WeatherPrompt: Multi-modality Representation Learning for All-Weather Drone Visual Geo-Localization

Jiahao Wen[1] **Hang Yu**[1*] **Zhedong Zheng**[2]
[1]School of Computer Engineering and Science, Shanghai University, China
[2]Faculty of Science and Technology and Institute of Collaborative Innovation, University of Macau, China
{wenjh,yuhang}@shu.edu.cn, zhedongzheng@um.edu.mo

Abstract

Visual geo-localization for drones faces critical degradation under weather perturbations, *e.g.*, rain and fog, where existing methods struggle with two inherent limitations: 1) Heavy reliance on limited weather categories that constrain generalization, and 2) Suboptimal disentanglement of entangled scene-weather features through pseudo weather categories. We present WeatherPrompt, a multi-modality learning paradigm that establishes weather-invariant representations through fusing the image embedding with the text context. Our framework introduces two key contributions: First, a Training-free Weather Reasoning mechanism that employs off-the-shelf large multi-modality models to synthesize multi-weather textual descriptions through human-like reasoning. It improves the scalability to unseen or complex weather, and could reflect different weather strength. Second, to better disentangle the scene and weather features, we propose a multi-modality framework with the dynamic gating mechanism driven by the text embedding to adaptively reweight and fuse visual features across modalities. The framework is further optimized by the cross-modal objectives, including image-text contrastive learning and image-text matching, which maps the same scene with different weather conditions closer in the representation space. Extensive experiments validate that, under diverse weather conditions, our method achieves competitive recall rates compared to state-of-the-art drone geo-localization methods. Notably, it improves Recall@1 by 13.37% under night conditions and by 18.69% under fog and snow conditions. Our code is available at https://github.com/Jahawn-Wen/WeatherPrompt.

1 Introduction

Drone visual geo-localization aims to match drone-view image with corresponding satellite views, supporting critical applications such as disaster response, urban surveillance, search-and-rescue, and environmental monitoring [1, 2, 3, 4, 5, 6]. However, variable weather conditions such as rain, fog and snow introduce noise, occlusions and low visibility, which severely distort image features [7, 8, 9] and leading conventional localization methods to suffer drastic performance degradation under extreme weather. Recent advances in cross-modal retrieval show that integrating natural language descriptions can substantially enhance the discrimination power of vision models, allowing better generalization in complex or ambiguous scenarios [10, 11, 12, 13]. Despite this progress, leveraging textual guidance for cross-weather drone geo-localization remains largely underexplored, especially considering the nuanced and dynamic nature of weather conditions encountered in the field. The ability of text to capture complex semantics and fine-grained details [14, 15] offers a promising

*Corresponding author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

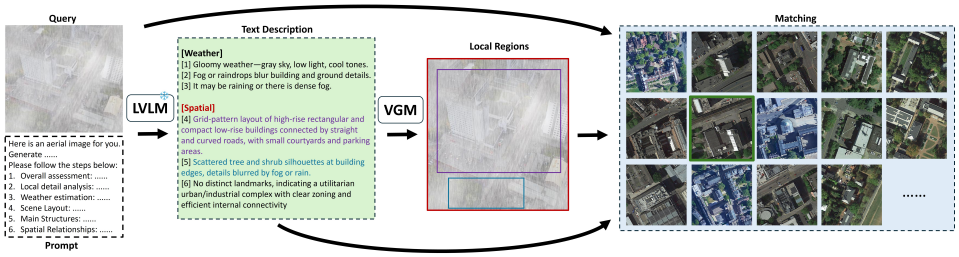

Figure 1: **Example of the proposed Chain-of-Thought description and matching.** Our framework generates structured weather and spatial Text Description via stepwise reasoning. We leverage Off-the-shelf Visual Grounding Model (VGM), *i.e.*, XVLM [16] to extract local region cues, which are integrated to further refine the matching process. Finally, we match images using weather description, global scene layout, and local region semantics to retrieve the corresponding satellite-view image.

avenue for cross-weather generalization. Addressing the all-weather visual geo-localization task presents two primary challenges: (1) Limited Weather Labeling. Existing approaches typically perform domain-specific fine-tuning on a limited set of predefined weather labels (*e.g.*, sunny, rainy). This closed-set paradigm fails to capture the continuous and combinatorial nature of real-world weather, thereby limiting model generalization to unseen or mixed conditions and preventing the exploitation of richer weather semantics. (2) Scene–Weather Feature Entanglement. Existing methods directly inject coarse pseudo-weather labels (*e.g.*, rain, fog) into visual representations during training, leading to a severe entanglement between scene semantics and weather disturbances. Consequently, the model learns suboptimal representations under mixed or unseen weather conditions, fails to disentangle scene content from weather noise, and is severely limited in cross-weather generalization.

For the first limitation, we propose a training-free weather reasoning mechanism leveraging off-the-shelf large multimodal models [17] (see Fig. 1). Specifically, we employ chain-of-thought (CoT) [18] prompting to automatically generate rich, step-by-step natural language weather descriptions for each geographical scene using a single randomly sampled drone-view image. This strategy circumvents manual description and expert controlled, enabling large-scale collection of high-quality and diverse multimodal samples at significantly reduced costs. Moreover, the stepwise reasoning introduced by the CoT prompts ensures semantic accuracy and formatting consistency across generated captions, further enhancing the reliability and usability of descriptions. This ultimately leads to improved generalization for complex and unseen weather conditions. To address the second challenge, we propose a multimodal framework equipped with a text embedding-driven dynamic gating mechanism to adaptively reweight and fuse visual features, effectively disentangling scene and weather attributes. Specifically, during training, the framework jointly optimizes multimodal objectives including image-text contrastive (ITC) loss and image-text matching (ITM) loss, aligning drone images with generated weather-aware captions. Additionally, a localized alignment loss is introduced to explicitly enforce consistency between annotated visual regions and textual descriptions across multiple granularities, encouraging the visual encoder to learn robust and weather-invariant scene representations. At inference, visual and textual embeddings are extracted in parallel, and the textual embeddings dynamically modulate visual features via the gating mechanism. The resulting multimodal representation is directly fed into a classification head for localization prediction, avoiding any additional online fine-tuning or dedicated parameter sets for different weather conditions. This design significantly reduces structural complexity and improves the deployment efficiency on resource-constrained platforms. The main contributions are as follows:

- **Training-Free Weather Reasoning**: We pioneer automatic weather semantics extraction through Large Vision Language Models (LVLMs) with chain-of-thought prompting, eliminating manual descriptions. Our hierarchical reasoning mechanism integrates continuous weather priors with spatial-object analysis, enabling weather-adaptive caption generation and scalable dataset construction.

- **Semantic Disentanglement via Language Guidance**: We devise a text-driven framework achieving scene-weather disentanglement through: (1) Multi-granularity alignment of visual features with continuous textual weather embeddings, (2) Region-level semantic consistency enforcement, (3) Dynamic textual gating for weather-invariant representations.

- **State-of-the-Art Generalization**: The proposed method has achieved average 87.72% Recall@1 on University-1652 and 83.1% on SUS-200 over 10 different weather conditions. For unseen weather combinations (*i.e.*, Dark+Rain+Fog), ours still arrives at 72.15% AP, validating unprecedented cross-domain generalization.

2 Related Work

Cross-view Geolocalization. Cross-view geo-localization aims to match images captured from different viewpoints with their corresponding geographic locations [1, 19, 20, 21]. Early approaches relied on hand-crafted local features such as SIFT [22] and SURF [23], as well as global descriptors like VLAD [24] and Fisher Vector [25], often combined with RANSAC [26] for geometric verification; however, they remain brittle under large viewpoint and illumination changes [27, 28]. With the advent of deep learning, deep metric learning frameworks based on global or part-based contrastive objectives have become dominant [29, 30, 31, 32, 33, 34]. These approaches employ triplet or InfoNCE losses to train end-to-end embeddings and integrate global pooling with spatial attention or multi-region partitioning strategies [35, 36]. Representative examples include the multi-part partitioning scheme *et al.* [2], the keypoint attention module [3], the content-aligned Transformer architecture [37], self-attention positional encoding by Yang *et al.* [35], the dual-path fusion network [36], and Bird's Eye View (BEV) [38], all of which substantially enhance cross-view feature alignment. Recent research has begun to address the impact of image degradations such as low-light, motion blur, and synthetic fog, often by using data augmentation, domain adaptation, or cross-modal transformers [39, 40, 41, 42, 43, 44]. However, most existing methods still depend on a limited set of discrete weather labels, restricting generalization to unseen or complex conditions. In contrast, our approach introduces a training-free, all-weather text-guided representation learning framework that leverages open-set weather descriptions to overcome these limitations.

Multi-modality Alignment. In this work, we address weather-aware text-guided representation learning, where the goal is to retrieve drone-view images based on fine-grained weather-related textual cues. Recent advances in vision–language alignment, such as CLIP [45], BLIP [46], and XVLM [16], have established powerful contrastive pre-training and cross-modal attention mechanisms, but previous studies mainly target static semantics or rigid spatial relations [47, 48, 49]. Additional efforts on adaptive fusion and region-word alignment [50, 51, 52, 53] have improved retrieval, but remain limited by closed vocabularies and overlook dynamic, fine-grained weather semantics. To address these gaps, we propose a framework that generates open-set weather descriptions via Chain-of-Thought prompting and applies text-driven dynamic gating for adaptive feature modulation, achieving robust cross-modal alignment under diverse weather conditions.

Large Vision Language Models for Vision via Prompting. Large Vision Language Models (LVLMs) such as GPT-3/4 [54, 55] and Qwen [17] have recently been applied to vision tasks using prompt engineering. Approaches like VisualGPT [56] and MM-CoT [57] leverage Chain-of-Thought (CoT) prompts [18] to elicit stepwise reasoning for visual question answering and captioning. However, scaling LVLMs to cross-view, multi-weather geo-localization remains challenging: free-form text often suffers from hallucinations and lacks semantic or structural consistency [58, 59, 60]. Prior works also overlook prompt design tailored for robust, multi-weather, multi-scale cross-modal alignment. To address this, we introduce the first CoT-driven description pipeline for multi-weather drone-to-satellite geo-localization. Our structured prompts regularize LVLMs' outputs, enabling scalable generation of high-quality, open-set weather descriptions to advance large-scale vision–language alignment.

3 Method

3.1 Open-Weather Description

As shown in Fig. 2, we present an overview of our multi-weather drone-view image captioning pipeline. To minimize description redundancy and mitigate overfitting, our approach begins by randomly sampling a single representative drone-view image from each geographical region, forming concise yet diverse image–text pairs. Notably, Large Vision Language Models (LVLMs) typically generate weather labels or captions intuitively, which may omit critical visual cues and exhibit semantic inconsistency. To address this limitation, we explicitly divide the captioning process into

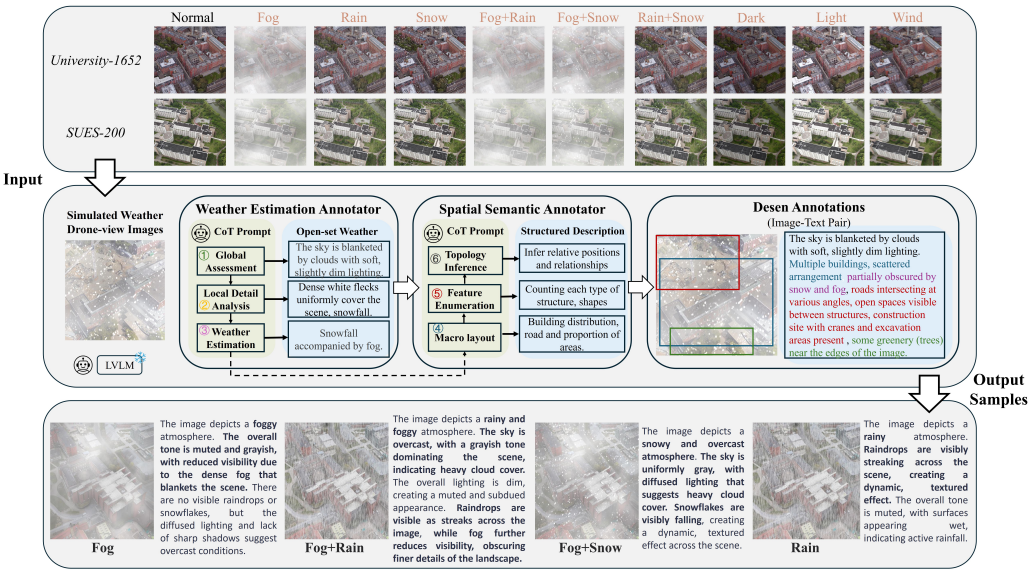

Figure 2: **The proposed training-free weather reasoning mechanism**. We synthesize drone-view images with diverse weather conditions based on the University-1652 and SUES-200 datasets, covering complex scenarios such as fog, rain, snow, and nighttime. For each synthesized image, we first employ stepwise Chain-of-Thought prompting to generate open-set weather descriptions, including global assessment, local detail analysis, and weather inference. Guided by the inferred weather prior, we then sequentially reason about the scene's macro layout, structural elements, and topological relationships, ultimately producing high-quality, structured image–text pairs.

two sequential phases: a weather estimation phase and a spatial semantics phase, guided by carefully designed Chain-of-Thought (CoT) prompts. Inspired by human hierarchical visual reasoning, our CoT prompting strategy enforces a rigorous three-step reasoning procedure: global perception, local analysis, and comprehensive synthesis.

In the weather estimation phase, the LVLMs first assesses global visibility to quantify observable range, subsequently identifies localized atmospheric indicators such as rain-streak reflections or fog diffusion patterns, and finally integrates these global and local cues to assign an accurate weather category. This structured inference process significantly alleviates the ambiguity inherent to single-shot captioning methods, establishing a reliable weather-conditioned prior for the following stage.

In the subsequent spatial semantics phase, conditioned explicitly on the inferred weather semantic, the model rapidly identifies macro-level layout features, including the spatial distribution of buildings, orientations of roads, and proportion of open spaces. Then, it captures micro-level structural details such as object counts, shapes, and local spatial arrangements. Ultimately, the pipeline synthesizes these cues by reasoning about relative positions and topological relationships to generate a structured textual description. This final output aligns comprehensively with weather semantics, spatial structure, and detailed scene attributes, ensuring robust, consistent, and high-quality multimodal descriptions.

Single-image Sampling. Since existing benchmarks (*e.g.*, University-1652) are partitioned by geographic regions, we randomly select only one drone-view image per region as a representative to avoid redundancy. Given the prohibitive cost of obtaining real drone-view imagery across diverse weather conditions, we utilize the imgaug [61] library to synthesize realistic weather variations. By parametrically adjusting visibility and occlusion effects such as rain, fog, and snow, we generate high-fidelity, diverse meteorological scenarios that closely resemble real-world conditions.

Weather Estimation Phase. Given a drone-view image with synthetically generated weather, we apply a pretrained large multimodal model [17] for automatic weather description. To mitigate hallucinations, vague phrasing, and inconsistent terminology typical of single-shot captioning, we adopt a Chain-of-Thought prompting strategy inspired by human visual perception. Specifically, we first quantify global visibility, then identify local meteorological cues such as rain-streak reflections or fog diffusion, and finally integrate global and local evidence to determine a reliable weather label. Any description that lacks explicit visibility information, meteorological cues, or contains

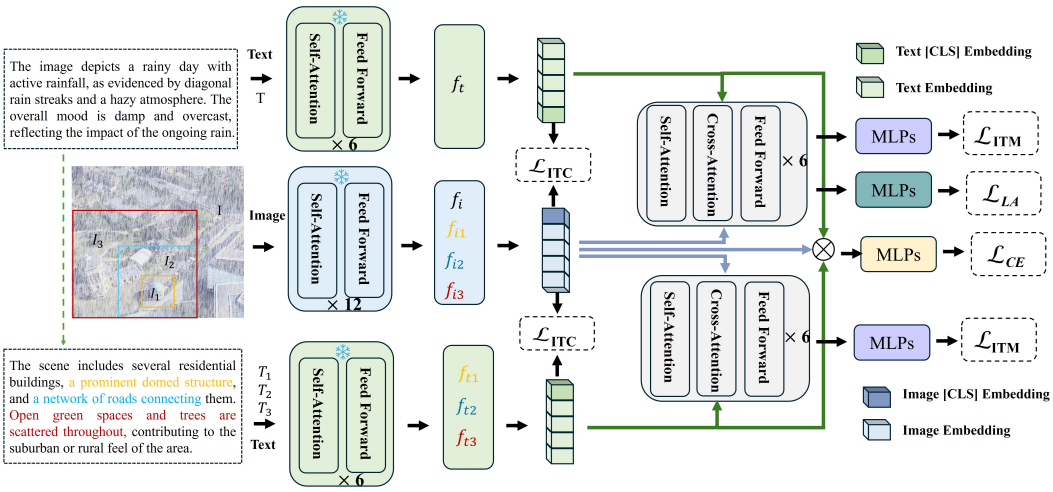

Figure 3: **The proposed multimodal alignment framework.** Our model extracts global and local features from drone images and multi-step weather captions, performs multi-granular image-text alignment, and dynamically fuses modalities via weather-driven gating for robust geo-localization.

uncertain terminology (*e.g.*, "possibly" or "uncertain") is automatically rejected and regenerated. This three-step reasoning pipeline ensures accurate, consistent, and reliable weather descriptions, thus providing a robust prior for the subsequent spatial semantics phase.

Spatial Semantics Phase. Conditioned on the inferred weather prior, we further generate detailed, scene-level textual descriptions essential for precise vision–language alignment. Initially, the model conducts a global scan to capture macro-level structures, including building distributions, road orientations, and proportions of open areas. Subsequently, it enumerates fine-grained elements, accurately counting structural entities, describing shapes, and detailing local spatial arrangements. Lastly, it synthesizes macro-layout and micro-level information to explicitly infer relative positions and topological relationships, producing structured, natural-language scene descriptions. These descriptions serve as fine-grained semantic supervision for downstream multimodal alignment.

Discussion. Existing cross-view geo-localization methods primarily rely on image-based retrieval. While some recent studies [21, 62] integrate textual descriptions, these approaches remain limited to coarse, scene-level labels and largely neglect weather semantics, resulting in compromised robustness under diverse weather conditions. In contrast, we introduce a training-free, CoT-driven pipeline for automatically generating multi-weather semantic descriptions. By employing large multimodal models guided by Chain-of-Thought prompting, we move beyond predefined discrete weather categories and achieve robust generalization to unseen and complex meteorological conditions. Furthermore, our framework systematically annotates both macro-level spatial layouts and micro-level structural attributes, embedding essential spatial cues into the image–text pairs. Relying solely on pretrained multimodal models without manual description, our approach efficiently produces large-scale, semantically consistent multimodal data, significantly facilitating downstream cross-weather multimodal representation learning.

3.2 Multimodal Alignment Model

We introduce a training-free, all-weather text-guided representation learning framework (see Fig. 3) that aligns weather semantics injected via text prompts with visual features to produce consistent, discriminative cross-view embeddings across diverse meteorological conditions. The framework consists of a pretrained visual encoder, a text encoder, and a cross-modal fusion module, capped by a lightweight classification head. During inference, the text embedding parametrizes a gating mechanism that adaptively modulates the visual feature channels; the fused multimodal representation is then passed through a single-layer MLP to predict the geographic location.

Weather-Driven Channel Gating. Let $f_{\mathrm{I}} \in \mathbb{R}^{B\times D}$ and $f_{\mathrm{T}} \in \mathbb{R}^{B\times D}$ represent the normalized visual and textual embeddings for a batch of B samples. We utilize text embeddings to generate an adaptive

channel-wise gating vector $g \in (0,1)^{B\times D}$:

$$z = \mathrm{ReLU}\big(f_{\mathrm{T}} W_1^{(gate)\top} + b_1^{(gate)}\big) \in \mathbb{R}^{B\times D/r}, \quad g \quad = \sigma\big(z W_2^{(gate)\top} + b_2^{(gate)}\big) \in (0,1)^{B\times D}, \tag{1}$$

where $W_1^{(gate)}$, $W_2^{(gate)}$, $b_1^{(gate)}$, $b_2^{(gate)}$ are learnable parameters, and r is the reduction ratio. The fused multimodal feature is computed by

$$f_{\text{fuse}} = g \odot f_{\mathrm{I}} + (1-g) \odot f_{\mathrm{T}} \quad \in \mathbb{R}^{B\times D}, \tag{2}$$

followed by a classification head for geo-localization. By dynamically modulating visual channels conditioned on textual weather semantics, our approach yields discriminative and weather-robust embeddings with minimal computational overhead.

Image-Text Contrastive. Given paired samples $\{(I_i, T_i)\}_{i=1}^{B}$, we compute the similarity matrix $S_{ij} = \frac{I_i^\top T_j}{\tau} \in \mathbb{R}^{B\times B}$ between normalized visual embeddings I_i and textual embeddings T_j, where τ is a learnable temperature parameter. We then convert the diagonal entries into retrieval probabilities by applying softmax over rows and columns:

$$p_{I\to T}^{(i)} = \frac{\exp(S_{ii})}{\sum_{j=1}^{B}\exp(S_{ij})}, \quad p_{T\to I}^{(i)} \quad = \frac{\exp(S_{ii})}{\sum_{j=1}^{B}\exp(S_{ji})}. \tag{3}$$

The contrastive loss is defined as:

$$\mathcal{L}_{\mathrm{ITC}} = -\frac{1}{2B}\sum_{i=1}^{B}\Big[\log p_{I\to T}^{(i)} + \log p_{T\to I}^{(i)}\Big], \tag{4}$$

enforcing global semantic alignment between the visual and textual modalities.

Image-Text Matching. To further enhance fine-grained discriminability, we introduce an image-text matching loss utilizing hard negatives. Given the similarity matrix S, for each image I_i and its paired textual T_i), we select the highest non–diagonal similarity entry as its hard negative: $T_i^- = \arg\max_{j\neq i} S_{ij}, I_i^- = \arg\max_{j\neq i} S_{ji}$. This forms $2B$ hard negative pairs (I_i, T_i^-) and (I_i^-, T_i), and B positive pairs (I_i, T_i), resulting in $3B$ pairs in total. Each pair is passed through the cross-modal encoder, and we extract the CLS embedding h_k which is fed to a binary classifier f_{match}. The matching probability is $p_k = \sigma\big(f_{\text{match}}(h_k)\big)$, where σ is the sigmoid function. Denoting the ground-truth label $y_k = 1$ for positive pairs and $y_k = 0$ for negatives, the matching loss is defined as

$$\mathcal{L}_{\mathrm{ITM}} = -\frac{1}{3B}\sum_{k=1}^{3B}\Big[y_k \log p_k + (1-y_k)\log(1-p_k)\Big], \tag{5}$$

which guides the model to hear subtle semantic distinctions between closely related pairs.

Localized Alignment Module. To enable text-driven fine-grained visual localization, the model responds to both global descriptions T and the region hints T_1, T_2, T_3. For the j-th concept , we extract its [CLS] embedding $x_{\text{cls}}^{j} \in \mathbb{R}^d$. from the cross-encoder, which is projected to normalized coordinate vector via a dedicated two-layer MLP:

$$\hat{l}^j = \sigma\big(W_2^{(loc)}\,\mathrm{GELU}(W_1^{(loc)}\,x_{\text{cls}}^{j} + b_1^{(loc)}) + b_2^{(loc)}\big) \in [0,1]^4 \tag{6}$$

where $W_1^{(loc)} \in \mathbb{R}^{2d\times d}$, $W_2^{(loc)} \in \mathbb{R}^{4\times 2d}$, $b_1^{(loc)} \in \mathbb{R}^{2d}$, $b_2^{(loc)} \in \mathbb{R}^4$. The resulting vector $\hat{l}^j = (\hat{c}_x, \hat{c}_y, \hat{w}, \hat{h})$ encodes the region center and size in normalized coordinates. The localized alignment loss supervises both overlap and regression accuracy:

$$\mathcal{L}_{\mathrm{LA}} = \mathbb{E}_{(I,T^j)\sim\mathcal{D}}\Big[\underbrace{1 - \mathrm{IoU}\big(l^j, \hat{l}^j\big)}_{\text{Overlap Loss}} + \underbrace{\|l^j - \hat{l}^j\|_1}_{\text{L1 Loss}}\Big], \tag{7}$$

where $l^j = (c_x, c_y, w, h)$ is the ground-truth region, $\hat{l}^j$ its prediction, and $\mathrm{IoU}(\cdot,\cdot)$ computes the Intersection-over-Union.

Classification Module. After multimodal fusion, we employ a single-layer MLP classification head to predict the geographic location. Let the fused feature for sample b be $z_b \in \mathbb{R}^D$, and classifier

Method	Normal		Fog		Rain		Snow		Fog+Rain		Fog+Snow		Rain+Snow		Dark		Over-exp		Wind		Mean	
	R@1	AP	R@1	AP	R@1	AP	R@1	AP	R@1	AP	R@1	AP	R@1	AP	R@1	AP	R@1	AP	R@1	AP	R@1	AP
Drone → Satellite																						
Zheng *et al.*[1] [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	57.61	62.03	39.70	44.65	51.85	56.75	58.28	62.83	55.17	59.71
ResNet-101 [68] [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	60.25	64.68	46.12	51.02	56.34	61.23	62.13	66.63	58.76	63.29
DenseNet121 [69] [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	60.21	64.57	48.57	53.41	54.04	58.88	60.74	65.14	59.01	63.44
Swin-T [64] [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	62.68	67.02	50.27	55.18	55.46	60.29	63.81	68.17	61.56	65.95
IBN-Net [70] [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	64.07	68.13	50.72	55.53	57.97	62.52	66.73	70.68	62.30	66.46
LPN [2] [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	65.38	69.29	53.68	58.10	60.90	65.27	66.46	70.35	64.16	68.14
Sample4Geo* [71] [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	38.19	42.33	46.34	49.91	75.77	78.90	**81.54**	**87.34**	61.10	64.32
Safe-Net* [72] [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	25.83	26.40	41.10	44.13	69.87	71.15	74.32	76.58	60.48	63.36
CCR* [73] [TCSVT'24]	92.54	93.78	85.57	87.13	67.46	68.82	55.16	59.14	63.11	60.97	27.74	31.48	23.06	46.85	51.10	54.19	**75.90**	**79.16**	81.31	87.22	62.30	66.87
MuSe-Net* [29][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	66.64	70.55	53.85	58.49	61.05	65.51	69.45	73.22	65.15	69.16
Ours	82.78	85.18	81.46	84.03	**80.34**	**83.11**	**77.60**	**80.67**	**78.75**	**81.69**	**73.38**	**76.94**	**78.41**	**81.40**	**67.22**	**71.06**	74.20	77.63	77.26	80.27	**77.14** (+11.99)	**80.20** (+11.04)
Satellite → Drone																						
Zheng *et al.*[1] [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	74.75	56.44	69.47	39.25	72.18	51.91	76.46	57.59	75.33	54.68
ResNet-101 [68] [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	80.17	62.61	75.32	45.37	79.60	58.21	82.31	64.67	80.13	60.74
DenseNet121 [69] [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	78.46	61.68	74.47	47.88	74.32	55.26	78.32	61.63	78.55	60.15
Swin-T [64] [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	78.89	63.49	75.61	48.43	76.60	56.57	78.74	64.45	78.53	61.83
IBN-Net [70] [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	83.02	65.74	78.46	50.77	79.46	58.64	84.02	67.94	82.27	63.36
LPN [2] [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	84.17	66.26	82.88	52.05	81.03	62.24	84.14	67.35	83.64	65.08
Sample4Geo* [71] [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	64.48	32.38	77.03	45.89	**91.58**	**77.04**	**93.30**	**81.42**	78.82	57.42
Safe-Net* [72] [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	61.07	29.86	73.15	43.08	88.54	74.65	90.02	78.21	75.69	58.80
CCR* [73] [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	61.11	32.21	64.80	46.28	86.01	71.23	92.67	76.55	77.02	64.02
MuSe-Net* [29] [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	84.88	67.75	80.74	53.01	81.60	62.09	86.31	70.03	84.68	65.75
Ours	89.16	81.80	88.73	80.58	**88.16**	**79.87**	**87.59**	**77.25**	**88.45**	**78.20**	**86.73**	**73.23**	**88.59**	**78.14**	**86.59**	**65.20**	85.31	73.25	87.88	76.33	**87.72** (+3.04)	**76.39** (+10.64)

Table 1: **Performance (R@1(%) and AP(%)) on University-1652** for Drone → Satellite and Satellite → Drone tasks. In both tasks, drone-view images are stylized 10 different weather conditions, and the satellite-view images are constant. Best results are highlighted in bold. * denotes the use of official pretrained weights.

output logits $o_b = W_{\text{clf}} z_b + b_{\text{clf}} \in \mathbb{R}^C$, where $W_{\text{clf}} \in \mathbb{R}^{C \times D}$ and $b_{\text{clf}} \in \mathbb{R}^C$. Given the ground-truth location label $y_b \in \{1, \ldots, C\}$, the softmax cross-entropy loss is:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{B} \sum_{b=1}^{B} \log \frac{\exp\left(o_{b,y_b}\right)}{\sum_{c=1}^{C} \exp\left(o_{b,c}\right)}. \tag{8}$$

Optimization Objectives. Our total loss L_{total} is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{ITC} + \mathcal{L}_{ITM} + \mathcal{L}_{LA} + \mathcal{L}_{CE}. \tag{9}$$

This unified objective enables the model to learn fine-grained, weather-invariant representations, substantially improving cross-weather generalization.

4 Experiment

Implementation Details. We adopt XVLM [16] as the backbone, which is pre-trained on 4M image–caption pairs, integrates BERT [63] as the text encoder and Swin Transformer [64] as the image encoder. The model is optimized using stochastic gradient descent (SGD) [65] with momentum 0.9 and weight decay 0.0005. Training 210 epochs, with the learning rate reduced by 0.1 at epoch 120 and by 0.01 at epoch 180. We resize input images to 384×384 pixels and divide them into 32×32 patches. During training, satellite-view images are augmented via random cropping and horizontal flipping, for drone-view images, we first apply style transformations using the imgaug [61] library and then perform the same random crop and flip augmentations. At test time, we compute the Euclidean distance between query and candidate embeddings to measure similarity. All experiments have been implemented in PyTorch [66] and conducted on a single NVIDIA RTX A6000 GPU, with an average inference time of 0.024s per query.

Dateset. University-1652 [1] is a large-scale cross-view geo-localization dataset comprising images from 1,652 university locations. Each location is represented by satellite, drone, and ground-level images, with 54 drone-view and 1 satellite-view images per building, as well as street-view imagery. The dataset is split into 701 training and 951 test buildings, with no overlap between train and test sets. **SUES-200** [67] contains multi-view drone and satellite images from 200 locations in Shanghai, encompassing diverse urban scenes, parks, lakes, and public buildings. Drone images are captured from multiple altitudes (150–300m) to simulate varied real-world conditions.

4.1 Comparison with Competitive Methods

The experimental results on the University-1652 dataset are shown in Tab 1. We compare the proposed WeatherPrompt with several competitive cross-view geo-localization methods. In the Drone → Satellite task, WeatherPrompt achieves a mean Recall@1 (R@1) accuracy of 77.14% (+11.99%) and a mean Average Precision (AP) of 80.20% (+11.04%), outperforming existing state-of-the-art methods for multi-weather geo-localization. Similarly, in the Satellite → Drone task, WeatherPrompt

Method	Normal		Fog		Rain		Snow		Fog+Rain		Fog+Snow		Rain+Snow		Dark		Over-exp		Wind		Mean	
	R@1	AP	R@1	AP	R@1	AP	R@1	AP	R@1	AP	R@1	AP	R@1	AP	R@1	AP	R@1	AP	R@1	AP	R@1	AP
Drone → Satellite																						
Zheng *et al.*[1] [backbone]	57.70	58.30	48.63	49.61	53.41	52.72	41.78	43.47	37.17	37.44	44.22	46.18	40.60	40.63	23.81	25.45	49.79	50.64	47.42	48.31	44.43	45.12
IBN-Net [70] [backbone]	65.34	63.78	56.03	56.57	55.73	58.55	47.80	49.53	43.45	44.98	50.04	51.00	45.51	45.92	29.61	30.93	56.01	56.96	57.36	58.10	50.69	51.63
Sample4Geo† [71] [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	20.05	25.95	34.18	38.99	38.40	43.68	**67.80**	**72.41**	41.95	47.22
Safe-Net* [72] [TIP'24]	76.31	75.35	**73.53**	73.44	54.15	55.05	48.94	50.10	45.12	47.92	40.05	40.18	25.95	26.12	29.74	31.48	54.86	58.68	58.10	58.95	50.68	51.63
CCR† [73] [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	31.25	32.94	31.03	34.36	59.97	61.07	52.02	53.33	52.06	53.46
MuSe-Net* [29] [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	49.42	50.87	29.34	31.03	55.02	56.36	59.97	61.05	52.02	53.33
Ours	**76.72**	75.51	68.49	68.87	**71.77**	**71.20**	**59.95**	**60.62**	**58.24**	**58.83**	**64.36**	**66.27**	**58.49**	**58.89**	**40.42**	**55.75**	**61.57**	**71.70**	65.19	67.00	**62.52** (+10.46)	**63.26** (+9.80)
Satellite → Drone																						
Zheng *et al.*[1] [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	59.72	37.55	45.49	25.28	52.11	43.40	56.62	45.31	61.21	42.01
IBN-Net [70] [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	68.46	45.95	54.72	31.53	65.64	53.77	73.48	57.03	67.79	49.79
Sample4Geo† [71] [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	26.25	16.44	56.25	29.75	58.75	30.38	**83.75**	**69.66**	54.75	37.76
Safe-Net* [72] [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	41.12	32.31	53.88	27.01	54.19	57.82	79.36	57.09	53.27	46.48
CCR† [73] [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	25.89	26.94	26.01	30.40	58.01	59.13	83.09	61.05	52.17	48.55
MuSe-Net* [29] [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	69.38	47.79	53.75	27.94	70.00	52.67	76.25	60.74	69.43	51.14
Ours	**90.61**	**81.24**	**86.14**	**71.15**	**83.94**	**73.80**	**71.03**	**60.19**	**84.41**	**58.49**	**79.16**	**64.93**	**77.28**	**60.15**	**56.75**	**47.85**	**81.65**	**74.04**	80.30	69.38	**80.73** (+11.30)	**66.12** (+14.98)

Table 2: **Performance (R@1(%) and AP(%)) on SUES-200** for Drone → Satellite and Satellite → Drone tasks. In both tasks, drone-view images are stylized in 10 weather conditions, while satellite-view images remain constant. Best results are highlighted in bold. * denotes the use of official pretrained weights. † denotes the use of official pretrained weights on University-1652.

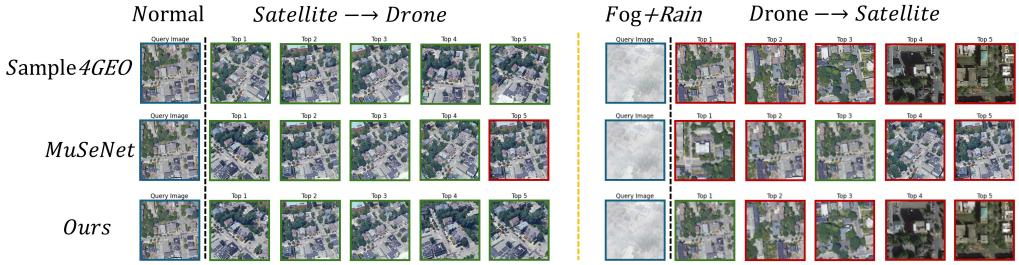

Figure 4: **Qualitative comparison under varying weather.** While existing methods perform reliably under clear weather, their accuracy drops markedly in adverse conditions. Our approach maintains superior localization performance, especially when drone images are severely affected by weather. Green boxes indicate correct matches, while images in red boxes represent incorrect matches.

attains a mean R@1 accuracy of 87.72% (+3.04%) and a mean AP of 76.39% (+10.64%). Our method validates competitive performance when handling drone view images under different conditions, especially in terms of cross-view geo-localization in challenging weather scenarios (*e.g.*, Dark, Fog+Snow, and Rain+Snow). In addition to quantitative comparisons, we present qualitative retrieval results in Figure 4. While existing methods can retrieve plausible matches in clear conditions, they suffer a substantial performance drop when encountering adverse weather conditions. In contrast, our approach consistently retrieves more accurate matches among the top-ranked candidates even in the presence of severe weather, highlighting its robustness and generalization capability. In the real-world captured SUES-200 dataset, we combine drone-view images captured at four altitudes (150m, 200m, 250m, and 300m) to simulate diverse operational heights in practical drone applications, enabling a comprehensive evaluation of our method. As shown in Tab 2, compared to the state-of-the-art multi-weather geolocalization method MuSe-Net, our method shows significant improvements in several metrics. Specifically, Ours improves mean R@1 accuracy in the Drone → Satellite task from 52.02% to 61.20% (+9.18%) and mean AP from 53.33% to 63.26% (+9.80%). In the Satellite → Drone task, our method increases mean R@1 accuracy from 69.43% to 80.73% (+11.30%) and mean AP from 51.14% to 66.12% (+14.98%).

4.2 Ablation Studies and Further Discussion

Impact of Weather-Driven Gating Mechanism. We conduct ablation experiments to quantify the impact of the weather-driven gating module on multimodal fusion. Table 3a summarizes three alternative fusion strategies under identical settings: (1) Concatenation, which concatenates visual and textual features directly; (2) Static Gate, which fuses modalities using fixed average weights; and (3) Dynamic Gate (Ours), which adaptively reweights visual channels based on weather semantics. We find that the dynamic gating mechanism improves mean AP by 2.3% over the no-gate baseline, with only 0.2M extra parameters and 0.4 ms added latency. This validates that weather-aware, channel-wise fusion enables more reliable feature integration, especially under adverse weather. The dynamic gate also reduces degradation from ambiguous weather descriptions, demonstrating its robustness for cross-weather geo-localization.

Impact of Text-Guided Weather Semantics. We conduct an ablation study to assess how multi-weather textual supervision affects cross-view geo-localization. Table 3b reports mean Recall@1

Method	D2S				S2D			
	Normal		Mean		Normal		Mean	
	R@1(%)	AP(%)	R@1(%)	AP(%)	R@1(%)	AP(%)	R@1(%)	AP(%)
+Concatenation	81.59	84.11	75.73	78.90	88.59	81.76	85.38	73.84
+Static Gate	82.35	84.75	75.26	78.45	88.45	**82.25**	85.89	74.61
+Dynamic Gate	**82.78**	**85.18**	**77.14**	**80.20**	**89.16**	81.80	**87.72**	**76.39**

(a) Impact of Weather-Driven Channel Gating on University-1652

CoT Step	D2S		S2D	
	Mean R@1(%)	Mean AP(%)	Mean R@1(%)	Mean AP(%)
NAN	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**

(b) Different CoT Setting on University-1652

	Dark+Rain+Fog				
		R@1(%)	R@5(%)	R@10(%)	AP(%)
D2S	Sample4Geo [71]	22.22	61.11	77.78	31.65
	MuSe-Net [29]	22.22	66.67	83.33	31.70
	Ours	**44.44**	**83.33**	**94.44**	**64.94**
S2D	Sample4Geo [71]	33.33	61.11	83.33	43.47
	MuSe-Net [29]	38.89	61.11	88.89	44.34
	Ours	**66.66**	**77.78**	**94.44**	**72.15**

(c) Evaluation on Real World Videos

Table 3: **Ablation Studies on University-1652 and Real-World Videos.** D2S denotes the Drone → Satellite and S2D denotes the Satellite → Drone.

and mean Average Precision (AP) across two tasks (Drone → Satellite and Satellite → Drone) on University-1652, under varying levels of Chain-of-Thought (CoT) guidance. The no-caption baseline (NAN) removes all text-based components, training the model with only the visual encoder and classification head. When progressively increasing the CoT prompt steps, we observe consistent performance gains for both tasks. Compared to the baseline, the best CoT setting (6-step) achieves improvements of +2.61% mean Recall@1 and +1.48% mean AP in the Drone → Satellite task, and +2.55% mean Recall@1 and +3.13% mean AP in the Satellite → Drone task. Notably, all models leveraging text-based weather guidance achieve higher mean Recall@1 and AP compared to the visual-only baseline, demonstrating that even minimal textual supervision provides substantial benefits for model robustness and accuracy.

Impact of Prompt Structuring on Multimodal Alignment. To comprehensively assess the impact of stepwise reasoning in prompt engineering on multimodal alignment, we generate captions using the same large vision–language model [17], varying only the number of reasoning steps prescribed in the prompt. The zero-step baseline produces a single-turn weather description without explicit intermediate reasoning. In contrast, the Chain-of-Thought (CoT) variants employ multi-stage reasoning prompts, including 2-step, 4-step, and 6-step configurations. The optimal 6-step CoT scheme decomposes the generation process into two reasoning chains: the first three steps sequentially estimate global visibility, local atmospheric cues, and fine-grained weather semantics; the latter three steps, conditioned on the inferred weather prior, successively infer macro scene layout, enumerate structural elements, and capture fine-grained spatial relationships. This structured decomposition compels the model to capture both meteorological variations and spatial semantics, yielding captions that are both semantically rich and consistent in format. As shown in Table 3b, model performance steadily improves as the number of reasoning steps increases, with the 6-step CoT achieving the best results in terms of both Recall@1 and mAP. These findings show that fine-grained, multi-stage reasoning significantly enhances cross-weather robustness and discriminative capacity, serving as a crucial ingredient for high-quality vision–language alignment in challenging multi-weather scenarios.

Real-World Performance under Adverse Weather Conditions. To validate real-world robustness, we collected 54 drone-satellite video pairs from YouTube to evaluate ours under three weather conditions, including Dark, Rain, and Fog. As shown in Tab 3c, our model consistently achieves superior results in terms of R@1, R@5, R@10, and Average Precision (AP). These results underscore the strong robustness and generalization capability of our model under challenging real-world scenarios with poor lighting or inclement weather conditions.

Limitations. Despite of significant advancements in cross-view geo-localization under diverse weather conditions, several limitations inherited from external components warrant discussion: (1) Dataset. The evaluation relies on existing datasets, which lack exhaustive geographic and weather diversity. Their limited scope in representing globally rare or region-specific weather phenomena may affect generalization to unseen environmental extremes. (2) Language Model Biases. The weather and spatial captions generated by off-the-shelf vision-language models inherit biases from

their pretraining corpora. Subtle inaccuracies in descriptor granularity, *e.g.*, "haze" vs. "fog", could propagate into the alignment process. Dataset and LVLM advances will mitigate these limitations.

5 Conclusion

In this paper, we propose a novel training-free, text-guided multi-modality alignment framework for robust cross-view geo-localization under complex and unseen weather conditions. By leveraging large vision–language models, we introduce a reasoning pipeline that automatically generates high-fidelity weather and spatial captions via chain-of-thought prompting, eliminating the need for costly manual descriptions or expert-controlled. Our multi-modal alignment model incorporates a dynamic channel-wise gating mechanism that adaptively fuses textual weather semantics with visual representations, achieving fine-grained disentanglement of scene and weather features. Extensive experiments on the University-1652 and SUES-200 benchmarks validate that our method consistently outperforms state-of-the-art approaches, particularly in challenging multi-weather scenarios. Our approach sets a new paradigm for leveraging language-driven priors in aerial geo-localization and offers a scalable path toward real-world deployment under diverse environmental conditions.

6 Acknowledgement

This work is supported by the Shanghai Committee of Science and Technology, China (Grant No.23ZR1423500), the National Natural Science Foundation of China under Grant No.62302287, University of Macau MYRG-GRG2024-00077-FST-UMDF and SRG2024-00002-FST, and the Science and Technology Development Fund (FDCT) 0043/2025/RIA1.

References

[1] Zhedong Zheng, Yunchao Wei, and Yi Yang. "University-1652: A Multi-view Multi-source Benchmark for Drone-based Geo-localization". In: *Proceedings of the 28th ACM International Conference on Multimedia (ACM MM)*. 2020, pp. 1395–1403.

[2] Tingyu Wang, Zhedong Zheng, Chenggang Yan, Jiyong Zhang, Yaoqi Sun, Bolun Zheng, and Yi Yang. "Each Part Matters: Local Patterns Facilitate Cross-view Geo-localization". In: *IEEE Transactions on Circuits and Systems for Video Technology* (Feb. 2022), pp. 867–879. DOI: 10.1109/TCSVT.2021.3061265.

[3] Jinliang Lin, Zhedong Zheng, Zhun Zhong, Zhiming Luo, Shaozi Li, Yi Yang, and Nicu Sebe. "Joint Representation Learning and Keypoint Detection for Cross-View Geo-Localization". In: *IEEE Transactions on Image Processing* (2022), pp. 3780–3792. DOI: 10.1109/TIP.2022.3175601.

[4] Zelong Zeng, Zheng Wang, Fan Yang, and Shin'ichi Satoh. "Geo-Localization via Ground-to-Satellite Cross-View Image Retrieval". In: *IEEE Transactions on Multimedia* 25 (2023), pp. 2176–2188. DOI: 10.1109/TMM.2022.3144066.

[5] Qiuming Zhu, Yi Zhao, Yang Huang, Zhipeng Lin, Lu HanJie Wang, Yunpeng Bai, Tianxu Lan, Fuhui Zhou, and Qihui Wu. "Demo Abstract: An UAV-based 3D Spectrum Real-time Mapping System". In: *IEEE INFOCOM 2022 – IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. 2022, pp. 1–2.

[6] Jie Wang, Qiuming Zhu, Zhipeng Lin, Junting Chen, Guoru Ding, Qihui Wu, Guochen Gu, and Qianhao Gao. "Sparse Bayesian Learning-based Hierarchical Construction for 3D Radio Environment Maps Incorporating Channel Shadowing". In: *IEEE Transactions on Wireless Communications* (2024).

[7] Tongtong Feng, Qing Li, Xin Wang, Mingzi Wang, Guangyao Li, and Wenwu Zhu. "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*. 2024, pp. 35–39.

[8] Vishwanath A. Sindagi, Poojan Oza, Rajeev Yasarla, and Vishal M. Patel. "Prior-based Domain Adaptive Object Detection for Hazy and Rainy Conditions". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2020.

[9] Wenyu Liu, Gaofeng Ren, Runsheng Yu, Shi Guo, Jianke Zhu, and Lei Zhang. "Image-Adaptive YOLO for Object Detection in Adverse Weather Conditions". In: *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*. Vol. 36. 2. 2022, pp. 1792–1800.

[10] Shuyu Yang, Yinan Zhou, Zhedong Zheng, Yaxiong Wang, Li Zhu, and Yujiao Wu. "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)*. 2023, pp. 4492–4501.

[11] Cairong Yan, Meng Ma, Yanting Zhang, and Yongquan Wan. "Dual-Path Multimodal Optimal Transport for Composed Image Retrieval". In: *Proceedings of the Asian Conference on Computer Vision (ACCV)*. 2024, pp. 1741–1755.

[12] Shuanglin Yan, Neng Dong, Liyan Zhang, and Jinhui Tang. "CLIP-Driven Fine-Grained Text-Image Person Re-identification". In: *IEEE Transactions on Image Processing* 32 (2023), pp. 6032–6046.

[13] Edward Vendrow, Omiros Pantazis, Alexander Shepard, Gabriel Brostow, Kate Jones, Oisin Mac Aodha, Sara Beery, and Grant Van Horn. "INQUIRE: A Natural World Text-to-Image Retrieval Benchmark". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 37. 2024, pp. 126500–126514.

[14] Juncheng Li, Xin He, Longhui Wei, Long Qian, Linchao Zhu, Lingxi Xie, Yueting Zhuang, Qi Tian, and Siliang Tang. "Fine-grained Semantically Aligned Vision-Language Pre-training". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 35. 2022, pp. 7290–7303.

[15] Wei Chow, Juncheng Li, Qifan Yu, Kaihang Pan, Hao Fei, Zhiqi Ge, Shuai Yang, Siliang Tang, Hanwang Zhang, and Qianru Sun. "Unified Generative and Discriminative Training for Multi-modal Large Language Models". In: *arXiv preprint* (2024). arXiv: 2411.00304 [cs.CV].

[16] Yan Zeng, Xinsong Zhang, and Hang Li. "Multi-Grained Vision Language Pre-Training: Aligning Texts with Visual Concepts". In: *arXiv preprint* (2021). arXiv: 2111.08276 [cs.CV].

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

[18] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V. Le, Denny Zhou, et al. "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 35. 2022, pp. 24824–24837.

[19] Gabriele Berton, Riccardo Mereu, Gabriele Trivigno, Carlo Masone, Gabriela Csurka, Torsten Sattler, and Barbara Caputo. "Deep Visual Geo-localization Benchmark". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 5396–5407.

[20] Hyo Jin Kim, Enrique Dunn, and Jan-Michael Frahm. "Learned Contextual Feature Reweighting for Image Geo-localization". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 2136–2145.

[21] Meng Chu, Zhedong Zheng, Wei Ji, Tingyu Wang, and Tat-Seng Chua. "Towards Natural Language-Guided Drones: GeoText-1652 Benchmark with Spatial Relation Matching". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2024, pp. 213–231.

[22] Javier Cruz-Mota, Iva Bogdanova, Benoıt Paquier, Michel Bierlaire, and Jean-Philippe Thiran. "Scale Invariant Feature Transform on the Sphere: Theory and Applications". In: *International Journal of Computer Vision* 98 (2012), pp. 217–241.

[23] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. "SURF: Speeded Up Robust Features". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2006.

[24] Jian Zhang, Yunyin Cao, and Qun Wu. "Vector of Locally and Adaptively Aggregated Descriptors for Image Feature Representation". In: *Pattern Recognition* 116 (2021), p. 107952.

[25] Jorge Sánchez, Florent Perronnin, Thomas Mensink, and Jakob Verbeek. "Image Classification with the Fisher Vector: Theory and Practice". In: *International Journal of Computer Vision* 105 (2013), pp. 222–245.

[26] Daniel Barath, Luca Cavalli, and Marc Pollefeys. "Learning to Find Good Models in RANSAC". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 15744–15753.

[27] Francesco Castaldo, Amir Zamir, Roland Angst, Francesco Palmieri, and Silvio Savarese. "Semantic Cross-View Matching". In: *2015 IEEE International Conference on Computer Vision Workshops (ICCVW)*. 2015. DOI: 10.1109/ICCVW.2015.137.

[28] Tsung-Yi Lin, Serge Belongie, and James Hays. "Cross-View Image Geolocalization". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2013. DOI: 10.1109/CVPR.2013.120.

[29] Tingyu Wang, Zhedong Zheng, Yaoqi Sun, Chenggang Yan, Yi Yang, and Tat-Seng Chua. "Multiple-environment Self-adaptive Network for Aerial-view Geo-localization". In: *Pattern Recognition* 152 (2024), p. 110363. DOI: 10.1016/j.patcog.2024.110363.

[30] Scott Workman, Richard Souvenir, and Nathan Jacobs. "Wide-Area Image Geolocalization with Aerial Reference Imagery". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2015.

[31] Tsung-Yi Lin, Yin Cui, Serge Belongie, and James Hays. "Learning Deep Representations for Ground-to-Aerial Geolocalization". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015.

[32] Arvind M. Vepa, Zukang Yang, Andrew Choi, Jungseock Joo, Fabien Scalzo, and Yizhou Sun. "Integrating Deep Metric Learning with Coreset for Active Learning in 3D Segmentation". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 37. 2024, pp. 71643–71671.

[33] Wonhyeok Choi, Mingyu Shin, and Sunghoon Im. "Depth-Discriminative Metric Learning for Monocular 3D Object Detection". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 36. 2023, pp. 80165–80177.

[34] Zhedong Zheng, Liang Zheng, and Yi Yang. "A Discriminatively Learned CNN Embedding for Person Re-identification". In: *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 14.1 (2017), pp. 1–20. DOI: 10.1145/3159171.

[35] Hongji Yang, Xiufan Lu, and Yingying Zhu. "Cross-View Geo-localization with Layer-to-Layer Transformer". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 34. 2021, pp. 29009–29020.

[36] Royston Rodrigues and Masahiro Tani. "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)*. 2022, pp. 3871–3879.

[37] Ming Dai, Jianhong Hu, Jiedong Zhuang, and Enhui Zheng. "A Transformer-based Feature Segmentation and Region Alignment Method for UAV-view Geo-localization". In: *IEEE Transactions on Circuits and Systems for Video Technology* 32.7 (2021), pp. 4376–4389.

[38] Hao Ju, Shaofei Huang, Si Liu, and Zhedong Zheng. "Video2BEV: Transforming Drone Videos to BEVs for Video-based Geo-localization". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2024.

[39] Yanjie Tan, Yifu Zhu, Zhaoyang Huang, Huailiang Tan, and Keqin Li. "MAPD: An FPGA-based Real-time Video Haze Removal Accelerator Using Mixed Atmosphere Prior". In: *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 42.12 (2023), pp. 4777–4790.

[40] Xianzheng Ma, Zhixiang Wang, Yacheng Zhan, Yinqiang Zheng, Zheng Wang, Dengxin Dai, and Chia-Wen Lin. "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)*. 2022, pp. 18922–18931.

[41] Xiaohan Zhang, Xingyu Li, Waqas Sultani, Yi Zhou, and Safwan Wshah. "Cross-View Geo-localization via Learning Disentangled Geometric Layout Correspondence". In: *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*. Vol. 37. 3. 2023, pp. 3480–3488.

[42] Sijie Zhu, Mubarak Shah, and Chen Chen. "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)*. 2022, pp. 1162–1171.

[43] Zefan Qu, Ke Xu, Gerhard Petrus Hancke, and Rynson W. H. Lau. "LuSh-NeRF: Lighting Up and Sharpening NeRFs for Low-light Scenes". In: *arXiv preprint* (2024). arXiv: 2411.06757 [cs.CV].

[44] Haoyu Chen, Wenbo Li, Jinjin Gu, Jingjing Ren, Sixiang Chen, Tian Ye, Renjing Pei, Kaiwen Zhou, Fenglong Song, and Lei Zhu. "RestoreAgent: Autonomous Image Restoration Agent via Multimodal Large Language Models". In: *arXiv preprint* (2024). arXiv: 2407.18035 [cs.CV].

[45] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. "Learning Transferable Visual Models from Natural Language Supervision". In: *Proceedings of the 38th International Conference on Machine Learning (ICML)*. Vol. 139. Proceedings of Machine Learning Research. PMLR, 2021, pp. 8748–8763.

[46] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation". In: *Proceedings of the 39th International Conference on Machine Learning (ICML)*. Vol. 162. Proceedings of Machine Learning Research. PMLR, 2022, pp. 12888–12900.

[47] Junjie Zhu, Yiying Li, Ke Yang, Naiyang Guan, Zunlin Fan, Chunping Qiu, and Xiaodong Yi. "MVP: Meta Visual Prompt Tuning for Few-shot Remote Sensing Image Scene Classification". In: *IEEE Transactions on Geoscience and Remote Sensing* 62 (2024), pp. 1–13.

[48] Keyan Chen, Chenyang Liu, Hao Chen, Haotian Zhang, Wenyuan Li, Zhengxia Zou, and Zhenwei Shi. "RSPrompter: Learning to Prompt for Remote Sensing Instance Segmentation Based on Visual Foundation Model". In: *IEEE Transactions on Geoscience and Remote Sensing* 62 (2024), pp. 1–17.

[49] Yuan Yuan, Yang Zhan, and Zhitong Xiong. "Parameter-efficient Transfer Learning for Remote Sensing Image–Text Retrieval". In: *IEEE Transactions on Geoscience and Remote Sensing* 61 (2023), pp. 1–14.

[50] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, Mingliang Xu, and Yi-Dong Shen. "Dual-Path Convolutional Image-Text Embeddings with Instance Loss". In: *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 16.2 (2020), pp. 1–23.

[51] Zihao Wang, Xihui Liu, Hongsheng Li, Lu Sheng, Junjie Yan, Xiaogang Wang, and Jing Shao. "CAMP: Cross-Modal Adaptive Message Passing for Text-Image Retrieval". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019, pp. 5764–5773.

[52] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. "UNITER: Universal Image-Text Representation Learning". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2020, pp. 104–120.

[53] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. "OSCAR: Object-Semantics Aligned Pre-training for Vision-Language Tasks". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2020, pp. 121–137.

[54] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. "Language Models are Few-Shot Learners". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 33. 2020, pp. 1877–1901.

[55] Josh Achiam, Steven Adler, Sandhini Agarwal, et al. "GPT-4 Technical Report". In: *arXiv preprint* (2023). arXiv: 2303.08774 [cs.CL].

[56] Jun Chen, Han Guo, Kai Yi, Boyang Li, and Mohamed Elhoseiny. "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)*. 2022, pp. 18030–18040.

[57] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. "Multimodal Chain-of-Thought Reasoning in Language Models". In: *arXiv preprint* (2023). arXiv: 2302.00923 [cs.CL].

[58] Gaurang Sriramanan, Siddhant Bharti, Vinu Sankar Sadasivan, Shoumik Saha, Priyatham Kattakinda, and Soheil Feizi. "LLM-Check: Investigating Detection of Hallucinations in Large Language Models". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 37. 2024, pp. 34188–34216.

[59] Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. "Hallucination Augmented Contrastive Learning for Multimodal Large Language Model". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024, pp. 27036–27046.

[60] Minchan Kim, Minyeong Kim, Junik Bae, Suhwan Choi, Sungkyung Kim, and Buru Chang. "Exploiting Semantic Reconstruction to Mitigate Hallucinations in Vision-Language Models". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2024, pp. 236–252.

[61] Alexander B. Jung, Kentaro Wada, Jon Crall, Satoshi Tanaka, Jake Graving, Christoph Reinders, Sarthak Yadav, Joy Banerjee, Gábor Vecsei, Adam Kraft, Zheng Rui, Jirka Borovec, Christian Vallentin, Semen Zhydenko, Kilian Pfeiffer, Ben Cook, Ismael Fernández, François-Michel De Rainville, Chi-Hung Weng, Abner Ayala-Acevedo, Raphael Meudec, Matias Laporte, et al. *imgaug*. `https://github.com/aleju/imgaug`. Accessed: 2020-02-01. 2020.

[62] Junyan Ye, Honglin Lin, Leyan Ou, Dairong Chen, Zihao Wang, Qi Zhu, Conghui He, and Weijia Li. "Where am I? Cross-View Geo-localization with Natural Language Descriptions". In: *arXiv preprint* (2024). arXiv: `2412.17007 [cs.CV]`.

[63] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*. 2019, pp. 4171–4186.

[64] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021, pp. 10012–10022.

[65] Herbert Robbins and Sutton Monro. "A Stochastic Approximation Method". In: *The Annals of Mathematical Statistics* (1951), pp. 400–407.

[66] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 32. 2019.

[67] Runzhe Zhu, Ling Yin, Mingze Yang, Fei Wu, Yuncheng Yang, and Wenbo Hu. "SUES-200: A Multi-height Multi-scene Cross-view Image Benchmark Across Drone and Satellite". In: *IEEE Transactions on Circuits and Systems for Video Technology* 33.9 (2023), pp. 4825–4839.

[68] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep Residual Learning for Image Recognition". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016. DOI: `10.1109/CVPR.2016.90`.

[69] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. "Densely Connected Convolutional Networks". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 4700–4808.

[70] Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. "Two at Once: Enhancing Learning and Generalization Capacities via IBN-Net". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 484–500.

[71] Fabian Deuser, Konrad Habel, and Norbert Oswald. "Sample4Geo: Hard Negative Sampling for Cross-View Geo-Localization". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2023, pp. 16847–16856.

[72] Jinliang Lin, Zhiming Luo, Dazhen Lin, Shaozi Li, and Zhun Zhong. "A Self-Adaptive Feature Extraction Method for Aerial-View Geo-Localization". In: *IEEE Transactions on Image Processing* (2024).

[73] Haolin Du, Jingfei He, and Yuanqing Zhao. "CCR: A Counterfactual Causal Reasoning-Based Method for Cross-View Geo-Localization". In: *IEEE Transactions on Circuits and Systems for Video Technology* (2024).

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims in the abstract and introduction accurately reflect the paper's core contributions and scope. The paper proposes a training-free weather reasoning approach for multi-weather geo-localization, introduces a text-guided feature alignment framework with dynamic gating, and yields consistent improvements on challenging benchmarks. All major claims are substantiated by method descriptions and experimental results throughout the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations of the work are discussed in Section 4.2.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (*e.g.*, independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, *e.g.*, if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [No]

Justification: The paper does not present formal theorems or proofs.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper provides all necessary details for experimental reproducibility. All modules, loss functions, and augmentation strategies are specified in the main text. To further support reproducibility, we will release our code and configuration files upon publication.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (*e.g.*, in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (*e.g.*, a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (*e.g.*, with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (*e.g.*, to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: As stated in the response above, we provided detailed instructions on how to replicate our experiment results in the main paper and further in the supplementary material. We will release our code and models upon paper acceptance.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (*e.g.*, for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (*e.g.*, data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provided detailed instructions on replicating the training and evaluation procedures in Section 4. We did not perform delicate tuning for the hyperparameters.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Error bars or statistical significance metrics are not reported, as our experimental protocol follows standard practice in the cross-view geo-localization field, where single-run, fixed split results are widely adopted for benchmarking. For direct comparability, we report results under the same evaluation protocol as prior work. We note that nearly all published baselines on these benchmarks also omit error bars.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (*e.g.*, Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (*e.g.*negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We described our compute setup as well as the training time and inference runtime in Section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (*e.g.*, preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer:[Yes]

   Justification: : Authors carefully read the NeurIPS Code of Ethics and preserved anonymity.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (*e.g.*, if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

Justification: Our paper primarily discusses the positive application prospects of the proposed method, such as urban management and disaster response, in the introduction and application sections. However, we do not specifically address potential negative societal impacts. While the method offers promising practical value, we acknowledge that issues such as privacy protection and ethical risks should be considered during real-world deployment. Future work may further analyze these aspects.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (*e.g.*, disinformation, generating fake profiles, surveillance), fairness considerations (*e.g.*, deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (*e.g.*, gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (*e.g.*, pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: Our study does not involve the release of any data or models with a high risk of misuse. All data used in this work are either standard public benchmarks or synthetically generated for research purposes, with no content that requires special safeguards. Additionally, all real-world validation images are collected from publicly available and safe content on Youtube, ensuring no sensitive or inappropriate material is included.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (*e.g.*, code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All assets used in this work, including code, datasets, and models, are properly credited in the text, with full compliance to their respective licenses and terms of use.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (*e.g.*, CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (*e.g.*, website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: The paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: We explicitly describe the use of a large language model (Qwen2.5-VL-32B) as a core component of our methodology. The LLM is leveraged to generate multi-step, chain-of-thought weather and spatial descriptions, which serve as supervisory signals for vision-language alignment in our framework. Detailed descriptions of the LLM's role, prompt design, and integration into the description pipeline are provided in Section 3 of the paper.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.