HCCM: Hierarchical Cross-Granularity Contrastive and Matching Learning for Cross-Modal Drone Navigation

Hao Ruan
Xiamen University
ruanhao@stu.xmu.edu.cn

Jinliang Lin
Xiamen University
jinlianglin@stu.xmu.edu.cn

Zhiming Luo*
Xiamen University
zhiming.luo@xmu.edu.cn

Yu Zang
Xiamen University
zangyu7@126.com

Cheng Wang
Xiamen University
cwang@xmu.edu.cn

Abstract

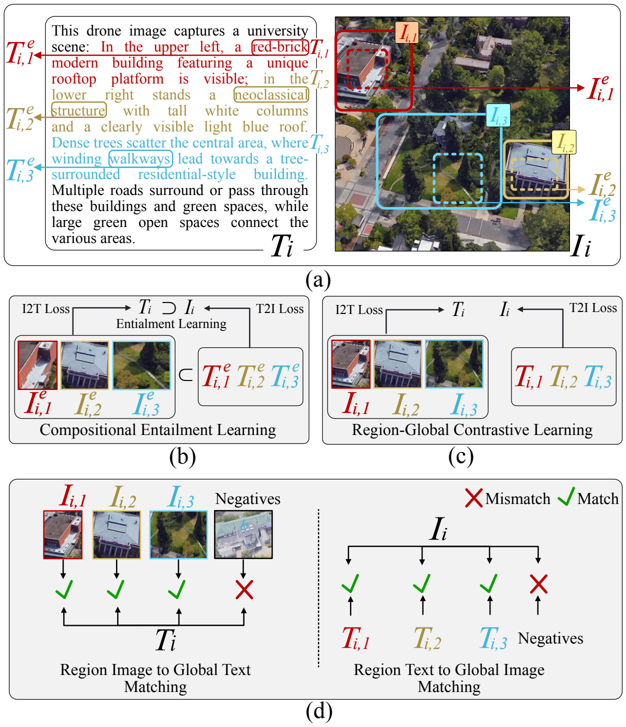
Natural Language-Guided Drones (NLGD) offer a novel and flexible interaction paradigm for tasks such as target matching and navigation. However, the wide field of view and complex compositional semantic relationships inherent in drone scenarios place greater demands on visual language understanding. First, mainstream Vision-Language Models (VLMs) primarily focus on global feature alignment and lack fine-grained semantic understanding. Second, existing hierarchical semantic modeling methods rely on precise entity partitioning and strict containment relationship constraints, which limits their effectiveness in complex drone environments. To address these challenges, we propose the Hierarchical Cross-Granularity Contrastive and Matching learning (HCCM) framework, comprising two core components: 1) Region-Global Image-Text Contrastive Learning (RG-ITC). Avoiding precise scene entity partitioning, RG-ITC models hierarchical local-to-global cross-modal semantics by contrasting local visual regions with global text semantics, and vice versa. 2) Region-Global Image-Text Matching Learning (RG-ITM). Instead of relying on strict relationship constraints, this component evaluates local semantic consistency within global cross-modal representations, improving the comprehension of complex compositional semantics. Furthermore, drone scenario textual descriptions are often incomplete or ambiguous, destabilizing global semantic alignment. To mitigate this, HCCM incorporates a Momentum Contrast and Momentum Distillation (MCD) mechanism, enhancing alignment robustness. Extensive experiments on the GeoText-1652 benchmark demonstrate HCCM significantly outperforms existing methods, achieving state-of-the-art Recall@1 scores of 28.8% (image retrieval) and 14.7% (text retrieval). Moreover, HCCM exhibits strong zero-shot generalization on the unseen ERA dataset, achieving 39.93% mean recall (mR), surpassing evaluated fine-tuned models. These results highlight the effectiveness and robustness of HCCM across diverse scenarios. Our implementation is available at https://github.com/rhao-hur/HCCM.

1. Introduction

In recent years, the application of Unmanned Aerial Vehicles (UAVs) has expanded from basic image acquisition to include complex tasks such as agricultural monitoring [2–5], target tracking [6–13], and cross-view target matching [14–16]. Among these, cross-view target matching has emerged as a crucial task, aiming to locate targets by matching images captured from different perspectives (*e.g.*, UAV, satellite, ground), which is often formulated as an image retrieval problem. However, relying solely on visual queries encounters challenges in cross-view target matching: performance is susceptible to variations in illumination, weather, and viewpoint changes, leading to degradation [17–20]. Furthermore, visual queries may not always be available in practical applications. Consequently, leveraging Natural Language-Guided Drones (NLGD) for target matching has emerged as a significant research direction, owing to their flexible querying approach and integrated language understanding capabilities [21].

Recent research in Natural Language-Guided Drones (NLGD) shows significant progress. To support research in the NLGD task, Chu et al. [21] introduced a large-scale NLGD dataset, GeoText-1652, and defined two core subtasks: **UAV Text Navigation** (text-guided UAV positioning) and **UAV View Target Localization** (matching descriptions to UAV views for target identification). They employed Vision-Language Models (VLMs) [22–27] with contrastive learning to align global image-text representa-

[*] Corresponding author



This drone image captures a university scene: In the upper left, a red-brick modern building featuring a unique rooftop platform is visible; in the lower right stands a neoclassical structure with tall white columns and a clearly visible light blue roof. Dense trees scatter the central area, where winding walkways lead towards a tree-surrounded residential-style building. Multiple roads surround or pass through these buildings and green spaces, while large green open spaces connect the various areas.

$T^e_{i,1}$
$T^e_{i,2}$
$T^e_{i,3}$
$T_{i,1}$
$T_{i,2}$
$T_{i,3}$
T_i
$I_{i,1}$
$I_{i,2}$
$I_{i,3}$
$I^e_{i,1}$
$I^e_{i,2}$
$I^e_{i,3}$
I_i
(a)
I2T Loss
$T_i \supset I_i$
T2I Loss
Entialment Learning
$I^e_{i,1}$ $I^e_{i,2}$ $I^e_{i,3}$ $\subset$ $T^e_{i,1}$ $T^e_{i,2}$ $T^e_{i,3}$
Compositional Entailment Learning
(b)
I2T Loss
T_i
I_i
T2I Loss
$I_{i,1}$ $I_{i,2}$ $I_{i,3}$
$T_{i,1}$ $T_{i,2}$ $T_{i,3}$
Region-Global Contrastive Learning
(c)
$I_{i,1}$ $I_{i,2}$ $I_{i,3}$ Negatives
✓ ✓ ✓ ✗
T_i
Region Image to Global Text Matching
✗ Mismatch ✓ Match
I_i
✓ ✓ ✓ ✗
$T_{i,1}$ $T_{i,2}$ $T_{i,3}$ Negatives
Region Text to Global Image Matching
(d)

Figure 1. Comparison of hierarchical vision-language modeling approaches. **(a)** Example input pair (I_i, T_i) demonstrating multiple semantic levels. The global (I_i, T_i), local region ($I_{i,k}$, solid box; $T_{i,k}$, sentence), and entity ($I^e_{i,k}$, dashed box; $T^e_{i,k}$, phrase). **(b)** Existing approach [1] models entity-level hierarchies $(I^e_{i,k}, T^e_{i,k})$ using strict part-whole entailment. **(c)** Proposed RG-ITC contrasts local region features ($I_{i,k}$ or $T_{i,k}$) against the complementary global representation (T_i or I_i) for cross-granularity semantic alignment. **(d)** Proposed RG-ITM assesses region-global cross-modal consistency by matching local features ($I_{i,k}$ or $T_{i,k}$) with the complementary global representation (T_i or I_i) (Match ✓ / Mismatch ×).

tions in a shared embedding space. Concurrently, Huang et al. [28] introduced the ERA and UDV datasets for NLGD, but explored non-VLM approaches. They utilized Convolutional Neural Networks (CNNs) [29–31] and Bidirectional Gated Recurrent Units (Bi-GRUs) [32] for visual-language encoding and developed the Text-Guided Visual Information Reasoning (TGVIR) mechanism for fine-grained cross-modal semantic alignment.

However, the NLGD task requires handling queries with compositional semantics, while current methods [21, 28] often exhibit **poor generalization for compositional understanding** and **fail to grasp cross-granularity semantic hierarchies**. As illustrated in Figure 1(a), let I_i, T_i be the global image/text, $I_{i,k}, T_{i,k}$ the local region image/text (solid box/sentence), and $I^e_{i,k}, T^e_{i,k}$ the fine-grained entity image/text (dashed box/phrase). A global image I_i typically contains multiple entity-level semantic regions (*e.g.*, $I^e_{i,1}, I^e_{i,2}, I^e_{i,3}$) corresponding to entity descriptions (*e.g.*, $T^e_{i,1}$, $T^e_{i,2}, T^e_{i,3}$) within the global text T_i. The compositional interplay of these regions defines the scene's semantics. Accurate scene distinction or instruction execution requires understanding the compositional roles of local regions within the global context. VLMs relying on global semantic alignment, often lack this fine-grained understanding. Furthermore, traditional sequence models [32] generalize poorly when interpreting complex compositional relationships, particularly in longer texts [33].

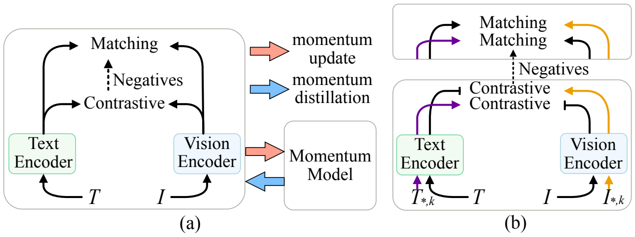


Figure 2. Comparison of contrastive and matching learning frameworks. (a) Momentum-enhanced framework with update and distillation mechanisms for improved stability, built upon the standard bidirectional global alignment (left). (b) Proposed region-global framework combining global (T, I) and regional $(T_{*,k}, I_{*,k})$ information with unidirectional region-to-global contrastive learning.

Recognizing the need to model relationships across different granularities, some VLM-based methods explore specific cross-modal interactions. For instance, Pal et al. [1] proposed Compositional Entailment Learning (Fig. 1(b)), modeling part-whole hierarchies via cross-modal contrastive learning ($I^e_{i,k} \to T_i$, $T^e_{i,k} \to I_i$) within hyperbolic space. This approach leverages semantic entailment learning [34], assuming more abstract entities $(I^e_{i,k}, T^e_{i,k})$ entail the global concrete concepts (I_i, T_i). Typically, textual descriptions tend to express more abstract concepts than images. Formally, A entails B is defined as $B \subset A$, implying intra-modal relations $T_i \subset T^e_{i,k}$ and $I_i \subset I^e_{i,k}$, and inter-modal relations $I^e_{i,k} \subset T^e_{i,k}$ and $I_i \subset T_i$. However, applying such strict, entailment-based hierarchies proves challenging for UAV bird's-eye views. UAV imagery frequently features complexly intertwined elements (*e.g.*, the Z-shaped road in Fig. 1(a)) and widely distributed similar elements (*e.g.*, trees), resisting clear delineation into discrete entities suitable for rigid decomposition. Moreover, UAV scene descriptions often prioritize element co-occurrence and composition over strict semantic entailment. Consequently, the geometric constraints imposed by entailment learning (*e.g.*, entailment cone loss [35]) may be overly restrictive for flexibly capturing the compositional semantics inherent in UAV scenarios.

To address the above issues, we propose the **H**ierarchical **C**ross-Granularity **C**ontrastive and **M**atching Learning (HCCM) method. Building upon the standard cross-modal contrastive and matching learning framework [26], HCCM introduces Region-Global Image Text Contrastive Learning

(RG-ITC) (Figure 1(c)). Unlike methods relying on entity partitioning, RG-ITC models semantic associations across granularities, specifically linking unimodal local information (image region $I_{i,k}$ or text fragment $T_{i,k}$) with the corresponding global representation of the other modality (text T_i and image I_i). This aims to capture the local-to-global cross-modal semantic hierarchical relationships within UAV scenarios. Furthermore, distinct from approaches modeling strict parent-child or part-whole relationships, we introduce Region-Global Image Text Matching Learning (RG-ITM) (Figure 1(d)) to enhance the model's ability to discern the semantic consistency between local details and the global context across modalities. Specifically, it assesses whether the semantic content derived from a unimodal local region ($I_{i,k}$ or $T_{i,k}$) is consistent with the corresponding global representation of the other modality (T_i or I_i). This process improves the model's comprehension and discrimination of complex spatial layouts and intertwined semantics.

However, directly applying this strategy in drone scenarios encounters certain limitations. Large-scale views often yield incomplete or ambiguous T_i, causing local-global alignment to amplify noise bias [36], impairing global performance. To mitigate this, we introduce Momentum Contrast and Momentum Distillation (MCD), employing negative queues and soft targets, respectively, to stabilize global alignment and enhance interference resistance. By combining RG-ITC, RG-ITM, and MCD, our proposed HCCM can effectively improve the performance of VLM in UAV scenarios.

In summary, the main contributions of this paper are as follows:

1. A Hierarchical Cross-Granularity Contrastive and Matching Learning (HCCM) framework is presented to address insufficient fine-grained feature alignment and difficulty in modeling hierarchical relationships in Natural Language-Guided Drone tasks.
2. Region-Global Image Text Contrastive Learning (RG-ITC) is designed to model cross-granularity hierarchies, and Region-Global Image Text Matching Learning (RG-ITM) is proposed to enhance composite semantic understanding.
3. A Momentum Contrast and Momentum Distillation (MCD) strategy is introduced to mitigate noise amplification from incomplete text descriptions.
4. Experiments on GeoText-1652 and ERA datasets validate the effectiveness and robustness of the proposed method.

2. Related Work

2.1. Vision and Language Navigation

Using natural language descriptions for positioning and navigation can enhance navigation efficiency, which has attracted the attention of researchers [21, 37]. For retrieving corresponding satellite images based on scene text descriptions, Ye et al. [38] proposed a text-based localization method, CrossText2Loc, which excels in handling long texts and interpretability. Xia et al. [39] proposed a Self-Attention Pooling (SAP) module to integrate data from multiple modalities, including natural language, images, and point clouds, to achieve cross-modal place recognition. To navigate drones through natural language commands, Chu et al. [21] introduced a natural language-guided UAV geolocalization benchmark, GeoText-1652, and proposed a blending spatial matching for region-level spatial relation matching. In addition, Huang et al. [28] utilized textual cues through Contextual Region Learning (CRL) and Consistency Semantic Alignment (CSA) mechanisms to guide the model in overcoming challenges related to context understanding and alignment in UAV images. Unlike existing methods, our approach primarily focuses on addressing the issue of insufficient fine-grained alignment in drone scenarios, which has been overlooked by existing methods.

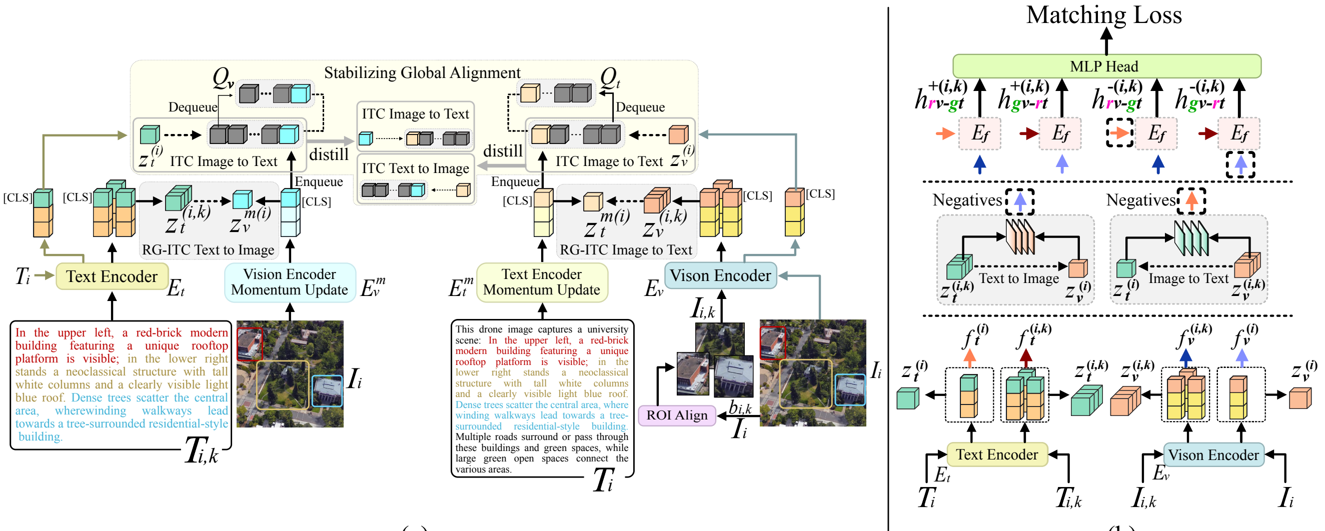
2.2. Visual Language Model for Feature Alignment

Vision-Language Models (VLMs) aim to learn joint representations of images and text [40–49]. CLIP [22] laid the groundwork using large-scale contrastive learning, followed by advancements like UNITER's image-text matching (ITM) [23], ALBEF's "align-before-fuse" strategy with hard negative mining [25], and X-VLM's focus on multi-level concept alignment [26]. Architecturally, METER [50] assessed various encoders and fusion strategies, while the BLIP series [24, 51] employed lightweight modules like Q-Former to integrate understanding and generation tasks efficiently.

Standard global image-text alignment fails to capture the hierarchical part-whole concepts inherent in visual and linguistic data. To overcome this, researchers have shifted towards fine-grained and hierarchical approaches. Early methods encoded semantic hierarchies in embedding spaces using partial order constraints or lexical entailment [52–54], while others utilized visual structures, aligning text with segmented image regions [55, 56] or focusing on object-level contrastive learning [57]. Recent developments have exploited hyperbolic geometric spaces for hierarchical representation. HyCoCLIP [1], for instance, models image-text relationships within this space, employing cross-modal contrastive learning between parts and wholes and using the entailment cone loss [35] to enforce hierarchical constraints both within and across modalities. However, its effectiveness depends on clear part-whole structures or explicit semantic relationships in the data.

3. Methodology

This section details the proposed Hierarchical Cross-granularity Contrastive and Matching Learning (HCCM) framework. We first outline the base vision-language encod-


Stabilizing Global Alignment
Q_v
Q_t
Dequeue
Dequeue
ITC Image to Text
ITC Text to Image
ITC Image to Text
ITC Image to Text
distill
distill
Enqueue
Enqueue
[CLS]
RG-ITC Text to Image
RG-ITC Image to Text
T_i
Text Encoder
E_t
Vision Encoder Momentum Update
E_v^m
E_t^m
Text Encoder Momentum Update
Vison Encoder
E_v
I_i
I_{i,k}
b_{i,k}
ROI Align
In the upper left, a red-brick modern building featuring a unique rooftop platform is visible; in the lower right stands a neoclassical structure with tall white columns and a clearly visible light blue roof. Dense trees scatter the central area, wherewinding walkways lead towards a tree-surrounded residential-style building.
T_{i,k}
This drone image captures a university scene: In the upper left, a red-brick modern building featuring a unique rooftop platform is visible; in the lower right stands a neoclassical structure with tall white columns and a clearly visible light blue roof. Dense trees scatter the central area, where winding walkways lead towards a tree-surrounded residential-style building. Multiple roads surround or pass through these buildings and green spaces, while large green open spaces connect the various areas.
T_i
Matching Loss
MLP Head
h_rv-gt^+(i,k)
h_gv-rt^+(i,k)
h_rv-gt^-(i,k)
h_gv-rt^-(i,k)
E_f
Negatives
Negatives
Text to Image
Image to Text
f_t^(i)
f_t^(i,k)
f_v^(i,k)
f_v^(i)
Text Encoder
Vison Encoder
E_t
E_v
T_i
T_{i,k}
I_{i,k}
I_i

(a) (b)

Figure 3. Overview of the HCCM architecture. (a) presents the integration of stabilizing global alignment ($\mathcal{L}_{ITC}$) and Region-Global Image Text Contrastive Learning ($\mathcal{L}_{RG-ITC}$), while (b) illustrates Region-Global Image Text Matching Learning ($\mathcal{L}_{RG-ITM}$) module with hard negative mining.

ing process (3.1). Then, we elaborate on the two core components: Region-to-Global Image-Text Contrastive Learning (RG-ITC) for hierarchical semantics learning (3.2) and Region-Global Image-Text Matching Learning (RG-ITM) for compositional semantics understanding (3.3). To enhance robustness, we introduce a mechanism for stabilizing global alignment (3.4). Finally, the overall training objective combining these elements is presented (3.5). Figure 2 illustrates the evolution from standard alignment frameworks to the proposed HCCM approach.

3.1. Vision-Language Encoding

The model processes data in batches of N samples, where each input consists of a global image-text pair (I_i, T_i). Each image I_i is associated with region patches $I_{i,k}$, defined by bounding boxes $b_{i,k}$ and extracted from I_i using ROI Align, along with corresponding text fragments $T_{i,k}$. We adopt XVLM [35] as the fundamental architecture, which comprises an image encoder E_v, a text encoder E_t, and a cross-modal fusion encoder E_f.

Global inputs I_i and T_i are encoded to produce feature sequences $f_v^{(i)}$ and $f_t^{(i)}$, with their [CLS] token embeddings $f_{v,\text{[CLS]}}^{(i)}$ and $f_{t,\text{[CLS]}}^{(i)}$ serving as global aggregated features. Regional patches $I_{i,k}$ and text fragments $T_{i,k}$ are similarly encoded to yield regional [CLS] embeddings $f_{v,\text{[CLS]}}^{(i,k)}$ and $f_{t,\text{[CLS]}}^{(i,k)}$.

For contrastive learning, all [CLS] embeddings are mapped through modality-specific projection layers (online: ϕ_v, ϕ_t; momentum: ϕ_v^m, ϕ_t^m) and L2-normalized to generate similarity embeddings, including global online $z_v^{(i)}, z_t^{(i)}$, global momentum $z_v^{m(i)}, z_t^{m(i)}$, and regional online $z_v^{(i,k)}, z_t^{(i,k)}$.

3.2. Region-to-Global Image-Text Contrastive Learning

We introduce Region-to-Global ITC (RG-ITC) to explicitly model part-to-whole hierarchical relationships. To achieve this, we first define the probability $p_{v\to t}^{(i,k)}$ of a visual region (i,k) matching its corresponding global text description i:

$$p_{v\to t}^{(i,k)} = \frac{\exp(s(z_v^{(i,k)}, z_t^{m(i)})/\tau)}{\sum_{j=1}^{N} \exp(s(z_v^{(i,k)}, z_t^{m(j)})/\tau)} . \tag{1}$$

The loss for this direction is the negative log-likelihood, averaged over all regions:

$$\mathcal{L}_{v\to t} = -\frac{1}{|\mathcal{R}_N|} \sum_{(i,k)\in\mathcal{R}_N} \log p_{v\to t}^{(i,k)} . \tag{2}$$

Symmetrically, we define the text-to-visual loss $\mathcal{L}_{t\to v}$. The final RG-ITC loss is the average of these two components:

$$\mathcal{L}_{RG-ITC} = \frac{1}{2}(\mathcal{L}_{v\to t} + \mathcal{L}_{t\to v}) , \tag{3}$$

where $s(\cdot,\cdot)$ is cosine similarity, τ is temperature, and $\mathcal{R}_N$ is the set of all valid region pairs in the batch.

3.3. Region-Global Image-Text Matching Learning

Image-Text Matching (ITM) is treated as a binary classification task. Its loss is the binary cross-entropy (BCE) over positive ($\mathcal{P}$) and negative ($\mathcal{N}$) pairs. For a single pair (I, T) with label y, the BCE term is:

$$\ell_{BCE}(y, p) = -\left[y \log p[1] + (1-y) \log p[0]\right] . \quad (4)$$

The total ITM loss is the average of these terms:

$$\mathcal{L}_{ITM} = \frac{1}{|\mathcal{P}| + |\mathcal{N}|} \sum_{(I,T)\in\mathcal{P}\cup\mathcal{N}} \ell_{BCE}(y, p_{match}) . \quad (5)$$

Our Region-Global ITM (RG-ITM) uses the same principle, averaging the BCE term over the set of all generated region-global pairs $\mathcal{H}_{\text{RG}}$:

$$\mathcal{L}_{\text{RG-ITM}} = \frac{1}{|\mathcal{H}_{\text{RG}}|} \sum_{h\in\mathcal{H}_{\text{RG}}} \ell_{BCE}(y^{(h)}, p^{(h)}_{match}) . \quad (6)$$

where $y^{(h)}$ is 1 for positive and 0 for negative examples.

3.4. Stabilizing Global Alignment with Momentum Contrast and Distillation

To mitigate noise from ambiguous descriptions, we stabilize global ITC learning with Momentum Contrast and Distillation (MCD).

Momentum Contrast. We use momentum queues (Q_v, Q_t) to create large, stable sets of negative features (Z^m_v, Z^m_t).

Momentum Distillation. We generate soft targets (q) by blending momentum model predictions with ground-truth labels (y):

$$q^{(i)}_{i2t} = \alpha \cdot \text{softmax}(z^{m(i)\top}_v Z^m_t/\tau) + (1-\alpha) \cdot y^{(i)}_{i2t} , \quad (7)$$

$$q^{(i)}_{t2i} = \alpha \cdot \text{softmax}(z^{m(i)\top}_t Z^m_v/\tau) + (1-\alpha) \cdot y^{(i)}_{t2i} . \quad (8)$$

The stabilized loss $\mathcal{L}_{ITC_{MCD}}$ is then defined using cross-entropy (H). Let's define the two directional loss components for each sample i:

$$\ell^{(i)}_{i2t} = H\left(q^{(i)}_{i2t}, \text{softmax}(z^{(i)\top}_v Z^m_t/\tau)\right) , \quad (9)$$

$$\ell^{(i)}_{t2i} = H\left(q^{(i)}_{t2i}, \text{softmax}(z^{(i)\top}_t Z^m_v/\tau)\right) . \quad (10)$$

The final loss is the average of these components over all samples:

$$\mathcal{L}_{ITC_{MCD}} = \frac{1}{2N} \sum_{i=1}^{N} \left(\ell^{(i)}_{i2t} + \ell^{(i)}_{t2i}\right) . \quad (11)$$

3.5. Overall Training Objective

Our full framework is trained with a multi-task objective, including a bounding box regression loss, $\mathcal{L}_{Box}$:

$$\begin{aligned}\mathcal{L}_{Box} = \sum_k \big[&\lambda_{L1}||\hat{b}_k - b_{i,k}||_1 \\ &+ \lambda_{GIoU}(1 - \text{GIoU}(\hat{b}_k, b_{i,k}))\big] .\end{aligned} \quad (12)$$

The complete training objective $\mathcal{L}_{total}$ is the weighted sum of all constituent losses:

$$\begin{aligned}\mathcal{L}_{total} = \;& w_{itc}\mathcal{L}_{ITC_{MCD}} \\ &+ w_{itm}\mathcal{L}_{ITM} \\ &+ w_{rg-itc}\mathcal{L}_{RG-ITC} \\ &+ w_{rg-itm}\mathcal{L}_{RG-ITM} \\ &+ w_{box}\mathcal{L}_{Box} .\end{aligned} \quad (13)$$

Minimizing this combined objective guides the HCCM model to learn robust, multi-granularity vision-language representations.

4. Experiments

This section evaluates the proposed HCCM method on the Natural Language-Guided Drone (NLGD) tasks of UAV text navigation (text-to-image retrieval) and UAV view target localization (image-to-text retrieval). We first outline the experimental setup, detailing the datasets, metrics (4.1), and implementation (4.2). We then present comparative results against state-of-the-art methods (4.3), followed by ablation studies (4.4) and an assessment of zero-shot generalization (4.5). Finally, visualization is provided (4.6).

4.1. Datasets and Evaluation Metrics

We use the official data provided by the *RoboSense Challenge 2025* [58] held at IROS 2025. This competition builds upon the legacy of the *RoboDepth Challenge 2023* [59, 60] at ICRA 2023 and the *RoboDrive Challenge 2024* [61, 62] at ICRA 2024, continuing the collective effort to advance robust and scalable robot perception. Each track in this competition is grounded on an established benchmark designed for evaluating real-world robustness and generalization [63–68]. Specifically, this task is built upon the **GeoText-1652** dataset [66] in **Track 4**, which benchmarks cross-modal image-text retrieval for language-guided drone navigation across drastically different viewpoints and real-world sensing conditions.

For assessing zero-shot generalization, we additionally evaluate performance on the ERA dataset [28]. Across all experiments, performance is measured using Recall@K metrics, specifically Recall@1 (R@1) and Recall@10 (R@10). In the zero-shot generalization evaluation, we additionally use mean Recall (mR).

4.2. Implementation Details

For fair comparison, we follow the setup from the GeoText-1652 [21], which uses a standard XVLM [26] model pre-trained on 16 million image-text pairs as the backbone. We fine-tune our model on the GeoText-1652 dataset [21] and employ the AdamW [69] optimizer (initial learning rate 3×10^{-5}, weight decay 0.01) for 6 epochs with batch size 24.

Table 1. Comparative performance evaluation of cross-modal retrieval methods on the GeoText-1652 benchmark. Results are presented using Recall@K (R@K) for both Image Query and Text Query tasks. † denotes results are reproduced by the provided source code. The best performances are in **bold**.

Method	Params	Pretrained Images	$Q_{Image}(\%)$ R@1	$Q_{Image}(\%)$ R@10	$Q_{Text}(\%)$ R@1	$Q_{Text}(\%)$ R@10
Zero-Shot Evaluation on GeoText-1652						
UNITER [23]	300M	4M	2.5	11.8	0.9	4.2
METER-Swin [50]	380M	4M	2.7	12.2	1.3	5.8
ALBEF [25]	210M	4M	2.9	12.4	1.8	7.1
ALBEF [25]	210M	14M	3.0	14.2	1.1	5.3
XVLM [26]	216M	4M	4.9	21.1	4.3	13.2
XVLM [26]	216M	16M	5.0	21.4	4.5	13.4
Fine-Tuned Evaluation on GeoText-1652						
HyCoCLIP [1]†	216M	16M	15.3	43.2	8.7	20.0
UNITER [23]	300M	4M	21.4	59.5	10.6	26.1
METER-Swin [50]	380M	4M	22.7	60.7	11.3	27.3
ALBEF [25]	210M	4M	22.9	62.3	12.3	28.6
ALBEF [25]	210M	14M	23.2	62.4	12.5	28.5
XVLM [26]	216M	4M	23.6	63.2	13.1	29.2
XVLM [26]	216M	16M	25.0	65.1	13.2	29.6
GeoText-1652 [21]	217M	16M	26.3	66.9	13.6	31.2
HCCM	216M	16M	**28.8**	**69.9**	**14.7**	**32.5**

Table 2. Ablation study on the GeoText-1652 dataset evaluating the contribution of individual components of our proposed HCCM method. *MC* and *MD* denote momentum contrast and momentum distillation, while *RG-ITC* and *RG-ITM* represent region-global contrastive and matching learning.

Components				$Q_{Image}(\%)$		$Q_{Text}(\%)$	
MC	*MD*	*RG-ITC*	*RG-ITM*	R@1	R@10	R@1	R@10
				25.51	65.54	12.84	29.27
✓				26.48	66.78	13.75	31.49
✓	✓			26.86	67.83	14.11	31.77
		✓		27.01	67.05	13.63	30.50
		✓	✓	27.04	67.52	14.04	30.97
✓	✓	✓		27.32	68.53	14.15	31.81
✓	✓		✓	26.89	67.72	14.41	32.28
✓	✓	✓	✓	**28.82**	**69.93**	**14.73**	**32.49**

Key hyperparameters for our HCCM method are as follows: momentum $\beta = 0.995$, queue size $Q = 57,600$, distillation $\alpha = 0.4$, and temperature $\tau = 0.07$. Loss weights, determined via preliminary search, are $w_{itc} = 0.25$, $w_{itm} = 1$, $w_{rg-itc} = 0.25$, $w_{rg-itm} = 0.5$, and $w_{box} = 0.1$.

4.3. Comparison with State-of-the-art Methods

In this experiment, we compare our HCCM with existing competitive methods on the GeoText-1652 dataset under both zero-shot and fine-tuned settings. We report the results of R@1 and R@10 across all methods, alongside the model parameters and the pretrained image size used. Notably, we reproduce HyCoCLIP [1] on the NLGD task using publicly available code.

From the results shown in Table 1, we can observe that: 1) In both Image Query and Text Query settings, our method achieves the best performance, significantly outperforming other methods across all metrics. Specifically, we reach 28.8% R@1 in the Image Query setting and 14.7% R@1 in the Text Query setting. 2) Compared to the state-of-the-art method Geotext-1652 [21] of the NLGD task, our method models the hierarchical relationships across modalities and performs fine-grained features alignment, leading to superior performance. 3) HyCoCLIP [1] employs compositional entailment learning to model the part-whole hierarchical relationships, which shows limitations in drone scenarios. In contrast, our method utilizes the proposed RG-ITC and RG-ITM learning strategies, which are better at extracting the complex, intertwined spatial semantic information present in drone scenes. 4) Compared to Text Query setting, all methods perform better in Image Query setting, indicating that the text retrieval task is more challenging.

4.4. Ablation Study

We perform extensive ablation studies on the GeoText-1652 [21] dataset to evaluate the effectiveness and characteristics of the proposed HCCM components. Following [21], the standard XVLM [26] is adopted as our baseline. Our re-implementation of the baseline yields an Image Query R@1 of 25.51%, which is consistent with the performance reported in the original paper (25.0%).

Contribution of Individual Components. As shown in Table 2, we evaluate the contribution of each component. Incorporating only momentum contrast (*MC*) and momentum distillation (*MD*) improves the baseline R@1 from 25.51% to 26.86% (+1.35), indicating the value of enhancing global representation stability. Similarly, integrating only the cross-granularity learning components (*RG-ITC* and *RG-ITM*) raises the R@1 to 27.04% (+1.53), confirming the benefit of fine-grained information. By combining all components, HCCM achieves the highest performance (Image Query R@1 of 28.82%), surpassing momentum-only and cross-granularity-only configurations by 1.96 and 1.78 points, respectively. This highlights a significant synergy: robust global alignment provides a stable foundation, while cross-granularity learning contributes essential fine-grained details.

Impact of Directional Losses. Further analysis in Table 3 investigates the impact of directional losses within RG-ITC and RG-ITM. Removing any single directional loss compo-

Table 3. Ablation study of internal components in RG-ITC and RG-ITM to explore the impact of removing cross-modal directional losses on the performance of models.

Method	Q_{Image} (%)		Q_{Text} (%)	
	R@1	R@10	R@1	R@10
HCCM	28.82	69.93	14.73	32.49
$-\mathcal{L}_{RG-ITC}(I_{i,k} \rightarrow T_i)$	27.62	68.90	14.43	32.02
$-\mathcal{L}_{RG-ITC}(T_{i,k} \rightarrow I_i)$	26.89	67.72	14.41	32.28
$-\mathcal{L}_{RG-ITM}(I_{i,k} \leftrightarrow T_i)$	27.19	68.33	14.47	31.80
$-\mathcal{L}_{RG-ITM}(T_{i,k} \leftrightarrow I_i)$	26.86	67.83	14.11	31.77

Table 4. Ablation on the complementarity of same-granularity (ITC+ITM) and cross-granularity (RG-ITC+RG-ITM) learning.

Components				Q_{Image}(%)		Q_{Text}(%)	
MC	*MD*	*ITC+ITM*	*RG-ITC+RG-ITM*	R@1	R@10	R@1	R@10
✓	✓	✓		26.86	67.83	14.11	31.77
✓	✓		✓	11.23	39.08	6.33	17.66
✓	✓	✓	✓	**28.82**	**69.93**	**14.73**	**32.49**

Table 5. Ablation study on the impact of the bounding box regression loss weight (w_{box}). R@5 metrics are also included for detailed comparison.

w_{box}	Q_{Image}(%)			Q_{Text}(%)		
	R@1	R@5	R@10	R@1	R@5	R@10
0.0	29.13	56.79	69.51	14.49	25.56	32.05
0.1	28.82	**57.30**	**69.93**	**14.73**	**25.98**	**32.49**
0.2	**29.14**	56.62	69.12	14.13	25.22	31.67
0.4	28.47	56.25	68.92	14.04	25.26	31.66
0.8	27.57	55.70	68.58	13.69	24.65	30.88

Table 6. Effectiveness of MCD under extreme text perturbations (Omission and Ambiguity). Performance is measured by R@1 (%).

Model	Omission	Ambiguity	R@1 (%)	
	25%	25%	Q_{Image}	Q_{Text}
HCCM	✓		23.12	10.79
HCCM-w/oMCD	✓		22.09	9.72
GeoText1652	✓		21.19	9.11
HCCM	✓	✓	19.73	9.69
HCCM-w/oMCD	✓	✓	18.64	9.40
GeoText1652	✓	✓	17.46	8.87

Table 7. Performance evaluation on rotated test images, reported as Text Query R@1 (%).

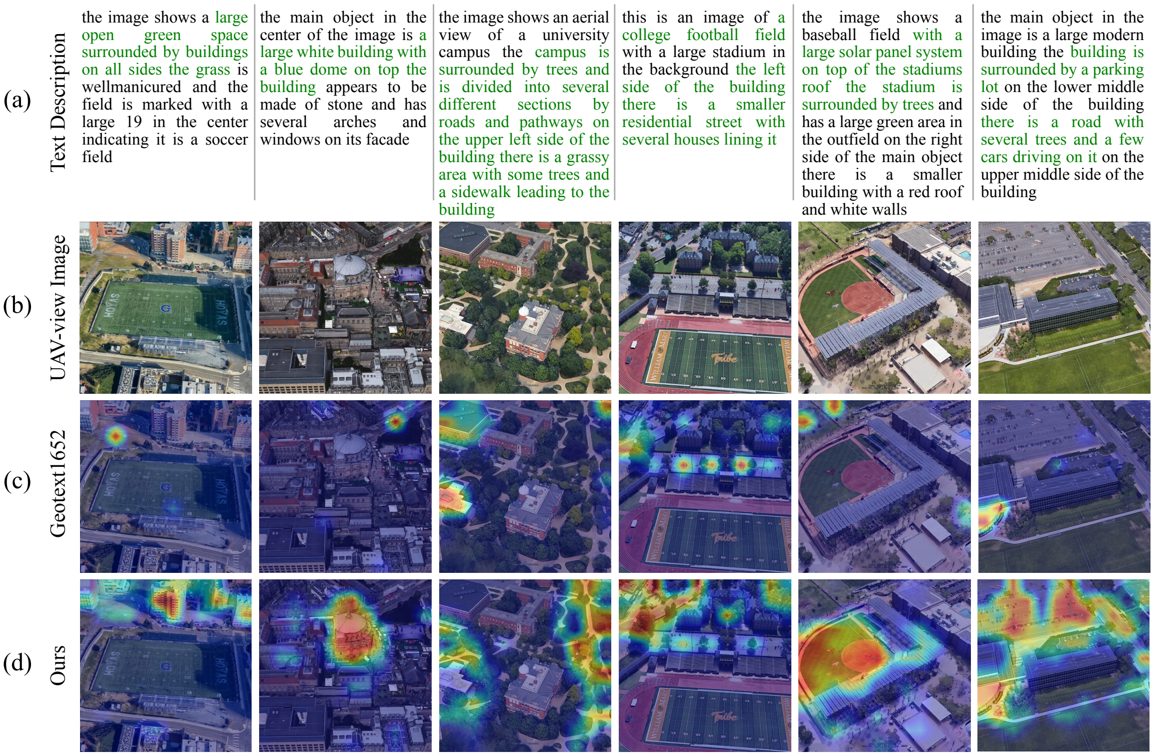
Rotate(°)	HCCM	GeoText1652
0	14.7	13.6
15	14.4	13.4
90	13.7	13.1
180	13.7	13.3
270	13.7	13.2

nent results in performance degradation. Notably, excluding the text-region to global image association in RG-ITC ($-\mathcal{L}_{RG-ITC}(T_{i,k} \rightarrow I_i)$) reduces Image Query R@1 by 1.93 points. Similarly, removing the corresponding matching loss in RG-ITM ($-\mathcal{L}_{RG-ITM}(T_{i,k} \leftrightarrow I_i)$) causes a drop of 1.96 points. This suggests that learning text-to-image associations is particularly crucial and confirms that all proposed bidirectional components contribute positively.

Complementarity of Learning Granularities. To validate the complementary nature of same- and cross-granularity approaches, we test a variant where cross-granularity learning (RG-ITC+RG-ITM) completely replaces the global modules (ITC+ITM). As shown in Table 4, this replacement leads to a severe performance degradation (Image R@1 drops from 26.86% to 11.23%). This result demonstrates that cross-granularity learning alone cannot substitute for global contextual understanding. Optimal performance is achieved only when both approaches work in synergy, confirming that cross-granularity learning effectively refines local-global relationships upon the stable foundation provided by same-granularity learning.

Impact of Bounding Box Regression Term. We analyze the impact of the bounding box regression weight, w_{box}, in Table 5. While removing the term ($w_{box} = 0.0$) boosts Image R@1 to 29.13%, a moderate weight of $w_{box} = 0.1$ yields the best overall results across Image R@10 and all Text Query metrics. This indicates that a moderate localization signal, while slightly constraining the primary image-level retrieval, is beneficial for enhancing fine-grained understanding and text-based querying.

Robustness Evaluation. We further test HCCM's robustness in two challenging scenarios. First, we evaluate the effectiveness of MCD under extreme text perturbations, where 50% of training text is corrupted by omission or ambiguity (Table 6). In both scenarios, the full HCCM model consistently outperforms its variant without MCD (HCCM-w/oMCD) and the GeoText1652 baseline, confirming that MCD provides significant resilience against noisy text. Second, we assess robustness to viewpoint changes by rotating test images (Table 7). HCCM maintains a stable and superior performance over GeoText1652 across all rotation angles, highlighting its enhanced rotational robustness.



the image shows a large open green space surrounded by buildings on all sides the grass is wellmanicured and the field is marked with a large 19 in the center indicating it is a soccer field

the main object in the center of the image is a large white building with a blue dome on top the building appears to be made of stone and has several arches and windows on its facade

the image shows an aerial view of a university campus the campus is surrounded by trees and is divided into several different sections by roads and pathways on the upper left side of the building there is a grassy area with some trees and a sidewalk leading to the building

this is an image of a college football field with a large stadium in the background the left side of the building there is a smaller residential street with several houses lining it

the image shows a baseball field with a large solar panel system on top of the stadiums roof the stadium is surrounded by trees and has a large green area in the outfield on the right side of the main object there is a smaller building with a red roof and white walls

the main object in the image is a large modern building the building is surrounded by a parking lot on the lower middle side of the building there is a road with several trees and a few cars driving on it on the upper middle side of the building

Figure 4. Visualization of activation maps. (a) Text descriptions with **key descriptions** highlighted in green. (b) Corresponding UAV-view images. (c) Activation maps from the SOTA method (GeoText-1652 [21]). (d) Activation maps from our proposed HCCM method.

Table 8. Comparison of fine-tuned results and zero-shot results on the ERA dataset. *mR* denotes mean Recall. The best performances are in **bold**.

Method	Q_{Image} (%)		Q_{Text} (%)		mR (%)
	R@1	R@10	R@1	R@10	
Reported Fine-tuned Results on ERA Dataset					
VSE++ [70]	10.13	53.91	9.79	42.90	30.39
PVSE K=2 [71]	11.04	51.65	11.31	46.95	31.52
PVSE K=1 [71]	11.14	53.75	9.96	47.97	32.14
CLIP [72]	12.73	51.52	11.31	43.91	31.45
PCME [73]	13.85	60.64	14.69	49.15	36.08
AMFMN-soft [74]	14.18	62.87	14.35	52.02	38.04
AMFMN-sim [74]	13.75	59.59	14.02	51.52	36.06
AMFMN-fusion [74]	11.62	60.51	15.20	50.33	36.15
GALR [75]	14.03	64.54	12.38	50.90	37.27
VCSR [28]	13.69	**66.37**	15.65	53.49	38.96
Zero-shot Results on ERA (Fine-tuned on GeoText-1652)					
HyCoCLIP [1]	7.77	23.31	8.04	21.96	16.22
XVLM [26]	14.19	50.34	14.05	53.99	34.70
GeoText-1652 [21]	17.91	54.73	17.09	56.76	38.00
HCCM	**19.93**	56.76	**18.58**	**59.93**	**39.93**

4.5. Zero-shot Generalization Evaluation

We assessed generalization via zero-shot cross-modal retrieval on the unseen ERA dataset [28], using models fine-tuned on GeoText-1652 [21]. Table 8 compares HCCM with other competitive methods fine-tuned on GeoText-1652 [1, 21, 26] and benchmark models fine-tuned directly on ERA [28, 70–75].

In zero-shot evaluation (Table 8, bottom), HCCM surpasses all methods fine-tuned only on GeoText-1652, achieving state-of-the-art R@1 (19.93% in image retrieval, 18.58% in text retrieval) and mR (39.93%). This notably exceeds the strong GeoText-1652 baseline [21] (+2.02% image R@1, +1.49% text R@1, +1.93% mR). Crucially, the zero-shot performance of HCCM on ERA even exceeds that of models fine-tuned directly on ERA (Table 8, top). The zero-shot mR (39.93%) of HCCM surpasses the best fine-tuned mR (38.96% by VCSR [28]). Likewise, the zero-shot R@1 scores of HCCM outperform the best respective fine-tuned R@1 results (14.18% image [74], 15.65% text [28]). The above results demonstrate that the hierarchical cross-granularity learning strategy of HCCM leads to exceptional zero-shot generalization capabilities. By generating robust transferable representations, HCCM effectively models compositional semantics to achieve powerful generalization, even surpassing models fine-tuned on the target domain.

4.6. Visualizing Attention for Semantic Grounding

We compare GradCAM [76] activation maps (Figure 4) for HCCM (row d) and the SOTA method GeoText-1652 [21] (row c) to assess fine-grained and compositional grounding against text descriptions (row a).

The SOTA method struggles to accurately ground fine-grained entities (e.g., ”blue dome” col 2, ”solar panels” col 5) and compositional relationships (e.g., ”building surrounded by” col 1, ”parking lot”/”road” layout col 6). Its diffuse attention in complex scenes (cols 3, 4) suggests difficulty parsing intricate arrangements, possibly from over-reliance on global matching. Conversely, HCCM shows significantly improved semantic grounding, localizing fine-grained details (”blue dome” col 2, ”solar panels” col 5) and better capturing compositional semantics: relating fields to ”surrounding buildings” (col 1), reflecting campus structure via ”roads and pathways” (col 3), linking fields to ”residential streets” (col 4), and grounding the building/”parking lot”/”road” context (col 6). This improved relational grounding, consistent with RG-ITC and RG-ITM objectives, is likely aided by RG-ITM’s consistency evaluation. Notably, in scenes with large uniform areas (e.g., fields in cols 1, 3), neither model strongly activates the primary object. This might occur because distinguishing such scenes relies more on grounding discriminative textual descriptions of contrasting features, boundaries, or relationships than on the homogenous central region.

In summary, HCCM’s hierarchical modeling (RG-ITC) and consistency evaluation (RG-ITM) yield more precise semantic grounding than global alignment methods alone. Its enhanced relational interpretation benefits complex NLGD tasks, especially in drone-view scenarios.

5. Conclusion

This paper introduces HCCM to enhance fine-grained and compositional understanding in Natural Language-Guided Drone (NLGD) tasks. By integrating Region-Global Contrastive (RG-ITC) and Matching (RG-ITM) learning, HCCM effectively models hierarchical cross-modal semantics and evaluates local-global consistency without requiring strict entity partitioning. Furthermore, a Momentum Contrast and Distillation (MCD) mechanism stabilizes global alignment against ambiguous descriptions common in drone views. Extensive experiments validate HCCM’s effectiveness, demonstrating state-of-the-art performance on the GeoText-1652 benchmark and superior zero-shot generalization on the ERA dataset, highlighting its robustness for complex drone vision-language applications.

References

[1] Avik Pal, Max van Spengler, Guido Maria D’Amely di Melendugno, Alessandro Flaborea, Fabio Galasso, and Pascal Mettes. Compositional entailment learning for hyperbolic vision-language models. 2025.

[2] Jorge Gago, Cyril Douthe, Rafael E Coopman, Pedro Pablo Gallego, Miquel Ribas-Carbo, Jaume Flexas, Jb Escalona, and Hb Medrano. Uavs challenge to assess water stress for sustainable agriculture. *Agricultural water management*, pages 9–19, 2015.

[3] 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. Sustainable agriculture by increasing nitrogen fertilizer efficiency using low-resolution camera mounted on unmanned aerial vehicles. *International Journal of Environmental Research and Public Health*, (20):3893, 2019.

[4] Paolo Tripicchio, Massimo Satler, Giacomo Dabisias, Emanuele Ruffaldi, and Carlo Alberto Avizzano. Towards smart farming and sustainable agriculture with drones. In *2015 international conference on intelligent environments*, pages 140–143. IEEE, 2015.

[5] E Wardihani, Magfur Ramdhani, Amin Suharjono, Thomas Agung Setyawan, Sidiq Syamsul Hidayat, Sarono Widodo Helmy, Eddy Triyono, and FIRDANIS Saifullah. Real-time forest fire monitoring system using unmanned aerial vehicle. *Journal of Engineering Science and Technology*, (6):1587–1594, 2018.

[6] Pengyu Zhang, Jie Zhao, Dong Wang, Huchuan Lu, and Xiang Ruan. Visible-thermal uav tracking: A large-scale benchmark and new baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8886–8895, 2022.

[7] Mei Chen, Xiaoyan Wang, Hong Wang, and Shufang Zhao. A uav-based energy-efficient and real-time object detection system with multi-source image fusion. *Journal of Circuits, Systems and Computers*, 31(09):2250166, 2022.

[8] Paraskevi Nousi, Ioannis Mademlis, Iason Karakostas, Anastasios Tefas, and Ioannis Pitas. Embedded uav real-time visual object detection and tracking. In *2019 IEEE International Conference on Real-time Computing and Robotics*, pages 708–713, 2019.

[9] Kenneth Chaney, Fernando Cladera, Ziyun Wang, Anthony Bisulco, M Ani Hsieh, Christopher Korpela, Vijay Kumar, Camillo J Taylor, and Kostas Daniilidis. M3ED: Multi-robot, multi-sensor, multi-environment event dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4016–4023, 2023.

[10] Lingdong Kong, Dongyue Lu, Xiang Xu, Lai Xing Ng, Wei Tsang Ooi, and Benoit R. Cottereau. EventFly: Event camera perception from ground to the sky. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1472–1484, 2025.

[11] Xiaoshuai Hao et al. SafeMap: Robust HD map construction from incomplete observations. In *International Conference on Machine Learning*, pages 22091–22102. PMLR, 2025.

[12] Xiaoshuai Hao, Mengchuan Wei, Yifan Yang, et al. Is your HD map constructor reliable under sensor corruptions? In *Advances in Neural Information Processing Systems*, volume 37, pages 22441–22482, 2024.

[13] Xiaoshuai Hao, Guanqun Liu, Yuting Zhao, et al. MSC-Bench: Benchmarking and analyzing multi-sensor corruption for driving perception. *arXiv preprint arXiv:2501.01037*, 2025.

[14] 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*, page 1395–1403, 2020.

[15] 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. *IEEE Transactions on Image Processing*, 31:3780–3792, 2022.

[16] Rong Li, Yuhao Dong, Tianshuai Hu, Ao Liang, et al. 3EED: Ground everything everywhere in 3D. *arXiv preprint arXiv:2511.01755*, 2025.

[17] Tingyu Wang, Zhedong Zheng, Yaoqi Sun, Chenggang Yan, Yi Yang, and Tat-Seng Chua. Multiple-environment self-adaptive network for aerial-view geo-localization. *Pattern Recognition*, 152:110363, 2024.

[18] Jinliang Lin, Zhiming Luo, Dazhen Lin, Shaozi Li, and Zhun Zhong. A self-adaptive feature extraction method for aerial-view geo-localization. *IEEE Transactions on Image Processing*, 2024.

[19] Ye Li et al. Optimizing LiDAR placements for robust driving perception in adverse conditions. *arXiv preprint arXiv:2403.17009*, 2024.

[20] Pengfei Wei et al. Unsupervised video domain adaptation for action recognition: A disentanglement perspective. In *Advances in Neural Information Processing Systems*, volume 36, pages 17623–17642, 2023.

[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 *European Conference on Computer Vision*, pages 213–231, 2024.

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 8748–8763, 2021.

[23] 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 *European conference on computer vision*, pages 104–120, 2020.

[24] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 12888–12900, 2022.

[25] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *Advances in Neural Information Processing Systems*, volume 34, pages 9694–9705, 2021.

[26] Yan Zeng, Xinsong Zhang, and Hang Li. Multi-grained vision language pre-training: Aligning texts with visual concepts. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 25994–26009, 2022.

[27] Kejia Zhang, Keda Tao, Jiasheng Tang, and Huan Wang. Poison as cure: Visual noise for mitigating object hallucinations in lvms. *arXiv preprint arXiv:2501.19164*, 2025.

[28] Jinghao Huang, Yaxiong Chen, Shengwu Xiong, and Xiaoqiang Lu. Visual contextual semantic reasoning for cross-modal drone image-text retrieval. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–12, 2024.

[29] Guoli Wang, Bin Fan, Shiming Xiang, and Chunhong Pan. Aggregating rich hierarchical features for scene classification in remote sensing imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10(9):4104–4115, 2017.

[30] Anamaria Radoi and Mihai Datcu. Multilabel annotation of multispectral remote sensing images using error-correcting output codes and most ambiguous examples. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2121–2134, 2019.

[31] Zhengyuan Zhang, Wenkai Zhang, Wenhui Diao, Menglong Yan, Xin Gao, and Xian Sun. Vaa: Visual aligning attention model for remote sensing image captioning. *IEEE Access*, 7:137355–137364, 2019.

[32] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Conference on Empirical Methods in Natural Language Processing*, pages 1724–1734, 2014.

[33] Qian Liu, Shengnan An, Jian-Guang Lou, Bei Chen, Zeqi Lin, Yan Gao, Bin Zhou, Nanning Zheng, and Dongmei Zhang. Compositional generalization by learning analytical expressions. *Advances in Neural Information Processing Systems*, 33:11416–11427, 2020.

[34] Matthew Le, Stephen Roller, Laetitia Papaxanthos, Douwe Kiela, and Maximilian Nickel. Inferring concept hierarchies from text corpora via hyperbolic embeddings. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3231–3241, 2019.

[35] Octavian Ganea, Gary Becigneul, and Thomas Hofmann. Hyperbolic entailment cones for learning hierarchical embeddings. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 1646–1655, 2018.

[36] Devansh Arpit, Stanisław Jastrzkebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, and Simon Lacoste-Julien. A closer look at memorization in deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 233–242, 2017.

[37] Zeying Gong et al. Stairway to success: An online floor-aware zero-shot object-goal navigation framework via LLM-driven coarse-to-fine exploration. *arXiv preprint arXiv:2505.23019*, 2025.

[38] Junyan Ye, Honglin Lin, Leyan Ou, Dairong Chen, Zihao Wang, Conghui He, and Weijia Li. Where am i? cross-view geo-localization with natural language descriptions. *arXiv preprint arXiv:2412.17007*, 2024.

[39] Yan Xia, Zhendong Li, Yun-Jin Li, Letian Shi, Hu Cao, João F. Henriques, and Daniel Cremers. Uniloc: Towards universal place recognition using any single modality. *arXiv preprint arXiv:2412.12079*, 2024.

[40] Jingyi Xu et al. Visual foundation models boost cross-modal unsupervised domain adaptation for 3D semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 26(11):20287–20301, 2025.

[41] Xidong Peng, Runnan Chen, Feng Qiao, et al. Learning to adapt SAM for segmenting cross-domain point clouds. In *European Conference on Computer Vision*, pages 54–71. Springer, 2024.

[42] Youquan Liu et al. Segment any point cloud sequences by distilling vision foundation models. In *Advances in Neural Information Processing Systems*, volume 36, pages 37193–37229, 2023.

[43] Runnan Chen et al. CLIP2Scene: Towards label-efficient 3D scene understanding by CLIP. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7020–7030, 2023.

[44] Runnan Chen et al. Towards label-free scene understanding by vision foundation models. In *Advances in Neural Information Processing Systems*, volume 36, pages 75896–75910, 2023.

[45] Lingdong Kong et al. LargeAD: Large-scale cross-sensor data pretraining for autonomous driving. *arXiv preprint arXiv:2501.04005*, 2025.

[46] Youquan Liu et al. Multi-space alignments towards universal LiDAR segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14648–14661, 2024.

[47] Lingdong Kong, Xiang Xu, Jiawei Ren, et al. Multi-modal data-efficient 3D scene understanding for autonomous driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):3748–3765, 2025.

[48] Lingdong Kong, Wesley Yang, Jianbiao Mei, Youquan Liu, Ao Liang, Dekai Zhu, Dongyue Lu, Wei Yin, Xiaotao Hu, Mingkai Jia, Junyuan Deng, Kaiwen Zhang, Yang Wu, Tianyi Yan, Shenyuan Gao, Song Wang, Linfeng Li, Liang Pan, Yong Liu, Jianke Zhu, Wei Tsang Ooi, Steven C. H. Hoi, and Ziwei Liu. 3D and 4D world modeling: A survey. *arXiv preprint arXiv:2509.07996*, 2025.

[49] Xiang Xu et al. 4D contrastive superflows are dense 3D representation learners. In *European Conference on Computer Vision*, pages 58–80, 2024.

[50] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, et al. 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*, pages 18166–18176, 2022.

[51] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 19730–19742, 2023.

[52] Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. Order-embeddings of images and language. *arXiv preprint arXiv:1511.06361*, 2015.

[53] Kim Anh Nguyen, Maximilian Köper, Sabine Schulte im Walde, and Ngoc Thang Vu. Hierarchical embeddings for hypernymy detection and directionality. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 233–243, 2017.

[54] Ivan Vulić and Nikola Mrkšić. Specialising word vectors for lexical entailment. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1134–1145, 2018.

[55] Pablo Arbeláez, Michael Maire, Charless Fowlkes, and Jitendra Malik. Contour detection and hierarchical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(5):898–916, 2011.

[56] Xiao Zhang and Michael Maire. Self-supervised visual representation learning from hierarchical grouping. In *Advances in Neural Information Processing Systems*, volume 33, pages 16579–16590, 2020.

[57] Jiahao Xie, Xiaohang Zhan, Ziwei Liu, Yew Soon Ong, and Chen Change Loy. Unsupervised object-level representation learning from scene images. In *Advances in Neural Information Processing Systems*, volume 34, pages 28864–28876, 2021.

[58] Lingdong Kong, Shaoyuan Xie, Zeying Gong, Ye Li, Meng Chu, Ao Liang, Yuhao Dong, Tianshuai Hu, Ronghe Qiu, Rong Li, Hanjiang Hu, Dongyue Lu, Wei Yin, Wenhao Ding, Linfeng Li, Hang Song, Wenwei Zhang, Yuexin Ma, Junwei Liang, Zhedong Zheng, Lai Xing Ng, Benoit R. Cottereau, Wei Tsang Ooi, Ziwei Liu, Zhanpeng Zhang, Weichao Qiu, Wei Zhang, Ji Ao, Jiangpeng Zheng, Siyu Wang, Guang Yang, Zihao Zhang, Yu Zhong, Enzhu Gao, Xinhan Zheng, Xueting Wang, Shouming Li, Yunkai Gao, Siming Lan, Mingfei Han, Xing Hu, Dusan Malic, Christian Fruhwirth-Reisinger, Alexander Prutsch, Wei Lin, Samuel Schulter, Horst Possegger, Linfeng Li, Jian Zhao, Zepeng Yang, Yuhang Song, Bojun Lin, Tianle Zhang, Yuchen Yuan, Chi Zhang, Xuelong Li, Youngseok Kim, Sihwan Hwang, Hyeonjun Jeong, Aodi Wu, Xubo Luo, Erjia Xiao, Lingfeng Zhang, Yingbo Tang, Hao Cheng, Renjing Xu, Wenbo Ding, Lei Zhou, Long Chen, Hangjun Ye, Xiaoshuai Hao, Shuangzhi Li, Junlong Shen, Xingyu Li, Hao Ruan, Jinliang Lin, Zhiming Luo, Yu Zang, Cheng Wang, Hanshi Wang, Xijie Gong, Yixiang Yang, Qianli Ma, Zhipeng Zhang, Wenxiang Shi, Jingmeng Zhou, Weijun Zeng, Kexin Xu, Yuchen Zhang, Haoxiang Fu, Ruibin Hu, Yanbiao Ma, Xiyan Feng, Wenbo Zhang, Lu Zhang, Yunzhi Zhuge, Huchuan Lu, You He, Seungjun Yu, Junsung Park, Youngsun Lim, Hyunjung Shim, Faduo Liang, Zihang Wang, Yiming Peng, Guanyu Zong, Xu Li, Binghao Wang, Hao Wei, Yongxin Ma, Yunke Shi, Shuaipeng Liu, Dong Kong, Yongchun Lin, Huitong Yang, Liang Lei, Haoang Li, Xinliang Zhang, Zhiyong Wang, Xiaofeng Wang, Yuxia Fu, Yadan Luo, Djamahl Etchegaray, Yang Li, Congfei Li, Yuxiang Sun, Wenkai Zhu, Wang Xu, Linru Li, Longjie Liao, Jun Yan, Benwu Wang, Xueliang Ren, Xiaoyu Yue, Jixian Zheng, Jinfeng Wu, Shurui Qin, Wei Cong, and Yao He. The RoboSense

challenge: Sense anything, navigate anywhere, adapt across platforms. https://robosense2025.github.io, 2025.

[59] Lingdong Kong, Yaru Niu, Shaoyuan Xie, Hanjiang Hu, Lai Xing Ng, Benoit Cottereau, Liangjun Zhang, Hesheng Wang, Wei Tsang Ooi, Ruijie Zhu, Ziyang Song, Li Liu, Tianzhu Zhang, Jun Yu, Mohan Jing, Pengwei Li, Xiaohua Qi, Cheng Jin, Yingfeng Chen, Jie Hou, Jie Zhang, Zhen Kan, Qiang Lin, Liang Peng, Minglei Li, Di Xu, Changpeng Yang, Yuanqi Yao, Gang Wu, Jian Kuai, Xianming Liu, Junjun Jiang, Jiamian Huang, Baojun Li, Jiale Chen, Shuang Zhang, Sun Ao, Zhenyu Li, Runze Chen, Haiyong Luo, Fang Zhao, and Jingze Yu. The RoboDepth challenge: Methods and advancements towards robust depth estimation. *arXiv preprint arXiv:2307.15061*, 2023.

[60] Lingdong Kong, Shaoyuan Xie, Hanjiang Hu, Lai Xing Ng, Benoit R. Cottereau, and Wei Tsang Ooi. RoboDepth: Robust out-of-distribution depth estimation under corruptions. In *Advances in Neural Information Processing Systems*, volume 36, pages 21298–21342, 2023.

[61] Lingdong Kong, Shaoyuan Xie, Hanjiang Hu, Yaru Niu, Wei Tsang Ooi, Benoit R. Cottereau, Lai Xing Ng, Yuexin Ma, Wenwei Zhang, Liang Pan, Kai Chen, Ziwei Liu, Weichao Qiu, Wei Zhang, Xu Cao, Hao Lu, Ying-Cong Chen, et al. The RoboDrive challenge: Drive anytime anywhere in any condition. *arXiv preprint arXiv:2405.08816*, 2024.

[62] Shaoyuan Xie, Lingdong Kong, Wenwei Zhang, Jiawei Ren, Liang Pan, Kai Chen, and Ziwei Liu. Benchmarking and improving bird's eye view perception robustness in autonomous driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):3878–3894, 2025.

[63] Shaoyuan Xie, Lingdong Kong, Yuhao Dong, Chonghao Sima, Wenwei Zhang, Qi Alfred Chen, Ziwei Liu, and Liang Pan. Are VLMs ready for autonomous driving? an empirical study from the reliability, data, and metric perspectives. In *IEEE/CVF International Conference on Computer Vision*, pages 6585–6597, 2025.

[64] Zeying Gong, Tianshuai Hu, Ronghe Qiu, and Junwei Liang. From cognition to precognition: A future-aware framework for social navigation. In *IEEE International Conference on Robotics and Automation*, pages 9122–9129, 2025.

[65] Ye Li, Lingdong Kong, Hanjiang Hu, Xiaohao Xu, and Xiaonan Huang. Is your LiDAR placement optimized for 3D scene understanding? In *Advances in Neural Information Processing Systems*, volume 37, pages 34980–35017, 2024.

[66] 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 *European Conference on Computer Vision*, pages 213–231, 2024.

[67] Ao Liang et al. Perspective-invariant 3D object detection. In *IEEE/CVF International Conference on Computer Vision*, pages 27725–27738, 2025.

[68] Lingdong Kong, Youquan Liu, Xin Li, Runnan Chen, Wenwei Zhang, Jiawei Ren, Liang Pan, Kai Chen, and Ziwei Liu. Robo3D: Towards robust and reliable 3D perception against corruptions. In *IEEE/CVF International Conference on Computer Vision*, pages 19994–20006, 2023.

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

[70] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. In *Proceedings of the British Machine Vision Conference*, 2018.

[71] Yale Song and Mohammad Soleymani. Polysemous visual-semantic embedding for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[72] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 8748–8763, 2021.

[73] Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio de Rezende, Yannis Kalantidis, and Diane Larlus. Probabilistic embeddings for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8415–8424, 2021.

[74] Zhiqiang Yuan, Wenkai Zhang, Kun Fu, Xuan Li, Chubo Deng, Hongqi Wang, and Xian Sun. Exploring a fine-grained multiscale method for cross-modal remote sensing image retrieval. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–19, 2022.

[75] Zhiqiang Yuan, Wenkai Zhang, Changyuan Tian, Xuee Rong, Zhengyuan Zhang, Hongqi Wang, Kun Fu, and Xian Sun. Remote sensing cross-modal text-image retrieval based on global and local information. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–16, 2022.

[76] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.