I_i; severity $s \in \{1, 3, 5\}$
1: Select box radius r_s from $(r_1, r_3, r_5) = (2, 5, 8)$ pixels.
2: Apply Gaussian smoothing with radius 0.2 pixels to I_i.
3: Apply box smoothing of radius r_s to obtain $I_i^{\text{blur},s}$, with window width $2r_s + 1$.

To evaluate retrieval with reduced spatial resolution, we downsample each image and then restore its dimensions using bicubic interpolation (Keys, 1981). This removes fine spatial information while keeping the input dimensions unchanged. Algorithm 6 varies the reduction factor with severity.

Algorithm 6 Low-Resolution Degradation

Require: Image I_i of size $H \times W$; severity $s \in \{1, 3, 5\}$
1: Select reduction factor d_s from $(d_1, d_3, d_5) = (2, 4, 8)$.
2: Bicubic-resize I_i to $\max\{1, \lfloor H/d_s \rfloor\} \times \max\{1, \lfloor W/d_s \rfloor\}$.
3: Bicubic-resize the reduced image back to $H \times W$ to obtain $I_i^{\text{lowres},s}$.

To evaluate retrieval under lossy compression, we apply JPEG compression at progressively lower quality settings (Wallace, 1991). This introduces compression artifacts while preserving image dimensions and text associations. Algorithm 7 specifies the quality settings used at each severity.

Algorithm 7 JPEG-Compression Degradation

Require: Image I_i; severity $s \in \{1, 3, 5\}$
1: Select JPEG quality q_s from $(q_1, q_3, q_5) = (60, 35, 15)$.
2: Apply JPEG compression at quality q_s and decode the result to obtain $I_i^{\text{JPEG},s}$.

A.7.2 Results and Analysis

Tables 17-19 report retrieval performance across the five degradation types and their severity levels.

Table 17: Full visual degradation robustness on RSICD-Test. Text and ground-truth correspondences are unchanged; only the I2T image queries and T2I image gallery are replaced. This table reports severity-1 corruptions.

Method	$\text{Recall}_{\text{I2T}}$ @1	@5	@10	$\text{Recall}_{\text{T2I}}$ @1	@5	@10	mR_{I2T}	mR_{T2I}	mR
Haze (s=1)									
AMFMN TGRS'22	7.41	18.39	27.63	5.29	16.72	27.78	17.81	16.60	17.20
GaLR TGRS'22	6.04	16.47	24.61	3.71	15.48	24.30	15.71	14.50	15.11
PIR ACMMM'23	6.77	21.04	32.20	5.73	21.21	34.78	20.00	20.57	20.29
PE TGRS'23	14.09	32.48	46.20	9.97	32.39	49.04	30.92	30.47	30.70
LAPS CVPR'24	8.69	24.43	38.33	7.03	24.04	38.57	23.82	23.21	23.52
FGVLA TGRS'25	16.74	35.13	47.85	11.78	34.49	51.51	33.24	32.59	32.92
EAPA ACMMM'25	14.82	32.30	46.11	10.48	32.72	48.00	31.08	30.40	30.74
CUP TNNLS'25	9.15	23.97	37.05	7.58	26.55	42.29	23.39	25.47	24.43
McRSA TMM'26	16.10	34.03	49.13	13.89	36.65	52.94	33.09	34.49	33.79
RRSITR CVPR'26	14.64	33.58	45.93	10.91	30.67	45.86	31.38	29.15	30.27
P^2Cal	**24.70**	**44.92**	**56.72**	**18.81**	**43.62**	**59.74**	**42.11**	**40.72**	**41.41**

Table 17 (continued)

Method	$\text{Recall}_{\text{I2T}}$ @1	@5	@10	$\text{Recall}_{\text{T2I}}$ @1	@5	@10	mR_{I2T}	mR_{T2I}	mR
Stripe Noise (s=1)									
AMFMN TGRS'22	3.20	9.97	15.74	2.34	11.44	18.74	9.64	10.84	10.24
GaLR TGRS'22	5.58	11.44	17.38	3.07	10.85	18.23	11.47	10.72	11.10
PIR ACMMM'23	3.48	10.61	17.93	3.77	16.96	28.49	10.67	16.41	13.54
PE TGRS'23	8.78	22.42	32.85	7.78	26.83	39.47	21.35	24.69	23.02
LAPS CVPR'24	4.21	14.27	20.95	4.61	16.80	27.85	13.14	16.42	14.78
FGVLA TGRS'25	11.99	26.81	37.79	9.75	29.08	45.27	25.53	28.03	26.78
EAPA ACMMM'25	8.87	22.60	32.66	8.84	25.91	39.30	21.38	24.68	23.03
CUP TNNLS'25	5.67	16.56	27.17	6.00	20.40	34.31	16.47	20.24	18.35
McRSA TMM'26	13.63	26.62	37.79	10.80	31.31	46.94	26.01	29.68	27.84
RRSITR CVPR'26	11.16	27.08	39.89	8.76	26.86	39.43	26.04	25.02	25.53
P^2Cal	**19.67**	**36.32**	**48.49**	**15.70**	**38.85**	**53.25**	**34.83**	**35.93**	**35.38**
Blur (s=1)									
AMFMN TGRS'22	3.02	12.99	19.85	3.28	12.96	21.67	11.95	12.64	12.29
GaLR TGRS'22	4.39	12.72	17.84	3.22	11.84	19.65	11.65	11.57	11.61
PIR ACMMM'23	5.31	17.75	27.26	4.59	19.25	32.86	16.77	18.90	17.84
PE TGRS'23	11.44	28.73	40.81	8.58	29.62	46.11	26.99	28.10	27.55
LAPS CVPR'24	6.04	19.21	30.74	5.01	21.26	35.39	18.66	20.55	19.61
FGVLA TGRS'25	14.27	31.84	44.28	10.36	30.81	48.75	30.13	29.97	30.05
EAPA ACMMM'25	10.80	26.08	38.70	9.13	28.49	42.40	25.19	26.67	25.93
CUP TNNLS'25	6.68	19.76	30.47	6.48	23.57	38.39	18.97	22.81	20.89
McRSA TMM'26	14.91	28.55	40.44	11.60	32.39	49.06	27.97	31.02	29.49
RRSITR CVPR'26	11.80	28.18	39.62	9.52	27.83	41.85	26.53	26.40	26.46
P^2Cal	**17.11**	**35.77**	**47.48**	**14.78**	**36.14**	**51.36**	**33.45**	**34.09**	**33.77**
Low Resolution (s=1)									
AMFMN TGRS'22	5.40	15.55	25.71	4.28	15.26	25.64	15.55	15.06	15.30
GaLR TGRS'22	4.94	15.65	23.70	3.92	14.46	24.04	14.76	14.14	14.45
PIR ACMMM'23	6.22	21.32	33.12	5.40	21.26	35.32	20.22	20.66	20.44
PE TGRS'23	13.17	30.56	43.18	9.64	31.44	47.25	28.97	29.44	29.20
LAPS CVPR'24	8.42	21.96	34.03	6.33	23.18	38.59	21.47	22.70	22.09
FGVLA TGRS'25	16.29	33.39	46.29	11.78	33.61	50.72	31.99	32.04	32.02
EAPA ACMMM'25	13.17	32.94	43.28	10.25	30.34	46.26	29.80	28.95	29.38
CUP TNNLS'25	7.69	22.78	35.04	7.61	26.72	41.24	21.84	25.19	23.52
McRSA TMM'26	15.19	32.48	45.75	13.25	35.57	51.77	31.14	33.53	32.34
RRSITR CVPR'26	14.00	31.84	44.74	10.50	30.23	45.00	30.19	28.58	29.38
P^2Cal	**20.49**	**39.62**	**51.42**	**15.70**	**40.37**	**55.46**	**37.18**	**37.18**	**37.18**
JPEG Compression (s=1)									
AMFMN TGRS'22	5.03	15.74	25.34	4.12	15.83	25.98	15.37	15.31	15.34
GaLR TGRS'22	6.77	15.83	23.15	4.46	14.82	24.63	15.25	14.64	14.95
PIR ACMMM'23	4.30	13.08	21.96	4.76	18.33	31.42	13.11	18.17	15.64
PE TGRS'23	13.54	30.38	42.27	10.25	32.08	47.17	28.73	29.83	29.28
LAPS CVPR'24	7.59	21.77	33.67	6.28	22.67	36.96	21.01	21.97	21.49
FGVLA TGRS'25	11.53	30.92	42.18	10.50	33.89	50.28	28.21	31.56	29.88
EAPA ACMMM'25	10.70	28.64	40.35	10.61	32.35	46.72	26.56	29.89	28.23
CUP TNNLS'25	8.51	23.15	35.32	8.34	25.42	40.99	22.33	24.92	23.62
McRSA TMM'26	14.64	31.56	44.92	11.99	35.54	51.33	30.37	32.95	31.66
RRSITR CVPR'26	13.72	32.20	44.65	10.21	31.49	45.14	30.19	28.95	29.57
P^2Cal	**22.42**	**40.44**	**50.78**	**16.83**	**41.26**	**56.03**	**37.88**	**38.04**	**37.96**

P^2Cal maintains a clear advantage across degradation types and retrieval directions. It leads every reported metric under mild degradation and retains the highest overall mR for every degradation type at intermediate severity. The advantage is particularly consistent under haze and stripe noise, where it leads all Recall metrics at every tested severity; it also maintains the highest overall mR across JPEG compression levels. Thus, its robustness is visible in individual retrieval ranks as well as aggregate performance. This pattern is consistent with learning discriminative embeddings from the reliable supervision provided by PSCM: MRA can exploit the scene and contextual cues that remain visible, while PSCF refines the ordering of plausible matches and hard negatives. Together, these mechanisms offer a plausible explanation for the sustained advantage when visual degradation weakens, but does not remove, useful matching evidence.

Table 18: Full visual degradation robustness on RSICD-Test. Text and ground-truth correspondences are unchanged; only the I2T image queries and T2I image gallery are replaced. This table reports severity-3 corruptions, which are used for the main visual robustness comparison.

Method	$\text{Recall}_{\text{I2T}}$			$\text{Recall}_{\text{T2I}}$			mR_{I2T}	mR_{T2I}	mR
	@1	@5	@10	@1	@5	@10			
Haze (s=3)									
AMFMN TGRS'22	4.76	17.11	24.15	4.19	15.52	23.51	15.34	14.41	14.88
GaLR TGRS'22	4.39	13.08	20.49	3.60	13.58	21.34	12.65	12.84	12.75
PIR ACMMM'23	6.95	20.68	30.74	5.53	20.48	34.42	19.46	20.14	19.80
PE TGRS'23	13.27	30.74	42.91	10.10	31.18	46.84	28.97	29.37	29.17
LAPS CVPR'24	6.22	19.85	29.00	5.76	20.38	33.38	18.36	19.84	19.10
FGVLA TGRS'25	14.36	33.94	46.11	11.24	33.34	50.94	31.47	31.84	31.66
EAPA ACMMM'25	13.54	31.75	44.37	10.27	32.19	47.28	29.89	29.91	29.90
CUP TNNLS'25	7.41	21.04	32.57	7.63	24.23	38.98	20.34	23.61	21.98
McRSA TMM'26	14.55	33.76	46.11	12.72	35.88	50.78	31.47	33.13	32.30
RRSITR CVPR'26	13.08	28.36	40.53	10.06	29.17	42.65	27.32	27.29	27.30
P^2Cal	**23.97**	**44.10**	**53.98**	**17.80**	**42.38**	**57.46**	**40.68**	**39.21**	**39.95**
Stripe Noise (s=3)									
AMFMN TGRS'22	1.10	3.02	5.22	1.57	5.49	10.01	3.11	5.69	4.40
GaLR TGRS'22	1.46	3.66	6.86	1.23	5.91	9.64	3.99	5.59	4.79
PIR ACMMM'23	1.83	7.04	11.99	3.75	14.58	25.03	6.95	14.45	10.70
PE TGRS'23	5.49	14.27	19.76	5.76	19.16	29.24	13.17	18.05	15.61
LAPS CVPR'24	1.65	4.12	7.32	1.87	7.90	13.52	4.36	7.76	6.06
FGVLA TGRS'25	7.14	18.21	26.90	6.66	23.11	35.79	17.42	21.85	19.64
EAPA ACMMM'25	4.48	12.81	19.30	6.08	18.76	27.69	12.20	17.51	14.86
CUP TNNLS'25	3.11	9.33	15.19	4.72	14.49	25.45	9.21	14.89	12.05
McRSA TMM'26	8.42	18.85	28.64	7.63	25.20	39.98	18.64	24.27	21.45
RRSITR CVPR'26	6.86	17.93	24.79	6.15	19.52	29.31	16.53	18.33	17.43
P^2Cal	**11.25**	**23.15**	**31.11**	**9.99**	**28.66**	**40.68**	**21.84**	**26.44**	**24.14**
Blur (s=3)									
AMFMN TGRS'22	1.46	4.03	7.59	1.68	5.67	9.77	4.36	5.71	5.04
GaLR TGRS'22	1.37	4.30	6.86	1.28	6.83	10.59	4.18	6.23	5.21
PIR ACMMM'23	2.29	6.68	11.62	3.11	13.05	22.03	6.86	12.73	9.79
PE TGRS'23	6.04	15.10	25.43	5.53	20.79	33.27	15.52	19.86	17.69
LAPS CVPR'24	4.85	15.28	23.60	4.78	16.65	27.78	14.58	16.40	15.49
FGVLA TGRS'25	7.04	20.49	**29.09**	7.25	23.59	36.76	18.87	22.53	20.70
EAPA ACMMM'25	6.59	16.56	24.98	6.20	20.62	31.09	16.04	19.30	17.67
CUP TNNLS'25	3.48	10.43	16.29	4.30	16.40	27.50	10.07	16.07	13.07
McRSA TMM'26	8.05	20.04	28.27	8.47	24.72	**38.59**	18.79	**23.93**	21.36
RRSITR CVPR'26	6.50	16.83	24.34	5.51	19.32	29.15	15.89	17.99	16.94
P^2Cal	**9.52**	**20.86**	28.82	**9.09**	**25.43**	36.27	**19.73**	23.60	**21.66**
Low Resolution (s=3)									
AMFMN TGRS'22	2.93	10.43	16.93	2.91	11.93	19.82	10.10	11.55	10.82
GaLR TGRS'22	4.03	10.34	15.74	2.96	10.39	17.13	10.04	10.16	10.10
PIR ACMMM'23	4.57	14.27	24.25	4.41	18.15	31.40	14.36	17.99	16.17
PE TGRS'23	9.42	25.34	35.22	7.90	26.11	41.23	23.33	25.08	24.20
LAPS CVPR'24	5.22	15.65	27.36	4.92	19.93	32.39	16.08	19.08	17.58
FGVLA TGRS'25	10.98	26.81	40.71	8.98	30.32	45.64	26.17	28.31	27.24
EAPA ACMMM'25	10.06	23.51	35.96	8.60	26.95	40.44	23.18	25.33	24.25
CUP TNNLS'25	6.22	17.38	28.36	5.95	22.63	36.18	17.32	21.59	19.45
McRSA TMM'26	11.34	26.72	38.88	9.92	30.43	45.16	25.65	28.50	27.07
RRSITR CVPR'26	11.44	24.98	35.04	8.91	26.26	38.46	23.82	24.54	24.18
P^2Cal	**14.91**	**31.20**	**41.08**	**12.10**	**32.02**	**46.28**	**29.06**	**30.13**	**29.59**
JPEG Compression (s=3)									
AMFMN TGRS'22	4.57	13.72	23.42	3.79	14.84	25.27	13.90	14.63	14.27
GaLR TGRS'22	5.03	14.36	21.50	4.34	14.88	23.59	13.63	14.27	13.95
PIR ACMMM'23	4.03	12.63	20.95	5.18	18.26	31.33	12.54	18.26	15.40
PE TGRS'23	11.80	29.46	40.99	10.45	30.83	47.63	27.42	29.64	28.53
LAPS CVPR'24	5.86	21.04	32.75	6.40	22.38	36.25	19.88	21.68	20.78
FGVLA TGRS'25	10.89	27.36	40.26	10.08	32.68	48.56	26.17	30.44	28.30
EAPA ACMMM'25	9.52	24.89	35.41	10.41	30.85	45.25	23.27	28.84	26.05
CUP TNNLS'25	8.33	22.69	34.49	6.84	24.83	39.14	21.84	23.60	22.72
McRSA TMM'26	13.27	32.02	43.18	12.46	34.05	49.46	29.49	31.99	30.74
RRSITR CVPR'26	12.99	29.37	41.45	10.34	30.05	44.17	27.94	28.19	28.07
P^2Cal	**18.48**	**36.87**	**48.49**	**15.65**	**38.04**	**53.34**	**34.61**	**35.68**	**35.14**

Performance generally declines as degradation becomes stronger, and the margin over competing methods narrows under severe stripe noise and JPEG compression. Strong blur and low resolution pose the clearest challenge: some baselines overtake P^2Cal in overall mR, although it still leads

I2T Recall@1 under strong blur. These conditions remove fine structures and object details needed to distinguish similar scenes. PSCM improves training correspondence reliability, but neither embedding adaptation nor posterior score calibration explicitly reconstructs missing visual information. The results therefore show strong robustness when sufficient semantic evidence survives, with greater sensitivity when the degradation removes that evidence.

Table 19: Full visual degradation robustness on RSICD-Test. Text and ground-truth correspondences are unchanged; only the I2T image queries and T2I image gallery are replaced. This table reports severity-5 corruptions.

Method	$\text{Recall}_{\text{I2T}}$			$\text{Recall}_{\text{T2I}}$			mR_{I2T}	mR_{T2I}	mR
	@1	@5	@10	@1	@5	@10			
Haze (s=5)									
AMFMN TGRS'22	3.48	8.87	15.83	3.46	11.00	17.38	9.39	10.61	10.00
GaLR TGRS'22	2.56	8.60	12.81	2.47	9.33	15.75	7.99	9.18	8.59
PIR ACMMM'23	5.95	17.20	27.17	4.78	19.58	33.10	16.77	19.15	17.96
PE TGRS'23	11.16	24.98	37.42	9.26	28.16	42.53	24.52	26.65	25.59
LAPS CVPR'24	3.66	10.89	19.21	3.77	14.82	24.26	11.25	14.28	12.77
FGVLA TGRS'25	12.81	30.38	41.08	10.47	32.10	47.43	28.09	30.00	29.05
EAPA ACMMM'25	11.53	28.82	39.43	9.35	29.33	43.53	26.59	27.40	26.99
CUP TNNLS'25	6.22	16.83	26.08	5.54	21.76	34.22	16.38	20.51	18.45
McRSA TMM'26	12.99	27.63	41.72	10.85	32.24	46.50	27.45	29.86	28.66
RRSITR CVPR'26	10.89	21.32	30.92	8.34	24.79	36.80	21.04	23.31	22.17
P^2Cal	**19.40**	**36.32**	**47.03**	**15.17**	**37.91**	**51.71**	**34.25**	**34.93**	**34.59**
Stripe Noise (s=5)									
AMFMN TGRS'22	0.27	0.73	1.19	0.70	2.38	3.75	0.73	2.28	1.50
GaLR TGRS'22	0.37	1.37	2.10	0.59	2.38	4.15	1.28	2.37	1.83
PIR ACMMM'23	1.83	5.03	8.78	2.89	12.31	22.10	5.21	12.43	8.82
PE TGRS'23	2.93	8.60	12.08	4.03	14.24	21.74	7.87	13.34	10.61
LAPS CVPR'24	0.55	1.83	3.84	0.77	3.42	6.02	2.07	3.40	2.73
FGVLA TGRS'25	3.48	8.05	13.17	4.67	15.57	23.17	8.23	14.47	11.35
EAPA ACMMM'25	1.92	5.22	8.97	3.17	10.34	16.65	5.37	10.05	7.71
CUP TNNLS'25	1.92	6.13	10.89	2.84	11.14	18.72	6.31	10.90	8.61
McRSA TMM'26	5.12	13.91	20.68	5.82	20.55	32.41	13.24	19.59	16.41
RRSITR CVPR'26	5.03	10.89	17.66	4.45	14.27	22.21	11.19	13.64	12.41
P^2Cal	**6.68**	**15.74**	**21.87**	**7.45**	**22.20**	**32.72**	**14.76**	**20.79**	**17.77**
Blur (s=5)									
AMFMN TGRS'22	0.46	2.56	4.39	0.62	2.78	5.11	2.47	2.84	2.66
GaLR TGRS'22	0.64	2.56	4.57	0.84	3.68	6.79	2.59	3.77	3.18
PIR ACMMM'23	1.37	4.48	8.14	2.40	9.84	17.88	4.66	10.04	7.35
PE TGRS'23	2.65	9.06	14.55	3.99	13.87	23.00	8.75	13.62	11.18
LAPS CVPR'24	3.57	10.34	16.29	2.74	12.39	20.86	10.07	12.00	11.04
FGVLA TGRS'25	4.94	**14.00**	**19.85**	**5.42**	17.29	26.55	**12.93**	16.42	**14.68**
EAPA ACMMM'25	4.21	10.34	16.47	4.61	13.92	20.93	10.34	13.15	11.75
CUP TNNLS'25	1.92	5.58	9.24	2.60	10.34	18.02	5.58	10.32	7.95
McRSA TMM'26	5.03	13.17	19.21	5.18	**17.58**	**27.87**	12.47	**16.88**	**14.68**
RRSITR CVPR'26	3.57	9.97	14.64	3.35	12.90	20.07	9.39	12.11	10.75
P^2Cal	**5.31**	13.17	18.21	5.27	16.98	25.09	12.23	15.78	14.00
Low Resolution (s=5)									
AMFMN TGRS'22	1.10	3.02	5.49	1.46	4.65	7.83	3.20	4.65	3.93
GaLR TGRS'22	1.28	3.66	6.31	1.37	5.53	8.80	3.75	5.23	4.49
PIR ACMMM'23	2.74	7.87	13.17	3.22	12.37	22.43	7.93	12.67	10.30
PE TGRS'23	4.39	12.44	20.31	4.48	17.16	27.70	12.38	16.45	14.41
LAPS CVPR'24	3.75	12.81	19.67	3.68	14.69	24.03	12.08	14.13	13.11
FGVLA TGRS'25	**6.77**	**16.29**	**25.16**	5.54	**20.57**	**31.78**	**16.07**	**19.30**	**17.69**
EAPA ACMMM'25	5.31	14.64	22.32	5.54	18.30	29.11	14.09	17.65	15.87
CUP TNNLS'25	2.10	7.78	13.17	3.93	14.69	24.37	7.68	14.33	11.00
McRSA TMM'26	5.58	14.82	22.87	**6.28**	19.89	31.29	14.42	19.15	16.79
RRSITR CVPR'26	5.22	14.27	20.68	4.61	16.16	25.89	13.39	15.55	14.47
P^2Cal	5.40	14.64	19.85	6.06	19.09	28.09	13.30	17.75	15.53

Table 19 (continued)

Method	$\text{Recall}_{\text{I2T}}$ @1	@5	@10	$\text{Recall}_{\text{T2I}}$ @1	@5	@10	mR_{I2T}	mR_{T2I}	mR
JPEG Compression (s=5)									
AMFMN TGRS'22	2.74	10.06	17.84	3.13	12.94	20.37	10.21	12.15	11.18
GaLR TGRS'22	4.03	11.34	16.83	3.20	11.55	20.09	10.73	11.61	11.17
PIR ACMMM'23	2.84	11.25	17.57	3.68	16.40	28.86	10.55	16.31	13.43
PE TGRS'23	10.34	24.61	<u>35.86</u>	<u>8.80</u>	<u>27.45</u>	**41.92**	23.60	**26.06**	<u>24.83</u>
LAPS CVPR'24	4.39	16.47	25.62	5.45	19.01	31.60	15.49	18.69	17.09
FGVLA TGRS'25	7.32	19.76	28.73	8.07	25.20	38.28	18.60	23.85	21.23
EAPA ACMMM'25	6.95	17.47	26.72	8.00	21.90	32.88	17.05	20.93	18.99
CUP TNNLS'25	5.95	17.11	26.72	5.36	19.85	32.13	16.59	19.11	17.85
McRSA TMM'26	10.16	24.70	**37.05**	8.67	26.55	40.84	23.97	25.35	24.66
RRSITR CVPR'26	<u>10.61</u>	**25.89**	35.59	8.20	26.42	38.81	<u>24.03</u>	24.48	24.26
P²Cal	**11.71**	<u>25.80</u>	35.59	**9.41**	**27.85**	<u>40.88</u>	**24.37**	<u>26.05</u>	**25.21**

A.8 RSICD Expanded-Gallery Retrieval with RS5M Distractors

To evaluate whether P²Cal can maintain accurate retrieval as the candidate database grows, we expand the RSICD test galleries with additional remote sensing image-text pairs from RS5M (Zhang et al., 2024). The queries and their original ground-truth correspondences remain fixed, so this experiment tests whether correct matches retain high ranks amid a larger pool of competing candidates.

A.8.1 Details of Experimental Settings

We use the original RSICD test split of 1,093 images and 5,465 texts and expand its galleries with 1K, 5K, or 10K RS5M image-text pairs. The corresponding gallery sizes are 2,093 images and 6,465 texts, 6,093 images and 10,465 texts, and 11,093 images and 15,465 texts. Each RS5M pair contributes one image and one text; larger galleries retain the distractors included at smaller scales. Algorithm 8 summarizes the construction, including the exclusion of duplicate images.

The original RSICD images query the expanded text gallery for I2T, and the original RSICD texts query the expanded image gallery for T2I. Queries and ground-truth associations remain unchanged: RS5M entries serve only as distractors. All methods use their RSICD checkpoints from the main experiments without additional training. We use the Recall and mean Recall metrics defined in Appendix A.4.1.

In the preceding experiments, the baseline methods were evaluated using their respective original codebases. To accommodate the enlarged galleries, we reimplemented the evaluation routines to support batched feature extraction and blockwise similarity computation. Depending on the method, these routines differ from the original implementations in text preprocessing, batch padding, and numerical precision, while the query sets, ground-truth correspondences, and definitions of Recall@K and mean Recall remain unchanged.

Algorithm 8 RSICD Expanded-Gallery Construction with RS5M Distractors

Require: Original RSICD test galleries and image-text associations; RS5M image-text pairs
1: Exclude RS5M pairs whose images exactly duplicate RSICD test images.
2: Retain the first 10K accepted image-text pairs in a fixed order.
3: **for** each distractor count $n \in \{1000, 5000, 10000\}$ **do**
4: Add the first n RS5M images and their texts to the corresponding RSICD galleries.
5: Keep the original RSICD queries and ground-truth associations unchanged.
6: **end for**

A.8.2 Results and Analysis

Table 20 shows that P²Cal consistently leads every I2T Recall metric and the I2T directional mean at all tested gallery sizes. Its overall mR ranks first for the smallest expanded gallery and second at the larger scales, showing that the framework remains competitive as more distractors are introduced. This sustained I2T advantage is consistent with the complementary roles of reliable supervision from PSCM, discriminative embedding adaptation through MRA, and candidate score refinement through

PSCF. Richer visual cues in image queries can support this discrimination, while the original paired texts provide multiple valid retrieval targets.

As the gallery grows, P^{2}Cal's I2T performance declines gradually while retaining its lead, whereas T2I performance decreases more noticeably. Its T2I Recall@1 remains second across gallery sizes, but the advantage is weaker at broader retrieval thresholds. Short text queries may not specify enough detail to distinguish their paired images from an increasing number of similar distractors, and the image gallery also grows more in relative terms. MRA and PSCF can improve discrimination and ordering within the candidate set, but cannot supply details absent from a query or recover a true match excluded during candidate selection. This provides a plausible explanation for the stronger I2T performance and the remaining T2I gap.

Table 20: Expanded-gallery retrieval comparison on RSICD with RS5M distractors. RS5M samples are used only as non-ground-truth distractors; RSICD test queries and ground-truth correspondences remain unchanged.

Method	$\text{Recall}_{\text{I2T}}$			$\text{Recall}_{\text{T2I}}$			mR_{I2T}	mR_{T2I}	mR
	@1	@5	@10	@1	@5	@10			
Original RSICD + 1K RS5M Distractors (2,093 images / 6,465 texts)									
AMFMN TGRS'22	5.95	17.29	27.63	3.88	14.55	23.24	16.96	13.89	15.43
GaLR TGRS'22	7.23	15.92	25.53	3.93	13.60	21.56	16.23	13.03	14.63
PIR ACMMM'23	8.97	24.52	34.86	5.89	20.38	33.12	22.78	19.80	21.29
PE TGRS'23	13.17	30.01	43.28	9.31	30.72	45.89	28.82	28.64	28.73
LAPS CVPR'24	9.52	23.06	35.77	6.15	21.74	34.82	22.78	20.90	21.84
FGVLA TGRS'25	15.28	32.66	47.48	9.84	30.19	45.93	31.81	28.65	30.23
EAPA ACMMM'25	14.18	30.74	43.46	9.61	26.79	40.55	29.46	25.65	27.55
CUP TNNLS'25	6.95	23.51	36.51	6.28	23.71	37.68	22.32	22.56	22.44
McRSA TMM'26	16.93	34.40	48.49	**13.34**	**35.74**	**50.89**	33.27	**33.32**	33.30
RRSITR CVPR'26	15.28	32.30	45.84	9.75	28.22	41.28	31.14	26.42	28.78
P^{2}Cal	**19.40**	**40.90**	**52.15**	12.37	32.22	46.68	**37.48**	30.42	**33.95**
Original RSICD + 5K RS5M Distractors (6,093 images / 10,465 texts)									
AMFMN TGRS'22	5.67	16.19	26.90	2.96	10.69	17.29	16.25	10.31	13.28
GaLR TGRS'22	7.50	15.65	24.34	3.18	10.58	16.93	15.83	10.23	13.03
PIR ACMMM'23	7.78	22.69	32.85	5.09	18.02	29.08	21.11	17.40	19.25
PE TGRS'23	12.99	30.65	43.09	8.34	26.77	41.37	28.91	25.49	27.20
LAPS CVPR'24	9.42	22.42	34.40	5.32	18.37	29.22	22.08	17.64	19.86
FGVLA TGRS'25	15.10	32.11	47.03	7.92	24.89	38.88	31.41	23.90	27.66
EAPA ACMMM'25	13.63	30.01	41.35	7.50	21.50	32.52	28.33	20.51	24.42
CUP TNNLS'25	6.22	21.50	32.48	5.45	19.60	30.54	20.07	18.53	19.30
McRSA TMM'26	16.93	34.13	47.76	**11.99**	**31.75**	**45.36**	32.94	**29.70**	**31.32**
RRSITR CVPR'26	15.00	32.02	44.65	8.20	23.55	33.98	30.56	21.91	26.23
P^{2}Cal	**19.12**	**40.07**	**50.78**	10.10	26.61	37.71	**36.66**	24.81	30.73
Original RSICD + 10K RS5M Distractors (11,093 images / 15,465 texts)									
AMFMN TGRS'22	5.49	15.92	25.34	2.63	8.97	14.18	15.58	8.59	12.09
GaLR TGRS'22	6.86	15.10	23.79	2.87	9.02	14.49	15.25	8.79	12.02
PIR ACMMM'23	7.32	21.13	31.02	4.72	16.51	26.37	19.82	15.87	17.84
PE TGRS'23	13.08	30.38	42.82	7.65	24.67	37.68	28.76	23.33	26.05
LAPS CVPR'24	9.88	20.95	33.94	4.89	16.72	26.22	21.59	15.94	18.77
FGVLA TGRS'25	15.00	32.11	46.20	7.10	21.96	33.89	31.10	20.98	26.04
EAPA ACMMM'25	13.36	28.82	39.89	6.39	18.70	27.70	27.36	17.60	22.48
CUP TNNLS'25	5.58	20.22	30.74	4.59	17.58	26.99	18.85	16.39	17.62
McRSA TMM'26	16.65	34.22	47.67	**11.14**	**29.50**	**42.36**	32.85	**27.67**	**30.26**
RRSITR CVPR'26	14.91	31.38	44.28	7.39	21.57	30.08	30.19	19.68	24.94
P^{2}Cal	**18.85**	**39.34**	**50.14**	8.45	23.18	33.03	**36.11**	21.55	28.83

The design also limits the added computation. Lightweight MRA adaptation and a compact PSCF network restrict the trainable components, while PSCM filtering reduces the number of pairs processed during subsequent training. When more noisy pairs are removed, a fixed training schedule therefore requires fewer optimization batches. At inference, the more expensive ITM and PSCF refinement is limited to the top-K_{cand} ITC candidates per query, although global scoring and candidate selection still scale with gallery size. These mechanisms support practical expansion of the retrieval gallery.