Learning Better UAV-Based Cross-View Object Geo-Localization from Multi-Modal Prompts: MoP-UAV Benchmark and MoPT Framework

Xiaohan Zhang[1*], Zhangkai Shen[1*], Si-Yuan Cao[1,2†] Xiaokai Bai[1], Yiming Li[1], Zheheng Han[1], Zhe Wu[1], Qi Ming[3], Hui-Liang Shen[1,4],

[1]College of Information Science and Electronic Engineering, Zhejiang University
[2]Ningbo Global Innovation Center, Zhejiang University
[3]College of Computer Science, Beijing University of Technology
[4]Key Laboratory of Airspace Sensing and Autonomous Unmanned Systems of Zhejiang Province
zhangxh2023@zju.edu.cn, 3200104864@zju.edu.cn, cao_siyuan@zju.edu.cn

Abstract

We present **MoP-UAV**, a new benchmark for UAV-based cross-view object geo-localization guided by multi-modal prompts. MoP-UAV supports fine-grained object-level cross-view localization under diverse prompt modalities, including natural language, bounding boxes, and click points. It offers potential for incorporating large foundation models like large language models (LLMs) and promotes the building of more flexible and intelligent UAV agents. Based on the benchmark, we propose **MoPT**, a **m**ulti-m**o**dal-**p**rompt-guided **t**ransformer that embeds prompts as token sequences and extract object location from UAV and satellite features via cross-attention. To enhance semantic consistency and performance, we further adopt a cross-view contrastive loss and propose a RefCOCOg-based pre-training strategy. Extensive experiments show that MoPT achieves robust localization under arbitrary prompt combinations. Notably, multi-modal-prompt training significantly boosts unimodal-prompt inference performance, highlighting the generalization benefits of multi-modal learning. MoPT trained with multi-modal prompts outperforms prior unimodal prompt works under the same setting.

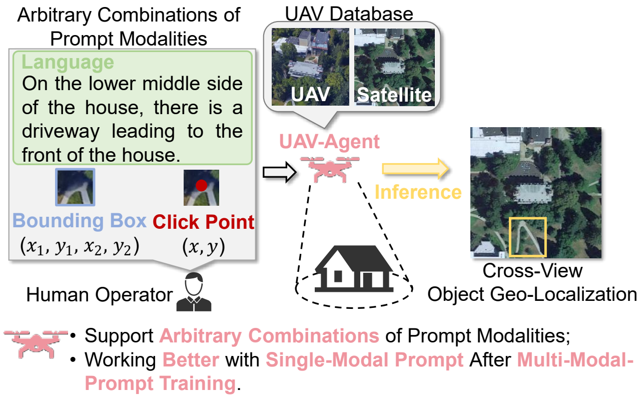
Arbitrary Combinations of Prompt Modalities
Language
On the lower middle side of the house, there is a driveway leading to the front of the house.
Bounding Box
(x_1, y_1, x_2, y_2)
Click Point
(x, y)
Human Operator
UAV Database
UAV
Satellite
UAV-Agent
Inference
Cross-View Object Geo-Localization
• Support Arbitrary Combinations of Prompt Modalities;
• Working Better with Single-Modal Prompt After Multi-Modal-Prompt Training.

Figure 1: Overview of the proposed benchmark MoP-UAV. The UAV-agent receives one or more types of prompts to indicate an object in a UAV-view image and localize it in the satellite image. Additionally, the UAV-agent trained with multi-modal prompts supports flexible prompt input and benefit from multi-modal training even under single-modal-prompt inference.

1 Introduction

Given a drone-view image and a human-provided prompt, unmanned-aerial-vehicle-based object geo-localization (UAV-OGL) aims to localize the indicated object within the corresponding satellite image of the same geographic region. This task is crucial for a variety of applications where GPS information for the object is unavailable (Sun et al. 2023; Li et al. 2025), such as smart city management (Yao et al. 2022), disaster response (Chini, Pierdicca, and Emery 2009; Kumar, Kim, and Hancke 2013), and autonomous navigation (Singamaneni et al. 2024; Zhai et al. 2024).

However, most previous UAV-based visual geo-localization benchmarks (Zheng, Wei, and Yang 2020; Xu et al. 2024; Ye et al. 2025) focus on estimating the UAV camera position using cross-view retrieval (Chu et al. 2024). These benchmarks fall short in supporting prompt understanding and object-level reasoning. To bridge this gap, DetGeo (Sun et al. 2023) introduces the first cross-view object geo-localization benchmark, which enables object-level localization from UAV to satellite imagery conditioned on click-point prompts. Nonetheless, click-points are not always semantically clear, as a single point may lie on only part of an object, making it difficult for the model to infer the intended full object. In contrast, multi-modal prompts such as natural language and bounding boxes have demonstrated superior effectiveness in vision-language tasks (Ye et al. 2024; Kirillov et al. 2023; Xia et al. 2024), as they provide explicit semantic cues (*e.g.*, category, attributes) and precise spatial constraints, significantly improving localization accuracy and interpretability. Incorporating these richer modalities offers a promising path toward applying multi-modal foundation models (Radford et al. 2021; Li

[*] Equal contribution.
[†] Corresponding author.

et al. 2023) to UAV-OGL, facilitating the development of UAV agents. Yet, the lack of a large-scale multi-modal UAV-OGL benchmark remains a major bottleneck.

Motivated by this, we propose a **m**ulti-m**o**dal-**p**rompt **UAV**-OGL benchmark (MoP-UAV), as shown in Fig. 1. Our benchmark repurposes GeoText-1652 (Chu et al. 2024) by leveraging the original annotations (language-box pairs) in UAV-view images to generate point prompts, and manually annotating corresponding object locations in the paired satellite images. As a result, our benchmark provides aligned language, box, and point prompts, enabling the study of UAV-based cross-view object geo-localization under diverse prompting modalities. Given the variability of user inputs across scenarios, our benchmark enables inference with arbitrary prompt combinations, while multi-modal training further enhances model generalization (Radford et al. 2021; Alayrac et al. 2022; Li et al. 2023). Therefore, our benchmark offers the exploration of enhancing unimodal-prompt inference by training with multi-modal prompts. These ensure flexibility and adaptability in practical UAV applications (Yao et al. 2022; Chini, Pierdicca, and Emery 2009; Singamaneni et al. 2024). Looking forward, the inclusion of high-level prompts such as natural language opens the door to integrating large language models (LLMs) into cross-view localization tasks. This provides a pathway toward more intelligent, interactive, and instruction-driven UAV systems capable of understanding complex user intentions through natural communication.

Based on the benchmark, we propose MoPT (a **m**ulti-m**o**dal-**p**rompt-guided cross-view object geo-localization **t**ransformer). MoPT flexibly incorporates multi-modal prompts by embedding each prompt as a token. The prompt tokens interact with UAV-view and satellite-view features through cross-attention, learning to align object-level semantics and extract location information. Trained with multi-modal prompts, MoPT not only adapts flexibly to various prompt combinations during inference, but also consistently outperforms single-modality-prompt-trained models when evaluated with unimodal prompts. Additionally, we introduce a cross-view contrastive loss to for better semantic alignment, while adopting a pre-training strategy using visual grounding dataset RefCOCOg (Yu et al. 2016) to boost the performance of MoPT.

In summary, the main contributions are as follows:

- In pursuit of facilitating UAV-based object geo-localization, we introduce MoP-UAV, a large-scale benchmark with multi-modal prompts (language, bounding boxes, and points). Beyond supporting multi-modal interaction, MoP-UAV provides the potential of leveraging multi-modal training to improve single-modality inference performance.
- We propose MoPT, a multi-modal-prompt-guided cross-view object geo-localization framework. MoPT supports inference under diverse prompt configurations. Notably, when trained with multi-modal prompts, MoPT outperforms under single-modality inference compared to the same model trained solely with the corresponding single-modality prompts.
- For better semantic alignment, we adapt a cross-view contrastive loss, which encourages MoPT to align semantically equivalent tokens between UAV-view and satellite-view features and between prompts from different modalities.
- To boost the performance of MoPT, we propose a RefCOCOg-based pre-training strategy to provide a stronger model initialization.

2 Related Work

Cross-View Geo-Localization. Cross-view geo-localization can be roughly divided into image geo-localization and object geo-localization. Cross-view image geo-localization aims to identify the geographic location of the camera location by finding the most correlated reference image (Deuser, Habel, and Oswald 2023; Zhang et al. 2024; Shi et al. 2019) or the most correlated position (Sarlin et al. 2023; Wang et al. 2023; Lentsch et al. 2023) on the reference image. However, the benchmarks in these works fall short in supporting object-level reasoning. In contrast, cross-view object geo-localization (Sun et al. 2023; Li et al. 2025; Huang et al. 2025) aims to determine the geographic location of an object indicated by a point prompt in a query image (ground-view or drone-view) on the satellite image. DetGeo (Sun et al. 2023) first proposes a benchmark that supports UAV-OGL. Nonetheless, this benchmark only support point prompts and suffers from limited scale. To address this, we propose the MoP-UAV to incorporate multi-modal prompts and extend current dataset scale.

Multi-Modal Learning. Multi-modal learning has shown strong potential in improving model generalization and robustness by jointly leveraging complementary information from different input sources (Radford et al. 2021; Alayrac et al. 2022; Li et al. 2023). In particular, models trained with multi-modal supervision often exhibit superior performance even when tested under unimodal conditions. Recent advances in vision-language pretraining (Li et al. 2023; Alayrac et al. 2022; Li et al. 2022) demonstrate that language prompts can effectively enhance spatial understanding when combined with visual inputs. Inspired by this, our work explore the multi-modal training and demonstrate that multi-modal-prompt training not only enables flexible inference under varied prompt conditions, but also improves performance under unimodal-prompt inference.

3 Dataset

We introduce MoP-UAV, a large-scale benchmark for UAV-based cross-view object geo-localization with multi-modal prompts. The dataset provides three types of prompts—language, bounding boxes, and points—for a total of 102,916 annotated objects across 31,944 UAV-satellite image pairs covering 698 unique scenarios. Each scenario contains a single satellite image paired with multiple UAV-view images captured from different viewpoints. All images are standardized to a resolution of 512×512 pixels. To mitigate data bias and prevent overfitting, we split the dataset by scenario. The training set comprises 598 scenarios, including 26,887 image pairs and 85,469 prompt-object pairs.

Benchmark	Task	**Object-Level Annotations**	Object-Level Prompt			Frames	Frames
		(Satellite)	Point	Bounding Box	Language	(UAV)	(Satellite)
DenseUAV (Dai et al. 2024)	IGL	–	–	–	–	9.1k	18.2k
University-1652 (Zheng, Wei, and Yang 2020)	IGL	–	–	–	–	89.2k	1.7k
GeoText-1652 (Chu et al. 2024)	IGL	–	–	113.3k	113.3k	89.2k	1.7k
CVOGL (Sun et al. 2023)	OGL	5.8k	5.3k	–	–	5.3k	5.8k
MoP-UAV (Ours)	OGL	**102.9k**	102.9k	102.9k	102.9k	32.0k	0.7k

Table 1: Comparison of the proposed MoP-UAV with existing UAV-based cross-view datasets. "IGL" and "OGL" represent image-level geo-localization and object-level geo-localization, respectively. Compared to GeoText-1652 (Chu et al. 2024) that is repurposed, MoP-UAV provides large-scale object-level annotations in the satellite view, which requires lots of human effort and opens potential for Multi-Modal-Guided UAV-OGL.

The test set contains the remaining 100 scenarios, with 5,066 image pairs and 17,447 prompt-object pairs. There is no overlap between the training and test sets in terms of either scenarios or images. Additionally, the previous dataset (Sun et al. 2023) suffers from limited object diversity, as most objects are buildings. This distribution bias may cause models to rely on category priors rather than accurately interpreting the prompts. In contrast, our benchmark includes distinct objects spanning a wide range of categories, promoting more robust object localization. Please refer to the supplementary material for more dataset details.

Dataset Collection and Annotation. We construct the MoP-UAV benchmark based on the training set of GeoText-1652 (Chu et al. 2024), which provides paired UAV-satellite images along with language-box annotations for each UAV-view image. We first extract the UAV images, satellite images, and their corresponding language-box annotations from GeoText-1652. Then we generate point prompts by computing the centroid of each bounding box in the UAV images. As a result, each UAV-view image is equipped with a set of three prompt types, *i.e.*, language, bounding box, and point. To establish the ground truth on the satellite side, we conduct extensive manual annotation to identify the corresponding object location in the paired satellite image, guided by the UAV image and its associated prompts. This annotation process required over 1,000 hours of human effort (6 experienced contributors). All annotations are cross-checked and verified by a senior reviewer.

Statistical Overview of the Dataset. Please refer to the supplementary material.

Challenges. Based on MoP-UAV, we identify two challenges critical for practical UAV applications (Chini, Pierdicca, and Emery 2009; Singamaneni et al. 2024; Yao et al. 2022). (1) models trained on this dataset should be able to perform joint reasoning over multi-modal prompts (*e.g.*, language, box, point) when available, while also supporting effective inference from a single modality. (2) Inspired by the success of multi-modal training in enhancing performance under unimodal inputs (Li et al. 2023; Radford et al. 2021), MoP-UAV explores whether multi-modal-prompt training can similarly improve single-modality inference accuracy in UAV-OGL.

Comparisons and Contributions. Table 1 compares our benchmark with representative existing datasets. Our benchmark focuses on a relatively new task, *i.e.*, cross-view object geo-localization from UAV perspectives under prompt guidance. The key difference from the existing object geo-localization dataset (Sun et al. 2023) lies in our support for multi-modal prompts. This not only improves prompt clarity but also enables more flexible interaction with UAV agents. Notably, to the best of our knowledge, MoP-UAV is the first benchmark in this domain to incorporate natural language prompts. This offers potential for leveraging multi-modal foundation models (Radford et al. 2021; Li et al. 2023; Liu et al. 2023) to push the boundaries of cross-view object geo-localization. More importantly, our benchmark enables the exploration of whether multi-modal-prompt training on MoP-UAV task can improve single-modal-prompt inference. This is particularly relevant for specific UAV-agent scenarios where rich prompts may not always be available due to limited computational resources (Cheng et al. 2024).

4 Methodology

Fig. 2 shows the architecture of our MoPT. MoPT encodes each input prompt modality independently and concatenates the resulting embeddings into a prompt token. This token undergoes a two-stage cross-attention process, *i.e.*, it first attends to the UAV-view features to extract context-relevant semantics, and then to the satellite-view features to perform object-level alignment. Finally, the updated prompt token sequence is passed through a linear regression head to predict the object location in the satellite image. During training, both the language encoder and the image encoder are frozen. Additionally, we introduce a cross-view contrastive loss and a pre-training strategy to enhance the performance of MoPT.

Framework

Motivation. MoPT is designed to address the demand for flexible prompt usage in UAV applications (Singamaneni et al. 2024; Chini, Pierdicca, and Emery 2009; Yao et al. 2022), where the availability of prompt modalities (e.g., language, box, or point) may vary across scenarios. This necessitates a model in which each prompt modality is processed independently, enabling adaptability to any combination of prompts during inference. However, existing cross-view object geo-localization approaches (Sun et al. 2023; Li et al. 2025; Huang et al. 2025) typically inject prompts by

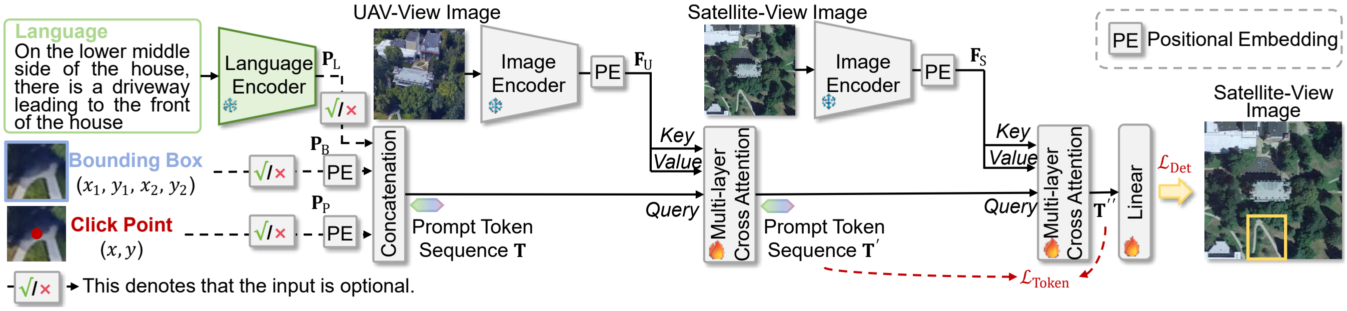
Language
On the lower middle side of the house, there is a driveway leading to the front of the house
Language Encoder
$\mathbf{P}_{\mathrm{L}}$
√/×
Bounding Box
(x_1, y_1, x_2, y_2)
$\mathbf{P}_{\mathrm{B}}$
PE
Click Point
(x, y)
$\mathbf{P}_{\mathrm{P}}$
PE
Concatenation
Prompt Token Sequence $\mathbf{T}$
UAV-View Image
Image Encoder
PE
$\mathbf{F}_{\mathrm{U}}$
Key
Value
Query
Multi-layer Cross Attention
Satellite-View Image
Image Encoder
PE
$\mathbf{F}_{\mathrm{S}}$
Prompt Token Sequence $\mathbf{T}'$
Key
Value
Query
Multi-layer Cross Attention
$\mathbf{T}''$
Linear
$\mathcal{L}_{\mathrm{Det}}$
$\mathcal{L}_{\mathrm{Token}}$
PE Positional Embedding
Satellite-View Image
√/× This denotes that the input is optional.

Figure 2: The overall architecture of our MoPT (a **m**ulti-m**o**dal-**p**rompt-guided cross-view object geo-localization **t**ransformer). During training, the language encoder (CLIP (Radford et al. 2021)) and image encoder (DINOv2 (Oquab et al. 2023)) are frozen. MoPT supports flexible inference under arbitrary combinations of prompt modalities. Moreover, MoPT trained with multi-modal prompts outperforms under single-modality inference than one trained solely with the corresponding single modality.

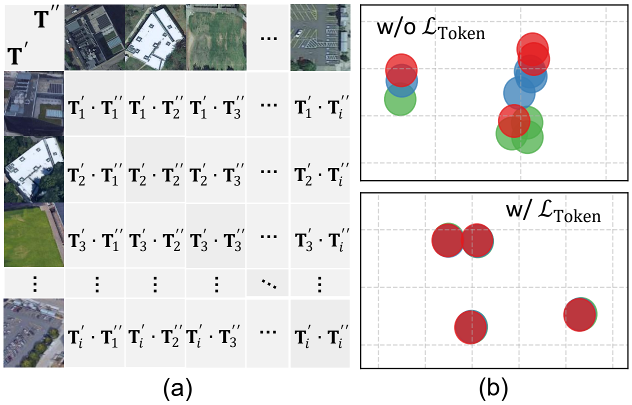

Figure 3: Illustration of $\mathcal{L}_{\mathrm{Token}}$. (a) Contrastive pairs are constructed between $\mathbf{T}'$ (from UAV view) and $\mathbf{T}''$ (from satellite view). (b) The t-SNE visualization of the token embeddings in $\mathbf{T}''$ for four sampled objects under three prompt modalities. With $\mathcal{L}_{\mathrm{Token}}$, embeddings of the three prompt types (text, box, and point) corresponding to the same object cluster tightly together, forming four distinct clusters. In contrast, without $\mathcal{L}_{\mathrm{Token}}$, the embeddings are less aligned and scattered across modalities.

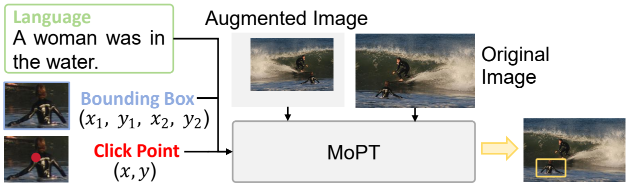

Figure 4: Illustration of our pre-training strategy. We leverage the existing language-box annotations of RefCOCOg (Yu et al. 2016) by treating an augmented image as the "UAV image" and the original image as the "satellite image". This allows the MoPT to be pretrained on existing large-scale dataset.

directly adding their embeddings to image features. Such fusion entangles prompts of different modalities and impairs both flexibility and generalization (Zhai et al. 2022; Li et al. 2019). In contrast, inspired by the attention mechanism in Transformers (Dosovitskiy et al. 2021), MoPT treats prompt embedding of each modality as a separate token, allowing them to independently interact with image features through cross-attention. This design retains modality-specific semantics and ensures robustness across different prompt combinations.

Structure. Given prompt embeddings from different modalities (*e.g.*, language $\mathbf{P}_{\mathrm{L}} \in \mathbb{R}^{1\times c}$, box $\mathbf{P}_{\mathrm{B}} \in \mathbb{R}^{2\times c}$, and point $\mathbf{P}_{\mathrm{P}} \in \mathbb{R}^{1\times c}$), we first concatenate them along the sequence dimension to form a prompt token sequence $\mathbf{T}$. Here, c is the feature dimension. To extract prompt-relevant semantics from the UAV image feature $\mathbf{F}_{\mathrm{U}} \in \mathbb{R}^{m_u\times c}$, the prompt tokens attend to the UAV features via a cross-attention mechanism, which can be expressed as

$$\mathrm{Attn}(\mathbf{T}, \mathbf{F}_{\mathrm{U}}) = \mathrm{Softmax}\left(\frac{\mathbf{T}\mathbf{F}_{\mathrm{U}}^{\top}}{\sqrt{d_k}}\right)\mathbf{F}_{\mathrm{U}}, \tag{1}$$

where m_u and d_k denote the spatial tokens of UAV features and the key/query dimension, respectively. After multiple cross-attention layers, we denote the updated prompt token sequence as $\mathbf{T}^{'}$.

Conditioned on $\mathbf{T}'$, we then extract object-level information from the satellite-view image feature $\mathbf{F}_{\mathrm{S}} \in \mathbb{R}^{m_s\times c}$ through another round of cross-attention. Here, m_s is the spatial tokens of satellite-view features. This can be expressed as

$$\mathrm{Attn}(\mathbf{T}', \mathbf{F}_{\mathrm{S}}) = \mathrm{Softmax}\left(\frac{\mathbf{T}'\mathbf{F}_{\mathrm{S}}^{\top}}{\sqrt{d_k}}\right)\mathbf{F}_{\mathrm{S}}. \tag{2}$$

Finally, we utilize a linear regression layer to predict the object location within the satellite image from the refined prompt token sequence $\mathbf{T}''$.

Loss Function

Motivation. MoP-UAV inherently involves two matching mechanisms: (1) cross-view object-level semantic alignment between UAV and satellite images, and (2) semantic consistency across different prompt modalities (*e.g.*, language, box, and point) that describe the same object. Therefore, we introduce a unified the cross-view contrastive loss $\mathcal{L}_{\text{Token}}$ that simultaneously strengthens cross-view and cross-prompt consistency. Motivated by the effectiveness of contrastive objectives in aligning semantic representations (Radford et al. 2021; Xia et al. 2024; Alayrac et al. 2022), we treat $\mathbf{T}'$ and $\mathbf{T}''$ as view-specific embeddings of the same object and enforce consistency between them, as shown in Fig. 3(a). This design not only facilitates accurate semantic matching across UAV and satellite views but also harmonizes information extracted from prompts of different modalities, mitigating modality imbalance and improving generalization (Li et al. 2023), as shown in Fig. 3(b).

Structure. We first apply average pooling across the spatial dimension of $\mathbf{T}'$ and $\mathbf{T}''$ to obtain a single embedding per instance. Then we treat each pair $(\mathbf{T}'_i, \mathbf{T}''_i)$ in a batch as positive, and all other pairs as negatives. Specifically, we adopt a symmetric InfoNCE loss (Radford et al. 2021) over cosine similarities $\text{sim}(\cdot,\cdot)$, encouraging view-specific embeddings of the same object to align while pushing apart mismatched ones. This can be expressed as

$$\mathcal{L}_{\text{Token}} = \frac{1}{2N}\sum_{i=1}^{N}\Bigg[-\log\frac{\exp(\text{sim}(\mathbf{T}'_i,\mathbf{T}''_i)/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(\mathbf{T}'_i,\mathbf{T}''_j)/\tau)} - \log\frac{\exp(\text{sim}(\mathbf{T}''_i,\mathbf{T}'_i)/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(\mathbf{T}''_i,\mathbf{T}'_j)/\tau)}\Bigg], \tag{3}$$

where N is the number of samples in a batch and τ denotes the temperature parameter. The $\mathcal{L}_{\text{Token}}$ is combined with a DETR-style detection loss $\mathcal{L}_{\text{Det}}$ to formulate the final training objective of MoPT. The overall loss is defined as

$$\mathcal{L} = \mathcal{L}_{\text{Det}} + \alpha\mathcal{L}_{\text{Token}}, \tag{4}$$

where α is a weighting coefficient and is experimentally set to 0.1 in our experiments.

How can training with multi-modal prompts benefit MoPT during single-modal-prompt inference? Recent studies in multi-modal learning have shown that exposure to diverse modalities during training enhances the robustness and generalizability of learned representations. For example, CLIP (Radford et al. 2021) demonstrates that contrastive training with image-text pairs improves visual representation quality, while BLIPv2 (Li et al. 2023) shows that multi-modal pretraining benefits unimodal downstream tasks such as visual question answering and captioning. In our case, each prompt modality is encoded independently and injected as a separate token. Through the use of cross-attention and our $\mathcal{L}_{\text{Token}}$, these tokens learn to interact with image features in a decoupled but complementary manner, encouraging MoPT to develop prompt-invariant semantic alignment mechanisms (Alayrac et al. 2022; Luo et al. 2020). As a result, MoPT achieves improved localization accuracy even when only a subset of prompt modalities is provided at inference time.

Pre-Training

Motivation. Transformers are generally more data-hungry than convolutional neural networks due to their lack of inductive bias (Dosovitskiy et al. 2021). However, MoP-UAV is a relatively new task, and there currently exists no large-scale dataset with rich multi-modal prompt annotations to support comprehensive pretraining. To address this, we extend the idea of unsupervised pre-training (Bulat et al. 2023; Dai et al. 2022) and leverage the existing visual grounding dataset RefCOCOg (Yu et al. 2016), which provides paired language and object-level annotations, to construct a pretraining strategy tailored for MoPT.

Structure. We simulate cross-view object localization using images from RefCOCOg (Yu et al. 2016) (Fig. 4). For each image, we treat the original image as the “satellite image” and generate a “UAV image” by applying augmentations such as random scaling and flipping. The prompts, in-

Prompt Modalities					
Text	Box	Point	Acc@25↑	Acc@50↑	Avg. IoU↑
✓			0.6106	0.4052	0.3653
	✓		0.6170	0.4074	0.3734
		✓	0.5983	0.4051	0.3685
✓	✓		0.6277	0.4387	0.3846
✓		✓	0.6208	0.4298	0.3780
	✓	✓	0.6192	0.4051	0.3747
✓	✓	✓	**0.6366**	**0.4436**	**0.3917**

Table 2: Ablation study on the effect of multi-modal prompt combinations. Performance is evaluated on the MoP-UAV test set using Acc@25↑, Acc@50↑, and Avg. IoU↑. Bold indicates the best and second-best results.

Prompt	Multi-Modal Training			Single-Modal Training		
	Acc@25↑	Acc@50↑	IoU↑	Acc@25↑	Acc@50↑	IoU↑
Text	0.6106	**0.4052**	**0.3653**	**0.6203**	0.3574	0.3498
Box	**0.6170**	**0.4074**	**0.3734**	0.5534	0.2082	0.2915
Point	**0.5983**	**0.4051**	**0.3685**	0.5588	0.2041	0.2920

Table 3: Ablation study on the effect of training with multi-modal prompts. We compare MoPT trained with multi-modal prompts versus MoPT trained with only the corresponding single-modal prompts. Performance is evaluated using Acc@25↑, Acc@50↑, and average IoU↑ on the MoP-UAV test set.

Prompt	with Pre-Training			without Pre-training		
	Acc@25↑	Acc@50↑	IoU↑	Acc@25↑	Acc@50↑	IoU↑
Text	**0.6106**	**0.4052**	**0.3653**	0.5544	0.2854	0.3136
Box	**0.6170**	**0.4074**	**0.3734**	0.5722	0.3003	0.3289
Point	**0.5983**	**0.4051**	**0.3685**	0.5739	0.3035	0.3302

Table 4: Ablation study on the effect of pretraining. We compare MoPT models trained with and without RefCOCOg-based pretraining across different prompt modalities. Evaluation is conducted on the MoP-UAV test set using Acc@25↑, Acc@50↑, and Avg. IoU↑.

Prompt Modalities			without $\mathcal{L}_{\text{Token}}$			with $\mathcal{L}_{\text{Token}}$		
Text	Box	Point	Acc@25	Acc@50	Avg. IoU	Acc@25	Acc@50	Avg. IoU
✓			0.6106	**0.4189**	**0.3725**	0.6106	0.4052	0.3653
	✓		**0.6217**	**0.4153**	**0.3755**	0.6170	0.4074	0.3734
		✓	0.5914	0.3876	0.3618	**0.5983**	**0.4051**	**0.3685**
	✓	✓	0.5696	0.2945	0.3253	**0.6192**	**0.4051**	**0.3747**
✓	✓	✓	0.6247	0.4421	0.3840	**0.6366**	**0.4436**	**0.3917**

Table 5: Ablation study on the effect of the proposed cross-view contrastive loss $\mathcal{L}_{\text{Token}}$. We report Acc@25↑, Acc@50↑, and Avg. IoU↑ on the MoP-UAV test set with and without this loss.

Method	Acc@25↑	Acc@50↑	IoU↑
DetGeo (Sun et al. 2023)	0.4830	0.3208	0.3273
OCGNet (Huang et al. 2025)	0.4970	0.3349	0.3373
MoPT (Ours)	**0.5983**	**0.4051**	**0.3685**

Table 6: Comparison with existing methods using point prompts only. All models are trained and tested with point prompts on the MoP-UAV test set.

cluding languages, boxes, and points, are derived from the original annotation and projected accordingly to the augmented image. MoPT is then trained to localize the referred object in the original image based on information from the prompt and augmented image. This strategy enables scalable pretraining using existing large-scale visual grounding datasets, providing a strong initialization for our MoPT.

5 Experiment

Pre-Training Dataset

RefCOCOg. The RefCOCOg dataset (Yu et al. 2016) is a large-scale benchmark for referring expression comprehension, with 95,010 language annotations on 49,822 object instances across 25,799 images. Compared to RefCOCO and RefCOCO+, it provides more descriptive expressions (8.43 words on average (Yu et al. 2016)), making it better suited for complex semantic grounding. Its language descriptions mainly focus on intrinsic attributes or actions of the objects (e.g., "An adult giraffe scratching its back with its horn") rather than spatial cues (e.g., "The giraffe on the left"), which reduces reliance on global layout and enables diverse data augmentations such as random flipping and cropping. This flexibility matches our pretraining strategy that uses data augmentation to simulate cross-view variations. In our experiments, all images are resized to 512×512 pixels.

Evaluation Metrics

We adopt the Intersection over Union (IoU) (Redmon and Farhadi 2018) as the basic measure of localization accuracy. Besides, following previous works (Sun et al. 2023; Li et al. 2025; Huang et al. 2025), we apply the Acc@0.25 and Acc@0.50 for evaluation. Higher IoU, higher Acc@0.25, and higher Acc@0.50 denote better performance. Please refer to the supplementary material for a detailed introduction.

Implementation Details

Please refer to the supplementary material.

Ablation Studies

Effect of Multi-Modal Prompt Combinations. Our MoPT flexibly handles arbitrary combinations of prompt modalities during inference (Table 2). Even with a single prompt type, it maintains competitive performance, while combining modalities consistently yields further gains. Models using two modalities outperform their unimodal counterparts, and the best results are obtained when all three modalities are jointly provided, showing that different prompts offer complementary information that enriches semantic understanding. Among unimodal settings, box prompts slightly outperform language and clearly surpass point prompts across all metrics. We attribute this to the intermediate nature of box prompts in UAV-OGL, which provide explicit spatial localization and also encode coarse semantic cues through object extent and shape, achieving a favorable balance between localization precision and semantic guidance (Li et al. 2022; Kirillov et al. 2023).

Effect of Training with Multi-Modal Prompts. To ensure a fair comparison, we constrain the single-modal training setting to use the same prompt modality in both the pretraining and main training stages. This prevents any latent influence from unseen modalities and isolates the effect of multi-modal training on final performance. As shown in Table 3, MoPT trained with multi-modal prompts consistently outperforms its single-modality counterpart across all prompt types during single-modal inference. This validates our analysis that multi-modal training enables the model to learn more generalizable and prompt-invariant representations, thereby enhancing localization performance under modality-limited scenarios. Interestingly, for the prompt of language modality, the single-modality model achieves slightly higher Acc@25, while falling short in Acc@50 and Average IoU. This suggests that although language prompts contain rich semantic cues, the model trained solely on textual inputs tend to make coarser predictions that satisfy lower IoU thresholds but lack precision. In contrast, the multi-modality trained model benefits from spatial signals provided by other modalities (*e.g.*, box and point), leading to more accurate localization (Yuan et al. 2024).

Effect of Pre-Training. As shown in Table 4, models trained with pretraining consistently outperform their

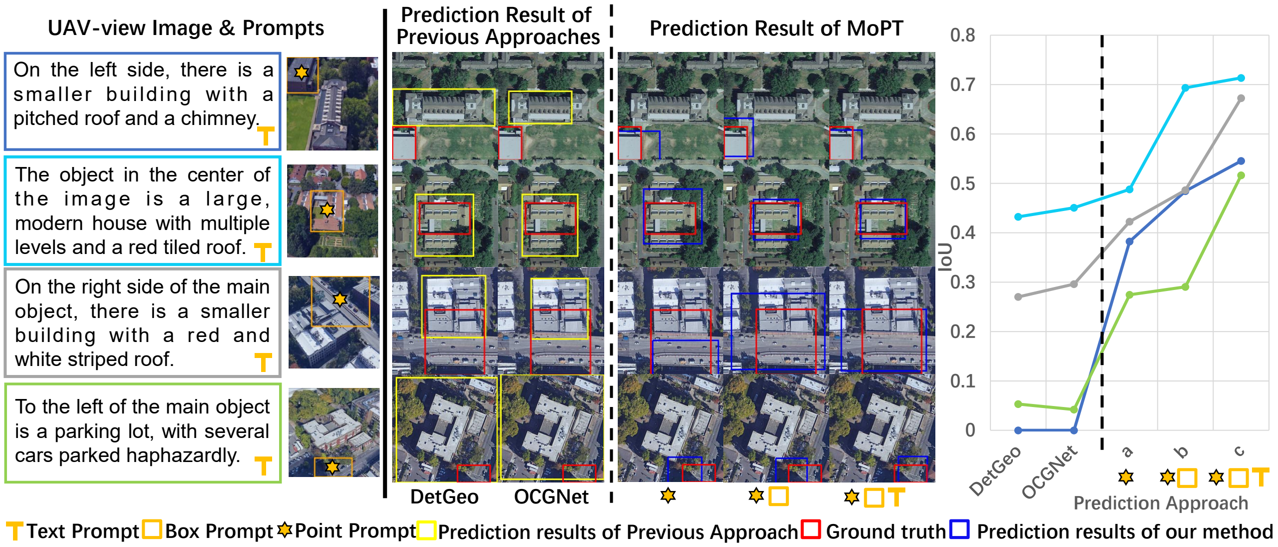

Figure 5: Representative results produced by MoPT and previous approaches (Sun et al. 2023; Huang et al. 2025). (Left) UAV-view images and corresponding prompts. (Middle) Prediction results of DetGeo (Sun et al. 2023) and OCGNet (Huang et al. 2025), both of which only support point prompts. (Right) Our MoPT results under varying prompt combinations, showing significant improvements even under single point prompts, with performance boosted as more prompt modalities are provided.

counterparts trained from scratch across all prompt modalities, indicating that pretraining provides stronger spatial understanding and localization precision. These results highlight the importance of large-scale pretraining. Interestingly, among the models trained without pretraining, the point prompt yields slightly better results than the box or language prompts. We attribute this to the inherent spatial inductive bias of point-based inputs, which directly encode coarse object locations without relying on semantic interpretation or boundary modeling. Similar observations have been reported in previous works such as SAM (Kirillov et al. 2023) and OV-SAM (Yuan et al. 2024), where point-based prompts often perform competitively due to their simplicity. However, prompts like boxes and language benefit more from pretraining, which equips the model to better align high-level semantics (Radford et al. 2021).

Effect of $\mathcal{L}_{\text{Token}}$. Table 5 shows that the cross-view contrastive loss $\mathcal{L}_{\text{Token}}$ consistently improves performance across all prompt settings. The gain is most pronounced for the box+point configuration, where performance drops markedly without $\mathcal{L}_{\text{Token}}$ but becomes comparable to other combinations once it is used. We argue that $\mathcal{L}_{\text{Token}}$ alleviates conflicts between modalities: point prompts provide coarse spatial cues, whereas box prompts impose finer geometric constraints. By aligning view-specific token representations and enforcing object-level semantic consistency across modalities (Radford et al. 2021; Alayrac et al. 2022; Xia et al. 2024), $\mathcal{L}_{\text{Token}}$ yields more robust multi-modal prompt inference.

Comparison Results

Quantitative Results. To validate the effectiveness of our MoPT, we select DetGeo (Sun et al. 2023) and OCGNet (Huang et al. 2025) as comparison methods, which are the most relevant open-source approaches designed for the closely related cross-view object geo-localization task. Given that these approaches are designed to operate with point prompts only, we preserve their original architecture and also evaluate our MoPT solely with point-prompt inference for fair comparison. As shown in Table 6, MoPT significantly outperforms the existing methods.

Qualitative Results. As shown in Fig. 5, our MoPT outperforms previous approaches (Sun et al. 2023; Huang et al. 2025) when using only point prompts, demonstrating its strong generalization and better spatial reasoning ability. Additionally, as additional prompt modalities (*e.g.*, text and box) are incorporated, predictions of our MoPT become more precise and robust. This confirms the effectiveness of our multi-modal prompt design.

6 Conclusion

In this paper, we introduce MoP-UAV, a new benchmark for UAV-OGL with multi-modal prompts. Based on this, we propose MoPT, which unifies diverse prompt types through tokenized embeddings. Extensive experiments demonstrate that MoPT not only adapts well to arbitrary prompt combinations but also benefits from multi-modal training to improve unimodal inference.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under grant 62301484, in part by the Ningbo Natural Science Foundation of China under grant 2024J454, and in part by the National Key Research and Development Program of China under grant 2023YFB3209800.

References

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; Ring, R.; Rutherford, E.; Cabi, S.; Han, T.; Gong, Z.; Samangooei, S.; Monteiro, M.; Menick, J.; Borgeaud, S.; Brock, A.; Nematzadeh, A.; Sharifzadeh, S.; Binkowski, M.; Barreira, R.; Vinyals, O.; Zisserman, A.; and Simonyan, K. 2022. Flamingo: a Visual Language Model for Few-Shot Learning. In *arXiv preprint arXiv:2204.14198*.

Bulat, A.; Guerrero, R.; Martinez, B.; and Tzimiropoulos, G. 2023. FS-DETR: Few-Shot DEtection TRansformer with prompting and without re-training. In *ICCV*, 11759–11768.

Cheng, T.; Song, L.; Ge, Y.; Liu, W.; Wang, X.; and Shan, Y. 2024. YOLO-World: Real-Time Open-Vocabulary Object Detection. In *CVPR*.

Chini, M.; Pierdicca, N.; and Emery, W. J. 2009. Exploiting SAR and VHR Optical Images to Quantify Damage Caused by the 2003 Bam Earthquake. *IEEE Transactions on Geoscience and Remote Sensing*, 47(1): 145–152.

Chu, M.; Zheng, Z.; Ji, W.; Wang, T.; and Chua, T.-S. 2024. Towards Natural Language-Guided Drones: GeoText-1652 Benchmark with Spatial Relation Matching. In *ECCV*.

Dai, M.; Zheng, E.; Feng, Z.; Qi, L.; Zhuang, J.; and Yang, W. 2024. Vision-Based UAV Self-Positioning in Low-Altitude Urban Environments. *IEEE Transactions on Image Processing*, 33: 493–508.

Dai, Z.; Cai, B.; Lin, Y.; and Chen, J. 2022. Unsupervised Pre-Training for Detection Transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–11.

Deuser, F.; Habel, K.; and Oswald, N. 2023. Sample4Geo: Hard Negative Sampling For Cross-View Geo-Localisation. In *ICCV*, 16847–16856.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv e-prints arXiv:2010.11929*.

Huang, Z.; Aryal, J.; Nahavandi, S.; Lu, X.; Lim, C. P.; Wei, L.; and Zhou, H. 2025. Object-Level Cross-View Geolocalization With Location Enhancement and Multihead Cross Attention. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 18: 22880–22890.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; Dollar, P.; and Girshick, R. 2023. Segment Anything. In *ICCV*, 4015–4026.

Kumar, A.; Kim, H.; and Hancke, G. P. 2013. Environmental Monitoring Systems: A Review. *IEEE Sensors Journal*, 13(4): 1329–1339.

Lentsch, T.; Xia, Z.; Caesar, H.; and Kooij, J. F. P. 2023. SliceMatch: Geometry-Guided Aggregation for Cross-View Pose Estimation. In *CVPR*, 17225–17234.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*.

Li, L. H.; Yatskar, M.; Yin, D.; Hsieh, C.-J.; and Chang, K.-W. 2019. VisualBERT: A Simple and Performant Baseline for Vision and Language. In *arXiv preprint arXiv:1908.03557*.

Li, L. H.; Zhang, P.; Zhang, H.; Yang, J.; Li, C.; Zhong, Y.; Wang, L.; Yuan, L.; Zhang, L.; Hwang, J.-N.; Chang, K.-W.; and Gao, J. 2022. Grounded Language-Image Pre-training. In *CVPR*.

Li, Z.; Yuan, X.; Liu, W.; and Xu, X. 2025. VAGeo: View-specific Attention for Cross-View Object Geo-Localization. In *arXiv preprint arXiv:2501.07194*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning. In *NeurIPS*.

Luo, H.; Ji, L.; Shi, B.; Huang, H.; Duan, N.; Li, T.; Li, J.; Bharti, T.; and Zhou, M. 2020. UniVL: A Unified Video and Language Pre-Training Model for Multimodal Understanding and Generation. *arXiv preprint arXiv:2002.06353*.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H. V.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; Howes, R.; Huang, P.-Y.; Xu, H.; Sharma, V.; Li, S.-W.; Galuba, W.; Rabbat, M.; Assran, M.; Ballas, N.; Synnaeve, G.; Misra, I.; Jegou, H.; Mairal, J.; Labatut, P.; Joulin, A.; and Bojanowski, P. 2023. DINOv2: Learning Robust Visual Features without Supervision. In *arXiv preprint arXiv:2304.07193*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, volume 139.

Redmon, J.; and Farhadi, A. 2018. YOLOv3: An incremental improvement. *arXiv e-prints arXiv:1804.02767*.

Sarlin, P.-E.; DeTone, D.; Yang, T.-Y.; Avetisyan, A.; Straub, J.; Malisiewicz, T.; Bulo, S. R.; Newcombe, R.; Kontschieder, P.; and Balntas, V. 2023. OrienterNet: Visual Localization in 2D Public Maps with Neural Matching. In *CVPR*, 21632–21642.

Shi, Y.; Liu, L.; Yu, X.; and Li, H. 2019. Spatial-Aware Feature Aggregation for Image based Cross-View Geo-Localization. In *NeurIPS*, volume 32.

Singamaneni, P. T.; Bachiller-Burgos, P.; Manso, L. J.; Garrell, A.; Sanfeliu, A.; Spalanzani, A.; and Alami, R. 2024. A survey on socially aware robot navigation: Taxonomy and future challenges. *The International Journal of Robotics Research*, 43(10): 1533–1572.

Sun, Y.; Ye, Y.; Kang, J.; Fernandez-Beltran, R.; Feng, S.; Li, X.; Luo, C.; Zhang, P.; and Plaza, A. 2023. Cross-View Object Geo-Localization in a Local Region With Satellite Imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 61: 1–16.

Wang, X.; Xu, R.; Cui, Z.; Wan, Z.; and Zhang, Y. 2023. Fine-Grained Cross-View Geo-Localization Using a Correlation-Aware Homography Estimator. *ArXiv*, abs/2308.16906.

Xia, Y.; Shi, L.; Ding, Z.; Henriques, J. F.; and Cremers, D. 2024. Text2Loc: 3D Point Cloud Localization from Natural Language. In *CVPR*.

Xu, W.; Yao, Y.; Cao, J.; Wei, Z.; Liu, C.; Wang, J.; and Peng, M. 2024. UAV-VisLoc: A Large-scale Dataset for UAV Visual Localization. *arXiv preprint arXiv:2405.11936*.

Yao, J.; Hong, D.; Gao, L.; and Chanussot, J. 2022. Multimodal Remote Sensing Benchmark Datasets for Land Cover Classification. In *IGARSS*, 4807–4810.

Ye, J.; Lin, H.; Ou, L.; Chen, D.; Wang, Z.; He, C.; and Li, W. 2024. Where am I? Cross-View Geo-localization with Natural Language Descriptions. *arXiv preprint arXiv:2412.17007*.

Ye, Y.; Teng, X.; Chen, S.; Li, Z.; Liu, L.; Yu, Q.; and Tan, T. 2025. Exploring the best way for UAV visual localization under Low-altitude Multi-view Observation Condition: a Benchmark. In *arXiv preprint arXiv:2503.10692*.

Yu, L.; Poirson, P.; Yang, S.; Berg, A. C.; and Berg, T. L. 2016. Modeling Context in Referring Expressions. In *arXiv preprint arXiv:1608.00272*.

Yuan, H.; Li, X.; Zhou, C.; Li, Y.; Chen, K.; and Loy, C. C. 2024. Open-Vocabulary SAM: Segment and Recognize Twenty-thousand Classes Interactively. In *ECCV*.

Zhai, R.; Zou, J.; Gan, V. J.; Han, X.; Wang, Y.; and Zhao, Y. 2024. Semantic enrichment of BIM with IndoorGML for quadruped robot navigation and automated 3D scanning. *Automation in Construction*, 166: 105605.

Zhai, X.; Wang, X.; Mustafa, B.; Steiner, A.; Keysers, D.; Kolesnikov, A.; and Beyer, L. 2022. LiT: Zero-Shot Transfer with Locked-image text Tuning. In *CVPR*, 18102–18112.

Zhang, X.; Li, X.; Sultani, W.; Chen, C.; and Wshah, S. 2024. GeoDTR: Toward Generic Cross-View Geolocalization via Geometric Disentanglement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12): 10419–10433.

Zheng, Z.; Wei, Y.; and Yang, Y. 2020. University-1652: A Multi-view Multi-source Benchmark for Drone-based Geo-localization. In *Proceedings of the 28th ACM International Conference on Multimedia*, 1395–1403.