Seeing with Words: Interpretable Language-Guided Drone Geo-localization via LLM-Enriched Semantic Attribute Alignment

Changsen Yuan, Yang-Hao Zhou*, *Student Member, IEEE*, Cunhan Guo, Danjie Han, Ge Shi, and Wenwu Wang, *Senior Member, IEEE*

***Abstract*—Natural language-guided drone geo-localization (DGL) provides an intuitive and scalable mode of human-drone interaction for tasks such as search, rescue, and surveillance. Recent Vision-Language Models (VLMs) can learn semantic correspondences between text and images during fine-tuning. However, their performance in DGL tasks remains constrained, as complex instructions and cluttered scenes often cause semantic dilution and granularity mismatch, leading to weak cross-modal alignment. Consequently, the models struggle with ambiguous targets and suffer from reduced localization accuracy. To address these challenges, we propose SAA-DGL, a framework for interpretable language-guided Drone Geo-Localization that enriches Semantic Attribute Alignment (SAA) with large language models (LLMs). It introduces two parameter-free cross-modal fusion modules: (1) the LLM-driven Cross-modal Semantic Attribute Enrichment (LCSAE) module, which extracts fine-grained attributes (e.g., color, shape, position) from text and embeds them into visual features as explicit semantic anchors, producing semantically enriched cross-modal representations; and (2) the Bidirectional Feature Alignment (BFA) module, which builds fusion relationships between visual and textual features via similarity-driven mechanisms, enabling effective integration of enriched visual and textual information. This design improves cross-modal consistency and interpretability while preserving pretrained alignment priors and enhancing training stability. Experiments on the GeoText-1652 benchmark show that SAA-DGL achieves state-of-the-art performance and strong robustness under complex visual and linguistic disturbances, validating its effectiveness for challenging geo-localization scenarios.**

***Index Terms*—Drone Geo-localization, Cross-modal Alignment, Cross-modal Feature Enhancement.**

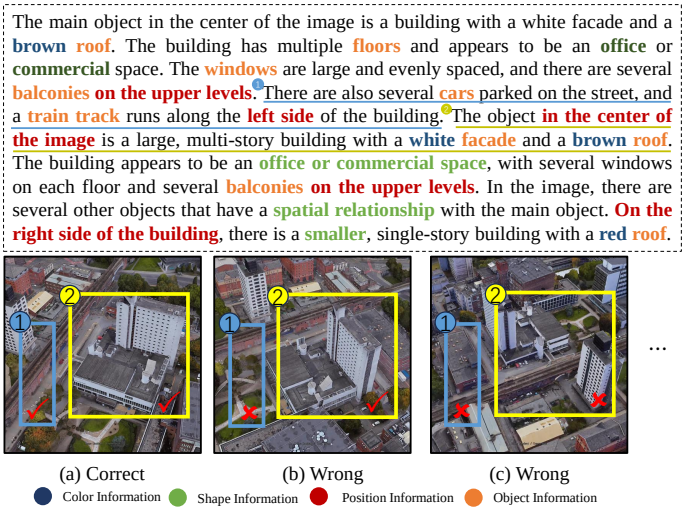


Fig. 1. Example of natural language-guided DGL. Orange words indicate object information, red words indicate location information, dark blue words indicate color information, and dark green words indicate shape information. The yellow line indicates the information of the object within the yellow box, while the dark blue line indicates the description of the blue box. ✘ denotes the wrong identification, and ✔ denotes the correct identification.

I. Introduction

NATURAL language-guided **d**rone **g**eo-**l**ocalization (DGL) is an emerging technology that translates human instructions into spatial semantic actions, enabling drones to comprehend and execute geo-localization tasks autonomously. This capability lays the foundation for more complex downstream applications and holds great promise in various fields, including disaster management [1], live search and rescue [2], and power line inspection [3]. Most existing DGL research is typically regarded as a sub-task of visual retrieval, with a primary focus on drone-view image-based search [4], [5]. In this setting, a query image is used to retrieve target images from a large-scale gallery, often involving cross-view matching across platforms such as drones and satellites. However, in real-world applications, obtaining a query image is often costly or impractical. In practice, users prefer to interact with drones via natural language, which serves as an effective modality for specifying and retrieving spatial targets in complex environments. In recent years, multimodal foundation models have made remarkable progress in text-related multimodal tasks, showcasing strong capabilities in cross-modal alignment [6]–[9]. To enhance human-drone interaction in natural language settings, Chu *et al.* [10] proposed GeoText-1652, a benchmark for natural language-guided drone geo-localization. Built on University-1652 [5], it incorporates fine-grained textual descriptions and spatial alignments across multi-view imagery, enabling effective instruction-to-region matching. However, GeoText-1652 struggles with real-world instructions that are

*Yang-Hao Zhou is the corresponding author.

Changsen Yuan is with Beijing University of Technology, Beijing, 100124, China. Email: yuanchangsen0@gmail.com.

Yang-Hao Zhou, Cunhan Guo and Ge Shi are with Beijing Institute of Technology, Beijing, 100811, China. Email: zhouyh77@bit.edu.cn;guocunhan@bit.edu.cn;gshi@bit.edu.cn.

Danjie Han is with Nanjing University of Science and Technology, Nanjing, 210094, China. Email: handanjie@njust.edu.cn.

Wenwu Wang is with the School of Computer Science and Electronic Engineering, University of Surrey, Guildford, GU2 7XH, UK. Email: w.wang@surrey.ac.uk.

rich in object semantics and compositional attributes. For instance, the instruction "`On the right side of the building, there is a smaller, single-story building with a red roof`" combines spatial cues with semantic details such as color and structure. These natural, attribute-based expressions require DGL systems to accurately parse and ground compositional semantics. As shown in Fig. 1, the system must retrieve the most relevant region from a large-scale drone image gallery based on a language query, demanding high precision and robust cross-modal alignment. Despite these advances, natural language-guided DGL still faces two key challenges: (1) **Dilution of key semantic information** within lengthy, attribute-rich instructions, and (2) **Mismatch of semantic granularity** between linguistic descriptions and visual representations. These issues weaken cross-modal alignment and ultimately limit localization performance.

Dilution of key semantic information. In drone navigation scenarios, natural language instructions often bury critical target cues within lengthy descriptions. As shown in Fig.1, the core references (highlighted) are typically surrounded by redundant background information, thereby introducing noise and weakening cross-modal discriminability. Existing methods such as keyword extraction [10] overlook compositional semantics, while fixed-length truncation [11] may discard crucial information, both of which limit precise alignment in real-world scenarios. **Mismatch of semantic granularity.** Language descriptions usually progress from high-level categories to fine-grained attributes (e.g., "building" $\rightarrow$ "brown roof"), whereas visual encoders tend to capture low-level features such as textures and edges. This inconsistency is particularly pronounced in multi-view scenarios: the same scene observed from different viewpoints may cause the model to incorrectly match text with semantically inaccurate but visually similar images, such as selecting (b) or (c) instead of the correct (a) in Fig. 1. This discrepancy significantly increases the difficulty of learning robust cross-modal correspondences. These challenges highlight the necessity of developing mechanisms capable of robustly identifying and aligning fine-grained semantic attributes within complex linguistic inputs.

To address the aforementioned challenges, we propose the **S**emantic-**A**ttribute **A**lignment approach for **D**rone **G**eo-**L**ocalization (SAA-DGL). To mitigate semantic dilution, we introduce a LLMs-driven Cross-modal Semantic Attribute Enrichment (LCSAE) module based on attribute-guided feature enhancement. Specifically, large language models (LLMs) are used to extract referential semantic attributes from natural language instructions, including color, shape, position, and object category. These attributes capture the visual intent embedded in textual descriptions and help reduce ambiguity and redundancy. By integrating them with visual features, the model achieves improved representational capacity and contextual grounding, enabling more precise alignment with drone-view observations. To address the mismatch in semantic granularity, we propose a Bidirectional Feature Alignment (BFA) module. This module first refines visual representations by aligning them with fine-grained semantic attributes from text, allowing the model to focus on spatially relevant regions. These enhanced visual features are then used to further guide textual representation, forming a dynamic bidirectional interaction. This process enhances cross-modal consistency and discriminability, effectively narrowing the semantic gap between language and vision. The main contributions of this paper are summarized as follows:

- We propose a novel framework, **SAA-DGL**, for natural language-guided DGL. It incorporates a LCSAE module that leverages LLMs to extract enriched target attributes from textual descriptions, which are explicitly used to guide fine-grained visual feature extraction and obtain more relevant visual semantics.
- We design a BFA module that establishes a closed-loop interaction between visual and textual modalities, improving cross-modal consistency and alleviating semantic granularity mismatches.
- Extensive experiments on natural language-guided DGL benchmarks demonstrate that our method achieves state-of-the-art performance and robustness on both text-to-image and image-to-text tasks.

II. Related Work

A. Drone Geo-localization

In recent years, intelligent drone agents equipped with advanced cameras have been increasingly deployed worldwide. Compared with traditional static surveillance systems, drones offer dynamic coverage, autonomous mobility, and flexible control of viewpoints and altitudes, enabling comprehensive area monitoring. These advantages have spurred diverse applications based on drone-view visual data [12], [13]. Among them, drone geo-localization [14] is a key capability for human-drone interaction and has attracted considerable research attention. Cross-view geo-localization has evolved from early CNN-based method [15] to Transformer-based approach incorporating attention and geometric reasoning [16], and more recently to CLIP-driven contrastive learning frameworks leveraging large-scale vision-language pretraining [17]. These developments have substantially enhanced the robustness and generalization of cross-view matching.

University-1652 [5] introduced drone-view imagery into the cross-view localization task, supporting image-to-image matching between drone and satellite views. DenseUAV [18] improves robustness by densely sampling urban scenes at low altitudes, while SUES-200 [19] increases viewpoint diversity through multi-altitude captures. GTA-UAV [20] synthesizes a large-scale, continuous drone dataset using a video game engine, facilitating evaluation under complex and realistic conditions. To better address dynamic environments, Video2BEV [21] transforms drone video sequences into Bird's Eye View (BEV) representations, capturing motion and spatial consistency for stable localization. GeoText-1652 [10] introduces a text-guided framework that integrates language, vision, and geolocation for interactive navigation. Although GeoText-1652 achieves promising text–image alignment, it still suffers from semantic dilution during cross-view transitions, which limits the extraction of key object attributes. To overcome this limitation, we propose SAA-DGL, a framework that

enriches and aligns semantic attributes (e.g., color, shape, and position) with visual features. In contrast to prior work that mainly emphasizes global alignment or contrastive objectives, our approach focuses on fine-grained attribute–image fusion, leading to improved robustness and interpretability in both text-to-image and image-to-text geo-localization.

B. Natural Language-guided Navigation

Natural language-guided navigation enables agents, such as robots or drones, to follow human instructions in unknown or partially known environments. Early research, particularly in indoor settings, was conducted under the Vision-and-Language Navigation (VLN) paradigm [22], where sequence-to-sequence models [23] were used to integrate linguistic, visual, and spatial information. Fried *et al.* [24] proposed the Speaker-Follower model using reinforcement learning. Recent advances leverage pre-trained language models to improve instruction grounding [25]. Vision-language pretraining [26], time-aware recurrent models [27], graph-based planning [28], and fine-grained alignment [29] have all shown promise in navigation tasks. Talk2Nav [30] incorporated dual attention and spatial memory to improve real-world performance. Xu *et al.* [31] proposed a factor graph-based sensor fusion framework for robust real-time navigation in unstructured environments. Zhou *et al.* [32] combined a frozen LLM with a policy network to enhance interpretability while maintaining competitive performance. In outdoor scenarios, DroneVLN [11] adapted the VLN-BERT model [27] to unmanned aerial vehicle (UAV) navigation using a pre-trained multimodal encoder to align language instructions with visual observations. With the advent of LLMs, recent efforts have explored their role in UAV navigation [33]. Gao *et al.* [34] enabled drones to follow instructions without task-specific data, improving adaptability in dynamic or data-scarce environments. Despite these advances, current methods still lack fine-grained alignment between textual semantics and drone-view perception, and remain fragile in complex outdoor settings where diverse viewpoints, environmental disturbances, and noisy instructions often lead to misalignment. To address these challenges, we explicitly model referential attributes such as color, shape, spatial position, and object identity, enabling more robust semantic grounding and reliable target localization under realistic conditions.

C. Visual-Language Modality Alignment

Visual-language modality alignment aims to establish precise correspondences between textual semantics and visual content, supporting various multimodal tasks such as image-text retrieval. Prior work has focused on enhancing fine-grained cross-modal alignment through adaptive gating [35], word-region matching [36], and object tag anchoring [37]. The advent of foundation models has led to large-scale cross-modal pretraining frameworks such as CLIP [38], which align image-text pairs in a shared embedding space. Jia *et al.* [6] proposed a dual-encoder architecture, and Li *et al.* [39] adopted pseudo-target-based self-training to enhance retrieval accuracy. BLIP [40] improved robustness by guiding the generation of high-quality captions, while Fei *et al.* [41] introduced unified scene graph representations to improve spatial and temporal alignment. Natural language-guided drone geo-localization is a fine-grained cross-modal task related to image-text retrieval. Chu *et al.* [10] were the first to apply the multi-granularity alignment model X-VLM [42] to this task. APTM [43] focuses on person retrieval with 27 manually predefined attributes and a parameterized cross-encoder for attribute–image alignment. In contrast, our work addresses drone geo-localization in large-scale environments by using LLMs to automatically extract open, fine-grained attributes from natural language without relying on fixed attribute sets. Despite recent progress, most methods overlook the influence of drone-viewpoint variations on object appearance and spatial semantics, often leading to misalignment. We also address this issue by introducing a semantic attribute alignment strategy that explicitly enhances cross-modal fusion without introducing additional trainable parameters. Different from existing approaches that rely on learnable query vectors or deep parameterized stacks, our modules perform single-pass, similarity-driven interactions. This design ensures efficiency and stability in limited-data scenarios while preserving the region–concept alignment priors already encoded in the pretrained backbone, ultimately improving robustness and interpretability in drone-view geo-localization.

III. Proposed Method

A. Problem Formulation and Task Definition

Natural language-guided DGL enables autonomous drones to interpret free-form language instructions and localize target regions within a geo-referenced drone-view image gallery. This task involves both semantic understanding and spatial reasoning, and is typically formulated as a cross-modal retrieval problem due to viewpoint variations and scene ambiguity. To support similarity-based matching, language and visual inputs are embedded in a shared space. We define two symmetric subtasks: **Text-to-Drone-view Image Retrieval (T2I)** and **Drone-view Image-to-Text Retrieval (I2T)**, where each instruction $t = (c, s)$ includes a global caption c and local sentences s for hierarchical alignment with corresponding drone-view images and subregions. Given an instruction t_i, the goal of **T2I** is to retrieve the most semantically and spatially aligned image from gallery $\mathcal{G}$:

$$f_{\text{T2I}}(t_i, \mathcal{G}) = \arg\max_{g \in \mathcal{G}} \text{Sim}(E_T(t_i), E_I(g)), \tag{1}$$

where E_T and E_I are the text and image encoders, respectively, and $\text{Sim}(\cdot, \cdot)$ denotes the similarity function in the shared space. Given a drone-view image g_j, the **I2T** system aims to retrieve the most relevant instruction from text $\mathcal{T}$:

$$f_{\text{I2T}}(g_j, \mathcal{T}) = \arg\max_{t \in \mathcal{T}} \text{Sim}(E_I(g_j), E_T(t)). \tag{2}$$

This dual retrieval setup enables intuitive verification of alignment between drone perception and human language, which is critical for interactive navigation tasks.

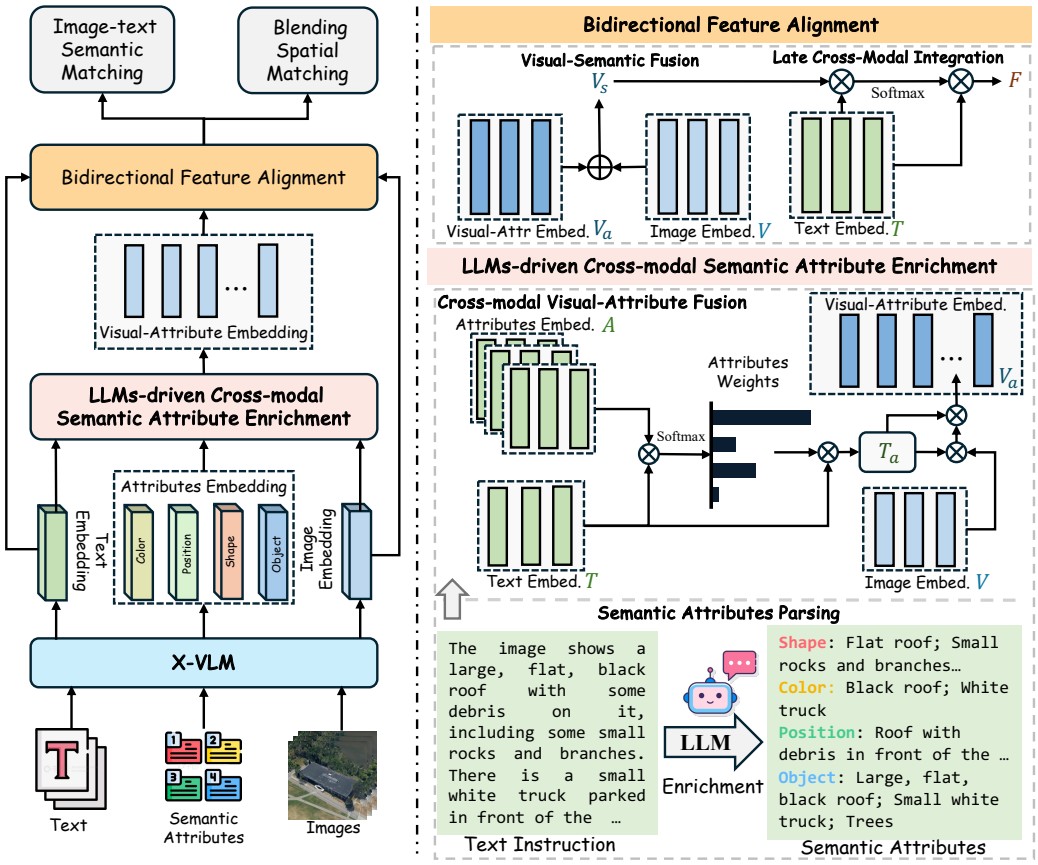


Fig. 2. Overview of the proposed SAA-DGL framework. The model consists of three main components: (1) **Semantic Attributes Parsing**, where an LLM extracts fine-grained semantic attributes (e.g., shape, color, location, object) from the input natural language description; (2) **LLMs-driven Cross-modal Semantic Attribute Enrichment (LCSAE)**, where semantic attributes are explicitly fused with textual and visual embeddings to enhance semantic grounding in the visual domain; and (3) **Bidirectional Feature Alignment (BFA)**, which establishes cross-modal feature interaction through visual-semantic fusion and late cross-modal integration, improving feature consistency across modalities. Notably, the figure distinguishes between the initial extraction of semantic attributes from text (bottom) and the subsequent attribute-guided feature enhancement (middle), although both involve semantic attribute representations.

B. Overview

Fig. 2 illustrates an overview of the proposed framework. During training, a natural language instruction $t \in \mathcal{T}$, paired with a set of drone-view images $g \in \mathcal{G}$ from multiple perspectives, is fed into a multimodal encoder to extract modality-specific features $\mathbf{T}$ and $\mathbf{V}$ for text and image, respectively. These features are further enriched through the LCSAE module, which prompts a large language model to extract object-centric semantic attributes $\mathbf{A}$ (e.g., shape, color, position, and content) from the instruction t. These attributes guide the refinement of visual features $\mathbf{V}$, resulting in representations $\mathbf{V_a}$ with instruction-aligned semantics. The BFA module then performs mutual refinement across modalities. The enriched visual feature $\mathbf{V_a}$, aligned with attribute embeddings $\mathbf{A}$, is fused with the original feature $\mathbf{V}$ to produce $\mathbf{V_s}$, which in turn enhanced by the textual representation $\mathbf{T}$ to yield $\mathbf{F}$. This bidirectional process facilitates fine-grained semantic integration across modalities. Finally, the Enriched Cross-modal Geolocation Matching (ECGM) module compares $\mathbf{V_s}$ and $\mathbf{F}$ to identify semantically and spatially aligned image-text pairs. Following the benchmark setting [10], ECGM jointly optimizes semantic and spatial alignment through a unified loss that integrates contrastive learning, binary relevance prediction, object grounding, and spatial relation modeling. These enriched features directly contribute to loss computation and supervision, enhancing the precision of natural language-guided drone geo-localization.

C. LLMs-driven Cross-modal Semantic Attribute Enrichment

The **LCSAE** module serves as a key component of our framework, designed to hierarchically extract and align semantic features across textual and visual modalities. It empowers the model to capture abstract semantic attributes, such as color, position, shape, and object identity, that are essential for establishing fine-grained correspondences between natural language descriptions and visual content. As illustrated in Fig. 2, the module consists of three synergistic components: Semantic Attribute Parsing, Multi-modal Enrichment Encoding, and Cross-modal Visual-Attribute Fusion.

1) Semantic Attributes Parsing: To bridge the semantic gap between the text and image, we leverage the LLM to parse input text and distill four visually salient attributes: color, position, shape, and object. These attributes are encoded as structured semantic representations that guide subsequent visual feature extraction, emulating human visual cognition

Prompt: `Description: c.`Please extract the following four types of information from the description above separately: 1. Color information; 2. Position information; 3.Shape information; 4. Object information. Summarize each type of information in one concise sentence.
Output:

An example of our prompt template

`Description:` The main object in the center of the image is a large, black building with a flat roof. The building appears to be made of concrete and has large windows on the upper levels. Please extract the following four types of information from the description above separately: 1. Color information; 2. Position information; 3.Shape information; 4. Object information. Summarize each type of information in one concise sentence.
Output: 1. The building is predominantly black. 2. The building is in the center of the image. 3. The building has a flat roof. 4. The main object is a large black building made of concrete with large windows on the upper levels.

Fig. 3. An illustration of the prompt design used for LLMs-driven attribute enrichment. The prompt is designed to guide Qwen2-72B [44] model to extract four types of contextual information—color, position, shape, and object—from textual scene descriptions. The upper part shows the abstract template, while the lower part provides a concrete example with both the input description and the LLM's structured output.

principles where salient features (e.g., color, position) serve as primary cues for scene understanding.

Formally, given an input caption t, we construct a specialized prompt (as shown in Fig. 3) and query the LLM to extract fine-grained semantic attributes, including color, position, shape, and object information. The resulting attribute set can be formulated as:

$$\mathcal{A}_t = \text{LLM}(\text{Prompt}(t)) = \{d_c, d_p, d_s, d_o\}, \tag{3}$$

where $\mathcal{A}_t$ denotes the set of textual descriptors guided by the input caption t, with d_c, d_p, d_s, and d_o corresponding to color, position, shape, and object-related semantics, respectively. These attributes serve as conceptual anchors that guide the image encoder to attend to text-relevant visual patterns (Fig. 1), thereby reducing cross-modal ambiguity and improving alignment precision.

2) Multi-modal Enrichment Encoding: We adopt X-VLM [42] as the backbone for multi-modal encoding, leveraging its strong capability in aligning visual and textual representations. X-VLM consists of a unified transformer architecture that jointly processes both image and text inputs, enabling effective cross-modal feature extraction and interaction. Given an input caption sequence t, a drone-view image g, and four attribute descriptors $\mathcal{A}_t = \{d_c, d_p, d_s, d_o\}$ representing color, position, shape, and object semantics, we feed them into X-VLM to obtain a set of enriched multi-modal embeddings:

$$\mathbf{T}, \mathbf{A}_c, \mathbf{A}_p, \mathbf{A}_s, \mathbf{A}_o, \mathbf{V} = \text{X-VLM}(t, d_c, d_p, d_s, d_o, g), \tag{4}$$

where $\mathbf{T}$ denotes the original texual embedding, $\mathbf{A} = \{\mathbf{A}_c, \mathbf{A}_p, \mathbf{A}_s, \mathbf{A}_o\}$ are the embeddings of the corresponding semantic attributes, and $\mathbf{V}$ is the visual representation.

3) Cross-modal Visual-Attribute Fusion: The goal of this component is to utilize the four visually salient attributes identified by the LLM to refine and enhance the alignment between textual and visual modalities. First, we employ the attribute embeddings $\mathbf{A}_c, \mathbf{A}_p, \mathbf{A}_s, \mathbf{A}_o$ as queries to attend over the caption embedding $\mathbf{T}$, computed as:

$$\begin{cases} \alpha_c = \mathbf{A}_c \mathbf{T}^T \\ \alpha_p = \mathbf{A}_p \mathbf{T}^T \\ \alpha_s = \mathbf{A}_s \mathbf{T}^T \\ \alpha_o = \mathbf{A}_o \mathbf{T}^T \\ \alpha = \text{Softmax}(\alpha_c + \alpha_p + \alpha_s + \alpha_o) \\ \mathbf{T_a} = \alpha \mathbf{T} \end{cases}. \tag{5}$$

Next, we use the attribute-aware textual representation $\mathbf{T_a}$ to guide attention over the original drone-view image embedding $\mathbf{V}$, formulated as:

$$\begin{cases} \beta = \text{Softmax}(\mathbf{V}{\mathbf{T_a}}^T) \\ \mathbf{V_a} = \beta \mathbf{V} \end{cases}. \tag{6}$$

Here, $\mathbf{V_a}$ forms a fine-grained, attribute-guided representation that emphasizes semantically aligned regions between the visual and textual modalities. This alignment enables more precise cross-modal fusion by focusing on the most relevant features as identified through attribute reasoning.

D. Bidirectional Feature Alignment

Inspired by the lens of information bottleneck theory [45], we propose a novel BFA module that establishes a co-refinement mechanism between textual and visual representations by simultaneously minimizing the cross-modal conditional entropy $\mathcal{H}(V|T)$ and $\mathcal{H}(T|V)$. This module aims to address the inherent semantic gap between heterogeneous modalities in vision-language representation learning. Unlike conventional unidirectional alignment methods, our framework employs an interactive architecture to enable mutual feature enhancement, effectively synchronizing cross-modal semantics while preserving modality-specific discriminative characteristics.

1) Visual-Semantic Fusion: In this stage, we employ residual connections [46] in line with the principle of information bottleneck theory. Specifically, we first formulate the visual feature refinement process. Let $\mathbf{V}$ denote the initial visual embedding extracted from the drone-view image, and $\mathbf{V_a}$ denote its attribute-aware counterpart obtained via attribute-guided attention. We compute the enhanced visual representation $\mathbf{V_a}$ using a residual formulation:

$$\mathbf{V_s} = \mathbf{V_a} + \mathbf{V}. \tag{7}$$

2) Late Cross-Modal Integration: To complete the bidirectional alignment, we design an asymmetric cross-attention mechanism that propagates visual contextual information back to the textual domain. Given the enhanced visual features $\mathbf{V_a}$ and the original textual embedding $\mathbf{T}$, we compute token-aware visual-textual attention scores:

$$\gamma = \text{Softmax}(\mathbf{V_a}\mathbf{T}^T), \tag{8}$$

where $\gamma \in \mathbb{R}^{N_g \times N_t}$ represents the attention distribution between the visual and textual modalities, with N_g and N_t

denoting the number of visual grid tokens and textual tokens, respectively. Each element of γ reflects the relevance between a visual patch (row of $\mathbf{V_a}$) and a textual token (column of $\mathbf{T}$). The refined textual representation is computed via:

$$\mathbf{F} = f(\mathbf{T}) + \mathbf{T}, \quad \text{where} \quad f(\mathbf{T}) = \gamma \mathbf{T}. \tag{9}$$

Here, $f(\mathbf{T})$ denotes the attention-weighted transformation of the caption features modulated by the visual cues. This residual formulation allows for stable and interpretable fusion, where the updated textual embedding $\mathbf{F}$ incorporates discriminative visual signals while retaining its original semantic intent. Finally, $\mathbf{V_s}$ and $\mathbf{F}$ form a fine-grained, attribute-guided cross-modal representation that emphasizes semantically aligned regions between vision and language. This mutual enhancement facilitates more fine-grained alignment and supports downstream geo-localization tasks.

E. Enriched Cross-modal Geolocation Matching

As part of our SAA-DGL framework, we introduce the ECGM module, which improves retrieval and localization by incorporating two complementary components: Image-Text Semantic Matching and Blending Spatial Matching. Built upon the CMG module [10], ECGM benefits from the enriched features, which enhance the semantic alignment and retrieval precision. In addition, we adopt the loss functions based on the GeoText-1652 baseline to ensure consistent evaluation and a fair comparison with prior work.

Image-Text Semantic Matching. Given an enriched cross-modal feature pair $(\mathbf{V_a}, \mathbf{F})$, where $\mathbf{V_a}$ is the attribute-aware image representation and $\mathbf{F}$ is the visual-aware text representation, we compute cosine similarity as:

$$\text{Sim}(g,t) = \frac{\mathbf{V_a}^\top \mathbf{F}}{\|\mathbf{V_a}\|_2 \|\mathbf{F}\|_2}, \tag{10}$$

where $\|\cdot\|_2$ denotes the L2 norm. Based on contrastive learning, all non-matching pairs within a batch of N samples are treated as negatives. The in-batch retrieval probabilities are:

$$\begin{cases} P_{g\to t} = \dfrac{\exp(\text{Sim}(g,t)/\tau)}{\sum_{i=1}^{N} \exp(\text{Sim}(g,t_i)/\tau)} \\ P_{t\to g} = \dfrac{\exp(\text{Sim}(g,t)/\tau)}{\sum_{i=1}^{N} \exp(\text{Sim}(g_i,t)/\tau)} \end{cases}, \tag{11}$$

where τ is a learnable temperature parameter. The In-batch Text-Image Contrastive (ITC) loss is defined as:

$$\mathcal{L}_{\text{ITC}} = -\frac{1}{2}\mathbb{E}\left[\log(P_{g\to t}) + \log(P_{t\to g})\right]. \tag{12}$$

We further incorporate an image-text matching loss $\mathcal{L}_{\text{ITM}}$, which uses a binary classifier to distinguish positive and hard negative pairs based on similarity. Together, $\mathcal{L}_{\text{ITC}}$ and $\mathcal{L}_{\text{ITM}}$ constitute the Image-Text Semantic Matching component.

Blending Spatial Matching. Following [10], we integrate spatially grounded objectives to enhance fine-grained localization. The grounding loss $\mathcal{L}_{\text{grounding}}$ uses text-guided cross-attention to regress bounding boxes of target objects. The spatial relation loss $\mathcal{L}_{\text{spatial}}$ classifies pairwise spatial configurations of regions into pre-defined categories. These two losses together form the Blending Spatial Matching component, supporting accurate object grounding and spatial reasoning.

Total Loss. The total loss integrates semantic and spatial objectives under a multi-task optimization strategy:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{ITC}} + \mathcal{L}_{\text{ITM}} + \lambda\left(\mathcal{L}_{\text{grounding}} + \mathcal{L}_{\text{spatial}}\right), \tag{13}$$

where $\lambda = 0.1$ balances the contribution of spatial objectives.

IV. Experiments

A. Datasets and Metrics

Dataset. We evaluate our model on the GeoText-1652 benchmark [10], which covers 1,652 buildings across 72 universities with images captured from satellite, drone, and ground views. Each image is annotated with fine-grained labels, including global and region-level descriptions, averaging 70.2 and 21.6 words respectively. The dataset provides a training set of 701 cases (50k images, 150k global and 452k region queries) and two testing splits: Full-Test (951 cases) and 24G-Test (190 cases with reduced case count but full data volume for efficient evaluation on limited GPUs). To further assess robustness, we construct AW-Test from 24G-Test, simulating adverse conditions in both modalities. Visual corruptions include brightness change, motion blur, rain, snow, and fog, while textual disturbances cover OCR errors, casual expressions, word deletion/repetition, and synonym substitution. This challenging test set enables systematic evaluation of model generalization under adverse weather and noisy instructions.

Evaluation Metrics. Following the standard evaluation protocol in [10], we adopt Recall at K (R@K) as the primary metric to assess retrieval performance, where K is set to 1, 5, and 10. Specifically, we report R@1, R@5, and R@10 for both T2I and I2T tasks. R@K indicates the percentage of queries for which the correct match is found within the top-K retrieved results. Higher R@K values reflect better retrieval performance and more accurate cross-modal alignment.

B. Implementation Details

In this study, we employ X-VLM [42] pretrained on 16 million (M) image-text pairs as our foundational architecture. The model incorporates a BERT-based [49] text encoder and a Swin-Transformer [25] visual encoder. For optimization, we utilize AdamW [50] with a weight decay of 0.01 and a learning rate of 3×10^{-5}. All input images are uniformly resized to 384×384 pixels with a patch size of 32 during training. Our data augmentation strategy is intentionally constrained to brightness adjustment and identity transformation, explicitly excluding random rotations and horizontal flips to preserve crucial spatial relationships in the visual data. During evaluation, we preprocess text queries derived from global descriptions by removing stop words to maintain semantic precision while improving computational efficiency. For other parameters, we follow the settings used in [10]. To provide a clearer sense of practicality, we report computational resources: the model was fine-tuned for three epochs on four NVIDIA A100 GPUs (80GB each), requiring approximately 12 hours of training. It is worth noting that the increase in

TABLE I
PERFORMANCE OF NATURAL LANGUAGE-GUIDED DRONE GEOLOCALIZATION ON GEOTEXT-1652 BENCHMARK AND AW-TEST (FOGGY CONDITIONS), EVALUATING TEXT-TO-DRONE-VIEW IMAGE RETRIEVAL (T2I) AND DRONE-VIEW IMAGE-TO-TEXT RETRIEVAL (I2T) TASKS ACROSS STANDARD AND ADVERSE-WEATHER SCENARIOS. TEXT QUERY: TEXT-TO-DRONE-VIEW IMAGE RETRIEVAL. IMAGE QUERY: DRONE-VIEW IMAGE-TO-TEXT RETRIEVAL.

Dataset	Methods	#Params	#Pretrained Images	Text Query (%)			Image Query (%)		
				R@1	R@5	R@10	R@1	R@5	R@10
Full-Test	UNITER [36]	300M	4M	0.9	2.7	4.2	2.5	7.4	11.8
	METWE-Swin [47]	380M	4M	1.3	3.9	5.8	2.7	8.0	12.2
	ALBEF [39]	210M	4M	1.8	4.8	7.1	2.9	8.1	12.4
	ALBEF [39]	210M	14M	1.1	3.5	5.3	3.0	9.1	14.2
	X-VLM [42]	216M	4M	4.3	9.9	13.2	4.9	14.2	21.1
	X-VLM [42]	216M	16M	4.5	9.9	13.4	5.0	14.4	21.4
	UNITER$_{Fine\text{-}tuned}$ [36]	300M	4M	10.6	20.4	26.1	21.4	43.4	59.5
	METWE-Swin$_{Fine\text{-}tuned}$ [47]	380M	4M	11.3	21.5	27.3	22.7	46.3	60.7
	ALBEF$_{Fine\text{-}tuned}$ [39]	210M	4M	12.3	22.8	28.6	22.9	49.5	62.3
	ALBEF$_{Fine\text{-}tuned}$ [39]	210M	14M	12.5	22.8	28.5	23.2	49.7	62.4
	X-VLM$_{Fine\text{-}tuned}$ [42]	216M	4M	13.1	23.5	29.2	23.6	50.0	63.2
	X-VLM$_{Fine\text{-}tuned}$ [42]	216M	16M	13.2	23.7	29.6	25.0	52.3	65.1
	CMG [10]	217M	16M	13.6	24.6	31.2	26.3	53.7	66.9
	SAA-DGL (Ours, X-VLM [42])	221M	16M	**14.9**	**28.4**	**35.8**	**28.2**	**56.5**	**69.0**
	SAA-DGL (Ours, X2-VLM [48])	261M	16M	**16.7**	**29.2**	**36.2**	**29.4**	**57.2**	**69.4**
24G-Test	CMG [10]	217M	16M	29.9	46.3	54.1	50.1	81.2	90.3
	SAA-DGL (Ours, X-VLM [42])	221M	16M	**30.7**	**49.2**	**57.6**	**52.1**	**83.3**	**91.4**
	SAA-DGL (Ours, X2-VLM [48])	261M	16M	**31.9**	**49.8**	**57.9**	**53.3**	**83.9**	**91.7**
AW-Test-Fog	CMG [10]	217M	16M	22.6	38.9	47.5	22.2	47.1	59.4
	SAA-DGL (Ours, X-VLM [42])	221M	16M	**23.5**	**40.5**	**49.5**	**23.7**	**50.8**	**64.8**
	SAA-DGL (Ours, X2-VLM [48])	261M	16M	**25.1**	**41.2**	**49.9**	**25.4**	**51.5**	**65.2**

parameters of our overall framework, compared to the baseline method, primarily arises from the incorporation of additional attribute features.

C. Main Results

We evaluate our model on Full-Test, 24G-Test, and AW-Test to assess its effectiveness and generalization, with results summarized in Table I. On the Full-Test set, our method achieves 14.9% R@1 for text-to-image retrieval and 28.2% R@1 for image-to-text retrieval, outperforming the best baseline CMG [10] by +1.3% and +1.9%, respectively. The gains are more pronounced in higher recall levels, e.g., +4.6% R@10 for text queries and +2.1% for image queries, demonstrating that explicitly modeling attributes such as color, position, shape, and object enhances robustness in complex matching scenarios. On the 24G-Test, our model continues to outperform CMG with notable improvements: R@1/5/10 reach 30.7%/49.2%/57.6% for text queries, and 52.1%/83.3%/91.4% for image queries. These results confirm the method's effectiveness under constrained memory settings. Under adverse-weather conditions (AW-Test), although overall performance drops due to fog-induced noise, our method still surpasses CMG, showing improved robustness by leveraging fine-grained attribute features that help mitigate visual degradation and preserve cross-modal alignment. Finally, our approach achieves these improvements with only a modest parameter increase of 4M (from 217M to 221M or 261M) on X-VLM and X2-VLM, highlighting its efficiency in computation and effectiveness in narrowing the semantic gap between modalities.

TABLE II
ABLATION STUDY OF OUR PROPOSED METHOD (SAA-DGL) ON THE 24G-TEST SET. WE EVALUATE THE CONTRIBUTION OF KEY COMPONENTS, INCLUDING THE LCSAE WITH ATTRIBUTES AND BFA MODULES. PERFORMANCE IS MEASURED USING RECALL@K (R@1, R@5, R@10) FOR BOTH T2I AND I2T GEO-LOCALIZATION TASKS.

Model	Text Query (%)			Image Query (%)		
	R@1	R@5	R@10	R@1	R@5	R@10
SAA-DGL (Ours)	**30.7**	49.2	**57.6**	**52.1**	**83.3**	**91.4**
w/o LCSAE	29.8	46.9	55.0	49.3	80.8	89.7
w/o Color	30.1	48.4	56.5	50.9	81.7	90.3
w/o Position	29.8	48.1	57.0	50.5	81.2	90.7
w/o Shape	30.8	49.0	57.3	51.6	83.0	90.8
w/o Object	30.5	**49.3**	56.9	51.0	82.8	91.7
w/o BFA	29.9	47.6	55.5	49.8	81.3	90.1
w/o VSF	30.6	48.2	56.3	51.2	82.7	91.0
w/o CMA	29.9	47.9	55.8	50.6	81.7	90.0

TABLE III
COMPARISON OF DIFFERENT FUSION STRATEGIES IN BFA ON THE 24G-TEST SET. WE EVALUATE ATTRIBUTE-ENHANCED VISUAL REPRESENTATIONS FUSED WITH GLOBAL TEXT FEATURES USING 1-LAYER CROSS ATTENTION (CA), 3-LAYER CA, AND OUR BFA DESIGN.

24G-Test		1-layer CA	3-layer CA	Ours
Text Query	R@1	30.1	29.6	30.7
	R@5	47.3	45.7	49.2
	R@10	55.0	53.4	57.6
Image Query	R@1	50.9	49.8	52.1
	R@5	81.7	80.6	83.3
	R@10	90.5	89.2	91.4

D. Ablation Study of SAA-DGL

To further assess the contribution of each component in our framework, we conduct an ablation study on the 24G-Test set,

as shown in Table II.

Effectiveness of LCSAE. We evaluate the role of the LCSAE module by progressively removing specific semantic attributes. **w/o LCSAE** denotes the removal of the entire module, while **w/o Color**, **w/o Position**, **w/o Shape**, and **w/o Object** represent models with specific attribute types excluded from the semantic feature extraction process. Removing the LCSAE module leads to a substantial performance drop, especially in image-to-text retrieval (R@1: 49.3% vs. 52.1%; R@5: 80.8% vs. 83.3%; R@10: 89.7% vs. 91.4%), confirming its importance in guiding visual feature refinement through textual semantics. Among individual attributes, excluding color and position results in larger performance degradation than removing shape or object features. This highlights that color and spatial cues are more visually salient and directly perceived, playing a critical role in grounding. Notably, the removal of positional information causes significant degradation (e.g., -0.9% R@1 for text query and -1.6% for image query), due to its importance in distinguishing visual content captured from varying angles. Without position cues, the model struggles to differentiate objects with similar appearances from different viewpoints. In contrast, excluding shape information yields the least impact, likely because shape-related descriptions in the dataset are less frequent and often ambiguous.

Effectiveness of BFA. Table II also evaluates the impact of the BFA module and its components. Specifically, **w/o BFA** removes the entire module, **w/o VSF** eliminates the Visual-Semantic Fusion branch, and **w/o CMA** excludes the Cross-Modal Attention mechanism. Removing BFA significantly degrades performance, confirming its critical role in aligning rich textual semantics with simpler visual cues. Textual descriptions often contain detailed semantic information that cannot be effectively grounded without feature-level alignment, especially in cross-view scenarios with ambiguous visuals. Excluding the VSF branch leads to moderate drops in both tasks (e.g., -0.1% R@1 on text query, -0.9% on image query), indicating that visual features derived solely from textual semantics are insufficient. Original visual embeddings provide complementary information that enhances representation when fused with semantic features. Similarly, removing CMA results in a noticeable performance decline, validating its importance in enabling deep interaction between modalities. CMA facilitates bidirectional attention between text and vision, helping the model capture fine-grained semantic correspondences essential for cross-modal understanding.

Effectiveness of Fusion Strategies in BFA. Table III summarizes the comparison of different fusion strategies on the 24G-Test set, including 1-layer cross-attention (CA), 3-layer CA, and our approach. For text-to-image retrieval, our method consistently achieves the highest performance, reaching 30.7% R@1, 49.2% R@5, and 57.6% R@10, clearly surpassing both 1-layer and 3-layer CA. A similar trend is observed in image-to-text retrieval, where our approach attains 52.1% R@1, 83.3% R@5, and 91.4% R@10. Interestingly, increasing the depth of CA from one to three layers fails to provide further improvements and can even cause slight degradation (e.g., R@1 decreases from 30.1% to 29.6% in text-to-image retrieval). This indicates that CA is not easily scalable in this setting, whereas our method avoids overfitting and ensures stable fusion. Overall, the results substantiate the effectiveness of our fusion strategy, showing that it not only competes with but also outperforms attention-based mechanisms while offering greater efficiency and robustness.

TABLE IV
DIFFERENT SEMANTIC INFORMATION. THE SPECIAL SEMANTIC INFORMATION IS GENERATED BY LLMS BASED ON THE CAPTION.

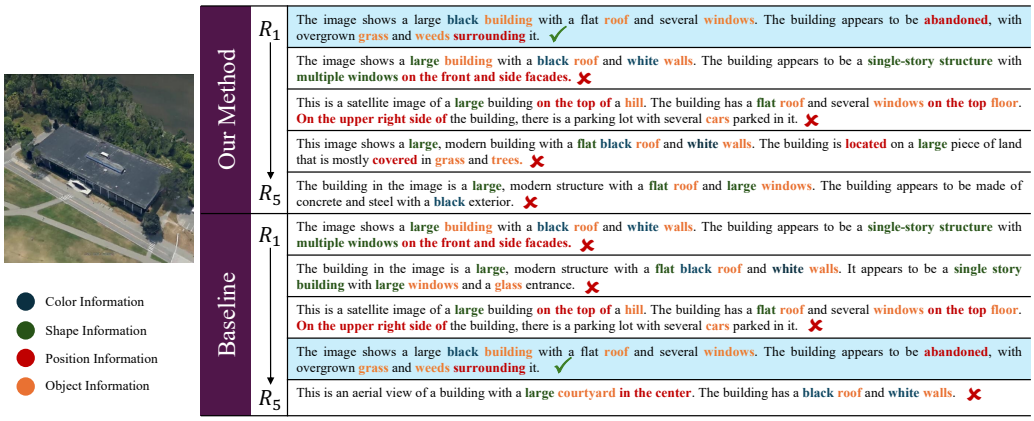
Attributes	Special Semantic Info	General Semantic Info
Color	e.g., "The building is black."	"Color Information"
Position	e.g., "The building is in the center."	"Position Information"
Shape	e.g., "Flat roof, rectangular shape."	"Shape Information"
Object	e.g., "A large black concrete building with upper windows."	"Object Information"

TABLE V
THE RESULTS OF DIFFERENT SEMANTIC INFORMATION ON 24G-TEST. PERFORMANCE COMPARISON OF SAA-DGL USING SPECIAL SEMANTIC INFORMATION (SSI) AND GENERAL SEMANTIC INFORMATION (GSI) AGAINST THE CMG BASELINE ON THE 24G-TEST DATASET.

24G-Test		CMG	SAA-DGL With SSI	SAA-DGL With GSI
Text Query	R@1	29.9	30.7	30.1
	R@5	46.3	49.2	48.5
	R@10	54.1	57.6	56.9
Image Query	R@1	50.1	52.1	51.6
	R@5	81.2	83.3	83.3
	R@10	90.3	91.4	91.0

E. Effect of Special and General Semantic Attributes

We investigate the effectiveness of general semantic information (GSI) and special semantic information (SSI), both derived from LLMs, in guiding drone-view navigation. This experiment evaluates whether GSI remains effective across varying textual attributes without depending on fine-grained LLM outputs. Results show that semantic information at different abstraction levels supports instruction-grounded localization and offers broader generalization to scenarios emphasizing specific semantics. We hypothesize that GSI benefits from the rich visual-semantic mappings learned by multimodal foundation models, allowing accurate alignment even without explicit fine-grained cues. As shown in Table IV and Table V, **for text queries**, the GSI-based method performs comparably to SSI (e.g., R@1: 30.1 vs. 30.7; R@10: 56.9 vs. 57.6), demonstrating the utility of general semantics in capturing task-relevant cues. Both GSI and SSI significantly outperform CMG, confirming the benefits of semantic-guided learning. **For image queries**, GSI also achieves strong performance (R@1: 51.6; R@10: 91.0; R@5: 83.3), matching SSI and highlighting the alignment advantages from general semantics. While SSI offers marginal gains, these benefits diminish with increased system complexity. In contrast, GSI provides practical advantages in low-resource settings, generalization to unseen attributes, and LLM-constrained deployments, enabling robust navigation without fine-grained semantic parsing.



(a) Image Query Retrieval. The texts are arranged from top to bottom in descending order of similarity scores (Top-5). Among them, a gray background with a green checkmark indicates a correct match, while the absence of a background with a red cross indicates a wrong match.

(b) Text Query Retrieval. The images are arranged from left to right in descending order of similarity scores (Top-5). The images marked with red borders indicate false matches, while those with green borders indicate true matches.

Fig. 4. Qualitative examples demonstrating the effectiveness of our method compared to prior work [10] on language-guided drone geo-localization.

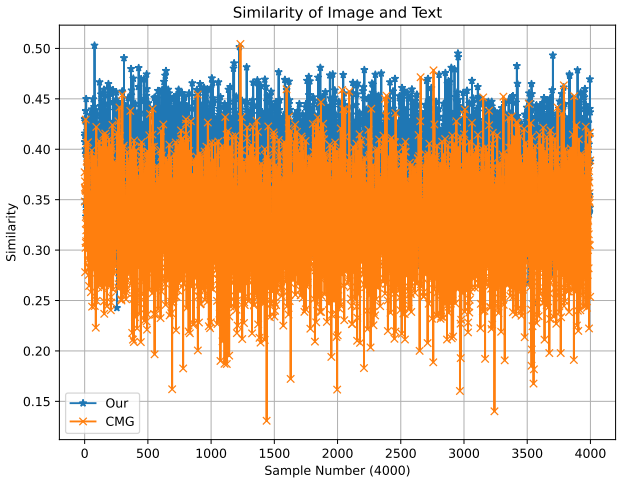


Fig. 5. Similarity between image and text on 24G-Test. Our method uses $(\mathbf{V_a})$ and $(\mathbf{T_a})$ to denote the hidden states of image and text. CMG employs $(\mathbf{V})$ and $(\mathbf{T})$ to denote the hidden states of image and text.

F. Similarity of Text and Visual Information

We employ cosine similarity to evaluate the alignment between text and image features, validating the effectiveness of our model. Since natural language-guided drone navigation is essentially a retrieval task, higher cross-modal similarity indicates stronger semantic alignment and improved navigation accuracy. We compute cosine similarity scores for 4000 randomly sampled pairs, with results shown in Fig. 5. Our method significantly outperforms CMG [10], with 40.4% of the samples achieving similarity values above 0.4, compared to only 5.2% for CMG in 4000 samples. Moreover, the similarity distribution in our method does not overly concentrate at high values, indicating that while semantic alignment is strong, the model also preserves feature discriminability, thus avoiding overfitting. This substantial improvement is primarily attributed to our semantic feature extraction and alignment strategy, which effectively captures core attributes from both modalities and enables precise cross-modal matching.

G. Qualitative Case Study

We present qualitative examples from the 24G-Test set to illustrate the effectiveness of our method in both image-to-text and text-to-image retrieval tasks. As shown in Fig. 4 (a), for an image query, our model ranks the correct text at the top-1 position, whereas CMG ranks it fourth. In Fig. 4 (b), for a text query, our method successfully retrieves the correct image, which CMG fails to retrieve within the top-5 results. These results highlight the advantage of our approach

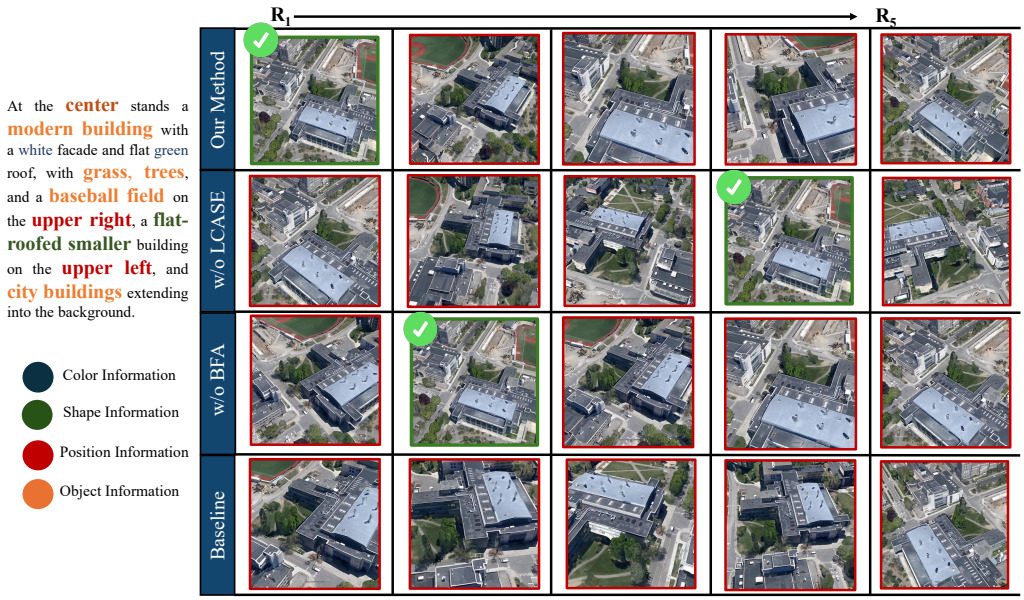


Fig. 6. Text to Image, which demonstrates the contribution of key components (LCASE and BFA). The images are arranged from left to right in descending order of similarity scores (Top-5). The images marked with red borders indicate false matches, while those with green borders indicate true matches.

in effectively leveraging enriched semantic attribute cues such as color, position, shape, and object identity, in combination with BFA, to improve retrieval accuracy. For instance, in text-to-drone-view image retrieval, spatial phrases like "on the middle left side," color attributes like "black," structural cues like "flat roof," and object references such as "trees" contribute to accurate image grounding. Conversely, in drone-view image-to-text retrieval, visual semantics such as "black," "windows," "grass," or "flat" allow for effective alignment with corresponding textual descriptions. These cases validate the effectiveness of the proposed model in semantic attribute extraction and cross-modal alignment, further demonstrating its advantages in interpretability and performance for natural language-guided geo-localization tasks.

H. Ablation study visualizations of retrieval results

To better understand the contribution of each component, we conduct ablation studies as illustrated in Fig. 6. The LCSAE module enriches the vision–language representation with attribute-level features, such as color, shape, position, and object cues. These attributes provide concept-level knowledge that enables the model to establish fine-grained visual–semantic alignment. In contrast, the BFA module focuses on integrating global semantics by performing late cross-modal fusion between textual instructions and visual features.

The qualitative results reveal distinct patterns. Without LCSAE, the model still preserves overall object identity and coarse spatial layout in the retrieved results, but fails to accurately capture positional details (e.g., distinguishing objects on the upper right versus upper left), leading to misalignment. Without BFA, the model can still retrieve relevant objects within the top-2 candidates, but insufficient global semantic fusion results in top-ranked errors. The baseline, lacking both attribute-level and global instruction–visual alignment, produces severe mismatches at both the local detail level and the global semantic level. These observations highlight the complementary nature of LCSAE and BFA: attribute-level enrichment provides robust concept grounding, while global fusion ensures coherent alignment with textual instructions.

I. Robustness Analysis on AW-Test

To further evaluate the robustness of our method under multimodal perturbations, simulating adverse weather and communication interference, we conduct experiments on AW-Test, a challenging benchmark constructed based on 24G-Test. AW-Test includes ten representative types of textual and visual corruptions with relatively high severity. Table. VI reports R@1 quantitative results under five textual and five visual query corruptions, while Fig. 7 presents representative qualitative comparisons. We adopt R@1 as the evaluation metric since it directly reflects the success or failure of drone geo-localization under severe disturbances. Experimental results show that both our method and the baseline suffer noticeable performance degradation under visual corruptions, particularly in challenging cases such as fog, motion blur, and snow, indicating that localization accuracy is severely affected in adverse weather or during rapid drone maneuvers. In textual corruptions, word-level perturbations cause weaker performance drops compared to character-level or sentence-level disturbances. Overall, our SAA-DGL method consistently outperforms the baseline across all corruptions, suggesting that enriching semantic attributes strengthens robustness, with less corrupted attributes playing a key role in maintaining retrieval accuracy. These results demonstrate that the proposed semantic attribute alignment significantly enhances robustness against adverse weather and complex scenarios in both text-to-image and image-to-text retrieval.

TABLE VI

ROBUSTNESS EVALUATION OF SAA-DGL WITH SPECIAL SEMANTIC INFORMATION (SSI) AND THE BASELINE ON AW-TEST. EACH MODALITY QUERY IS EVALUATED ONLY UNDER ITS OWN CORRUPTION TYPES; **NO CORR.** DENOTES THE CLEAN SET. WE REPORT R@1 AS THE PRIMARY METRIC SINCE IT DIRECTLY MEASURES THE ACCURACY OF THE TOP-RANKED RETRIEVAL RESULT UNDER QUERY CORRUPTIONS, WHICH IS MOST CRITICAL IN DGL.

R@1 under Text Query Corruptions						
Methods	**No Corr.**	**OCR Errors**	**Casual Var.**	**Word Delet.**	**Word Repet.**	**Synonym Subst.**
CMG	29.9	23.4	24.7	26.2	28.7	24.9
SAA-DGL	30.7	25.2	26.1	27.0	30.3	25.8
R@1 under Image Query Corruptions						
Methods	**No Corr.**	**Brightness**	**Fog**	**Motion Blur**	**Rain**	**Snow**
CMG	50.1	25.1	22.2	18.6	23.5	21.3
SAA-DGL	52.1	27.4	23.7	19.3	25.3	23.9

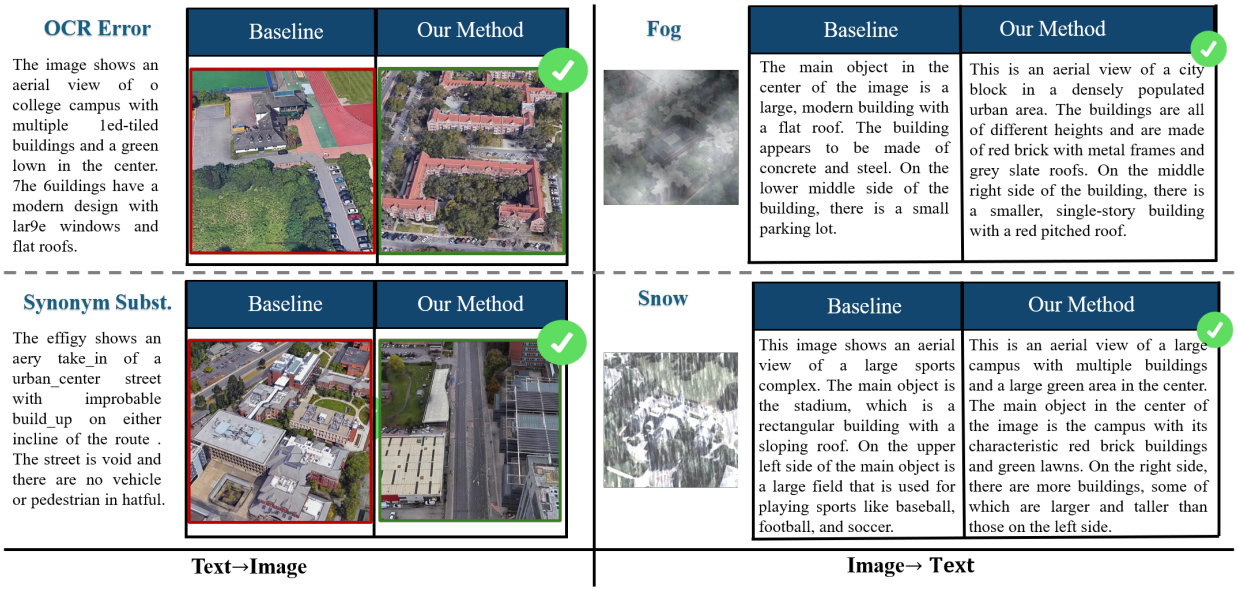


Fig. 7. Comparisons of our method with the baseline on AW-Test under corruption scenarios across both text and image modalities regarding the R@1 metric.

V. CONCLUSION

In this paper, we have presented SAA-DGL, a framework for natural language-guided drone geo-localization. Specifically, our approach incorporates an LLMs-driven Cross-modal Semantic Attribute Enrichment module, wherein Large Language Models extract enriched target attributes (e.g., color, shape) from textual commands. These attributes are then explicitly used to achieve fine-grained fusion between visual features and enriched attributes. Furthermore, we introduce a Bidirectional Feature Alignment module that effectively fuses these attribute-enriched visual representations with textual information. Within BFA, visual features interactively enhance language representations via visual-semantic fusion and cross-modal alignment, thereby reinforcing mutual consistency and reducing modality gaps. Experiments conducted on the GeoText-1652 benchmark and a specialized dataset featuring adverse weather conditions demonstrate that our method achieves state-of-the-art performance in text-to-image and image-to-text drone geo-localization tasks. This provides robust support for precise drone geo-localization, consequently enhancing navigation capabilities.

VI. FUTURE WORK

While SAA-DGL achieves robust retrieval and localization, some limitations remain. The framework focuses on still-image queries without modeling temporal dynamics, and it is evaluated in single-agent settings without considering collaborative localization. In addition, efficiency under resource-constrained UAV platforms is not fully explored. Future work may enrich semantic attributes with more diverse cues (e.g., texture, material, or affordance) to capture fine-grained distinctions, and extend the framework to video-based geo-localization by leveraging temporal consistency. Exploring multi-agent scenarios with communication-efficient cooperation is another promising direction. Finally, lightweight designs or knowledge distillation could improve real-time deployment, while incorporating external priors such as maps, 3D information, or scene graphs may further enhance robustness in complex environments.

ACKNOWLEDGMENT

This work has been partially supported by Natural Science Foundation of Fujian Province (2025J01297) and Technology Innovation Fund Project for SMEs, Yunnan Province (202404AP110047).

REFERENCES

[1] V. K. Patki, A. Mehbodniya, J. L. Webber, A. Kuppusamy, M. A. Haq, A. Kumar, and S. Karupusamy, "Improving the geo-drone-based route for effective communication and connection stability improvement in the emergency area ad-hoc network," *Sustain. Energy Technol. Assess.*, 2022.

[2] J.-I. Meguro, K. Ishikawa, T. Hasizume, J.-I. Takiguchi, I. Noda, and M. Hatayama, "Disaster information collection into geographic information system using rescue robots," in *IROS*, 2006, pp. 3514–3520.

[3] J. Xing, G. Cioffi, J. Hidalgo-Carrió, and D. Scaramuzza, "Autonomous power line inspection with drones via perception-aware mpc," in *IROS*, 2023, pp. 1086–1093.

[4] M. He, J. Liu, P. Gu, and Z. Meng, "Leveraging map retrieval and alignment for robust uav visual geo-localization," *IEEE Trans. Instrum. Meas.*, 2024.

[5] Z. Zheng, Y. Wei, and Y. Yang, "University-1652: A multi-view multi-source benchmark for drone-based geo-localization," in *ACM MM*, 2020, pp. 1395–1403.

[6] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. V. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *ICML*, 2021, pp. 4904–4916.

[7] T. Wang, Z. Zheng, Y. Sun, C. Yan, Y. Yang, and T.-S. Chua, "Multiple-environment self-adaptive network for aerial-view geo-localization," *Pattern Recognit.*, vol. 152, p. 110363, 2024.

[8] X. Shen, D. Li, J. Zhou, Z. Qin, B. He, X. Han, A. Li, Y. Dai, L. Kong, M. Wang *et al.*, "Fine-grained audible video description," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 10 585–10 596.

[9] J. Zhou, D. Guo, R. Guo, Y. Mao, J. Hu, Y. Zhong, X. Chang, and M. Wang, "Towards open-vocabulary audio-visual event localization," *arXiv preprint arXiv:2411.11278*, 2024.

[10] M. Chu, Z. Zheng, W. Ji, T. Wang, and T.-S. Chua, "Towards natural language-guided drones: Geotext-1652 benchmark with spatial relation matching," in *ECCV*, 2024, pp. 213–231.

[11] S. Liu, H. Zhang, Y. Qi, P. Wang, Y. Zhang, and Q. Wu, "Aerialvln: Vision-and-language navigation for uavs," in *ICCV*, 2023, pp. 15 338–15 348.

[12] Z. Liu, Y. Shang, T. Li, G. Chen, Y. Wang, Q. Hu, and P. Zhu, "Robust multi-drone multi-target tracking to resolve target occlusion: A benchmark," *IEEE Trans. Multim.*, vol. 25, pp. 1462–1476, 2023.

[13] N. Yin, C. Liu, R. Tian, and X. Qian, "Sdpdet: Learning scale-separated dynamic proposals for end-to-end drone-view detection," *IEEE Trans. Multim.*, vol. 26, pp. 7812–7822, 2024.

[14] Z. Zeng, Z. Wang, F. Yang, and S. Satoh, "Geo-localization via ground-to-satellite cross-view image retrieval," *IEEE Trans. Multim.*, vol. 25, pp. 2176–2188, 2023.

[15] M. Zhai, Z. Bessinger, S. Workman, and N. Jacobs, "Predicting ground-level scene layout from aerial imagery," in *CVPR*, 2017, pp. 4132–4140.

[16] X. Zhang, X. Li, W. Sultani, Y. Zhou, and S. Wshah, "Cross-view geo-localization via learning disentangled geometric layout correspondence," in *AAAI*, 2023, pp. 3480–3488.

[17] F. Deuser, K. Habel, and N. Oswald, "Sample4geo: Hard negative sampling for cross-view geo-localisation," in *ICCV*, 2023, pp. 16 801–16 810.

[18] M. Dai, E. Zheng, Z. Feng, L. Qi, J. Zhuang, and W. Yang, "Vision-based uav self-positioning in low-altitude urban environments," *IEEE Trans. Image Process.*, vol. 33, pp. 493–508, 2024.

[19] R. Zhu, L. Yin, M. Yang, F. Wu, Y. Yang, and W. Hu, "Sues-200: A multi-height multi-scene cross-view image benchmark across drone and satellite," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 9, pp. 4825–4839, 2023.

[20] Y. Ji, B. He, Z. Tan, and L. Wu, "Game4loc: A uav geo-localization benchmark from game data," *CoRR*, vol. abs/2409.16925, 2024.

[21] H. Ju, S. Huang, S. Liu, and Z. Zheng, "Video2bev: Transforming drone videos to bevs for video-based geo-localization," *CoRR*, 2025.

[22] A. Ku, P. Anderson, R. Patel, E. Ie, and J. Baldridge, "Room-across-room: Multilingual vision-and-language navigation with dense spatiotemporal grounding," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 4392–4412.

[23] Y. Kim, "Convolutional neural networks for sentence classification," in *EMNLP*, 2014, pp. 1746–1751.

[24] D. Fried, R. Hu, V. Cirik, A. Rohrbach, J. Andreas, L.-P. Morency, T. Berg-Kirkpatrick, K. Saenko, D. Klein, and T. Darrell, "Speaker-follower models for vision-and-language navigation," *NeurIPS*, vol. 31, 2018.

[25] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *ICCV*, 2021, pp. 9992–10 002.

[26] P.-L. Guhur, M. Tapaswi, S. Chen, I. Laptev, and C. Schmid, "Airbert: In-domain pretraining for vision-and-language navigation," in *ICCV*, 2021, pp. 1614–1623.

[27] Y. Hong, Q. Wu, Y. Qi, C. Rodriguez-Opazo, and S. Gould, "Vln-bert: A recurrent vision-and-language bert for navigation," in *CVPR*, 2021, pp. 1643–1653.

[28] J. Krantz, E. Wijmans, A. Majumdar, D. Batra, and S. Lee, "Beyond the nav-graph: Vision-and-language navigation in continuous environments," in *ECCV*, 2020, pp. 104–120.

[29] K. He, Y. Huang, Q. Wu, J. Yang, D. An, S. Sima, and L. Wang, "Landmark-rxr: Solving vision-and-language navigation with fine-grained alignment supervision," in *NeurIPS*, 2021, pp. 652–663.

[30] A. B. Vasudevan, D. Dai, and L. Van Gool, "Talk2nav: Long-range vision-and-language navigation with dual attention and spatial memory," *Int. J. Comput. Vis.*, vol. 129, no. 1, pp. 246–266, 2021.

[31] J. Xu, G. Yang, Y. Sun, and S. Picek, "A multi-sensor information fusion method based on factor graph for integrated navigation system," *IEEE Access*, vol. 9, pp. 12 044–12 054, 2021.

[32] G. Zhou, Y. Hong, Z. Wang, X. E. Wang, and Q. Wu, "Navgpt-2: Unleashing navigational reasoning capability for large vision-language models," in *ECCV*, 2024, pp. 260–278.

[33] Y. Tian, F. Lin, Y. Li, T. Zhang, Q. Zhang, X. Fu, J. Huang, X. Dai, Y. Wang, C. Tian, B. Li, Y. Lv, L. Kovács, and F.-Y. Wang, "Uavs meet llms: Overviews and perspectives towards agentic low-altitude mobility," *Inf. Fusion*, vol. 122, p. 103158, 2025.

[34] Y. Gao, Z. Wang, L. Jing, D. Wang, X. Li, and B. Zhao, "Aerial vision-and-language navigation via semantic-topo-metric representation guided llm reasoning," *CoRR*, vol. abs/2410.08500, 2024.

[35] Z. Wang, X. Liu, H. Li, L. Sheng, J. Yan, X. Wang, and J. Shao, "Camp: Cross-modal adaptive message passing for text-image retrieval," in *ICCV*, 2019, pp. 5763–5772.

[36] Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, "Uniter: Universal image-text representation learning," in *ECCV*, 2020, pp. 104–120.

[37] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, Y. Choi, and J. Gao, "Oscar: Object-semantics aligned pre-training for vision-language tasks," in *ECCV*, 2020, pp. 121–137.

[38] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*, 2021, pp. 8748–8763.

[39] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," *NeurIPS*, vol. 34, pp. 9694–9705, 2021.

[40] J. Li, D. Li, C. Xiong, and S. C. H. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *ICML*, 2022, pp. 12 888–12 900.

[41] H. Fei, S. Wu, M. Zhang, M. Zhang, T.-S. Chua, and S. Yan, "Enhancing video-language representations with structural spatio-temporal alignment," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 12, pp. 7701–7719, 2024.

[42] Y. Zeng, X. Zhang, and H. Li, "Multi-grained vision language pre-training: Aligning texts with visual concepts," in *ICML*, 2022, pp. 25 994–26 009.

[43] S. Yang, Y. Zhou, Z. Zheng, Y. Wang, L. Zhu, and Y. Wu, "Towards unified text-based person retrieval: A large-scale multi-attribute and language search benchmark," in *Proceedings of the 31st ACM international conference on multimedia*, 2023, pp. 4492–4501.

[44] A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei *et al.*, "Qwen2. 5 technical report," *arXiv preprint arXiv:2412.15115*, 2024.

[45] N. Tishby and N. Slonim, "Data clustering by markovian relaxation and the information bottleneck method," in *NeurIPS*, 2000, pp. 640–646.

[46] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[47] Z.-Y. Dou, Y. Xu, Z. Gan, J. Wang, S. Wang, L. Wang, C. Zhu, P. Zhang, L. Yuan, N. Peng, Z. Liu, and M. Zeng, "An empirical study of training end-to-end vision-and-language transformers," in *CVPR*, 2022, pp. 18 145–18 155.

[48] Y. Zeng, X. Zhang, H. Li, J. Wang, J. Zhang, and W. Zhou, "X ˆ{2}2-vlm: All-in-one pre-trained model for vision-language tasks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 46, no. 5, pp. 3156–3168, 2023.
[49] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT*, 2019, pp. 4171–4186.
[50] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *ICLR*, 2019.

Changsen Yuan received the Ph.D. degree in computer science from the Beijing Institute of Technology in 2023. He is currently an associate researcher in Beijing University of Technology, China. His current research interests mainly focus on knowledge graphs and information extraction, with a focus on bridging cross-modal perception, cognitive reasoning, and intelligent interpretation across visual, linguistic, and sensory data.

Yang-Hao Zhou (Student Member, IEEE) received the Bachelor's Degree from Jilin University, China, in 2018 and received the Master's Degree from The University of Sheffield, in 2020. He is currently pursuing the PhD degree in the Department of Computer Science and Technology, Beijing Institute of Technology, China. His research interests are in Robust Multi-Modal Learning and and its Application, Multimodal Large Language Models, AIGC.

Cunhan Guo received the PhD degree from the School of Emergency Management Science and Engineering, University of Chinese Academy of Sciences, in 2025. He is currently a postdoctoral at the School of Computer Science and Technology at Beijing Institute of Technology. His research interests include Multimodal Cognition and Computer Vision, and Large Language Model Agents.

Danjie Han received the M.S. degree from the College of Computer and Information Engineering, Henan Normal University, Xinxiang, China, 2016. Currently, she is a Ph.D. student at Nanjing University of Science and Technology. Her current research interests mainly focus on Multi-Modal Learning, Knowledge Graphs and Information Extraction.

Ge Shi received the Ph.D. degree in Computer Science from Beijing Institute of Technology in 2020. He is currently an associate professor at the School of Computer Science, Beijing Institute of Technology, China. His main research interests include large-scale model pre-training, knowledge graph, and multimodal learning.

Wenwu Wang (Senior Member, IEEE, M'02-SM'11) is currently a Professor in Signal Processing and Machine Learning, and an Associate Head in External Engagement, School of Computer Science and Electronic Engineering, University of Surrey, UK. He is also an AI Fellow at the Surrey Institute for People Centred Artificial Intelligence. His current research interests include signal processing, machine learning and perception, artificial intelligence, machine audition, human-AI collaboration. He has more than 400 papers in these areas.