

GRASP: Granularity-Aware Region Alignment and Semantic Prototype Learning for Fine-Grained Cross-Modal Understanding in Drone Views

Jiahui Cui*

Yan Zhao*

Kan Wei*

cuijiahui24@mails.ucas.ac.cn

zhaoyan242@mails.ucas.ac.cn

weikan24@mails.ucas.ac.cn

Aerospace Information Research
Institute, Chinese Academy of
Sciences

Beijing, China

Key Laboratory of Target Cognition
and Application Technology (TCAT)
Beijing, China

University of Chinese Academy of
Sciences

Beijing, China

School of Electronic, Electrical and
Communication Engineering,

University of Chinese Academy of
Sciences

Beijing, China

Enze Zhu

zhuenze23@mails.ucas.ac.cn

Aerospace Information Research
Institute, Chinese Academy of
Sciences

Beijing, China

Key Laboratory of Target Cognition
and Application Technology (TCAT)
Beijing, China

University of Chinese Academy of
Sciences

Beijing, China

School of Electronic, Electrical and
Communication Engineering,

University of Chinese Academy of
Sciences

Beijing, China

Peirong Zhang

zhangpeirong22@mails.ucas.ac.cn

Aerospace Information Research
Institute, Chinese Academy of
Sciences

Beijing, China

Key Laboratory of Target Cognition
and Application Technology (TCAT)
Beijing, China

University of Chinese Academy of
Sciences

Beijing, China

School of Electronic, Electrical and
Communication Engineering,

University of Chinese Academy of
Sciences

Beijing, China

Lei Wang

wanglei002931@aircas.ac.cn

Aerospace Information Research
Institute, Chinese Academy of
Sciences

Beijing, China

Key Laboratory of Target Cognition
and Application Technology (TCAT)
Beijing, China

Yiru Wang[✉]

wangyiru@aircas.ac.cn

Aerospace Information Research
Institute, Chinese Academy of
Sciences

Beijing, China

Key Laboratory of Target Cognition
and Application Technology (TCAT)
Beijing, China

Abstract

Fine-grained cross-modal understanding in drone views is essential for aerial vision-language navigation. However, the inherent wide field of view and overhead perspective of drone scenarios impose dual challenges on vision-language understanding. At the macro level, overwhelming background clutter in visual representations leads to *Cross-Modal Focus Misalignment*, where the model prioritizes global environmental similarities over specific object details.

At the micro level, *Visual Isomorphism* creates ambiguity, where candidates share similar geometric structures yet differ only in subtle attributes. To address these challenges, we propose the Granularity-Aware Region Alignment and Semantic Prototype (GRASP) learning framework, enhancing discriminative capability through two synergistic strategies. Specifically, we introduce Region-Focused Alignment (RFA) to promote object-centric cross-modal alignment while suppressing background interference. Concurrently, to tackle visual isomorphism, we propose Semantic Perturbation Enhanced Matching (SPEM), which leverages a foreground-purified Semantic Prototype Codebook (SPC) to construct semantically perturbed negatives for fine-grained semantic discrimination. Extensive experiments on the GeoText-1652 benchmark and the unseen ERA dataset demonstrate that GRASP achieves competitive performance in drone-view fine-grained image-text retrieval, validating its effectiveness for

*These authors contributed equally to this research.



This work is licensed under a Creative Commons Attribution 4.0 International License.
MM '26, Rio de Janeiro, Brazil

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2213-4/2026/11

<https://doi.org/10.1145/3767308.3835872>

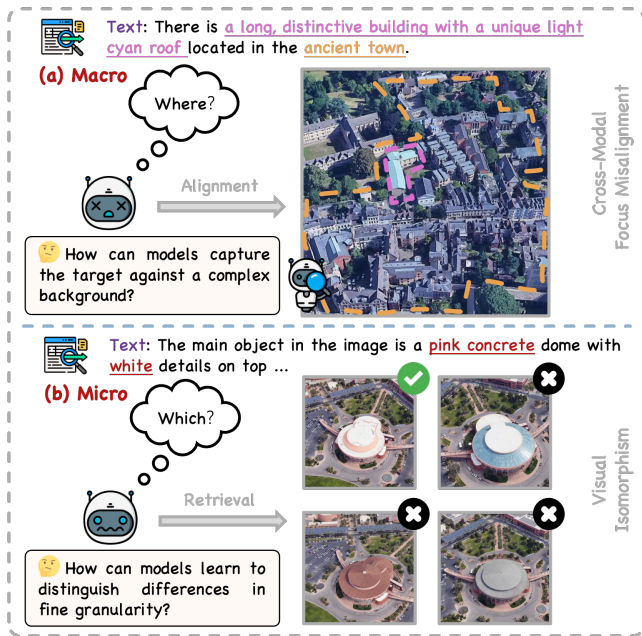


Figure 1: Dual challenges in drone-view understanding: (a) background-dominated alignment causes Cross-Modal Focus Misalignment; (b) geometrically similar objects with fine-grained attribute differences cause Visual Isomorphism.

cross-modal understanding in aerial scenarios. Our code implementation is available at <https://github.com/UCAS-JC/GRASP>.

CCS Concepts

• Computing methodologies → Visual content-based indexing and retrieval; • Information systems → Image search.

Keywords

Natural Language-Guided Drones, Cross-Modal Retrieval, Fine-Grained Understanding

ACM Reference Format:

Jiahui Cui, Yan Zhao, Kan Wei, Enze Zhu, Peirong Zhang, Lei Wang, and Yiru Wang. 2026. GRASP: Granularity-Aware Region Alignment and Semantic Prototype Learning for Fine-Grained Cross-Modal Understanding in Drone Views. In *Proceedings of the 34th ACM International Conference on Multimedia (MM '26)*, November 10–14, 2026, Rio de Janeiro, Brazil. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3767308.3835872>

1 Introduction

Unmanned Aerial Vehicles (UAVs) have become essential platforms across diverse real-world applications. In the realm of sustainable agriculture, UAVs facilitate precision crop monitoring and targeted intervention [10, 16, 29]. In public safety contexts, they are indispensable for disaster response and search-and-rescue missions [21, 22, 27]. Similarly, within smart city frameworks, UAVs provide scalable solutions for traffic surveillance and facility inspection [33]. Propelled by recent strides in vision-language models, Natural Language-Guided Drones (NLGD) have emerged as a flexible paradigm for human-machine interaction [4, 5, 14]. Central to this

paradigm is the ability to interpret the semantic correspondence between natural language and aerial imagery, which necessitates deep cross-modal understanding within drone-view contexts.

Existing approaches use pre-trained vision-language models as feature extractors and align modalities through contrastive learning and image-text matching [18, 19, 26]. GeoText-1652 [8] shifts drone-view retrieval from coarse alignment toward discriminative and object-centric reasoning. However, achieving fine-grained cross-modal understanding in drone views poses distinctive challenges at both macro and micro levels.

At the macro level, there exists a significant *Cross-Modal Focus Misalignment*. Unlike ground-level photography where subjects are naturally framed, drone cameras capture expansive areas where targets are dwarfed by complex, noisy environments. Consequently, while textual descriptions provide elaborate attributes of specific targets, these objects visually occupy only a minute fraction of the wide field-of-view aerial imagery. As illustrated in Figure 1 (a), this semantic-visual disparity causes existing global image-text alignment methods to be inevitably overwhelmed by background noise, failing to ground the specific targets defined in the text [8].

At the micro level, *Visual Isomorphism* is uniquely pronounced in aerial scenes due to the top-down perspective. Because drones observe the world from above, critical vertical geometric cues—such as building facades, doors, and relative object heights that standard models rely heavily upon—are severely compressed or obscured. As a result, distinct objects often project deceptively similar structural layouts (Figure 1 (b)). This heightened visual ambiguity forces the matching process to rely heavily on fine-grained discriminative attributes, including materials and colors. Unfortunately, standard contrastive and matching frameworks adopt passive in-batch negative sampling, which rarely encounters such visually isomorphic negative pairs in a natural batch, thus limiting the model's ability to learn these critical fine-grained differences [8, 37].

To address these challenges, we propose the Granularity-Aware Region Alignment and Semantic Prototype (GRASP) learning framework. GRASP enhances fine-grained cross-modal understanding through two synergistic strategies carefully tailored for drone views: Region-Focused Alignment (RFA) and Semantic Perturbation Enhanced Matching (SPEM). Specifically, to overcome the overwhelming background clutter, the RFA mechanism utilizes region-level image-text contrastive learning and a cross-modal ranking loss to explicitly enhance local, object-centric understanding while suppressing visual background interference. Concurrently, to tackle the severe visual homogenization caused by the overhead view, we propose SPEM. As illustrated in Figure 2, unlike standard frameworks that rely on passive easy negatives, SPEM actively constructs semantically perturbed hard negatives via Textual Semantic Perturbation (TSP) and Visual Semantic Perturbation (VSP). In detail, TSP deterministically substitutes fine-grained attribute words to generate hard textual negatives while preserving the exact sentence context. Meanwhile, VSP establishes a foreground-purified Semantic Prototype Codebook (SPC) and leverages it to synthesize visual negatives that are strictly bounded within plausible attribute distributions, effectively averting out-of-distribution noise. TSP together with VSP explicitly forces the model to focus on the subtle attribute differences that distinguish visually isomorphic targets. Notably, RFA and SPEM are employed exclusively during the training phase.

By bolstering the model’s cross-modal understanding capabilities, they achieve significant performance gains with absolute zero additional computational overhead during inference.

Our main contributions are summarized as follows:

- A Granularity-Aware Region Alignment and Semantic Prototype (GRASP) learning framework is presented, which systematically addresses the distinct cross-modal focus misalignment and visual isomorphism challenges in overhead drone-view understanding, introducing zero inference cost.
- A Region-Focused Alignment (RFA) mechanism is designed, which introduces region-level image-text contrastive learning and a cross-modal ranking loss to recalibrate alignment across both full-image and object-centric levels, mitigating background distraction.
- A Semantic Perturbation Enhanced Matching (SPEM) strategy is proposed, which explicitly constructs counterfactual hard negatives via controlled fine-grained attribute word replacement and visual prototype feature perturbation. This effectively forces the model to learn fine-grained semantic discrimination under severe visual isomorphism.
- Experiments on the GeoText-1652 benchmark and the unseen ERA dataset demonstrate that GRASP achieves competitive performance in drone-view fine-grained image-text retrieval, validating its robustness and generalizability for cross-modal understanding in aerial scenarios.

2 Related Work

2.1 Vision-Language Cross-Modal Alignment

Vision Language Models (VLMs) learn joint representations of images and text through large-scale pre-training. Early works like CLIP [26] employ contrastive learning for global image-text alignment. Subsequent dual-stream architectures improve upon this foundation: ALBEF [19] introduces hard negative mining with Image-Text Contrastive (ITC) loss, UNITER [7] proposes word-region alignment, and XVLM [37] incorporates multi-level alignment tasks. METER [9] systematically evaluates encoder and fusion strategies, while BLIP [17, 18] unifies understanding and generation via lightweight modules. However, global alignment often fails to capture fine-grained semantic distinctions. To address this, hierarchical approaches encode semantic relationships using partial order constraints [23, 30] or hyperbolic spaces [11, 24]. Region-based methods align text with segmented image regions [2, 38] or dynamically reweight hierarchical features [34]. Recent work [32] explores constructing negative samples to foster fine-grained alignment. However, this approach faces two limitations in aerial scenarios. First, it is easily overwhelmed by severe background noise, failing to produce targeted perturbations on the specific object. Second, its token-level manipulation disrupts region-level integrity, as it fragments targets that naturally span multiple tokens. Consequently, achieving fine-grained cross-modal understanding in drone views remains a significant challenge.

2.2 Cross-Modal Localization and Navigation

Cross-view localization [3, 39] addresses viewpoint-invariant visual matching, with transformer-based methods [6] attending to

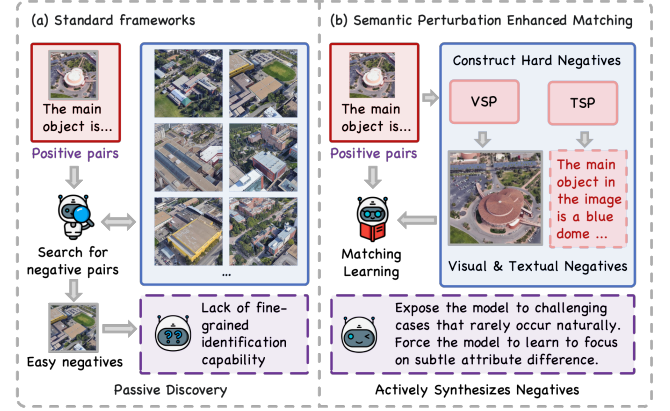


Figure 2: Standard passive in-batch sampling mainly yields easy negatives (a), whereas SPEM actively synthesizes hard textual and visual negatives via TSP and VSP (b) to improve attribute discrimination.

salient regions and partitioning strategies [28, 31] enriching representations. However, these purely visual approaches lack language grounding capabilities. Vision Language Navigation (VLN) [1] extends cross-modal understanding to embodied agents, requiring them to ground natural language instructions in visual observations for sequential decision-making. Indoor VLN methods achieve this through cross-modal attention [20, 25], episodic memory [13], and structured spatial representations [12]. Recent efforts integrate natural language for aerial cross-modal understanding. CrossText2Loc [36] and UniLoc [35] enable text-based satellite retrieval. GeoText-1652 [8] establishes a drone-view benchmark with spatial matching, while VCSR [15] employs textual cues to guide semantic reasoning within the drone-view visual context.

3 Methodology

3.1 Vision-Language Encoding

As illustrated in Figure 3(a), we first perform vision-language encoding to obtain visual and textual embeddings that serve as the input representations for subsequent RFA and SPEM. Specifically, the Vision Encoder employs a Swin Transformer to encode an input image I into a sequence of patch embeddings

$$\mathbf{V} = \{\mathbf{v}_{\text{cls}}, \mathbf{v}_1, \dots, \mathbf{v}_N\} \in \mathbb{R}^{(N+1) \times d_v},$$

where N denotes the number of patches and $d_v = 1024$. In parallel, the Text Encoder adopts BERT to encode an input text T into a sequence of token embeddings $\mathbf{W} = \{\mathbf{w}_{\text{cls}}, \mathbf{w}_1, \dots, \mathbf{w}_L\} \in \mathbb{R}^{(L+1) \times d_t}$, where $d_t = 768$.

For global contrastive learning in RFA, the global features $\mathbf{v}_{\text{cls}}$ and $\mathbf{w}_{\text{cls}}$ are projected to a joint embedding space $\mathbb{R}^{256}$ via linear projections, followed by L2 normalization for cosine similarity computation. To support fine-grained matching in SPEM, we partition the BERT encoder: the early layers are utilized for unimodal textual encoding, while the last six layers are dedicated to deep cross-modal interaction via cross-attention to capture complex visual-linguistic alignments. Structurally, RFA operates on the embeddings produced

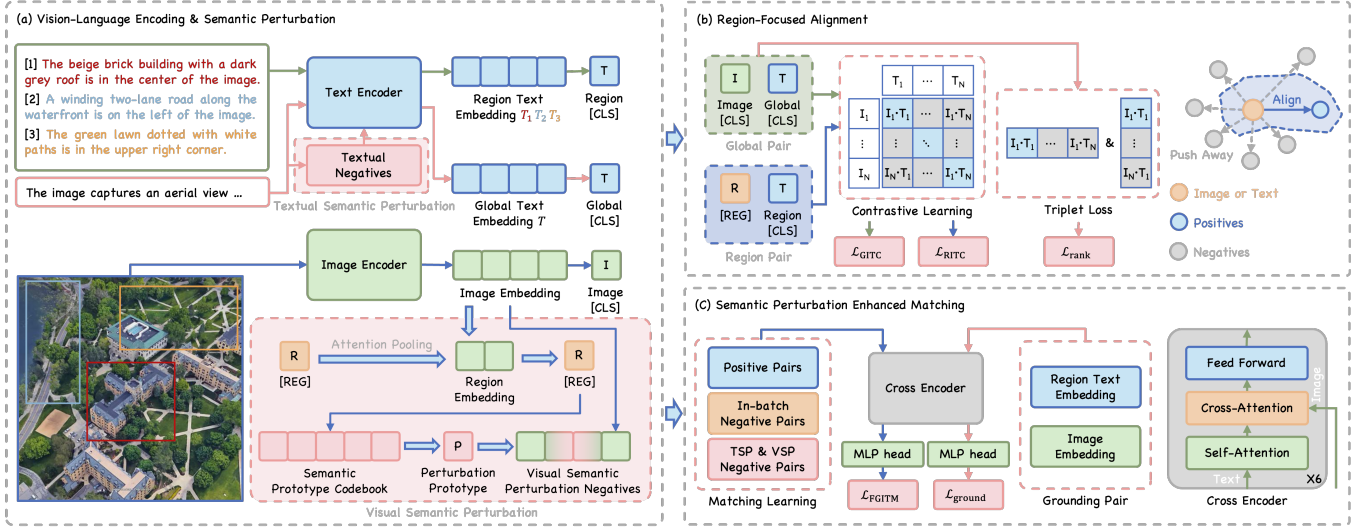


Figure 3: Overview of GRASP. (a) Encoders produce image and text embeddings, while TSP and SPC-based VSP synthesize hard negatives. (b) RFA combines global and region-level contrastive losses with ranking loss. (c) SPEM trains fine-grained matching with in-batch and synthesized negatives, together with grounding supervision.

by the vision and text encoders, while SPEM further relies on the fused cross-modal representations.

3.2 Region-Focused Alignment

The RFA module takes the visual and textual embeddings from Section 3.1 as input and establishes multi-granularity cross-modal correspondences by performing global/region image-text contrastive alignment and margin-based ranking against the hardest in-batch negatives, as depicted in Figure 3(b).

Global Representation Learning. Given a batch of B image-text pairs $\{(I_i, T_i)\}_{i=1}^B$, we first project the global features $\mathbf{v}_{\text{cls}}$ and $\mathbf{w}_{\text{cls}}$ into a shared embedding space and apply L2 normalization:

$$\hat{\mathbf{v}}_{\text{cls}} = \frac{\phi_v(\mathbf{v}_{\text{cls}})}{\|\phi_v(\mathbf{v}_{\text{cls}})\|_2}, \quad \hat{\mathbf{w}}_{\text{cls}} = \frac{\phi_t(\mathbf{w}_{\text{cls}})}{\|\phi_t(\mathbf{w}_{\text{cls}})\|_2}, \quad (1)$$

where $\phi_v : \mathbb{R}^{d_v} \rightarrow \mathbb{R}^{256}$ and $\phi_t : \mathbb{R}^{d_t} \rightarrow \mathbb{R}^{256}$ are learnable linear projections. The global similarity matrix $\mathbf{S} \in \mathbb{R}^{B \times B}$ is computed as:

$$s_{ij} = \cos(\hat{\mathbf{v}}_{\text{cls}}^{(i)}, \hat{\mathbf{w}}_{\text{cls}}^{(j)}) = \hat{\mathbf{v}}_{\text{cls}}^{(i)\top} \hat{\mathbf{w}}_{\text{cls}}^{(j)}. \quad (2)$$

We optimize global alignment via a symmetric InfoNCE objective:

$$\mathcal{L}_{\text{GITC}} = -\frac{1}{2B} \sum_{i=1}^B \left[\log \frac{\exp(s_{ii}/\tau)}{\sum_{j=1}^B \exp(s_{ij}/\tau)} + \log \frac{\exp(s_{ii}/\tau)}{\sum_{j=1}^B \exp(s_{ji}/\tau)} \right] \quad (3)$$

where τ denotes the temperature.

Region-Level Representation Learning. To capture fine-grained target semantics, we introduce a learnable region query vector $\mathbf{q}_{\text{reg}} \in \mathbb{R}^{d_v}$ that dynamically aggregates visual patch features within the target bounding box. Let $\mathbf{V}_{\text{bbox}} = \{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_M\} \in \mathbb{R}^{M \times d_v}$ denote the subset of M patch features spatially located within the

bounding box. The region representation is computed via scaled dot-product attention pooling:

$$\mathbf{a} = \text{Softmax} \left(\frac{\mathbf{q}_{\text{reg}} \mathbf{V}_{\text{bbox}}^\top}{\sqrt{d_v}} \right), \quad \mathbf{r} = \mathbf{a} \mathbf{V}_{\text{bbox}}, \quad (4)$$

where $\mathbf{a} \in \mathbb{R}^M$ represents the attention weights over patches. The aggregated region feature $\mathbf{r}$ is projected and normalized as $\hat{\mathbf{r}} = \phi_r(\mathbf{r}) / \|\phi_r(\mathbf{r})\|_2$ with $\phi_r : \mathbb{R}^{d_v} \rightarrow \mathbb{R}^{256}$. We then compute the region-text similarity $s_{ij}^r = \cos(\hat{\mathbf{r}}^{(i)}, \hat{\mathbf{w}}_{\text{cls}}^{(j)})$ and optimize:

$$\mathcal{L}_{\text{RITC}} = -\frac{1}{2B} \sum_{i=1}^B \left[\log \frac{\exp(s_{ii}^r/\tau)}{\sum_{j=1}^B \exp(s_{ij}^r/\tau)} + \log \frac{\exp(s_{ii}^r/\tau)}{\sum_{j=1}^B \exp(s_{ji}^r/\tau)} \right] \quad (5)$$

Hard Negative Mining with Margin Constraint. To further enhance discriminability, we enforce a margin-based ranking constraint that explicitly separates positive pairs from the hardest in-batch negatives. For each sample, we define such negatives as:

$$s_{\text{neg}}^{\text{i2t}} = \max_{j \neq i} s_{ij}, \quad s_{\text{neg}}^{\text{t2i}} = \max_{j \neq i} s_{ji} \quad (6)$$

The ranking objective ensures that the positive similarity $s_{\text{pos}} = s_{ii}$ exceeds these hard negatives by at least margin m :

$$\mathcal{L}_{\text{rank}} = \frac{1}{2} \left[\max(0, m - s_{\text{pos}} + s_{\text{neg}}^{\text{i2t}}) + \max(0, m - s_{\text{pos}} + s_{\text{neg}}^{\text{t2i}}) \right] \quad (7)$$

3.3 Semantic Perturbation Enhanced Matching

SPEM actively synthesizes semantically perturbed hard negatives in both textual and visual modalities, as illustrated in Figure 2 and Figure 3(a). Unlike passive in-batch mining that is bottlenecked by dataset distribution, SPEM constructs counterfactual samples

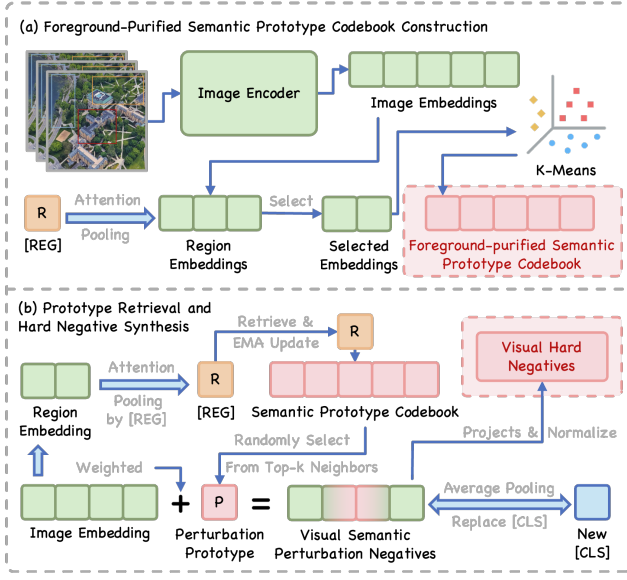


Figure 4: Visual Semantic Perturbation (VSP). (a) Foreground-Purified Semantic Prototype Codebook Construction: K-Means clusters attention-selected foreground patches into semantic prototypes. **(b) Hard Negative Synthesis:** a top- k neighbor prototype is interpolated with the region embedding, while EMA updates the codebook.

that preserve overall structure while perturbing discriminative attributes, exposing the model to isomorphic challenges.

Textual Semantic Perturbation. TSP generates hard textual negatives by systematically replacing attribute words in the input text T . Instead of relying on generative language models for text perturbation, which might introduce unpredictable syntactic changes and hallucinations, we adopt a deterministic, hierarchical dictionary-based substitution. From a causal representation learning perspective, this guarantees *strict control variates*: it modifies *only* the targeted discriminative attribute while preserving the exact grammatical structure and contextual semantics. This precision is crucial for constructing high-quality counterfactuals. Given a text with L tokens $\{w_1, w_2, \dots, w_L\}$, we construct a perturbed text $\tilde{T}$ through three priority levels:

Level 1 (Primary Attributes): We first identify color words $\mathcal{W}_c$ and orientation words $\mathcal{W}_o$ that serve as primary discriminative attributes for building objects in aerial views. For each $w_i \in \mathcal{W}_c \cup \mathcal{W}_o$, we retrieve its semantic opposite from predefined dictionaries:

$$\tilde{w}_i = \mathcal{D}_{\text{opposite}}(w_i), \quad \forall w_i \in \mathcal{W}_c \cup \mathcal{W}_o \quad (8)$$

This targets the most salient attributes that distinguish visually similar buildings.

Level 2 (Antonym Substitution): When the number of replaced words $|\tilde{W}| < \eta$ falls below threshold η , we expand to general antonyms. For remaining tokens, we query an antonym dictionary:

$$\tilde{w}_j = \mathcal{D}_{\text{antonym}}(w_j), \quad \text{if } w_j \in \text{dom}(\mathcal{D}_{\text{antonym}}) \quad (9)$$

Level 3 (Random Perturbation): If still insufficient, random words with length greater than three are selected for replacement to reach

the preset threshold η :

$$\tilde{w}_k = \mathcal{R}(\mathcal{V}), \quad \text{where } |w_k| > 3 \quad (10)$$

where $\mathcal{R}(\mathcal{V})$ samples from the vocabulary. This hierarchical design ensures that perturbations prioritize semantically meaningful attributes while maintaining sufficient diversity.

Visual Semantic Perturbation. VSP constructs hard visual negatives in feature space through attribute-level perturbation, as illustrated in Figure 4. Unlike prior methods that construct visual dictionaries from generic scenes, which introduce severe background noise irrelevant to the structural perturbations of the target objects, we propose a Semantic Prototype Codebook (SPC) that captures discrete, foreground-purified semantic prototypes.

Foreground-Purified Codebook Construction. As depicted in Figure 4 (a), the SPC $C = \{c_1, c_2, \dots, c_K\}$ is a discrete set of K semantic attribute prototypes, where $c_k \in \mathbb{R}^{d_v}$. To construct foreground-purified prototypes, we leverage the attention mechanism from RFA. For each training image, the attention weights over bounding box patches are computed as:

$$a = \text{Softmax} \left(\frac{\mathbf{q}_{\text{reg}} \mathbf{V}_{\text{bbox}}^T}{\sqrt{d_v}} \right) \quad (11)$$

We retain only patches with attention weights in the top $\rho\%$ to form the foreground feature set, filtering out background clutter:

$$\mathcal{F} = \{v_i \mid a_i > \text{Percentile}(a, 100 - \rho)\} \quad (12)$$

The foreground features from all training samples $\bigcup \mathcal{F}$ are clustered via K-Means, producing K cluster centroids as semantic prototypes that represent pure textural attributes of target objects.

Prototype Retrieval and Perturbation. As shown in Figure 4 (b), for a region feature r , we first retrieve its nearest neighbor prototype from SPC via cosine similarity:

$$c_{\text{nn}} = \arg \max_{c \in C} \frac{r^T c}{\|r\| \|c\|} \quad (13)$$

To generate semantically similar yet distinct perturbations, we sample a perturbation prototype c_{pert} from the top- k nearest neighbors of c_{nn} , excluding c_{nn} itself:

$$c_{\text{pert}} \sim \text{Uniform}(\mathcal{N}_k(c_{\text{nn}}) \setminus \{c_{\text{nn}}\}) \quad (14)$$

where $\mathcal{N}_k(c_{\text{nn}}) = \{c \in C : \text{rank}(\cos(c_{\text{nn}}, c)) \leq k\}$. This sampling strategy ensures the perturbation remains within a plausible semantic neighborhood.

Feature-Space Hard Negative Synthesis. For the foreground patch set $\mathcal{F}$ identified by attention weights, we generate perturbed features via weighted interpolation:

$$v'_i = (1 - \alpha)v_i + \alpha c_{\text{pert}}, \quad \forall v_i \in \mathcal{F} \quad (15)$$

where $\alpha \in [0, 1]$ is the mixing coefficient controlling perturbation strength. Background patches remain fixed to preserve scene context, ensuring that perturbations are exclusively applied to the target attributes. The perturbed patch sequence is used to recompute the global representation:

$$v'_{\text{cls}} = \text{AvgPool}(\{v'_1, v'_2, \dots, v'_N\}) \quad (16)$$

which is projected to the joint space $\mathbb{R}^{256}$ and serves as a visual hard negative.

Manifold-Constrained Perturbation. While feature-level perturbation often raises concerns about generating out-of-distribution (OOD) features, VSP intrinsically avoids this risk. Given that SPC is constructed exclusively from in-distribution foreground features (Eq. 12), and $\mathbf{c}_{\text{pert}}$ is strictly constrained within the top- k semantic neighborhood (Eq. 14), the convex combination (Eq. 15) guarantees that the synthesized negative features remain safely bounded within the local in-distribution manifold.

Dynamic Codebook Update. During training, SPC is dynamically updated to track the evolving feature space. For each region feature $\mathbf{r}$, its nearest prototype is updated via Exponential Moving Average (EMA):

$$\mathbf{c}_{\text{nn}} \leftarrow \lambda \mathbf{c}_{\text{nn}} + (1 - \lambda) \text{sg}(\mathbf{r}) \quad (17)$$

where λ is the EMA momentum and $\text{sg}(\cdot)$ stops gradients.

Cross-Modal Matching with Synthesized Negatives. As shown in Figure 3(c), the cross-modal fusion encoder performs fine-grained matching by receiving five types of image-text pairs:

- (1) positive pairs (I, T) ,
- (2) in-batch hard image negatives (I_{hard}^-, T) where

$$I_{\text{hard}}^- = \arg \max_{j \neq i} s_{ji},$$

- (3) in-batch hard text negatives (I, T_{hard}^-) where

$$T_{\text{hard}}^- = \arg \max_{j \neq i} s_{ij},$$

- (4) TSP-generated text negatives $(I, \tilde{T})$, and
- (5) VSP-generated visual negatives $(\tilde{I}, T)$.

The matching score $p(y | I, T)$ is computed through deep cross-attention, and the model is optimized via binary cross-entropy:

$$\mathcal{L}_{\text{FGITM}} = -\mathbb{E}[y \log p + (1 - y) \log(1 - p)] \quad (18)$$

where $y = 1$ for positive pairs (I, T) , and $y = 0$ for negative pairs $(I', T') \in \{(I_{\text{hard}}^-, T), (I, T_{\text{hard}}^-), (I, \tilde{T}), (\tilde{I}, T)\}$. Unlike standard InfoNCE contrastive losses that operate on globally pooled vectors and are inherently insensitive to micro-level perturbations (e.g., altering a single color word), we specifically employ BCE for $\mathcal{L}_{\text{FGITM}}$. Supported by the fusion encoder’s deep cross-attention, BCE effectively supervises dense, token-to-patch interactions, explicitly penalizing the model for localized semantic mismatches.

3.4 Training Objective

Grounding Loss. For consistency with standard practice, we retain the grounding loss to supervise region-text alignment. Given a text segment T_k and its corresponding ground-truth bounding box b_k , a bounding box regression head H_{box} predicts $\hat{b}_k$ from the fused features. The grounding loss penalizes the discrepancy between predicted and ground-truth boxes through a combination of L1 distance and Generalized IoU loss:

$$\mathcal{L}_{\text{ground}} = \sum_k [\lambda_{\text{L1}} \|b_k - \hat{b}_k\|_1 + \lambda_{\text{giou}} \mathcal{L}_{\text{GIoU}}(b_k, \hat{b}_k)] \quad (19)$$

where λ_{L1} and λ_{giou} are the weights for L1 and GIoU losses.

Table 1: Zero-shot and fine-tuned retrieval on GeoText-1652 benchmark. Results are presented using Recall@K (R@K) and mean Recall (mR) for both Image Query and Text Query tasks. The best performances are in bold.

Method	Params	Pretrained Images	Image Query (%)			Text Query (%)			mR (%)
			R@1	R@5	R@10	R@1	R@5	R@10	
Zero-Shot Evaluation on GeoText-1652									
UNITER[7]	300M	4M	2.5	7.4	11.8	0.9	2.7	4.2	4.92
METER-Swin[9]	380M	4M	2.7	8.0	12.2	1.3	3.9	5.8	5.65
ALBEF[19]	210M	4M	2.9	8.1	12.4	1.8	4.8	7.1	6.18
ALBEF[19]	210M	14M	3.0	9.1	14.2	1.1	3.5	5.3	6.03
XVLM[37]	216M	4M	4.9	14.2	21.1	4.3	9.1	13.2	11.13
XVLM[37]	216M	16M	5.0	14.4	21.4	4.5	9.9	13.4	11.43
Fine-Tuned Evaluation on GeoText-1652									
UNITER[7]	300M	4M	21.4	43.4	59.5	10.6	20.4	26.1	30.23
METER-Swin[9]	380M	4M	22.7	46.3	60.7	11.3	21.5	27.3	31.63
ALBEF[19]	210M	4M	22.9	49.5	62.3	12.3	22.8	28.6	33.07
ALBEF[19]	210M	14M	23.2	49.7	62.4	12.5	22.8	28.5	33.18
XVLM[37]	216M	4M	23.6	50.0	63.2	13.1	23.5	29.2	33.77
XVLM[37]	216M	16M	25.0	52.3	65.1	13.2	23.7	29.6	34.82
GeoText-1652[8]	217M	16M	26.3	53.7	66.9	13.6	24.6	31.2	36.05
NAS[32]	217M	16M	26.8	54.0	66.8	13.4	24.6	31.3	36.15
GRASP(ours)	217M	16M	29.6	57.0	69.7	14.5	25.4	31.8	38.04

Overall Objective. The complete training objective is a weighted sum of all component losses:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{GITC}} + \lambda_2 \mathcal{L}_{\text{RITC}} + \lambda_3 \mathcal{L}_{\text{rank}} + \lambda_4 \mathcal{L}_{\text{FGITM}} + \lambda_5 \mathcal{L}_{\text{ground}} \quad (20)$$

where $\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5$ are the weight coefficients for each loss term.

4 Experiments

4.1 Dataset and Evaluation Metrics

Dataset. We evaluate GRASP on two benchmarks: GeoText-1652 [8] for fine-tuned evaluation and ERA [15] for zero-shot transfer evaluation. GeoText-1652 is designed for fine-grained drone-view geo-localization with natural language. Following the official protocol, we use 701 locations for training, 351 for validation, and 600 for testing. ERA [15] is an aerial image-text retrieval benchmark containing 2,864 drone-view images. We use this dataset for zero-shot evaluation to assess cross-domain generalization capability.

Evaluation Metrics. We report Recall@K (R@1, R@5, and R@10) and mean Recall (mR) for image-to-text and text-to-image retrieval; mR averages the six Recall@K scores.

4.2 Implementation Details

For fair comparison, we follow the GeoText-1652 setup [8], which uses a standard XVLM [37] model pre-trained on 16M image-text pairs as the backbone. We fine-tune our model on GeoText-1652 dataset. The SPC contains $K = 2048$ prototypes. For SPC initialization, we retain the top $\rho = 95\%$ patches by attention weights. The EMA momentum coefficient λ is set to 0.995. For VSP, the mixing coefficient α is 0.7 and we sample perturbation prototypes from top-3 neighbors. For TSP, we replace up to 7 attribute words per sentence. The loss weights are set to $\lambda_1 = \lambda_2 = 1.0$, $\lambda_3 = \lambda_4 = 0.6$ and $\lambda_5 = 0.1$. The margin m in ranking loss is 0.25. We use AdamW optimizer with learning rate 3×10^{-5} and batch size 32.

Two-Stage Training Policy & Zero Inference Cost. Fine-tuning adopts a two-stage procedure. The first stage performs warm-up

Table 2: Ablation Study on the GeoText-1652 benchmark.

Components				Image Query(%)			Text Query(%)			mR(%)
$\mathcal{L}_{\text{rank}}$	$\mathcal{L}_{\text{RITC}}$	TSP	VSP	R@1	R@5	R@10	R@1	R@5	R@10	
		✓	✓	26.3	53.7	66.9	13.6	24.6	31.2	36.05
✓		✓	✓	28.4	55.7	68.2	13.9	25.0	31.2	37.01
	✓	✓	✓	28.5	55.8	68.7	14.1	25.2	31.3	37.26
✓	✓			27.8	54.8	67.9	14.0	24.8	31.2	36.75
✓	✓	✓		28.4	56.0	68.9	14.0	25.2	31.5	37.33
✓	✓		✓	28.6	56.1	69.1	14.2	25.4	31.8	37.53
✓	✓	✓	✓	29.6	57.0	69.7	14.5	25.4	31.8	38.04

training for 1 epoch, utilizing all five loss functions with partial FGITM that includes only positive pairs and in-batch negatives, without TSP/VSP-generated hard negatives. After this stage, foreground features are extracted using the trained model to initialize SPC. The second stage introduces SPEM and trains for 2 epochs, incorporating TSP and VSP-generated hard negatives into FGITM for fine-grained discrimination refinement. Crucially, SPC construction and hard-negative synthesis are used only during training and add no inference cost.

4.3 Main Results

Table 1 presents comparative results on the GeoText-1652 dataset under both zero-shot and fine-tuned settings. In the zero-shot setting, all methods exhibit limited performance due to the significant domain gap between natural images and drone-view aerial imagery. XVLM achieves the best zero-shot performance among baselines, benefiting from its multi-grained vision-language pretraining.

In the fine-tuned setting, GRASP achieves competitive performance across all metrics. Compared to the GeoText-1652 baseline [8], GRASP improves R@1 by a significant +3.3 % for Image Query and +0.9 % for Text Query. The larger improvement on Image Query demonstrates that GRASP successfully overcomes the visual isomorphism bottleneck inherent in overhead drone views. While baseline models frequently retrieve visually similar but semantically incorrect buildings (due to identical top-down geometric shapes), the integration of SPEM forces the model to heavily penalize mismatching fine-grained attributes (e.g., roof color, localized materials). This directly addresses the core challenge of aerial imagery. We also compare against NAS [32], which explores feature-space negative construction. While effective in general vision tasks, its performance gains are limited in drone-view scenarios (+0.1 % mR over baseline). In contrast, GRASP leverages the proposed RFA and SPEM for fine-grained drone-view cross-modal understanding, yielding substantially better results (+1.99 % mR over baseline).

4.4 Ablation Studies

To analyze the contribution of each component, we conduct ablation experiments by enabling different combinations of RFA and SPEM components (Table 2). Starting from the baseline (mR 36.05 %), applying SPEM with semantic perturbations (TSP+VSP) improves performance to 37.01 %, validating that perturbation-driven hard negatives and prototype-based augmentation enhance fine-grained discrimination even without explicit RFA. In parallel, enabling RFA

Table 3: Fine-tuned and zero-shot retrieval on ERA.

Method	Image Retrieval (%)			Text Retrieval (%)			mR (%)
	R@1	R@5	R@10	R@1	R@5	R@10	
Fine-tuned Results on ERA Dataset							
VSE++	10.13	35.20	53.91	9.79	30.40	42.90	30.39
PVSE K=2	11.04	35.57	51.65	11.31	32.60	46.95	31.52
PVSE K=1	11.14	36.08	53.75	9.96	33.95	47.97	32.14
CLIP	12.73	37.33	51.52	11.31	31.92	43.91	31.45
PCME	13.85	42.87	60.64	14.69	35.30	49.15	36.08
AMFMN-soft	14.18	46.79	62.87	14.35	38.01	52.02	38.04
AMFMN-sim	13.75	43.41	59.59	14.02	34.12	51.52	36.06
AMFMN-fusion	11.62	42.26	60.51	15.20	36.99	50.33	36.15
GALR	14.03	45.15	64.54	12.38	36.59	50.90	37.27
VCSR	13.69	46.31	66.37	15.65	38.28	53.49	38.96
Zero-shot Results on ERA Dataset (Fine-tuned on GeoText-1652)							
XVLM	14.19	36.15	50.34	14.05	39.46	53.99	34.70
GeoText-1652	17.91	39.19	54.73	17.09	42.30	56.76	38.00
GRASP (Ours)	18.61	42.87	55.23	19.46	44.39	56.25	39.47

via $\mathcal{L}_{\text{rank}}$ and $\mathcal{L}_{\text{RITC}}$ yields 36.75 %, indicating region-aware constraints help suppress background noise and strengthen separability.

When combined, RFA further complements SPEM (e.g., 37.30 % / 37.26 % with additional ranking or region-level contrastive constraints), and under RFA, VSP provides a larger gain than TSP (37.53 % vs. 37.33 %), highlighting that explicitly modeling the visual attribute manifold is particularly vital when vertical geometric cues are lost in aerial perspectives. Ultimately, integrating RFA with both TSP and VSP achieves the best result (38.04 %), confirming the powerful synergy between suppressing background clutter (RFA) and enforcing micro-level attribute discrimination (SPEM).

4.5 Zero-Shot Generalization Evaluation

To evaluate the generalization capability of GRASP, we conduct zero-shot experiments on the ERA dataset [15]. Models are trained exclusively on GeoText-1652 and directly evaluated on ERA without any fine-tuning, simulating real-world deployment scenarios where the model encounters novel geographic regions and scene types.

As shown in Table 3, GRASP obtains mR of 39.47 %, outperforming the best fine-tuned method VCSR by +0.51 %. Compared to the GeoText-1652 baseline under the same zero-shot setting, GRASP improves mR by +1.47 %. The substantial improvements in R@1 for both Image Retrieval (+0.70 %) and Text Retrieval (+2.37 %) demonstrate that GRASP learns transferable representations that generalize across different aerial image domains. This can be attributed to two factors: (1) RFA promotes domain-robust cross-modal grounding; and (2) SPEM improves transferability by learning domain-invariant texture prototypes.

4.6 Visualizing Attention for Semantic Grounding

As shown in Figure 5, to qualitatively evaluate the effectiveness of our GRASP framework in addressing the dual challenges of drone-view retrieval, we visualize and compare the gradient-based class activation maps of GRASP against the baseline GeoText-1652 [8].

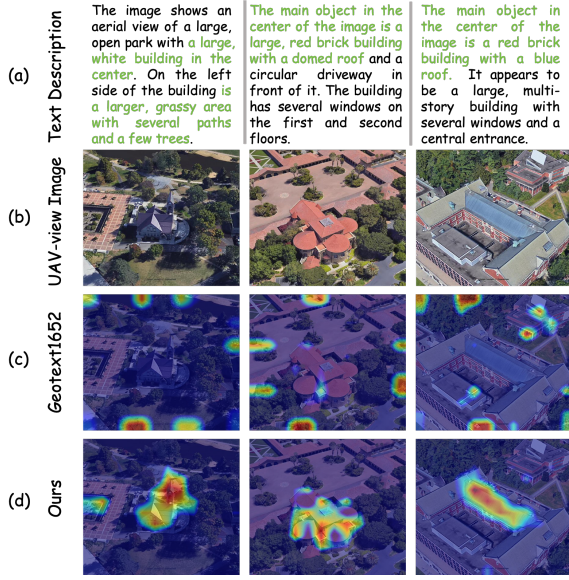


Figure 5: Visualization of activation maps. (a) Text descriptions with key descriptions highlighted in green. (b) UAV images. (c) GeoText-1652 baseline [8]. (d) GRASP.

As observed in row c, the baseline frequently exhibits diffuse and inaccurate attention. At the macro level, it suffers from *Cross-Modal Focus Misalignment*, where its activation maps are easily dominated by surrounding environmental noise and often spill over into irrelevant background regions. At the micro level, when confronted with *Visual Isomorphism*, the baseline struggles to ground the specific descriptive attributes. For example, it fails to accurately isolate the specific "red brick building with a domed roof" (col 2) from neighboring buildings of similar size and shape, indicating an over-reliance on global structural matching rather than precise semantic comprehension.

In contrast, GRASP demonstrates highly localized and accurate semantic grounding (row d), and most crucially, GRASP successfully captures fine-grained textural attributes, overcoming the severe visual homogenization inherent in UAV views. As shown in the visualizations, our method precisely localizes micro-level semantic cues such as a "white building" (col 1) and a "red brick building" (cols 2 and 3). This superior performance can be attributed to the following factors:

First, the integration of RFA effectively suppresses background interference. By explicitly introducing region-level contrastive learning and cross-modal ranking, RFA recalibrates the model's focus from coarse global scenes to object-centric target areas. This allows GRASP to tightly concentrate its activation on the actual foreground subjects rather than the expansive, noisy terrain surrounding them. Second, this powerful discriminative capability is the direct result of our SPEM strategy. By actively constructing counterfactual hard negatives via TSP and VSP, SPEM heavily penalizes subtle attribute mismatches. This mechanism forces the network to look beyond identical geometric shapes and strictly ground the exact fine-grained attributes (e.g., roof color, localized materials) specified in the text.

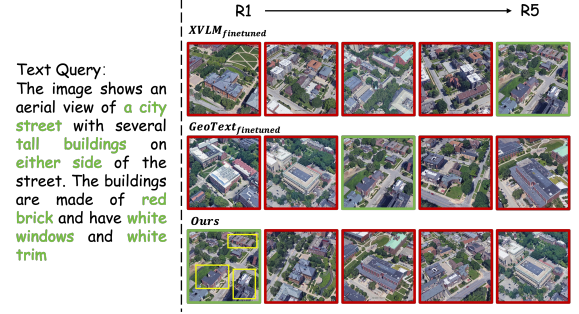


Figure 6: Qualitative text-to-image retrieval results ranked by similarity. Red and green boxes indicate incorrect and correct matches; query attributes are highlighted in green.

In summary, the visualizations confirm the synergistic combination of RFA and SPEM effectively transitions the model from background-dominated global alignment to precise, attribute-aware semantic grounding, proving GRASP's robustness in drone views.

4.7 Retrieval Case Study

To demonstrate the efficacy of our proposed framework in real-world scenarios, we conduct a detailed case study on text-image retrieval. The text query requires the model to identify specific architectural features ("red brick", "white windows", "white trim") arranged in a specific spatial layout ("tall buildings on either side of the street"). Figure 6 visualizes the retrieval rankings of our method compared to strong baselines.

Baseline: The baseline models fail to retrieve the ground truth immediately, locating it at Rank 5 and Rank 3, respectively. While the incorrect images retrieved by these models share global contextual similarities (e.g., aerial views of urban areas), they fail to capture the fine-grained features corresponding to the query due to the aforementioned visual isomorphism.

Ours: Our method achieves a higher recall, retrieving the correct image at Rank 1. The visualization confirms that our model strictly captures both the localized textural attributes and the spatial relations. This aligns with our claim that integrating SPEM allows the model to filter out geometrically similar but attribute-mismatched candidates, a capability missed by baseline.

5 Conclusion

In this paper, we propose GRASP framework for fine-grained cross-modal understanding in drone views. To tackle *Cross-Modal Focus Misalignment*, we introduce RFA, enabling dual-granularity cross-modal alignment via region-level image-text contrastive learning and cross-modal ranking loss. To address *Visual Isomorphism*, we propose SPEM, which leverages a foreground-purified SPC to construct hard negatives, forcing the model to discriminate fine-grained attribute differences. Employed exclusively during training, RFA and SPEM boost performance at no additional inference cost. Extensive experiments demonstrate that GRASP achieves competitive performance on the GeoText-1652 benchmark and remarkable zero-shot capability on the ERA dataset, validating its effectiveness for fine-grained cross-modal understanding in aerial scenarios.

Acknowledgments

This work was supported by the Key Laboratory of Target Recognition and Application under Grant E53712020F, the Key Program of the Chinese Academy of Sciences under Grant KGFZD-145-25-38 and RCJJ-145-24-13 and the Science and Disruptive Technology Program of Aerospace Information Research Institute under Grant AIRCAS2024-AIRCAS-SDTP-03.

References

- [1] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. 2018. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3674–3683.
- [2] Pablo Arbelaez, Michael Maire, Charles Fowlkes, and Jitendra Malik. 2010. Contour detection and hierarchical image segmentation. *IEEE transactions on pattern analysis and machine intelligence* 33, 5 (2010), 898–916.
- [3] Gabriele Berton, Riccardo Mereu, Gabriele Trivigno, Carlo Masone, Gabriela Csurka, Torsten Sattler, and Barbara Caputo. 2022. Deep visual geo-localization benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5396–5407.
- [4] Valts Blukis, Yannick Terme, Eyvind Niklasson, Ross A Knepper, and Yoav Artzi. 2019. Learning to map natural language instructions to physical quadcopter control using simulated flight. *arXiv preprint arXiv:1910.09664* (2019).
- [5] Meghan Chandarana, Erica L Meszaros, Anna Trujillo, and B Danette Allen. 2017. 'Fly Like This': Natural Language Interface for UAV Mission Planning. In *International Conference on Advances in Computer-Human Interactions*. Paper NF1676L-26108.
- [6] Guanlin Chen, Pengfei Zhu, Bing Cao, Xing Wang, and Qinghua Hu. 2023. Cross-drone transformer network for robust single object tracking. *IEEE transactions on circuits and systems for video technology* 33, 9 (2023), 4552–4563.
- [7] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*. Springer, 104–120.
- [8] Meng Chu, Zhedong Zheng, Wei Ji, Tingyu Wang, and Tat-Seng Chua. 2024. Towards natural language-guided drones: GeoText-1652 benchmark with spatial relation matching. In *European Conference on Computer Vision*. Springer, 213–231.
- [9] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, et al. 2022. An empirical study of training end-to-end vision-and-language transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 18166–18176.
- [10] Jorge Gago, Cyril Douthe, Rafael E Coopman, Pedro Pablo Gallego, Miquel Ribas-Carbo, Jaume Flexas, Jb Escalona, and Hb Medrano. 2015. UAVs challenge to assess water stress for sustainable agriculture. *Agricultural Water Management* 153 (2015), 9–19.
- [11] Octavian Ganea, Gary Bécigneul, and Thomas Hofmann. 2018. Hyperbolic entailment cones for learning hierarchical embeddings. In *International conference on machine learning*. PMLR, 1646–1655.
- [12] Georgios Georgakis, Karl Schmeckpeper, Karan Wanchoo, Soham Dan, Eleni Miltisakaki, Dan Roth, and Kostas Daniilidis. 2022. Cross-modal map learning for vision and language navigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 15460–15470.
- [13] Yicong Hong, Qi Wu, Yuankai Qi, Cristian Rodriguez-Opazo, and Stephen Gould. 2021. Vln bert: A recurrent vision-and-language bert for navigation. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*. 1643–1653.
- [14] Baichuan Huang, Deniz Bayazit, Daniel Ullman, Nakul Gopalan, and Stefanie Tellex. 2019. Flight, camera, action! using natural language and mixed reality to control a drone. In *2019 International conference on robotics and automation (ICRA)*. IEEE, 6949–6956.
- [15] Jinghao Huang, Yaxiong Chen, Shengwu Xiong, and Xiaoqiang Lu. 2024. Visual Contextual Semantic Reasoning for Cross-Modal Drone Image-Text Retrieval. *IEEE Transactions on Geoscience and Remote Sensing* (2024).
- [16] Dong-Wook Kim, Tae-Sun Min, Yoonha Kim, Renato Rodrigues Silva, Hae-Nam Hyun, Ju-Sung Kim, Kyung-Hwan Kim, Hak-Jin Kim, and Yong Suk Chung. 2019. Sustainable agriculture by increasing nitrogen fertilizer efficiency using low-resolution camera mounted on unmanned aerial vehicles. *International Journal of Environmental Research and Public Health* 16, 20 (2019), 3893.
- [17] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*. PMLR, 19730–19742.
- [18] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*. PMLR, 12888–12900.
- [19] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in Neural Information Processing Systems* 34 (2021), 9694–9705.
- [20] Arjun Majumdar, Ayush Shrivastava, Stefan Lee, Peter Anderson, Devi Parikh, and Dhruv Batra. 2020. Improving vision-and-language navigation with image-text pairs from the web. In *European Conference on Computer Vision*. Springer, 259–274.
- [21] Jun-Ichi Meguro, Kiitirou Ishikawa, Takumi Hasizume, Jun-Ichi Takiguchi, Itsuki Noda, and Mitinori Hatayama. 2006. Disaster information collection into geographic information system using rescue robots. In *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 3514–3520.
- [22] Abolfazl Mehdodniya, Julian L Webber, Sathishkumar Karupusamy, et al. 2022. Improving the geo-drone-based route for effective communication and connection stability improvement in the emergency area ad-hoc network. *Sustainable Energy Technologies and Assessments* 53 (2022), 102558.
- [23] Kim-Anh Nguyen, Maximilian Köper, Sabine Schulte im Walde, and Ngoc Thang Vu. 2017. Hierarchical embeddings for hypenymy detection and directionality. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 233–243.
- [24] Avik Pal, Max van Spengler, Guido Maria D'Amely di Melendugno, Alessandro Flaborea, Fabio Galasso, and Pascal Mettes. 2024. Compositional entailment learning for hyperbolic vision-language models. *arXiv preprint arXiv:2410.06912* (2024).
- [25] Yuankai Qi, Zizheng Pan, Yicong Hong, Ming-Hsuan Yang, Anton Van Den Hengel, and Qi Wu. 2021. The road to know-where: An object-and-room informed sequential bert for indoor vision-language navigation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1655–1664.
- [26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.
- [27] Md Tahmid Rashid, Daniel Yue Zhang, and Dong Wang. 2020. Socialdrone: An integrated social media and drone sensing system for reliable disaster response. In *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*. IEEE, 218–227.
- [28] Royston Rodrigues and Masahiro Tani. 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*. 3871–3879.
- [29] Paolo Tripicchio, Massimo Satler, Giacomo Dabisias, Emanuele Ruffaldi, and Carlo Alberto Avizzano. 2015. Towards smart farming and sustainable agriculture with drones. In *2015 international conference on intelligent environments*. IEEE, 140–143.
- [30] Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. 2015. Order-embeddings of images and language. *arXiv preprint arXiv:1511.06361* (2015).
- [31] Tingyu Wang, Zhedong Zheng, Chenggang Yan, Jiyong Zhang, Yaoqi Sun, Bolun Zheng, and Yi Yang. 2021. Each part matters: Local patterns facilitate cross-view geo-localization. *IEEE Transactions on Circuits and Systems for Video Technology* 32, 2 (2021), 867–879.
- [32] Yeyuan Wang, Dehong Gao, Lei Yi, Linbo Jin, Jinxia Zhang, Libin Yang, and Xiaoyan Cai. 2025. Enhancing fine-grained vision-language pretraining with negative augmented samples. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 8060–8068.
- [33] E Wardihani, Magfur Ramdhani, Amin Suharjono, Thomas Agung Setyawan, Sidiq Syamsul Hidayat, Sarono Widodo Helmy, Eddy Triyono, and FIRDANIS Saifullah. 2018. Real-time forest fire monitoring system using unmanned aerial vehicle. *Journal of Engineering Science and Technology* 13, 6 (2018), 1587–1594.
- [34] Hui Wu, Haoquan Zhai, Yuchen Li, Hengyi Cai, Peirong Zhang, Yidan Zhang, Lei Wang, Chunle Wang, Yingyan Hou, Shuaiqiang Wang, et al. 2025. MARA: A Multimodal Adaptive Retrieval-Augmented Framework for Document Question Answering. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 4329–4338.
- [35] Yan Xia, Zhendong Li, Yun-Jin Li, Letian Shi, Hu Cao, João F Henriques, and Daniel Cremers. 2024. UniLoc: Towards universal place recognition using any single modality. *arXiv preprint arXiv:2412.12079* (2024).
- [36] Junyan Ye, Honglin Lin, Leyan Ou, Dairong Chen, Zihao Wang, Qi Zhu, Conghui He, and Weijia Li. 2025. Where am i? cross-view geo-localization with natural language descriptions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5890–5900.
- [37] Yan Zeng, Xinsong Zhang, and Hang Li. 2021. Multi-grained vision language pre-training: Aligning texts with visual concepts. *arXiv preprint arXiv:2111.08276* (2021).
- [38] Xiao Zhang and Michael Maire. 2020. Self-supervised visual representation learning from hierarchical grouping. *Advances in Neural Information Processing Systems* 33 (2020), 16579–16590.

- [39] Zhedong Zheng, Yunchao Wei, and Yi Yang. 2020. University-1652: A multi-view multi-source benchmark for drone-based geo-localization. In *Proceedings of the*

28th ACM international conference on Multimedia. 1395–1403.