International Journal of Computer Vision (2026) 134:533
https://doi.org/10.1007/s11263-026-03064-y

Weather4Geo: Language-Guided Weather–Content Disentangling for All-Weather Drone Visual Geo-Localization

Hang Yu[1] · **Jiahao Wen**[1] · **Junjie Zhu**[2] · **Zhedong Zheng**[3]

Received: 11 December 2025 / Accepted: 18 September 2026
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2026

Abstract

Variable weather introduces appearance and semantic perturbations that significantly degrade drone visual geo-localization. Existing cross-view geo-localization methods exhibit two inherent limitations in modeling weather: (1) they typically reduce weather to a few discrete labels, lacking explicit representations of continuous and fine-grained patterns, and (2) even when textual descriptions are introduced, they are usually injected as auxiliary prompts into a unified embedding space, leaving weather and content entangled. To address these issues, we propose Weather4Geo, a multi-modal framework that learns weather-robust scene representations by fusing image embeddings with language-guided weather–content priors. First, we leverage off-the-shelf large vision–language models to generate multi-weather textual descriptions in a training-free manner, explicitly characterizing complex, continuous weather patterns and their intensity variations without manual annotations. Second, we construct weather–content prototypes as semantic anchors for cross-weather alignment and introduce a Weather-Aligner module that adaptively maps features under adverse conditions into a weather-invariant content subspace. We optimize the framework in two stages with cross-modal objectives that combine image–text matching, classification supervision, and discriminative constraints, encouraging embeddings of the same scene under different weather conditions to cluster while preserving geographic discriminability. Extensive experiments show that, under diverse weather conditions, our method achieves competitive recall rates compared to state-of-the-art drone geo-localization methods. Notably, Recall@1 is improved by 17.82% at night and by 21.16% under fog and snow conditions.

Keywords Cross-view geolocalization · Drone navigation · Content disentanglement · All-weather robustness

1 Introduction

Drone visual geo-localization takes the precise matching between drone-view image and its corresponding satellite views as a central objective, providing crucial spatial localization capability for tasks such as urban monitoring, environmental sensing, and disaster emergency response (Zheng et al., 2020; Wang et al., 2022; Lin et al., 2022; Zeng et al., 2023; Zhu et al., 2022; Wang et al., 2024a). However, in real-world deployments, variable weather conditions such as rain, fog, and snow introduce noise, occlusions, and reduced visibility, which severely distort image features (Feng et al., 2024b; Sindagi et al., 2020; Liu et al., 2022) and lead conventional localization methods to suffer severe performance degradation under extreme weather.

Recent advances in cross-modal retrieval have shown that introducing natural language as additional supervision can effectively alleviate the failure modes of purely visual features under severe occlusion, texture degradation, or

Communicated by Hong Liu.

✉ Junjie Zhu
csezhujunjie@ust.hk

Hang Yu
yuhang@shu.edu.cn

Jiahao Wen
wenjh@shu.edu.cn

Zhedong Zheng
zhedongzheng@um.edu.mo

[1] School of Computer Engineering and Science, Shanghai University, Shanghai, China

[2] Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong, China

[3] Faculty of Science and Technology and Institute of Collaborative Innovation, University of Macau, Macau, China

Published online: 25 November 2026

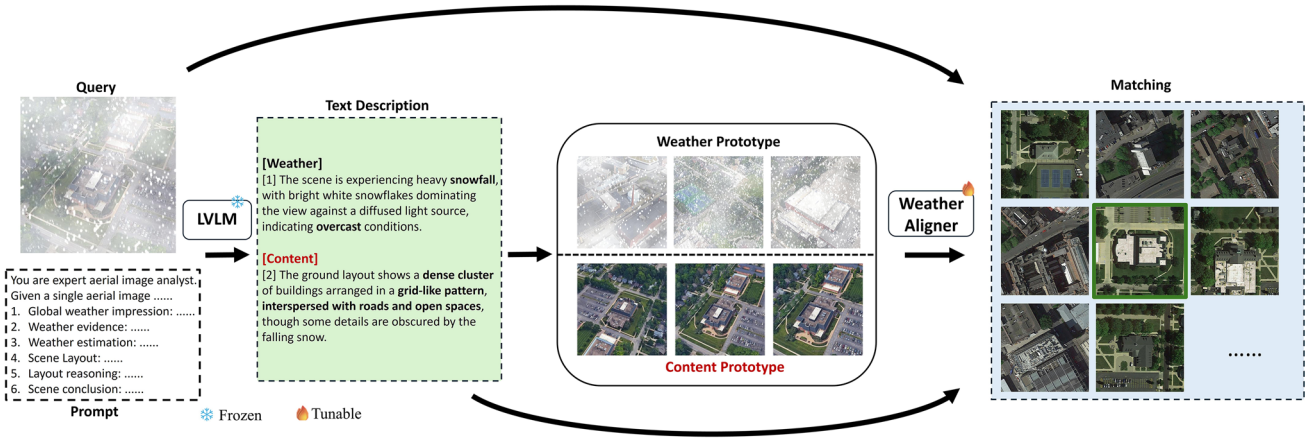

Fig. 1 Examples of the proposed CoT description and prototype-based matching. Our framework performs stepwise reasoning to generate structured weather and scene textual descriptions, which are encoded and aggregated into weather and content prototypes serving as semantic anchors for cross-weather alignment. Conditioned on the weather prototypes, the WeatherAligner adaptively rectifies drone visual features and aligns them to the content subspace, enabling robust matching to the corresponding satellite-view image

semantically complex scenes, thereby substantially enhancing the discriminative power and generalization ability of the model (Yang et al., 2023; Yan et al., 2024, 2023; Vendrow et al., 2024). Compared with modeling appearance solely in the image space, textual descriptions are better suited to encode semantics and fine-grained local details (Li et al., 2022; Chow et al., 2024), providing explicit characterizations of scene structure and its environmental conditions, and thus offering richer priors for learning robust representations.

Addressing the all-weather visual geo-localization task presents two primary challenges: **(1) Limited Weather Labeling.** Existing research on cross-view geo-localization and weather robustness still remains at a coarse level in terms of weather semantic modeling. Weather is typically reduced to a few discrete labels (*e.g.*, rain, snow), lacking explicit representations of continuous weather changes, composite weather conditions, and their intensity variations. Such label-level modeling of weather can hardly capture the rich and complex weather distributions encountered in real-world scenarios, and fails to provide stable and reliable semantic priors for subsequent all-weather representation learning. **(2) Weather-Content entanglement.** In existing methods, the limited weather descriptions are typically injected into a unified embedding space as additional prompt tokens and compressed together with image features into a single latent representation. This causes the two factors to remain entangled in the representation space: on the one hand, aggressively enforcing weather invariance tends to erase the fine-grained geometric and layout cues required for cross-view matching, on the other hand, if these details are preserved, weather-specific appearance patterns leak back into the feature space, making them difficult to isolate and analyze independently.

To address the first limitation, we design a training-free weather reasoning module built on top of an off-the-shelf large vision-language model (Bai et al., 2025) (see Fig. 1). For each geographic location, a single randomly sampled drone-view image is fed into the large vision-language models (LVLMs) together with a chain-of-thought (CoT) prompt (Wei et al., 2022), guiding the model to decompose the scene and infer detailed weather properties in a stepwise manner. In this way, we avoid manual description and expert-crafted rules, and can harvest large quantities of diverse, high-quality multimodal samples at very low annotation cost. The CoT-style reasoning further enforces semantic fidelity and formatting consistency across the generated captions, providing reliable, structured weather descriptions that serve as textual priors for subsequent all-weather representation learning. To address the second challenge, we propose Weather4Geo, a language-guided weather–content disentangling framework for all-weather drone visual geo-localization. Specifically, we construct weather–content prototypes from multi-weather textual descriptions, compressing scene structure and weather states into separate sets of content and weather prototypes that serve as semantic anchors for cross-weather alignment. A lightweight WeatherAligner residual module, modulated by the weather prototypes, is introduced during training to provide weather-conditioned residual alignment constraints, encouraging the retrieval encoder to learn a shared weather-invariant content representation. Our training strategy follows a two-stage paradigm that progressively refines the model. In the first stage, we jointly optimize image–text contrastive learning, image–text matching, and classification supervision to learn geographically discriminative vision–language aligned representations. Building on Stage I, the second stage introduces

image–prototype alignment and cross-weather consistency objectives to regularize the prototype memory and WeatherAligner, encouraging embeddings of the same scene under different weather conditions to cluster in a weather-invariant content subspace while preserving cross-view discriminability. At inference time, retrieval is performed using only the visual encoder, while WeatherAligner, textual prototypes, LVLM-based captioning, and weather-style embeddings are discarded. The model does not require online text generation, weather labels, or residual-offset computation during testing, making the inference pipeline compact and efficient.

The main contributions are as follows:

- **Multi-Weather Semantics Modeling**: We employ an LVLM with CoT prompts to progressively infer weather attributes from drone images without manual descriptions. The hierarchical reasoning mechanism fuses continuous weather priors with spatial object semantics to generate multi-weather textual descriptions.
- **Language-Guided Weather–Content Disentangling Framework**: We propose Weather4Geo, a two-stage vision–language framework that achieves weather–content disentanglement through: (1) Constructing weather–content prototypes from multi-weather textual descriptions as semantic anchors for cross-weather alignment, (2) Designing a lightweight WeatherAligner residual module modulated by the weather prototypes, (3) Adopting a two-stage training strategy to preserve cross-view matching performance.
- **Robust All-Weather Generalization**: Under 10 different weather conditions, our method achieves average Recall@1 scores of 89.04% and 84.31% on University-1652 and SUES-200, respectively. Even under a weather pattern unseen during training (snow+rain+fog), it still reaches 71.52% AP, further validating its robust cross-weather generalization capability.

This article is organized as follows. Section 2 introduces related work, Sect. 3 presents the proposed method, Sect. 4 reports experimental results on standard benchmarks, and Sect. 6 concludes the paper.

2 Related Work

2.1 Cross-View Geolocalization

Cross-view geo-localization aims to match images captured from different viewpoints with their corresponding geographic locations (Zheng et al., 2020; Berton et al., 2022; Kim et al., 2017; Chu et al., 2024). Early approaches relied on hand-crafted local features such as SIFT (Cruz-Mota et al., 2012) and SURF (Bay et al., 2006), as well as global descriptors like VLAD (Zhang et al., 2021) and Fisher Vector (Sánchez et al., 2013), often combined with RANSAC (Barath et al., 2022) for geometric verification, however, they remain brittle under large viewpoint and illumination changes (Castaldo et al., 2015; Lin et al., 2013).

With the rise of deep learning, cross-view geo-localization has been largely driven by deep metric learning models that optimize contrastive objectives at both global and local levels (Wang et al., 2024b; Workman et al., 2015; Lin et al., 2015; Vepa et al., 2024; Choi et al., 2023; Zheng et al., 2017; Wen et al., 2026). These methods typically learn end-to-end embeddings with triplet or InfoNCE-style losses, and couple global pooling with spatial attention mechanisms or explicit multi-region decomposition strategies (Yang et al., 2021; Rodrigues & Tani, 2022). Representative instances include the multi-part partitioning scheme (Wang et al., 2022), the keypoint-guided attention module (Lin et al., 2022), the content-aligned Transformer architecture (Dai et al., 2021), the self-attention positional encoding of Yang et al. (2021), the dual-path fusion network (Rodrigues & Tani, 2022), and the Bird's-Eye-View (BEV) representation (Ju et al., 2024), all of which have significantly improved cross-view feature alignment. Recent efforts have addressed real-world image degradations, including low-light conditions and motion blur, in order to improve the robustness of cross-view matching (Qu et al., 2024; Chen et al., 2024b). Tan et al. (2023) synthesize haze-degraded data to simulate fog conditions, Ma et al. (2022) introduce domain adaptation from drone to satellite image, GeoDTR (Zhang et al., 2023) employs heterogeneous augmentations and non-uniform cropping to enhance local discriminability, and TransGeo (Zhu et al., 2022) leverage a cross-modal Transformer to model long-range spatial correlations.

2.2 Multi-modality Alignment

In this work, we study weather-aware, text-guided representation learning for cross-view geo-localization, which can be cast as a fine-grained text-to-image retrieval problem with dynamic environmental attributes. Foundation vision–language models such as CLIP (Radford et al., 2021) and BLIP (Li et al., 2022) learn transferable cross-modal embeddings from large-scale image–text pairs, while XVLM (Zeng et al., 2021) further leverages cross-modal attention and learnable prompts for fine-grained alignment. At the architectural level, subsequent work explores dual-path encoders, adaptive gating, graph-based reasoning, and large-scale word–region alignment (Zheng et al., 2020; Wang et al., 2019; Li et al., 2019; Chen et al., 2020; Li et al., 2020; Yang et al., 2023; Feng et al., 2025; Gao et al., 2024) to refine vision–language matching in complex scenes. To adapt such models to specific domains and tasks, prompt-tuning techniques have been investigated in person re-identification and remote-

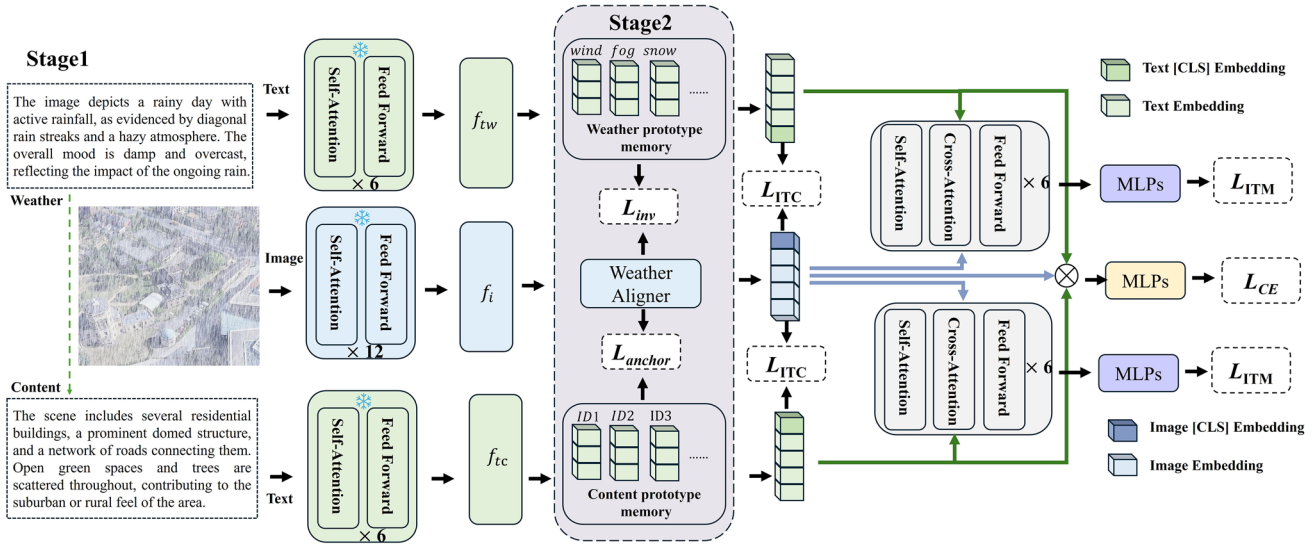

Fig. 2 The proposed language-guided weather-content disentangling framework. Stage 1 learns a weather-aware vision–language backbone on multi-weather image–text pairs using contrastive, matching, and classification objectives. Stage 2 introduces a text prototype memory together with a WeatherAligner module: weather condition embeddings derived from the multi-weather descriptions are used to conditionally modulate drone features under adverse weather and align them with the corresponding textual prototypes, thereby forming a canonical content subspace in which drone-to-satellite geo-localization remains stable across diverse weather conditions

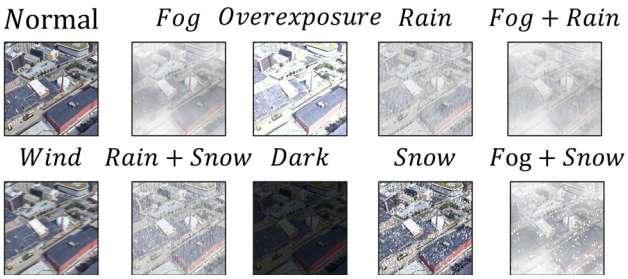

Fig. 3 Examples of different weather conditions in University-1652 benchmark

sensing retrieval (Zhu et al., 2024; Chen et al., 2024a; Yuan et al., 2023; Lu et al., 2023; Yu et al., 2025). However, these methods mainly focus on static semantics or relatively rigid spatial relations, rather than dynamic environmental factors such as weather. More recently, Large Vision–Language Models (LVLMs) such as GPT-4 and Qwen (Achiam et al., 2023; Bai et al., 2025) have been exploited via prompt engineering to produce richer textual descriptions, such as VisualGPT and MM-CoT (Chen et al., 2022; Zhang et al., 2023a) adopt chain-of-thought prompts (Wei et al., 2022) to perform stepwise reasoning in vision–language tasks.

2.3 Disentangled Representation Learning

Early works on disentangled representation learning aim to separate latent factors such as content versus style or identity versus pose, often through structured latent variables and encoder–decoder architectures (Higgins et al., 2017; Bhalla et al., 2024). In domain adaptation and generalization, approaches such as Domain Separation Networks explicitly split features into domain-invariant and domain-specific components to improve robustness across domains (Bousmalis et al., 2016; Tzeng et al., 2017; Hoyer et al., 2023). Related ideas have also been applied to factorizing object- and attribute-related features, as well as other conditional factors, to enable more controllable generation and recognition (Zhang et al., 2023b; Yang et al., 2024; Ahn et al., 2024; Dai et al., 2025). These studies suggest that explicitly structuring the feature space helps isolate nuisance variations while preserving task-relevant information (Feng et al., 2024a). However, most existing disentanglement approaches operate in single-view or category-level settings and rarely consider cross-view geo-localization under diverse and continuously varying weather conditions. In this work, we follow the factorized perspective but target the drone–satellite retrieval scenario, treating scene content and weather as two explicit factors. We learn separate subspaces for content and weather, and use LVLM-generated open-set weather descriptions to regularize this decomposition, yielding cross-view representations that remain content-consistent yet weather-aware.

3 Method

Our method follows a two-stage hierarchical alignment strategy. In the first stage, we perform global multimodal pre-alignment using LVLM-generated multi-weather captions, where each weather-augmented drone image is asso-

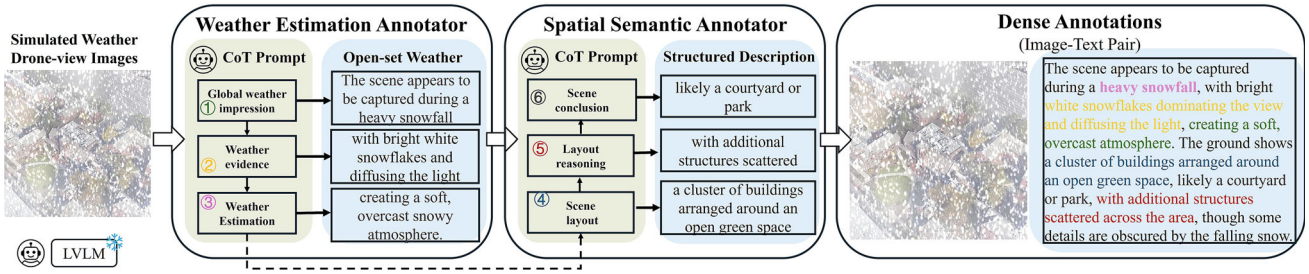

Fig. 4 Examples of Open-Weather Captioning. We generate drone-view images under a wide spectrum of weather conditions on the benchmarks, covering challenging cases such as night, wind, and fog+snow scenes. For each synthesized image, we first apply a step-wise CoT prompting scheme to obtain open weather descriptions that encompass global impressions, fine-grained local evidence, and explicit weather inference. Conditioned on the inferred weather prior, we further reason about the scene's macro layout, structural components, and topological organization in sequence, thereby constructing high-quality, structured image–text pairs (Best viewed when zooming in)

ciated with a corresponding textual description, standard image–text contrastive and region classification objectives are employed to learn a unified cross-view semantic representation. In the second stage, we introduce a textual prototype memory and a lightweight WeatherAligner module, which compress appearance discrepancies across different weather conditions through image–text prototype constraints, and cross-weather consistency losses, thereby enhancing the stability and fine-grained accuracy of cross-weather retrieval while preserving overall scene understanding and geographic discriminability. Fig. 2 illustrates an overview of the proposed framework.

3.1 Problem Statement and Open-Weather Captioning

Problem Definition. Following prior work (Wang et al., 2024b), we assume that satellite views, under fixed illumination and weather, provide spatially consistent observations, whereas drone views undergo significant appearance changes as lighting and weather vary. We consider geo-localization datasets composed of drone–satellite image pairs collected from multiple regions. Each sample is associated with a region identity indicating its geographic location and contains at least one satellite image and one corresponding drone-view image for that region. As shown in Fig. 3, we employ data augmentation (Jung et al., 2020) to generate multiple weather-stylized variants for each drone-view image, where each variant simulates a specific type of weather (*e.g.*, fog, rain, snow, or night). We formulate aerial geo-localization as a cross-view image retrieval task. During training, the model learns discriminative and robust feature embeddings under region-ID supervision. During inference, given a query image from either the drone or satellite view and a gallery of images from the opposite view, the model maps all images into a shared feature space and ranks the gallery according to feature similarity to accomplish cross-view matching.

Open-Weather Captioning. As shown in Fig. 4, we first randomly select one representative drone-view image from each geographic region and apply data augmentations (Jung et al., 2020) to these representative samples, thereby obtaining drone images under a variety of weather conditions. We then generate textual descriptions only for these augmented images, which preserves rich diversity in the weather distribution while avoiding redundant annotation of many nearly identical viewpoints within the same region and significantly reducing both computational and labeling cost. Specifically, we design a chain-of-thought prompting scheme on a large vision–language model (Bai et al., 2025) that organizes caption generation into a unified weather-first, structure-second reasoning process: the model first forms an internal global impression of the atmospheric state by jointly considering sky brightness, overall colour tone, contrast, and scene clarity, and then incrementally refines this judgment using local visual cues such as rain-streak reflections and fog diffusion, any response that fails to characterize visibility, lacks explicit meteorological evidence, or contains vague qualifiers such as possibly or uncertain is automatically discarded and regenerated until a semantically stable and self-consistent weather prior is obtained. Conditioned on this prior, the model produces a subsequent scene description that focuses on spatial structure: it summarises, from an overhead perspective, the coarse layout of building clusters, road axes, and open areas, then refines this view with the counts, shapes, and local arrangements of key entities, and explicitly encodes the relative positions and topological relations among regions and objects. The resulting weather-aware scene descriptions are both stable across different weather conditions and consistent with the underlying geometry, thereby providing fine-grained supervision for downstream vision–language alignment and cross-view retrieval.

3.2 Training Objective

Our training proceeds in two stages: Stage I learns cross-view vision–language alignment, and Stage II introduces weather-aware fine-grained constraints to refine weather consistency and content alignment.

In Stage I (Sect. 3.3), we only optimize global image-text alignment and region-level classification objectives, enabling the network to learn robust cross-modal and cross-view representations. The base loss is defined as:

$$\mathcal{L}_{\text{base}} = \mathcal{L}_{\text{ITC}} + \mathcal{L}_{\text{ITM}} + \mathcal{L}_{\text{CE}}. \tag{1}$$

In Stage II (Sect. 3.4), we introduce the WeatherAligner modules and the textual prototype memory as auxiliary training-time alignment components. In addition to the base loss, we introduce a weather-consistency regularizer to refine the representations learned in Stage I. We denote the weather term as:

$$\mathcal{L}_{\text{weather}} = \mathcal{L}_{\text{anchor}} + \lambda_{\text{inv}}\mathcal{L}_{\text{inv}}, \tag{2}$$

where $\mathcal{L}_{\text{anchor}}$ aligns adverse-weather samples with canonical image and text anchors, and $\mathcal{L}_{\text{inv}}$ encourages compact representations of the same region across different weather conditions in the discriminative space. The total loss in Stage II is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{base}} + \mathcal{L}_{\text{weather}}. \tag{3}$$

3.3 Global Cross-View Alignment

In Stage I, we first learn a global representation for cross-view retrieval. We train the shared visual and textual encoders with standard image–text contrastive and matching objectives, together with region-level classification supervision.

Image–Text Contrastive Learning. Given a mini-batch of B paired samples $\{(x_i, t_i)\}_{i=1}^{B}$, where x_i and t_i denote the ℓ_2-normalized global embeddings of a drone image and its caption, respectively, we first construct a similarity matrix

$$S_{ij} = \frac{x_i^\top t_j}{\tau}, \quad S \in \mathbb{R}^{B\times B}, \tag{4}$$

with a learnable temperature parameter τ. We then interpret each row of S as the logits for retrieving text given an image, and each column as the logits for retrieving image given text, and obtain the corresponding softmax probabilities $p_{I\to T}^{(i)}$ and $p_{T\to I}^{(i)}$ on the diagonal. The image–text contrastive loss is defined as:

$$\mathcal{L}_{\text{ITC}} = -\frac{1}{2B}\sum_{i=1}^{B}\Big[\log p_{I\to T}^{(i)} + \log p_{T\to I}^{(i)}\Big], \tag{5}$$

which encourages matched image–text pairs to be close in the joint embedding space, while pushing apart mismatched combinations across the batch.

Image–Text Matching Learning. To further enhance fine-grained discrimination, we complement the global contrastive objective with a hard-negative image–text matching loss. Given the similarity matrix S, for each positive pair (x_i, t_i) we identify the most confusing negative caption and image by:

$$t_i^- = \arg\max_{j\neq i} S_{ij}, \quad x_i^- = \arg\max_{j\neq i} S_{ji}, \tag{6}$$

which yields B positive pairs (x_i, t_i) and $2B$ hard negative pairs (x_i, t_i^-) and (x_i^-, t_i). Each pair is fed into the cross-modal encoder to obtain a [CLS] representation h_k, which is passed through a binary classifier to produce a matching probability $p_k = \sigma(f_{\text{match}}(h_k))$. With $y_k = 1$ for positive pairs and $y_k = 0$ for negatives, the matching loss is defined as:

$$\mathcal{L}_{\text{ITM}} = -\frac{1}{3B}\sum_{k=1}^{3B}\Big[y_k \log p_k + (1-y_k)\log(1-p_k)\Big], \tag{7}$$

which drives the model to distinguish subtle semantic differences between closely related image–text pairs.

Region Classification Learning. To preserve discriminative power with respect to geographic regions, we attach a region classification head on top of the drone-view embedding. We make $f_b \in \mathbb{R}^D$ denote the embedding of the b-th sample, and define the classifier logits over C regions as

$$o_b = W_{\text{cls}} f_b + b_{\text{cls}} \in \mathbb{R}^C \tag{8}$$

where $W_{\text{cls}} \in \mathbb{R}^{C\times D}$ and $b_{\text{cls}} \in \mathbb{R}^C$ are learnable parameters. Given the ground-truth region label $y_b \in \{1, \ldots, C\}$ for each sample, the region-level supervision is implemented with a standard softmax cross-entropy loss:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{B}\sum_{b=1}^{B}\log\frac{\exp(o_{b,y_b})}{\sum_{c=1}^{C}\exp(o_{b,c})}, \tag{9}$$

which encourages samples from the same region to cluster together while separating different regions in the embedding space.

3.4 Weather-Consistent Alignment

In Stage II, we refine the representations produced by the backbone to be consistent across different weather conditions while preserving the cross-view alignment learned in Stage I. To this end, we introduce the WeatherAligner and textual

prototype memory to impose additional weather-consistent alignment constraints during training.

WeatherAligner. We introduce a weather-aware residual adapter on the image branch to apply lightweight corrections to the global representations of samples captured under adverse weather conditions. The global embedding of a drone image is $f_{\mathrm{i}} \in \mathbb{R}^D$. For each synthetic weather image, we associate a learnable weather-style embedding, which is used as a conditioning input to the adapter.

Specifically, WeatherAligner encodes the weather-style embedding f_{tw} into a low-dimensional vector $e_{\mathrm{w}} \in \mathbb{R}^{d_w}$ through a projection head, then concatenates this vector with the image feature, and feeds the result into a two-layer MLP to predict a residual offset:

$$\Delta f = \mathrm{MLP}\big([f_{\mathrm{i}}; e_{\mathrm{w}}]\big) \in \mathbb{R}^D, \tag{10}$$

from which the weather-adjusted representation is obtained as:

$$f_{\mathrm{align}} = f_{\mathrm{i}} + \Delta f. \tag{11}$$

During training, WeatherAligner is applied only to the drone branch and is activated exclusively for samples with synthetic weather perturbations, while backbone features under normal conditions are kept as canonical references to constrain cross-weather alignment.

Textual prototype memory. To provide stable textual anchors for both scene content and weather conditions, we maintain textual prototype memories during training. In practice, we instantiate a region-wise content prototype memory and optionally a weather-type prototype memory, while these memories are not accessed at inference time. We set C to denote the number of geographic regions in the training set. For the c-th region, we maintain a D-dimensional textual prototype vector $p_c \in \mathbb{R}^D$ together with an update counter N_c. Whenever the model processes a normal-weather sample from region c in the forward pass of Stage II, we obtain a normalized textual feature $f_{\mathrm{tc}} \in \mathbb{R}^D$ from the text encoder and projection head, and update the prototype in a gradient-free exponential moving average (EMA) (Izmailov et al., 2018) manner:

$$\tilde{p}_c = m\, p_c + (1-m)\, f_{\mathrm{tc}}, \quad p_c = \frac{\tilde{p}_c}{\|\tilde{p}_c\|_2}, \tag{12}$$

where $m \in [0, 1)$ is a momentum coefficient. If region c has not yet acquired a prototype ($N_c = 0$), we simply set $p_c = f_{\mathrm{tc}}$ and increment N_c. The Weather prototypes are updated in the same EMA fashion from weather-text features f_{tw} and absorbed into the conditioning vector e_{w} through the projection head of WeatherAligner.

After sufficient training iterations, each p_c can be regarded as the aggregated textual semantics of region c under normal-weather descriptions. In Stage II, when we use WeatherAligner to correct adverse-weather drone features from the same region, we additionally pull these features towards the corresponding canonical prototype p_c, yielding a joint image–text anchoring constraint that explicitly reduces cross-weather dispersion for each region in the embedding space. Importantly, the prototype memory is only involved in defining training losses, while cross-view retrieval at test time relies solely on the visual encoder, without accessing the prototype bank, WeatherAligner, so its generalization ability is not restricted for unseen regions or novel weather combinations.

Weather-consistent regularization. We impose a unified weather-consistency regularization term to constrain the representations of the same region under different weather conditions. This regularizer consists of two components: the first aligns adverse-weather features with canonical image and text anchors, and the second tightens cross-weather discriminability in the classification space.

Canonical anchor alignment. Let $f_{\mathrm{norm}}^{(c)} \in \mathbb{R}^D$ denote the drone feature of region c under normal weather, and let $p_c \in \mathbb{R}^D$ its canonical textual prototype. For an adverse-weather condition $w > 0$, the WeatherAligner produces a corrected drone feature $f_{\mathrm{align}}^{(c,w)}$. We jointly exploit the image and text anchors to constrain adverse-weather samples:

$$\begin{aligned}\mathcal{L}_{\mathrm{anchor}} = \mathbb{E}_{c,\, w>0}\big[\| f_{\mathrm{align}}^{(c,w)} - f_{\mathrm{norm}}^{(c)} \|_2^2 \\ - \lambda_{\mathrm{txt}}\, \cos\big(\hat{f}_{\mathrm{align}}^{(c,w)},\, p_c\big)\big].\end{aligned} \tag{13}$$

where $\hat{f}_{\mathrm{align}}^{(c,w)}$ denotes the ℓ_2-normalized aligned feature, p_c is the normalized textual prototype maintained by Eq. 12, and λ_{txt} controls the strength of the prototype-based constraint. This term encourages all adverse-weather views to cluster around the joint anchor defined by the normal-weather image and its canonical textual prototype, thereby explicitly compensating for the shifts introduced by weather variations.

Cross-weather invariance in the classification space. In parallel, we further reduce cross-weather discrepancies in the classification space. Let z_i be the logits (or the input to the linear classifier) of the drone branch, with region label y_i and weather indicator w_i. We first apply ℓ_2 normalisation to z_i, and then enforce pairwise consistency over all sample pairs from the same region but different weather conditions:

$$\mathcal{L}_{\mathrm{inv}} = \mathbb{E}_{i,j:\, y_i = y_j,\, w_i \neq w_j}\big[1 - \cos(\hat{z}_i, \hat{z}_j)\big], \tag{14}$$

which encourages the discriminative representations of the same geographic region to remain consistent across different weather patterns.

4 Experiment

In our experiments, we construct Drone → Satellite and Satellite → Drone retrieval tasks on multi-weather benchmarks such as University-1652 and SUES-200, and systematically evaluate retrieval performance and cross-weather generalization under different settings.

4.1 Setup

Implementation Details. We adopt XVLM (Zeng et al., 2021), pre-trained on about 4 M image–text pairs, as the unified cross-modal backbone, with BERT (Devlin et al., 2019) as the text encoder and a Swin Transformer (Liu et al., 2021) as the image encoder. Training follows a two-stage scheme under a shared optimization setup: we use stochastic gradient descent (SGD) (Robbins & Monro, 1951) with momentum 0.9 and weight decay 5×10^{-4} for 210 epochs, where epochs 1–180 correspond to the first stage and epochs 181–210 to the second stage. The learning rate is decayed by a factor of 0.1 at epoch 120 and by 0.01 at epoch 180. Drone-view and satellite-view images are resized to 384×384 and partitioned into 32×32 patches before being fed into the visual encoder. Satellite images are augmented by random cropping and horizontal flipping, drone images are first processed by the imgaug library (Jung et al., 2020) to synthesize fog, rain, snow, and low-illumination conditions, and then undergo the same crop-and-flip operations, yielding diverse multi-weather views. In Stage I, we jointly optimize the $\mathcal{L}_{\mathrm{ITC}}$, the $\mathcal{L}_{\mathrm{ITM}}$, and the $\mathcal{L}_{\mathrm{CE}}$ on multi-weather image–text pairs, with all three losses weighted by 1. In Stage II, to avoid training instability caused by abruptly activating strong regularizers, we adopt a linear ramp-up schedule over 20 epochs for the weather-consistency terms, gradually increasing their effective weights from 0 to the target values. After ramp-up, we set the prototype weight in Eq. (13) to $\lambda_{\mathrm{txt}} = 0.03$ and the cross-weather invariance coefficient in Eq. (2) to $\lambda_{\mathrm{inv}} = 0.01$. At test time, we extract L_2-normalized global embeddings from drone and satellite branches, compute Euclidean distances for similarity, and rank candidates accordingly. All experiments are implemented in PyTorch (Paszke et al., 2019) and run on a single NVIDIA RTX A6000 GPU.

Dataset. *University-1652* (Zheng et al., 2020) is a large-scale dataset for cross-view geo-localization and multi-view retrieval. It contains images from 1,652 building locations in universities, captured from satellite, drone, and ground views. Each building has 54 drone images, 1 satellite image, and additional street-view images. The training set includes 701 buildings, while the test set has 951 buildings, without overlap. *SUES-200* (Zhu et al., 2023) is a multi-view dataset comprising drone and satellite images from 200 locations in Shanghai. It covers diverse scenes such as urban areas, parks, lakes, and public buildings. Drone images are taken from various altitudes (150 m, 200 m, 250 m, 300 m) in real-world scenes.

Evaluation protocol. We report Recall@K (R@K) and Average Precision (AP) in our experiments. R@K measures the proportion of correctly localized images in the top-K list. AP is the area under the precision-recall curve.

4.2 Comparison with Competitive Methods

The experimental results on the University-1652 dataset are shown in Table 1. We compare the proposed method with several competitive cross-view geo-localization methods. On the Drone → Satellite retrieval task, Ours attains a mean Recall@1 (R@1) of 79.18% (+14.03%) and a mean Average Precision (AP) of 82.02% (+12.86%), establishing a clear margin over existing methods tailored for multi-weather geo-localization. For the Satellite → Drone direction, our approach further achieves a mean R@1 of 89.04% (+4.36%) and a mean AP of 78.35% (+12.60%), confirming consistent gains across both retrieval directions. These improvements indicate that the learnt representations remain stable when the drone-view appearance is severely degraded, particularly under challenging conditions such as Dark, Over-exposure, Rain+Snow and Fog+Snow. In addition to quantitative comparisons, Fig. 5 visualizes representative retrieval cases. While existing methods behave reliably in clear-weather scenarios, their retrieval quality deteriorates significantly in the presence of adverse weather.

On the real-world SUES-200 benchmark, we aggregate drone-view image captured at four altitudes (150 m, 200 m, 250 m, and 300 m) to emulate the range of operational heights encountered in practical deployments, thereby enabling a more comprehensive assessment of our approach. As shown in Table 2, the performance on the SUES-200 dataset is consistent with the observations on University-1652, and our approach achieves clear advantages over the state-of-the-art multi-weather geo-localization method MuSe-Net across all evaluation metrics. Specifically, in the drone → satellite retrieval task, our method improves the mean R@1 from 52.62% to 68.86% and the mean AP from 53.93% to 73.39%, in the satellite→drone task, the mean R@1 is further increased from 69.43% to 84.31% and the mean AP is further increased from 51.14% to 70.14%, confirming the stable gains of the proposed approach across different datasets and cross-view settings.

4.3 Ablation Studies

In this section, we analyze several aspects of our approach, including the importance of text guidance and the effectiveness of each component.

Impact of Text-Guided Alignment. We conduct an ablation study within the Stage I to evaluate the impact of multi-

Table 1 Retrieval performance on University-1652 in terms of R@1(%) and AP(%) for Drone → Satellite and Satellite → Drone tasks

Method	Normal		Fog		Rain		Snow		Fog+Rain		Fog+Snow	
	R@1	AP	R@1	AP	R@1	AP	R@1	AP	R@1	AP	R@1	AP
Drone → Satellite												
Zheng et al. (Zheng et al., 2020) [backbone]	67.83	71.74	60.97	65.23	60.29	64.61	55.58	60.09	54.75	59.40	44.85	49.78
ResNet-101 (He et al., 2016) [backbone]	70.07	73.04	63.87	68.22	63.34	67.59	59.75	64.15	57.45	62.12	48.31	53.28
DenseNet121 (Huang et al., 2017) [backbone]	69.48	73.26	64.25	68.47	63.47	67.64	59.29	63.70	59.68	64.13	50.41	55.20
Swin-T (Liu et al., 2021) [backbone]	69.27	73.18	66.46	70.52	65.44	69.60	61.79	66.23	63.96	68.21	56.44	61.07
IBN-Net (Pan et al., 2018) [backbone]	72.35	75.85	66.68	70.64	67.95	71.73	62.77	66.85	62.64	66.84	51.09	55.79
LPN (Wang et al., 2022) [TCSVT'21]	74.33	77.60	69.31	72.95	67.96	71.72	64.90	68.85	64.51	68.52	54.16	58.73
Sample4Geo* (Deuser et al., 2023) [ICCV'23]	**92.70**	**93.85**	**88.70**	**90.55**	62.44	66.17	52.76	57.24	52.70	56.77	19.79	23.16
Safe-Net* (Lin et al., 2024) [TIP'24]	86.98	88.85	82.12	86.10	67.13	68.90	60.50	63.01	54.80	58.73	32.12	39.77
CCR* (Du et al., 2024) [TCSVT'24]	92.54	93.78	85.57	87.13	67.46	68.82	55.16	59.14	63.11	66.97	27.74	31.48
MuSe-Net* (Wang et al., 2024b)[PR'24]	74.48	77.83	69.47	73.24	70.55	74.14	65.72	69.70	65.59	69.64	54.69	59.24
Ours	84.54	86.74	83.57	85.89	**81.89**	**84.43**	**79.09**	**81.96**	**80.51**	**83.25**	**75.85**	**79.04**
Satellite → Drone												
Zheng et al. (Zheng et al., 2020) [backbone]	83.45	67.94	79.60	61.12	77.60	59.73	73.18	55.07	75.89	54.45	70.76	43.26
ResNet-101 (He et al., 2016) [backbone]	85.73	71.79	82.45	66.46	81.46	65.68	79.74	61.72	79.74	60.59	74.75	50.31
DenseNet121 (Huang et al., 2017) [backbone]	83.74	70.34	82.31	66.32	81.17	65.23	78.60	60.33	79.46	61.66	74.61	51.14
Swin-T (Liu et al., 2021) [backbone]	80.74	68.94	81.03	67.46	81.17	66.39	78.46	61.33	79.17	64.65	74.89	56.57
IBN-Net (Pan et al., 2018) [backbone]	86.31	73.54	84.59	67.61	84.74	69.03	80.88	64.44	83.31	63.71	77.89	52.14
LPN (Wang et al., 2022) [TCSVT'21]	87.02	75.19	86.16	71.34	83.88	69.49	82.88	65.39	84.59	66.28	79.60	55.19
Sample4Geo* (Deuser et al., 2023) [ICCV'23]	**95.29**	91.42	**93.87**	**87.46**	73.04	50.27	76.18	47.58	71.18	44.53	52.21	16.21
Safe-Net* (Lin et al., 2024) [TIP'24]	91.22	86.06	90.04	85.43	71.12	68.56	73.26	45.62	68.23	41.78	49.32	34.72
CCR* (Du et al., 2024) [TCSVT'24]	95.15	**91.80**	90.93	80.62	81.83	73.89	69.92	65.41	76.92	70.53	50.89	31.64
MuSe-Net* (Wang et al., 2024b) [PR'24]	88.02	75.10	87.87	69.85	87.73	71.12	83.74	66.52	85.02	67.78	80.88	54.26
Ours	90.01	84.16	89.73	82.57	**89.73**	**81.99**	**89.59**	**78.47**	**90.01**	**80.03**	**88.16**	**74.11**

Method	Rain+Snow		Dark		Over-exp		Wind		Mean	
	R@1	AP	R@1	AP	R@1	AP	R@1	AP	R@1	AP
Drone → Satellite										
Zheng et al. (Zheng et al., 2020) [backbone]	57.61	62.03	39.70	44.65	51.85	56.75	58.28	62.83	55.17	59.71
ResNet-101 (He et al., 2016) [backbone]	60.25	64.68	46.12	51.02	56.34	61.23	62.13	66.63	58.76	63.29
DenseNet121 (Huang et al., 2017) [backbone]	60.21	64.57	48.57	53.41	54.04	58.88	60.74	65.14	59.01	63.44
Swin-T (Liu et al., 2021) [backbone]	62.68	67.02	50.27	55.18	55.46	60.29	63.81	68.17	61.56	65.95
IBN-Net (Pan et al., 2018) [backbone]	64.07	68.13	50.72	55.53	57.97	62.52	66.73	70.68	62.30	66.46

Table 1 continued

Method	Rain+Snow		Dark		Over-exp		Wind		Mean	
	R@1	AP	R@1	AP	R@1	AP	R@1	AP	R@1	AP
LPN (Wang et al., 2022) [TCSVT'21]	65.38	69.29	53.68	58.10	60.90	65.27	66.46	70.35	64.16	68.14
Sample4Geo* (Deuser et al., 2023) [ICCV'23]	38.19	42.33	46.34	49.91	75.77	78.90	**81.54**	**87.34**	61.10	64.62
Safe-Net* (Lin et al., 2024) [TIP'24]	25.83	26.40	41.10	44.13	69.87	71.15	74.32	76.58	59.48	62.36
CCR* (Du et al., 2024) [TCSVT'24]	23.06	46.85	51.10	54.19	75.90	79.16	81.31	87.22	62.30	67.47
MuSe-Net* (Wang et al., 2024b)[PR'24]	66.64	70.55	53.85	58.49	61.05	65.51	69.45	73.22	65.15	69.16
Ours	**79.96**	**82.75**	**71.67**	**75.15**	**76.18**	**79.41**	78.58	81.56	**79.18***(+14.03)*	**82.02** (+12.86)
Satellite → Drone										
Zheng et al. (Zheng et al., 2020) [backbone]	74.75	56.44	69.47	39.25	72.18	51.91	76.46	57.59	75.33	54.68
ResNet-101 (He et al., 2016) [backbone]	80.17	62.61	75.32	45.37	79.60	58.21	82.31	64.67	80.13	60.74
DenseNet121 (Huang et al., 2017) [backbone]	78.46	61.68	74.47	47.88	74.32	55.26	78.32	61.63	78.55	60.15
Swin-T (Liu et al., 2021) [backbone]	78.89	63.49	75.61	48.43	76.60	56.57	78.74	64.45	78.53	61.83
IBN-Net (Pan et al., 2018) [backbone]	83.02	65.74	78.46	50.77	79.46	58.64	84.02	67.94	82.27	63.36
LPN (Wang et al., 2022) [TCSVT'21]	84.17	66.26	82.88	52.05	81.03	62.24	84.14	67.35	83.64	65.08
Sample4Geo* (Deuser et al., 2023) [ICCV'23]	64.48	32.38	77.03	45.89	**91.58**	**77.04**	**93.30**	**81.42**	78.82	57.42
Safe-Net* (Lin et al., 2024) [TIP'24]	61.07	29.86	73.15	43.08	88.54	74.65	90.02	78.21	75.69	58.80
CCR* (Du et al., 2024) [TCSVT'24]	61.11	32.21	64.80	46.28	86.01	71.23	92.67	76.55	77.02	64.02
MuSe-Net* (Wang et al., 2024b) [PR'24]	84.88	67.75	80.74	53.01	81.60	62.09	86.31	70.03	84.68	65.75
Ours	**88.73**	**79.29**	**87.30**	**69.29**	87.16	75.40	90.01	78.19	**89.04** *(+4.36)*	**78.35** *(+12.60)*

Drone-view images are rendered under 10 different weather conditions, while satellite-view images remain unchanged. The best results are highlighted in bold, and * indicates models using officially released pretrained weights

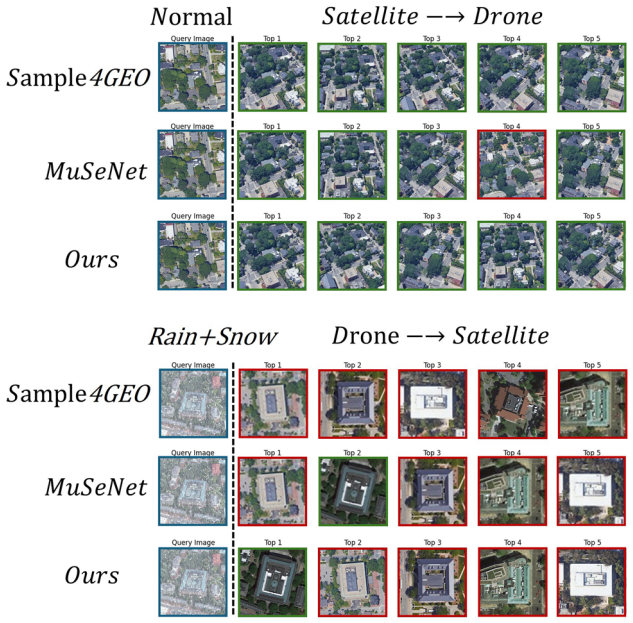

Fig. 5 Qualitative comparison across weather conditions. Existing methods behave robustly in clear-weather scenes but exhibit a pronounced drop in localization accuracy under adverse conditions. In contrast, our approach sustains more reliable retrieval performance, particularly when drone images are heavily degraded by weather. Green boxes denote correct matches, whereas red boxes mark incorrect ones

weather textual guidance on cross-view geo-localization. Specifically, we construct a vision-only baseline model by removing all text-related components, retaining only the visual encoder and the region classification head, and training this model from scratch under the same optimization settings. We then re-enable the text branch and introduce CoT weather descriptions with different numbers of reasoning steps as guidance. In addition to whether CoT is used and how many reasoning steps are adopted, all other architectural designs and training configurations are kept identical to the vision-only baseline, and all Stage II modules are disabled in this experiment.

Table 3 reports the mean Recall@1 and mean AP on the University-1652 dataset for both Drone → Satellite and Satellite → Drone retrieval tasks under different numbers of CoT steps. As the number of CoT steps increases, both retrieval tasks exhibit a consistent performance improvement. Relative to the vision-only baseline, the setting with 6 CoT steps improves the Drone → Satellite task by 2.61% in mean Recall@1 and 1.48% in mean AP, and boosts the Satellite → Drone task by 2.55% in mean Recall@1 and 3.13% in mean AP. Furthermore, all models leveraging text-based weather guidance consistently surpass the baseline on both mean Recall@1 and mean AP, which confirms that even weak weather-aware text cues can noticeably strengthen robustness and retrieval accuracy across diverse weather conditions.

Impact of Weather-Consistent Alignment. We first compare the full two-stage method with a Stage I-only baseline to quantify the gain brought by the proposed weather-consistent alignment stage. For a fair comparison, the Stage I baseline is trained for the same number of epochs as the full model: we remove the entire Stage II and continue to optimize only the Stage I objectives ($\mathcal{L}_{\text{ITC}}$, $\mathcal{L}_{\text{ITM}}$, and $\mathcal{L}_{\text{CE}}$), while keeping all other settings (data, optimizer, and learning rate schedule) unchanged.

As shown in Fig. 6, the full two-stage model achieves higher mean Recall@1 and mean AP than the Stage I-only baseline on University-1652 for both Drone → Satellite and Satellite → Drone retrieval. On University-1652, compared with the Stage I only baseline, adding the weather-consistent alignment stage leads to improvements of 2.04% and 1.82% in mean Recall@1 and mean AP for the Drone → Satellite task, and 1.32% and 1.96% in mean Recall@1 and mean AP for the Satellite → Drone task. Similar gains in Recall@1 and AP are observed on the SUES-200 dataset. Moreover, the improvements are particularly pronounced under adverse or compound weather conditions. For example, under the Dark condition in the Drone → Satellite task, both R@1 and AP increase by around 4%. This indicates that the weather-consistent alignment stage effectively suppresses weather-induced appearance shifts while preserving cross-view discriminability.

Impact of WeatherAligner. To evaluate the effect of the weather-adaptive residual branch and its anchor-based alignment constraint, we conduct an ablation study focusing on the role of $\mathcal{L}_{\text{anchor}}$. Specifically, we remove the anchor alignment loss in Stage II and retain only the cross-weather invariance regularization term $\mathcal{L}_{\text{inv}}$, while keeping all other network components outside WeatherAligner and all training configurations unchanged.

As shown in Table 4, removing $\mathcal{L}_{\text{anchor}}$ leads to a decrease of about 0.79% and 0.69% in mean Recall@1 and AP on the D2S task, respectively. A similar degradation is observed for S2D, where Recall@1 and AP decrease by 0.29% and 0.67%, respectively. The gap becomes more pronounced under challenging weather conditions such as Dark and Fog+Snow, indicating that relying solely on cross-weather invariance is insufficient to fully compensate for weather-induced appearance changes. These observations show that the anchor-based alignment loss provides crucial region-level geometric and semantic constraints within WeatherAligner, and plays an essential role in stabilizing and strengthening cross-weather retrieval performance.

Impact of Cross-Weather Invariance. To evaluate the effect of the cross-weather invariance constraint, we perform an ablation on the loss term $\mathcal{L}_{\text{inv}}$. In this setting, the cross-weather invariance loss is disabled (i.e., no similarity constraint is imposed between samples from the same region under different weather), while WeatherAligner and

Table 2 Retrieval performance on SUES-200 in terms of R@1(%) and AP(%) for Drone → satellite and satellite → Drone tasks

Method	Normal		Fog		Rain		Snow		Fog+Rain		Fog+Snow	
	R@1	AP	R@1	AP	R@1	AP	R@1	AP	R@1	AP	R@1	AP
Drone → Satellite												
Zheng etal (Zheng et al., 2020) [backbone]	57.70	58.30	48.63	49.61	53.41	54.72	41.78	43.47	37.17	37.44	44.22	46.18
IBN-Net (Pan et al., 2018) [backbone]	65.34	66.78	56.03	56.57	55.73	58.55	47.80	49.53	43.45	44.98	50.04	51.00
Sample4Geo† (Deuser et al., 2023) [ICCV'23]	74.93	78.76	72.58	76.44	34.60	41.56	28.95	35.02	35.10	41.47	12.95	17.90
Safe-Net* (Lin et al., 2024) [TIP'24]	**76.31**	76.35	**73.53**	74.44	54.15	55.05	48.94	50.10	45.12	47.92	40.05	40.18
CCR† (Du et al., 2024) [TCSVT'24]	73.22	74.53	70.95	73.14	60.14	64.95	50.31	53.12	45.87	49.14	45.80	47.87
MuSe-Net* (Wang et al., 2024b) [PR'24]	66.07	67.02	58.49	59.65	58.94	60.14	54.85	56.12	44.31	45.82	49.81	51.26
Ours	75.80	**79.81**	73.52	**77.74**	**73.92**	**77.84**	**65.10**	**69.87**	**68.16**	**72.82**	**58.22**	**63.53**
Satellite → Drone												
Zheng etal (Zheng et al., 2020) [backbone]	70.20	57.98	63.77	46.90	68.72	50.85	61.72	39.70	62.10	32.75	71.70	40.39
IBN-Net (Pan et al., 2018) [backbone]	73.68	62.91	67.41	55.75	72.30	56.44	64.07	47.69	66.98	39.54	71.10	47.32
Sample4Geo† (Deuser et al., 2023) [ICCV'23]	87.50	79.57	83.75	71.14	42.50	25.24	40.00	21.59	38.75	23.22	30.00	10.58
Safe-Net* (Lin et al., 2024) [TIP'24]	88.31	80.35	81.33	68.60	40.21	41.04	36.43	37.50	33.12	35.45	24.78	27.65
CCR† (Du et al., 2024) [TCSVT'24]	**90.59**	**80.45**	82.99	70.62	43.39	45.90	39.81	40.88	42.63	39.46	29.32	30.65
MuSe-Net* (Wang et al., 2024b) [PR'24]	76.56	66.02	72.19	57.87	72.19	58.11	68.38	51.22	66.56	42.25	69.06	46.80
Ours	86.56	80.04	**88.44**	**76.18**	**84.69**	**74.37**	**82.19**	**67.51**	**83.44**	**69.07**	**80.31**	**57.31**

Method	Rain+Snow		Dark		Over-exp		Wind		Mean	
	R@1	AP	R@1	AP	R@1	AP	R@1	AP	R@1	AP
Drone → Satellite										
Zheng etal (Zheng et al., 2020) [backbone]	40.60	40.63	23.81	25.45	49.79	50.64	47.42	48.31	44.43	45.48
IBN-Net (Pan et al., 2018) [backbone]	45.51	45.92	29.61	30.93	56.01	56.96	57.36	58.10	50.69	51.93
Sample4Geo† (Deuser et al., 2023) [ICCV'23]	20.05	25.95	34.18	38.99	38.40	43.68	67.80	72.41	41.95	47.22
Safe-Net* (Lin et al., 2024) [TIP'24]	25.95	26.12	29.74	31.48	54.86	58.68	58.10	58.95	50.68	51.93
CCR† (Du et al., 2024) [TCSVT'24]	31.25	32.94	31.03	34.36	59.97	61.07	52.02	53.33	52.06	54.45
MuSe-Net* (Wang et al., 2024b) [PR'24]	49.42	50.87	29.34	31.03	55.02	56.36	59.97	61.05	52.62	53.93
Ours	**70.84**	**74.99**	**63.84**	**68.73**	**64.52**	**69.83**	**74.73**	**78.75**	**68.86** *(+16.24)*	**73.39** *(+19.46)*
Satellite → Drone										
Zheng etal (Zheng et al., 2020) [backbone]	59.72	37.55	45.49	25.28	52.11	43.40	56.62	45.31	61.21	42.01
IBN-Net (Pan et al., 2018) [backbone]	68.46	45.95	54.72	31.53	65.64	53.77	73.48	57.03	67.79	49.79
Sample4Geo† (Deuser et al., 2023) [ICCV'23]	26.25	16.44	56.25	29.75	58.75	30.38	83.75	69.66	54.75	37.76
Safe-Net* (Lin et al., 2024) [TIP'24]	41.12	32.31	53.88	27.01	54.19	57.82	79.36	57.09	53.27	46.48
CCR† (Du et al., 2024) [TCSVT'24]	25.89	26.94	26.01	30.40	58.01	59.13	83.09	61.05	52.17	48.55
MuSe-Net* (Wang et al., 2024b) [PR'24]	69.38	47.79	53.75	27.94	70.00	52.67	76.25	60.74	69.43	51.14
Ours	**85.31**	**70.40**	**82.50**	**62.02**	**82.81**	**69.85**	**86.88**	**74.62**	**84.31** *(+14.88)*	**70.14** *(+19.00)*

Drone-view images are rendered under 10 different weather conditions, whereas satellite-view images remain unchanged. Best results are highlighted in bold. *Indicates models using officially released pretrained weights, and †Indicates models pretrained on University-1652

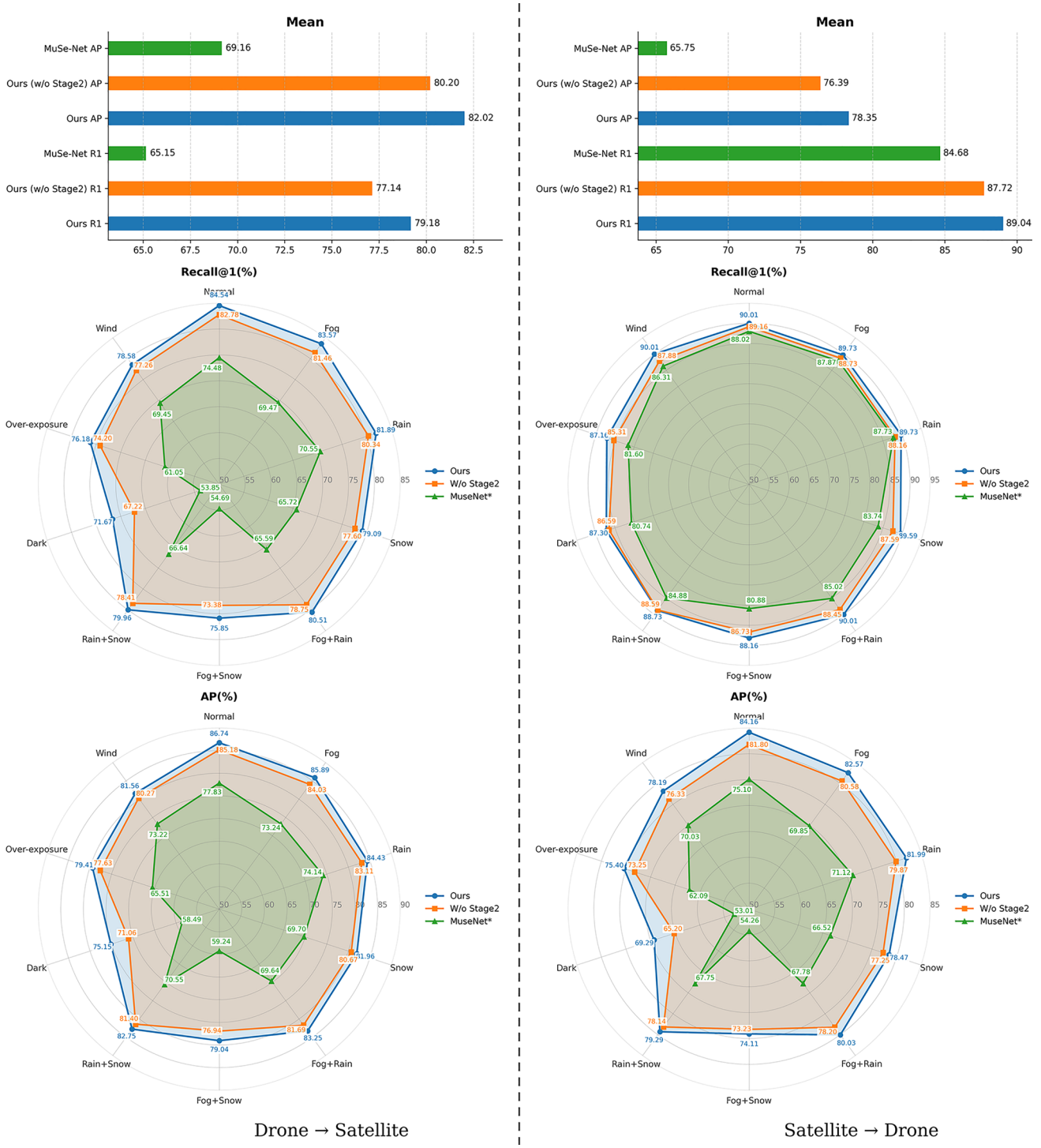

Fig. 6 The effectiveness of the Weather-Consistent Alignment on University-1652. The radar chart details results under different weather conditions and the bar chart shows the mean result. (Best viewed when zooming in.)

Table 3 Different CoT settings on University-1652

Step	D2S		S2D	
	Mean R@1(%)	Mean AP(%)	Mean R@1(%)	Mean AP(%)
N/A	74.53	78.72	85.17	73.26
0	75.10	79.15	85.45	73.65
2	75.35	79.10	85.85	74.40
4	76.05	79.80	86.60	75.20
6	**77.14**	**80.20**	**87.72**	**76.39**

D2S denotes the Drone → Satellite and S2D denotes the Satellite → Drone

its anchor loss $\mathcal{L}_{\text{anchor}}$ are retained and all other architectural and training settings remain unchanged.

As reported in Table 4, suppressing the cross-weather invariance term consistently degrades performance on both directions, mean Recall@1 and AP on the D2S task decline, and the S2D task exhibits the same downward trend with drops of up to 0.58% in Recall@1 and 0.49% in AP. More importantly, we observe substantially larger variance across the 10 weather conditions, indicating that, without explicit cross-weather invariance constraints, features of the same region undergo non-negligible drift under different weather conditions, making the retrieval results much more sensitive to changes in the underlying weather distribution.

Impact of textual prototype memory. In Fig. 7a, the w/o prototypes setting disables the textual prototype memory: textual prototypes are no longer updated and the prototype-based alignment term is removed from the training objective, while WeatherAligner, the image-anchor loss, and the cross-weather invariance regularizer $\mathcal{L}_{\text{inv}}$ remain unchanged. Compared with the full configuration, turning off the textual prototype memory consistently degrades performance: mean Recall@1 and AP on the D2S task drop by 1.11% and 0.87%, and the S2D task exhibits similar declines of 1.03% and 1.52%, respectively. These results indicate that image anchors alone do not provide a sufficiently stable semantic reference, textual prototypes supply complementary semantic anchors that reinforce cross-weather alignment and contribute noticeably to overall robustness.

4.4 More Discussion

Robustness to Captioning LVLM. To probe whether the performance gains are tied to a particular LVLM, we replaced the original caption generator with GLM-4.1V-9B-Thinking (Team et al., 2025), while keeping the prompts and structured caption format unchanged. We regenerate the weather-aware captions, retrain the network from scratch, and repeat the evaluation. As shown in Fig. 7b, the resulting R@1 and mAP are essentially unchanged compared to the original setup, indicating that the improvements mainly originate from the normalized caption schema and training objectives rather than the specific LVLM.

Table 5 Cross-dataset retrieval performance

Method	D2S Mean		S2D Mean	
	R@1	AP	R@1	AP
Sample4Geo*	41.95	47.22	54.75	37.76
Safe-Net*	21.78	26.56	42.25	22.83
CCR*	52.06	54.45	52.17	48.55
MuSe-Net*	52.99	57.91	58.50	44.40
Ours	**63.56**	**67.43**	**75.30**	**62.77**

Models are trained on University-1652 and evaluated on SUES-200. * denotes the use of official pretrained on University-1652 weights. Best results are highlighted in bold

Cross-City Transfer Results. We further assess the generalization ability of our approach in a cross-dataset setting. The model is trained exclusively on the University-1652 dataset (Zheng et al., 2020) and then directly evaluated on SUES-200 (Zhu et al., 2023), without any additional retraining or fine-tuning. Apart from the dataset itself, all architectural and training configurations are kept identical to those described in the main experiments. For all competing methods, we use the official pre-trained weights released by the authors. The SUES-200 test split is constructed by randomly sampling and combining images from four different

Table 4 Ablation study of different loss functions

$\mathcal{L}_{\text{base}}$	$\mathcal{L}_{\text{anchor}}$	$\mathcal{L}_{\text{inv}}$	D2S		S2D	
			Mean R@1	Mean AP	Mean R@1	Mean AP
✓			77.14	80.20	87.72	76.39
✓	✓		78.60	81.53	88.55	78.07
✓		✓	78.39	81.33	88.75	77.68
✓	✓	✓	**79.18**	**82.02**	**89.04**	**78.35**

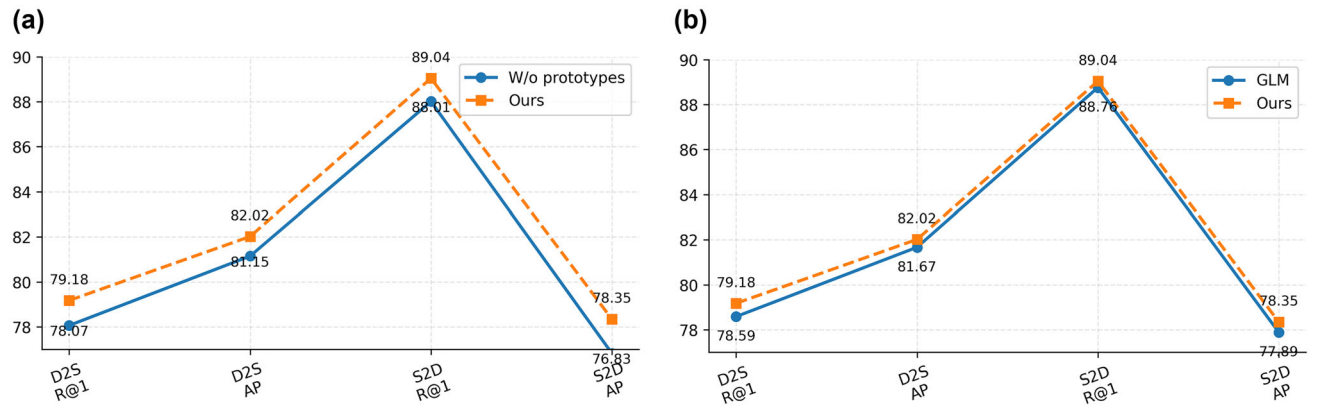

Fig. 7 Impact of textual prototype memory and caption generator.

altitude levels, thereby providing a challenging evaluation protocol with substantial viewpoint variation.

As summarized in Table 5, the proposed method achieves the best performance on both Drone → Satellite and Satellite → Drone retrieval. In the Drone → Satellite task, our method achieves a mean R@1 of 63.56%, representing an improvement of 10.57% over the previous best method MuSe-Net, whose mean R@1 is 52.99%. The mean AP is also increased by 9.52%. In the Satellite → Drone task, our method reaches a mean R@1 of 75.30%, which is 16.80% higher than the 58.50% achieved by MuSe-Net, and the mean AP likewise gains an improvement of 18.37%. It is also evident that prior methods such as Sample4Geo (Deuser et al., 2023) and Safe-Net (Lin et al., 2024) maintain satisfactory R@1 under clear or mildly perturbed conditions (*e.g.*, Normal and Wind), but their robustness degrades noticeably in more challenging scenarios, including Fog+Rain, Rain+Snow, and Dark, where performance drops sharply and becomes highly weather-dependent. In contrast, our method preserves consistently strong performance across diverse weather combinations and complex scenes, indicating markedly improved stability and generalization across both datasets and weather conditions.

We further analyze top-1 failure cases under the University-1652 → SUES-200 transfer setting. As shown in Fig. 8, failures mainly occur in scenes with repetitive water/bridge structures, large open areas, campus-style building layouts, dense public-building regions, or inconsistent visible regions between drone and satellite views. These scenarios often lack distinctive local landmarks and contain hard negatives with highly similar coarse spatial layouts. The incorrect top-1 results are still close to the query in global layout, indicating that Weather4Geo can transfer scene-level semantic and structural cues across cities. However, fine-grained place identity discrimination remains challenging when repetitive geographic structures, weather degradation, viewpoint changes, and cross-city distribution shifts coexist.

Physics-Guided Weather Stress Test. We further supplement a physics-guided drone-side stress test to examine whether Weather4Geo overly depends on the original imgaug-based weather simulation. For fog/haze degradation, we use a Koschmieder-style scattering model with pseudo-depth approximation to simulate depth-dependent visibility attenuation. For rain degradation, we combine near-camera rain streaks with a mild atmospheric veil to simulate close-range rain streaks, local occlusions, and slight visibility reduction that commonly occur in low-altitude drone imaging. All degradations are applied only to the drone query side, while the satellite gallery remains unchanged, and no additional training or fine-tuning is performed. As shown in Table 6, Weather4Geo still achieves stable and competitive performance under these physics-guided degradations, confirming its robustness beyond fixed augmentation templates.

Generalization to Unseen Weather. To evaluate the model's generalization ability under truly out-of-distribution scenarios, we further conduct experiments on three extreme weather conditions that never appear during training, namely *Winter Night*, *Freezing Rain*, and *Blizzard*. The results are reported in Table 7.

On the Drone → Satellite task, existing methods exhibit a clear performance drop under these unseen extreme conditions, revealing limited cross-weather robustness. Taking MuSe-Net as a representative baseline, its macro-averaged Recall@1 and AP over the three weather types are 38.99% and 43.96%, respectively, whereas our method maintains these metrics at 52.03% and 56.66%, corresponding to absolute gains of 13.04% and 12.70%. The advantage is particularly pronounced in the most severely degraded scenario, Winter Night, where our method attains 57.73% Recall@1 and 62.17% AP, whereas MuSe-Net only achieves 39.45% Recall@1 and 44.36% AP, underscoring its much more stable cross-weather retrieval performance under extreme illumination changes. A similar trend is observed on the Satellite → Drone task. The proposed method achieves the

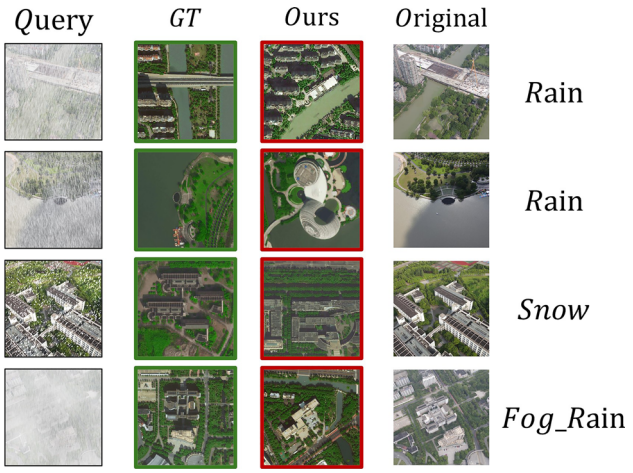

Fig. 8 Failure case analysis under University-1652 → SUES-200 cross-dataset transfer. The green box indicates the ground-truth satellite image, and the red box indicates the incorrect top-1 retrieval result

Table 6 Direct evaluation under physics-guided drone-side weather degradations

Weather	R@1	R@5	R@10	AP
Fog/Haze	84.63	94.05	95.87	86.80
Ours (Fog)	83.57	93.59	95.50	85.89
Rain	83.36	93.63	95.75	85.72
Ours (Rain)	81.89	93.05	95.17	84.43

best Recall@1 and AP across all three extreme weather conditions, with mean Recall@1 and AP reaching 80.27% and 55.17%, respectively. This performance is markedly superior to existing approaches, among which MuSe-Net attains a mean Recall@1 of 73.04% and CCR achieves a mean AP of 44.50%, yielding absolute improvements of 7.23% and 10.67%, respectively.

Overall, these results validate that the proposed semantic-guided framework can effectively cope with complex and unseen weather distributions: even when other methods suffer severe performance degradation under extreme weather, our approach consistently maintains high retrieval accuracy.

Variance-based Analysis of Weather Robustness. Figure 9 depicts the weather-wise variance of Recall@1 (R@1) and Average Precision (AP) for both Drone → Satellite and Satellite → Drone tasks. Existing approaches exhibit highly unstable behavior across weather types: for the Drone → Satellite direction, the R@1 variance of Sample4Geo and CCR reaches 498.45 and 494.02, respectively, and even the most stable competitor, MuSe-Net, still shows a variance of 41.00. In contrast, our method reduces the R@1 variance to 13.46 for Drone → Satellite and 1.15 for Satellite → Drone. A similar pattern is observed for AP, where our approach attains the smallest variance among all methods (10.82 and 17.73, respectively), while previous methods typically exceed 300 and in some cases even 500. These substantial gaps indicate that our method not only improves average accuracy but also delivers markedly more consistent retrieval performance across diverse weather conditions, effectively alleviating the instability and sharp weather-induced performance drops observed in prior work.

Table 7 Performance under extreme weather conditions on the University-1652 dataset

Mean R@1 (%)				
Method	Task	Winter night	Freezing rain	Blizzard
Drone → Satellite				
CCR*	D2S	36.05	47.29	25.73
Safe-Net*	D2S	3.25	2.64	11.46
Sample4Geo*	D2S	25.40	37.83	36.00
MuSe-Net*	D2S	39.45	44.09	33.44
Ours	D2S	**57.73**	**60.53**	**37.83**
Satellite → Drone				
CCR*	S2D	69.97	66.16	60.88
Safe-Net*	S2D	56.92	41.08	57.92
Sample4Geo*	S2D	58.35	65.62	66.90
MuSe-Net*	S2D	73.32	76.03	69.76
Ours	S2D	**82.31**	**83.17**	**75.32**
		mean AP (%)		
		Drone → Satellite		
CCR*	D2S	38.79	51.70	30.07
Safe-Net*	D2S	4.87	4.30	15.43
Sample4Geo*	D2S	28.60	41.95	40.42
MuSe-Net*	D2S	44.36	49.01	38.51
Ours	D2S	**62.17**	**64.89**	**42.92**
Method	Task	Winter Night	Freezing Rain	Blizzard
		Satellite → Drone		
CCR*	S2D	44.08	55.20	34.23
Safe-Net*	S2D	8.28	8.15	22.05
Sample4Geo*	S2D	21.33	32.76	32.47
MuSe-Net*	S2D	40.19	43.51	35.21
Ours	S2D	**63.20**	**58.39**	**43.91**

* denotes the use of official pretrained weights

Caption quality and robustness to noisy textual priors. We first evaluate the reliability of LVLM-generated captions as soft semantic priors. We randomly sample 1,000 image-caption pairs from University-1652 and SUES-200, with balanced coverage over 10 weather conditions. The evaluation is conducted using GPT−5.5-based automatic assessment together with manual spot-checking to reduce the bias of a single evaluation source. We assess the captions from four aspects: weather-label consistency, weather-evidence plausibility, scene-content grounding, and coarse-layout consistency. The weather-related dimensions achieve 88.9% Acceptable Acc., while scene-content grounding and coarse-

Table 8 Robustness to noisy textual priors on University-1652

Setting	Normal		Fog		Rain		Snow		Fog+Rain		Fog+Snow	
	R@1	AP	R@1	AP	R@1	AP	R@1	AP	R@1	AP	R@1	AP
Ours	90.01	84.16	89.73	82.57	**89.73**	**81.99**	**89.59**	**78.47**	**90.01**	80.03	**88.16**	**74.11**
10%	**91.01**	**84.22**	**90.58**	**82.59**	89.02	**81.99**	87.59	78.40	89.02	**80.29**	87.59	74.09
20%	89.44	83.47	89.02	82.19	87.73	80.77	88.59	77.88	87.16	79.15	87.02	73.79
30%	89.87	83.78	90.16	82.27	88.30	81.33	88.16	77.62	89.02	79.57	87.59	73.57

Setting	Rain+Snow		Dark		Over-exp		Wind		Mean	
	R@1	AP	R@1	AP	R@1	AP	R@1	AP	R@1	AP
Ours	88.73	**79.29**	87.30	69.29	**87.16**	**75.40**	**90.01**	78.19	**89.04**	78.35
10%	**89.02**	79.24	87.45	**69.86**	85.59	74.71	87.73	**78.73**	88.46	**78.41**
20%	88.02	78.57	**89.30**	69.29	85.31	74.22	88.02	77.73	87.96	77.71
30%	88.73	78.68	87.02	69.12	86.59	74.67	89.02	78.28	88.45	77.89

We randomly replace a certain proportion of generated captions with captions from other geographic regions and weather conditions, and report R@1(%) and AP(%) under different weather conditions for the S2D task

layout consistency both reach 100.0% Acceptable Acc. These results indicate that the generated captions are not perfect for fine-grained weather evidence, but remain reliable for describing major geographic content and spatial layout.

We further assess the sensitivity of Weather4Geo to caption noise by randomly replacing a proportion of generated captions with captions from other geographic regions and weather conditions. To isolate the influence of noisy textual priors on the alignment stage, all settings are initialized from the same Stage-I checkpoint, and the Stage-II alignment process is retrained with different caption-noise ratios while keeping the architecture and other training settings unchanged. As shown in Table 8, increasing the caption noise ratio from 10% to 30% results in only minor performance fluctuations, with the mean R@1/AP changing from 88.46/78.41 to 88.45/77.89. This indicates that Weather4Geo is not brittle to moderate caption noise, since noisy captions can be mitigated by region-level semantic aggregation, textual prototype memory, and cross-weather consistency.

Comparison with general disentanglement strategies. We compare Weather4Geo with two general feature disentanglement strategies, GRL and DSN. As shown in Table 9, GRL and DSN provide limited gains, suggesting that simply suppressing or separating weather-related variations is insufficient for all-weather cross-view geo-localization. These methods lack task-specific geographic semantic anchors, such as normal-weather textual prototypes and cross-weather geographic consistency constraints. In contrast, Weather4Geo achieves the best mean R@1/AP in both D2S and S2D directions by performing language-guided weather-content alignment. This indicates that reducing weather-induced feature shifts while preserving discriminative geographic cues is more effective than generic feature-level disentanglement.

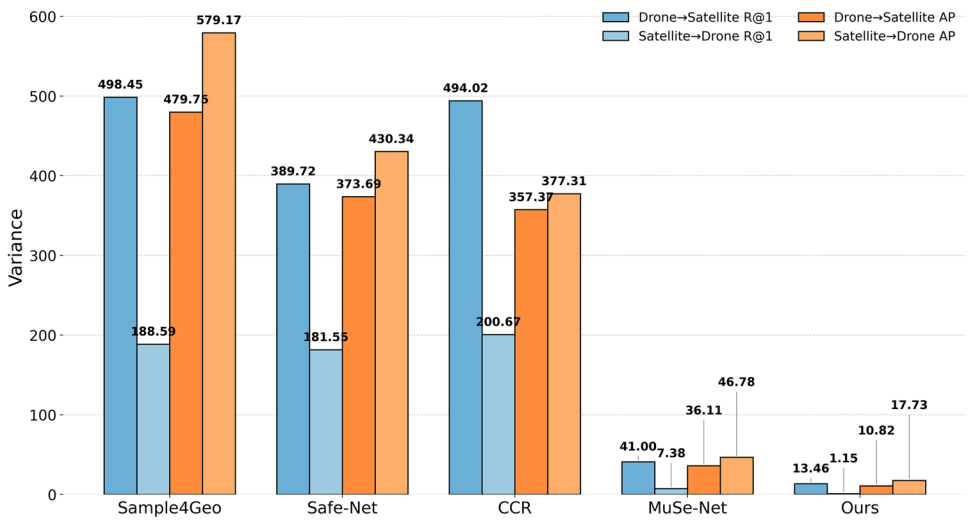

Fig. 9 Variance comparison of R@1 and AP across different models for cross-view localization on the University-1652 dataset

Table 9 Ablation comparison of different training variants on University-1652

Method	Normal		Fog		Rain		Snow		Fog+Rain		Fog+Snow	
	R@1	AP	R@1	AP	R@1	AP	R@1	AP	R@1	AP	R@1	AP
Drone → Satellite												
Stage-I text-only	82.78	85.18	81.46	84.03	80.34	83.11	77.60	80.67	78.75	81.69	73.38	76.94
GRL	83.69	86.03	82.71	85.14	80.82	83.49	77.49	80.61	79.22	82.11	74.27	77.67
DSN	83.77	86.06	82.72	85.12	80.95	83.60	77.65	80.74	79.33	82.19	74.63	77.97
Ours	**84.54**	**86.74**	**83.57**	**85.89**	**81.89**	**84.43**	**79.09**	**81.96**	**80.51**	**83.25**	**75.85**	**79.04**
Satellite → Drone												
Stage-I text-only	89.16	81.80	88.73	80.58	88.16	79.87	87.59	77.25	88.45	78.20	86.73	73.23
GRL	89.30	83.38	90.01	82.13	89.02	81.50	88.30	78.14	**90.30**	79.64	88.30	74.12
DSN	89.44	83.60	**90.16**	82.25	**89.73**	81.78	88.59	78.37	**90.30**	79.92	**88.59**	**74.14**
Ours	**90.01**	**84.16**	89.73	**82.57**	**89.73**	**81.99**	**89.59**	**78.47**	90.01	**80.03**	88.16	74.11

Method	Rain+Snow		Dark		Over-exp		Wind		Mean	
	R@1	AP	R@1	AP	R@1	AP	R@1	AP	R@1	AP
Drone → Satellite										
Stage-I text-only	78.41	81.40	67.22	71.06	74.20	77.63	77.26	80.27	77.14	80.20
GRL	78.51	81.52	70.54	74.11	75.34	78.63	77.34	80.45	77.99	80.98
DSN	78.65	81.62	70.20	73.69	75.51	78.76	77.15	80.30	78.06	81.01
Ours	**79.96**	**82.75**	**71.67**	**75.15**	**76.18**	**79.41**	**78.58**	**81.56**	**79.18**	**82.02**
Satellite → Drone										
Stage-I text-only	88.59	78.14	86.59	65.20	85.31	73.25	87.88	76.33	87.72	76.39
GRL	87.87	79.03	87.02	69.78	86.02	74.81	89.16	77.97	88.53	78.05
DSN	87.87	79.19	86.45	**69.99**	86.16	75.14	89.16	78.17	88.64	78.26
Ours	**88.73**	**79.29**	**87.30**	69.29	**87.16**	**75.40**	**90.01**	**78.19**	**89.04**	**78.35**

We report R@1(%) and AP(%) under 10 weather conditions for Drone → Satellite and Satellite → Drone tasks

Computational cost of LVLM-based captioning. The LVLM is only used once during offline data preparation. The generated captions and textual prototypes are cached and reused, and no LVLM call, CoT reasoning, or text generation is required during inference. As shown in Table 10, for University-1652, generating 7,010 captions with Qwen2.5-VL-32B-Int on 4× RTX A6000 GPUs takes 1.55 h, and the caption bank is only 2.17 MB. The WeatherAligner contains only 0.149M parameters and is used only as a training-time auxiliary module. Thus, the LVLM-related cost is a one-time offline overhead, and the online inference cost remains close to that of a standard visual retrieval model.

Table 10 Settings and statistics of LVLM-based caption generation

Setting	Value
Model	Qwen2.5-VL-32B
GPU	4× RTX A6000 48GB
Captioned Images	7,010
Max New Tokens	128
Generation Time	1.55 h
Avg. Caption Length	47 words
Avg. Input Tokens	790.0
Avg. Generated Tokens	56.35

Sensitivity to weather-disentanglement loss weights. To analyze the influence of different disentanglement constraint strengths on the final localization performance, we keep the network architecture, training data, training epochs, and other hyperparameters unchanged, and only adjust the weather-disentanglement loss weights λ_{weather} and λ_{inv}. Specifically, we evaluate three settings: a weaker disentanglement constraint of 0.01/0.003, the default setting of 0.03/0.01, and a stronger disentanglement constraint of 0.09/0.03. We then evaluate the S2D and D2S retrieval performance under 10 weather conditions on the full University-1652 test set. As shown in Table 11, the three settings achieve very similar average performance within the tested weight range. These results indicate that Weather4Geo does not rely on a precisely tuned disentanglement weight to obtain stable performance. The weather-disentanglement branch serves more as a stable auxiliary regularizer for suppressing weather-related perturbations, while the final geo-localization performance is still mainly dominated by

Table 11 Sensitivity to weather-disentanglement loss weights

Setting	Normal		Fog		Rain		Snow		Fog+Rain		Fog+Snow	
	R@1	AP	R@1	AP	R@1	AP	R@1	AP	R@1	AP	R@1	AP
Drone → Satellite												
Official (0.03/0.01)	84.38	86.61	**83.57**	**85.89**	**81.97**	**84.49**	**79.06**	81.94	80.57	83.28	75.79	78.99
Weak (0.01/0.003)	84.40	86.63	83.47	85.80	81.88	84.43	79.04	**81.96**	80.59	83.29	75.87	79.05
Strong (0.09/0.03)	**84.42**	**86.64**	83.56	85.88	81.94	84.47	78.97	81.89	**80.64**	**83.34**	**75.88**	**79.08**
Satellite → Drone												
Official (0.03/0.01)	90.01	84.19	**90.30**	82.58	89.44	81.93	**89.16**	78.54	89.44	80.08	88.02	74.12
Weak (0.01/0.003)	**90.16**	84.19	**90.30**	**82.63**	**89.87**	82.00	**89.16**	**78.55**	**89.59**	80.17	88.16	74.30
Strong (0.09/0.03)	**90.16**	**84.20**	**90.30**	**82.63**	**89.87**	**82.01**	89.02	**78.55**	**89.59**	**80.18**	**88.45**	**74.31**

Setting	Rain+Snow		Dark		Over-exp		Wind		Mean	
	R@1	AP	R@1	AP	R@1	AP	R@1	AP	R@1	AP
Drone → Satellite										
Official (0.03/0.01)	**79.90**	**82.72**	**71.94**	**75.33**	76.14	79.39	**78.56**	**81.53**	**79.19**	**82.02**
Weak (0.01/0.003)	79.84	82.68	71.56	74.97	**76.25**	**79.49**	78.39	81.43	79.13	81.97
Strong (0.09/0.03)	79.89	82.71	71.78	75.16	76.19	79.43	78.51	81.49	79.18	82.01
Satellite → Drone										
Official (0.03/0.01)	**88.30**	79.23	86.73	69.44	86.45	75.24	**89.30**	**78.19**	88.72	78.35
Weak (0.01/0.003)	**88.30**	**79.29**	**88.02**	**70.09**	86.45	75.33	89.16	78.18	**88.92**	**78.47**
Strong (0.09/0.03)	88.16	**79.29**	87.30	69.46	**87.02**	**75.38**	89.16	**78.19**	88.90	78.42

We report R@1(%) and AP(%) under 10 weather conditions for Drone → Satellite and Satellite → Drone tasks. Each setting is denoted by $\lambda_{\text{weather}}/\lambda_{\text{inv}}$. Best results are highlighted in bold

Table 12 Zero-shot cross-dataset retrieval performance on AnyVisLoc and UAV-VisLoc

Method	Recall (%)			PDM (%)		
	@1↑	@3↑	@5↑	@1↑	@3↑	@5↑
AnyVisLoc (Ye et al., 2025)						
Ours	**39.14**	**59.83**	**68.23**	**67.56**	**61.78**	**57.48**
CCR	38.27	56.70	65.34	63.58	57.44	53.11
Sample4Geo	33.56	52.61	62.03	60.52	55.55	51.86
SafeNet	28.76	44.42	52.81	53.49	47.87	44.17
MuseNet	30.03	48.87	57.74	56.78	52.06	48.40
UAV-VisLoc (Xu et al., 2024)						
Ours	**18.61**	**33.51**	**41.49**	**38.66**	**34.78**	**32.06**
CCR	18.22	32.75	40.42	36.82	32.38	29.65
Sample4Geo	15.76	28.95	36.59	34.12	30.41	28.10
SafeNet	9.78	19.07	25.35	25.07	22.49	21.00
MuseNet	10.51	20.97	27.71	26.39	24.07	22.60

PDM@K measures the rank-weighted spatial proximity of the top-K retrieved results to the ground-truth location. Higher values indicate more accurate localization. Best results are highlighted in bold

cross-view geographic semantic matching and identity discrimination objectives.

Cross-Dataset Evaluation on Real-World UAV Data. To further examine whether our method can generalize from synthetic-weather training distributions to complex real-world environments, we conduct target-training-free cross-dataset evaluations on two real-world UAV-to-Satellite datasets, AnyVisLoc (Ye et al., 2025) and UAV-VisLoc (Xu et al., 2024). AnyVisLoc was collected using real UAVs under diverse scenes, weather conditions, seasons, and illumination conditions, and provides corresponding satellite reference maps. UAV-VisLoc contains query images captured along real UAV flight trajectories and their corresponding large-scale satellite maps, covering seasonal variations between summer and autumn. Since neither dataset provides official per-image weather-condition annotations, we do not manually divide the query images into weather subsets, thereby avoiding pseudo-weather labels introduced by subjective judgment or automatic classification. Instead, we use all publicly available and valid UAV queries as a unified test set to evaluate the overall cross-dataset generalization ability of the model under real-world scene distributions. For both target datasets, no target-domain images are used to retrain or fine-tune the model. All network weights and inference configurations are determined solely on the University-1652 source domain. For AnyVisLoc, we adopt a unified satellite-reference retrieval protocol and use the same query set, gallery construction procedure, and evaluation metrics for all compared methods. For UAV-VisLoc, each UAV query is retrieved against the complete satellite gallery of its corresponding flight region. GPS information is used only to

determine the ground-truth location and compute the evaluation metrics, rather than to restrict the candidate gallery or generate image similarity scores.

As shown in Table 12, on AnyVisLoc, our method achieves R@1/3/5 scores of 39.14/59.83/68.23% and PDM@1/3/5 scores of 67.56/61.78/57.48%. Compared with CCR, our method improves R@1/3/5 by 0.87, 3.13, and 2.89 percentage points, respectively, and PDM@1/3/5 by 3.98, 4.35, and 4.37 percentage points, respectively. On UAV-VisLoc, our method achieves R@1/3/5 scores of 18.61/33.51/41.49% and PDM@1/3/5 scores of 38.66/34.78/32.06%, also demonstrating competitive cross-dataset retrieval performance. These results show that a model trained only on University-1652 and synthetic weather conditions can be transferred to two independent real-world UAV-to-Satellite datasets without any target-domain training or fine-tuning. Therefore, these experiments provide further evidence of the ability of our method to generalize from synthetic-weather training distributions to complex real-world environments. Compared with evaluations conducted solely on the original synthetic-weather test sets, these real-world results reduce the likelihood that the observed performance gains arise only from synthetic degradation artifacts or overfitting to a specific simulated distribution.

It should be noted that existing real-world datasets do not yet provide fully paired observations of the same locations captured under multiple strictly controlled weather conditions, nor do they provide unified per-image weather annotations. Therefore, the real-world cross-dataset evaluations and the original synthetic-weather evaluations are complementary rather than interchangeable. The synthetic-weather evaluations enable controlled analyses of the effects of different weather degradations on retrieval performance, whereas the real-world evaluations examine cross-dataset generalization under naturally occurring environmental variations. In future work, we will further investigate a large-scale real-world multi-weather UAV-to-Satellite benchmark with location-paired observations, explicit weather annotations, and temporally synchronized data.

5 Limitation and Future Work

The current evaluation relies on controllable query-side weather simulation and physics-guided stress tests. Although this setting enables reproducible robustness evaluation under representative adverse weather conditions, it cannot fully cover real-world variations in acquisition time, seasonal appearance, atmospheric visibility, sensor characteristics, and complex weather mixtures. Future work will further construct or introduce real multi-weather and multi-temporal drone-satellite datasets for more rigorous evaluation under open-world all-weather conditions. In addition, Weather4Geo currently focuses on single-frame image-based geo-localization. For continuous drone observations, temporal context and inter-frame consistency may provide useful cues when single-frame evidence is degraded by severe weather or viewpoint changes. Future work will explore extending language-guided weather-content alignment to video-based all-weather geo-localization through temporal aggregation and sequence-level matching.

6 Conclusion

In this paper, we introduced Weather4Geo, a language-guided framework for all-weather drone visual geo-localization. We first construct multi-weather image descriptions using a large vision-language model, and then use these rich textual descriptions to supervise complex weather patterns without any manual annotation. We devise a two-stage alignment strategy based on these descriptions. In the first stage, a text-guided image–text objective is used to learn weather-aware cross-view representations. In the second stage, a WeatherAligner module further refines these representations through anchor-based alignment, cross-weather invariance regularization, and a region-level textual prototype memory, so that the learned features remain discriminative for geographic identity while being markedly less sensitive to adverse and unseen weather. Extensive experiments on University-1652 and SUES-200, including cross-dataset evaluation and tests under unseen extreme weather, demonstrate that Weather4Geo not only achieves substantial gains over state-of-the-art methods in terms of Recall@1 and AP, but also significantly reduces performance variability across different weather conditions.

Funding This work was supported by the National Key R&D Program of China (Grant No. 2025YFF0514701).

Data Availability All the datasets used in this paper are publicly available.

References

Achiam, J., Adler, S., & Agarwal, S. (2023). GPT-4 technical report. arXiv preprint cs.CL arXiv:2303.08774.

Ahn, W.-J., Yang, G.-Y., Choi, H.-D., & Lim, M.-T. (2024). Style blind domain generalized semantic segmentation via covariance alignment and semantic consistence contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 3616–3626).

Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., & Lin, J. (2025). Qwen2.5-VL technical report. arXiv preprint arXiv:2502.13923 [cs.CV].

Barath, D., Cavalli, L., & Pollefeys, M. (2022). Learning to find good models in RANSAC. In *Proceedings of the IEEE/CVF conference*

on computer vision and pattern recognition (CVPR) (pp. 15744–15753).

Berton, G., Mereu, R., Trivigno, G., Masone, C., Csurka, G., Sattler, T., & Caputo, B. (2022). Deep visual geo-localization benchmark. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 5396–5407).

Bhalla, U., Oesterling, A., Srinivas, S., Calmon, F., & Lakkaraju, H. (2024). Interpreting clip with sparse linear concept embeddings (splice). *Advances in Neural Information Processing Systems, 37*, 84298–84328.

Bousmalis, K., Trigeorgis, G., Silberman, N., Krishnan, D., & Erhan, D. (2016). Domain separation networks. *Advances in neural information processing systems*, 29.

Bay, H., Tuytelaars, T., & Van Gool, L. (2006). SURF: Speeded up robust features. In *Proceedings of the European conference on computer vision (ECCV)*.

Chen, J., Guo, H., Yi, K., Li, B., & Elhoseiny, M. (2022). Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 18030–18040).

Chen, K., Liu, C., Chen, H., Zhang, H., Li, W., Zou, Z., & Shi, Z. (2024). RSPrompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model. *IEEE Transactions on Geoscience and Remote Sensing, 62*, 1–17.

Chen, H., Li, W., Gu, J., Ren, J., Chen, S., Ye, T., Pei, R., Zhou, K., Song, F., & Zhu, L. (2024). Restoreagent: Autonomous image restoration agent via multimodal large language models. arXiv preprint arXiv:2407.18035 [cs.CV]

Chen, Y.-C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., & Liu, J. (2020). UNITER: Universal image-text representation learning. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 104–120).

Chow, W., Li, J., Yu, Q., Pan, K., Fei, H., Ge, Z., Yang, S., Tang, S., Zhang, H., & Sun, Q. (2024). Unified generative and discriminative training for multi-modal large language models. arXiv preprint cs.CV arXiv:2411.00304.

Cruz-Mota, J., Bogdanova, I., Paquier, B., Bierlaire, M., & Thiran, J.-P. (2012). Scale invariant feature transform on the sphere: Theory and applications. *International Journal of Computer Vision, 98*, 217–241.

Choi, W., Shin, M., & Im, S. (2023). Depth-discriminative metric learning for monocular 3d object detection. In *Advances in neural information processing systems (NeurIPS)* (Vol. 36, pp. 80165–80177).

Castaldo, F., Zamir, A., Angst, R., Palmieri, F., & Savarese, S. (2015). Semantic cross-view matching. In *2015 IEEE international conference on computer vision workshops (ICCVW)*https://doi.org/10.1109/ICCVW.2015.137

Chu, M., Zheng, Z., Ji, W., Wang, T., & Chua, T. S. (2024). Towards natural language-guided drones: GeoText-1652 benchmark with spatial relation matching. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 213–231).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)* (pp. 4171–4186).

Dai, Y., Feng, Y., Ma, N., Zhao, X., & Gao, Y. (2025). Cross-modal 3d shape retrieval via heterogeneous dynamic graph representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 47*(4), 2370–2387. https://doi.org/10.1109/TPAMI.2024.3524440

Deuser, F., Habel, K., & Oswald, N. (2023). Sample4geo: Hard negative sampling for cross-view geo-localization. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)* (pp. 16847–16856).

Du, H., He, J., & Zhao, Y. (2024). CCR: A counterfactual causal reasoning-based method for cross-view geo-localization. *IEEE Transactions on Circuits and Systems for Video Technology, 34*, 11630–11643.

Dai, M., Hu, J., Zhuang, J., & Zheng, E. (2021). A transformer-based feature segmentation and region alignment method for UAV-view geo-localization. *IEEE Transactions on Circuits and Systems for Video Technology, 32*(7), 4376–4389.

Feng, Y., Huang, J., Du, S., Ying, S., Yong, J.-H., Li, Y., Ding, G., Ji, R., & Gao, Y. (2025). Hyper-yolo: When visual object detection meets hypergraph computation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 47*, 2388–2401.

Feng, Y., Ji, S., Liu, Y.-S., Du, S., Dai, Q., & Gao, Y. (2024). Hypergraph-based multi-modal representation for open-set 3d object retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 46*(4), 2206–2223. https://doi.org/10.1109/TPAMI.2023.3332768

Feng, T., Li, Q., Wang, X., Wang, M., Li, G., & Zhu, W. (2024). Multi-weather cross-view geo-localization using denoising diffusion models. In *Proceedings of the 2nd workshop on UAVs in multimedia: Capturing the world from a new perspective* (pp. 35–39).

Gao, Y., Lu, J., Li, S., Li, Y., & Du, S. (2024). Hypergraph-based multi-view action recognition using event cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 46*(10), 6610–6622. https://doi.org/10.1109/TPAMI.2024.3382117

Hoyer, L., Dai, D., Wang, H., & Van Gool, L. (2023). Mic: Masked image consistency for context-enhanced domain adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11721–11732).

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 4700–4708).

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., & Lerchner, A. (2017). beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International conference on learning representations*

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)* https://doi.org/10.1109/CVPR.2016.90

Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D., & Wilson, A. G. (2018). Averaging weights leads to wider optima and better generalization. arXiv preprint arXiv:1803.05407.

Ju, H., Huang, S., Liu, S., & Zheng, Z. (2024). Video2bev: Transforming drone videos to BEVs for video-based geo-localization. In *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*.

Jung, A.B., Wada, K., Crall, J., Tanaka, S., Graving, J., Reinders, C., Yadav, S., Banerjee, J., Vecsei, G., Kraft, A., Rui, Z., Borovec, J., Vallentin, C., Zhydenko, S., Pfeiffer, K., Cook, B., Fernández, I., De Rainville, F.-M., Weng, C.-H., Ayala-Acevedo, A., Meudec, R., Laporte, M., et al. (2020) imgaug. https://github.com/aleju/imgaug. Accessed 01 Feb 2020.

Kim, H. J., Dunn, E., & Frahm, J.-M. (2017). Learned contextual feature reweighting for image geo-localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 2136–2145).

Lin, T.-Y., Belongie, S., & Hays, J. (2013). Cross-view image geolocalization. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)* https://doi.org/10.1109/CVPR.2013.120.

Lin, T.-Y., Cui, Y., Belongie, S., & Hays, J. (2015). Learning deep representations for ground-to-aerial geolocalization. In *Proceedings of*

the IEEE conference on computer vision and pattern recognition (*CVPR*).

Li, J., He, X., Wei, L., Qian, L., Zhu, L., Xie, L., Zhuang, Y., Tian, Q., & Tang, S. (2022). Fine-grained semantically aligned vision-language pre-training. In *Advances in neural information processing systems* (*NeurIPS*) (vol. 35, pp. 7290–7303).

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). SWIN transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision* (*ICCV*), pp. 10012–10022.

Lin, J., Luo, Z., Lin, D., Li, S., & Zhong, Z. (2024). A self-adaptive feature extraction method for aerial-view geo-localization. *IEEE Transactions on Image Processing, 34*, 126–139.

Li, J., Li, D., Xiong, C., & Hoi, S. C. H.(2022). BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th international conference on machine learning (ICML). Proceedings of machine learning research (PMLR)* (vol. 162, pp. 12888–12900).

Liu, W., Ren, G., Yu, R., Guo, S., Zhu, J., & Zhang, L. (2022). Image-adaptive YOLO for object detection in adverse weather conditions. In *Proceedings of the AAAI conference on artificial intelligence (AAAI)* (Vol. 36, pp. 1792–1800).

Lu, J., Wan, H., Li, P., Zhao, X., Ma, N., & Gao, Y. (2023). Exploring high-order spatio-temporal correlations from skeleton for person re-identification. *IEEE Transactions on Image Processing, 32*, 949–963. https://doi.org/10.1109/TIP.2023.3236144

Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., & Wei, F. (2020). OSCAR: Object-semantics aligned pre-training for vision-language tasks. In *Proceedings of the European Conference on Computer Vision* (*ECCV*) (pp. 121–137).

Li, K., Zhang, Y., Li, K., Li, Y., & Fu, Y.(2019). Visual semantic reasoning for image-text matching. In *Proceedings of the IEEE/CVF international conference on computer vision* (*ICCV*), pp. 4654–4662.

Lin, J., Zheng, Z., Zhong, Z., Luo, Z., Li, S., Yang, Y., & Sebe, N. (2022). Joint representation learning and keypoint detection for cross-view geo-localization. *IEEE Transactions on Image Processing, 31*, 3780–3792. https://doi.org/10.1109/TIP.2022.3175601

Ma, X., Wang, Z., Zhan, Y., Zheng, Y., Wang, Z., Dai, D., & Lin, C.-W. (2022). Both style and fog matter: Cumulative domain adaptation for semantic foggy scene understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 18922–18931.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., & Antiga, L. (2019). PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems* (*NeurIPS*) (Vol. 32).

Pan, X., Luo, P., Shi, J., & Tang, X. (2018). Two at once: Enhancing learning and generalization capacities via IBN-net. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 484–500).

Qu, Z., Xu, K., Hancke, G. P., & Lau, R.W.H. (2024) Lush-NeRF: Lighting up and sharpening NeRFs for low-light scenes. arXiv preprint arXiv:2411.06757 [cs.CV].

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., & Clark, J. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748–8763).

Robbins, H., & Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics, 22*(3), 400–407.

Rodrigues, R., & Tani, M. (2022). Global assists local: Effective aerial representations for field-of-view constrained image geo-localization. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision (WACV)* (pp. 3871–3879).

Sindagi, V. A., Oza, P., Yasarla, R., & Patel, V. M. (2020). Prior-based domain adaptive object detection for hazy and rainy conditions. In *Proceedings of the European conference on computer vision* (*ECCV*).

Sánchez, J., Perronnin, F., Mensink, T., & Verbeek, J. (2013). Image classification with the fisher vector: Theory and practice. *International Journal of Computer Vision, 105*, 222–245.

Tzeng, E., Hoffman, J., Saenko, K., & Darrell, T. (2017). Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7167–7176).

Team, V., Hong, W., Yu, W., Gu, X., Wang, G., Gan, G., Tang, H., Cheng, J., Qi, J., Ji, J., Pan, L., Duan, S., Wang, W., Wang, Y., Cheng, Y., He, Z., Su, Z., Yang, Z., Pan, Z., Zeng, A., Wang, B., Chen, B., Shi, B., Pang, C., Zhang, C., Yin, D., Yang, F., Chen, G., Xu, J., Zhu, J., Chen, J., Chen, J., Chen, J., Lin, J., Wang, J., Chen, J., Lei, L., Gong, L., Pan, L., Liu, M., Xu, M., Zhang, M., Zheng, Q., Yang, S., Zhong, S., Huang, S., Zhao, S., Xue, S., Tu, S., Meng, S., Zhang, T., Luo, T., Hao, T., Tong, T., Li, W., Jia, W., Liu, X., Zhang, X., Lyu, X., Fan, X., Huang, X., Wang, Y., Xue, Y., Wang, Y., Wang, Y., An, Y., Du, Y., Shi, Y., Huang, Y., Niu, Y., Wang, Y., Yue, Y., Li, Y., Zhang, Y., Wang, Y., Wang, Y., Zhang, Y., Xue, Z., Hou, Z., Du, Z., Wang, Z., Zhang, P., Liu, D., Xu, B., Li, J., Huang, M., Dong, Y., & Tang, J. (2025) *GLM-4.5V and GLM-4.1V-Thinking: Towards versatile multimodal reasoning with scalable reinforcement learning*.

Tan, Y., Zhu, Y., Huang, Z., Tan, H., & Li, K. (2023). MAPD: An FPGA-based real-time video haze removal accelerator using mixed atmosphere prior. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 42*(12), 4777–4790.

Vendrow, E., Pantazis, O., Shepard, A., Brostow, G., Jones, K., Mac Aodha, O., Beery, S., & Van Horn, G. (2024). INQUIRE: A natural world text-to-image retrieval benchmark. In *Advances in neural information processing systems (NeurIPS)* (Vol. 37, pp. 126500–126514).

Vepa, A. M., Yang, Z., Choi, A., Joo, J., Scalzo, F., & Sun, Y. (2024). Integrating deep metric learning with coreset for active learning in 3d segmentation. In *Advances in neural information processing systems (NeurIPS)* (Vol. 37, pp. 71643–71671).

Wang, Z., Liu, X., Li, H., Sheng, L., Yan, J., Wang, X., & Shao, J. (2019). CAMP: Cross-modal adaptive message passing for text-image retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision* (*ICCV*), (pp. 5764–5773).

Workman, S., Souvenir, R., & Jacobs, N. (2015). Wide-area image geolocalization with aerial reference imagery. In *Proceedings of the IEEE international conference on computer vision* (*ICCV*).

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In *Advances in neural information processing systems (NeurIPS)* (Vol. 35, pp. 24824–24837).

Wen, J., Yu, H., & Zheng, Z. (2026). FANet: Fovea attention network for robust aerial geo-localization across diverse weather conditions. *IEEE Transactions on Image Processing*. https://doi.org/10.1109/TIP.2026.3690315

Wang, J., Zhu, Q., Lin, Z., Chen, J., Ding, G., Wu, Q., Gu, G., & Gao, Q. (2024). Sparse Bayesian learning-based hierarchical construction for 3d radio environment maps incorporating channel shadowing. *IEEE Transactions on Wireless Communications, 23*, 14560–14574.

Wang, T., Zheng, Z., Sun, Y., Yan, C., Yang, Y., & Chua, T.-S. (2024). Multiple-environment self-adaptive network for aerial-view geo-localization. *Pattern Recognition, 152*, Article 110363. https://doi.org/10.1016/j.patcog.2024.110363

Wang, T., Zheng, Z., Yan, C., Zhang, J., Sun, Y., Zheng, B., & Yang, Y. (2022). Each part matters: Local patterns facilitate cross-view geo-localization. *IEEE Transactions on Circuits and Systems for*

Video Technology, 32, 867–879. https://doi.org/10.1109/TCSVT.2021.3061265

Xu, W., Yao, Y., Cao, J., Wei, Z., Liu, C., Wang, J., & Peng, M. (2024). UAV-VisLoc: A large-scale dataset for UAV visual localization arXiv preprint. arXiv:2405.11936

Yan, S., Dong, N., Zhang, L., & Tang, J. (2023). CLIP-driven fine-grained text-image person re-identification. *IEEE Transactions on Image Processing, 32*, 6032–6046.

Yang, Y., Jiang, P.-T., Hou, Q., Zhang, H., Chen, J., & Li, B. (2024). Multi-task dense prediction via mixture of low-rank experts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 27927–27937).

Yang, H., Lu, X., & Zhu, Y. (2021). Cross-view geo-localization with layer-to-layer transformer. In *Advances in Neural Information Processing Systems (NeurIPS)* (Vol. 34, pp. 29009–29020).

Yan, C., Ma, M., Zhang, Y., & Wan, Y. (2024). Dual-path multimodal optimal transport for composed image retrieval. In *Proceedings of the Asian Conference on Computer Vision (ACCV)* (pp. 1741–1755).

Ye, Y., Teng, X., Chen, S., Li, Z., Liu, L., Yu, Q., & Tan, T. (2025). Exploring the best way for UAV visual localization under low-altitude multi-view observation condition: A benchmark. arXiv preprint arXiv:2503.10692

Yu, H., Wen, J., & Zheng, Z. (2025). Camel: Cross-modality adaptive meta-learning for text-based person retrieval. *IEEE Transactions on Information Forensics and Security, 20*, 4651–4663.

Yuan, Y., Zhan, Y., & Xiong, Z. (2023). Parameter-efficient transfer learning for remote sensing image-text retrieval. *IEEE Transactions on Geoscience and Remote Sensing, 61*, 1–14.

Yang, S., Zhou, Y., Zheng, Z., Wang, Y., Zhu, L., & Wu, Y. (2023). Towards unified text-based person retrieval: A large-scale multi-attribute and language search benchmark. In *Proceedings of the 31st ACM International Conference on Multimedia (ACM MM)* (pp. 4492–4501).

Zhang, J., Cao, Y., & Wu, Q. (2021). Vector of locally and adaptively aggregated descriptors for image feature representation. *Pattern Recognition, 116*, Article 107952.

Zhang, X., Li, X., Sultani, W., Zhou, Y., & Wshah, S. (2023). Cross-view geo-localization via learning disentangled geometric layout correspondence. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)* (Vol. 37, pp. 3480–3488)

Zhu, J., Li, Y., Yang, K., Guan, N., Fan, Z., Qiu, C., & Yi, X. (2024). MVP: Meta visual prompt tuning for few-shot remote sensing image scene classification. *IEEE Transactions on Geoscience and Remote Sensing, 62*, 1–13.

Zhu, S., Shah, M., & Chen, C. (2022). TransGeo: Transformer is all you need for cross-view image geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 1162–1171).

Zheng, Z., Wei, Y., & Yang, Y. (2020). University-1652: A multi-view multi-source benchmark for drone-based geo-localization. *Proceedings of the 28th ACM International Conference on Multimedia (ACM MM)* (pp. 1395–1403).

Zeng, Z., Wang, Z., Yang, F., & Satoh, S. (2023). Geo-localization via ground-to-satellite cross-view image retrieval. *IEEE Transactions on Multimedia, 25*, 2176–2188. https://doi.org/10.1109/TMM.2022.3144066

Zhu, R., Yin, L., Yang, M., Wu, F., Yang, Y., & Hu, W. (2023). SUES-200: A multi-height multi-scene cross-view image benchmark across drone and satellite. *IEEE Transactions on Circuits and Systems for Video Technology, 33*(9), 4825–4839.

Zheng, Z., Zheng, L., Garrett, M., Yang, Y., Xu, M., & Shen, Y.-D. (2020). Dual-path convolutional image-text embeddings with instance loss. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 16*(2), 1–23.

Zhu, Q., Zhao, Y., Huang, Y., Lin, Z., Wang, L. H., Bai, Y., Lan, T., Zhou, F., & Wu, Q. (2022). Demo abstract: An UAV-based 3d spectrum real-time mapping system. *IEEE INFOCOM 2022—IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)* (pp. 1–2).

Zeng, Y., Zhang, X., & Li, H. (2021). Multi-grained vision language pre-training: Aligning texts with visual concepts. arXiv preprint cs.CV arXiv:2111.08276

Zhang, Z., Zhang, A., Li, M., Zhao, H., Karypis, G., & Smola, A. (2023). Multimodal chain-of-thought reasoning in language models. arXiv preprint arXiv:2302.00923 [cs.CL]

Zhang, W., Zhai, G., Wei, Y., Yang, X., & Ma, K. (2023). Blind image quality assessment via vision-language correspondence: A multitask learning perspective. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 14071–14081).

Zheng, Z., Zheng, L., & Yang, Y. (2017). A discriminatively learned CNN embedding for person re-identification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 14*(1), 1–20. https://doi.org/10.1145/3159171

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.