Hierarchy-Aware Intent Recognition and Task-Oriented Text Generation for Non-Expert Satellite Instructions

Li Xinyue[1], Wang Chunhui[2,*], Li Hongwei[2], Jiang Zifeng[3]

[1] School of Aeronautics and Astronautics, Zhejiang University – surprisemiluo@163.com
[2] College of Information Science and Electronic Engineering, Zhejiang University – hytgwch@zju.edu.cn; lhw1992@zju.cn
[3] STAR.VISION Aerospace Group Limited, Hangzhou, China – jiang.zifeng@star.vision
[*] Corresponding author: Wang Chunhui

Keywords: Intent Recognition, Hierarchical Annotation, Negative Sample, Parameter-Efficient Fine-tuning, Intelligent Remote Sensing Task Understanding.

Abstract

With the rapid advancement of large language models, natural-language-based understanding of satellite task requests is becoming increasingly important for improving the accessibility of remote-sensing services. However, satellite commands issued by non-expert users are often conversational, ambiguous, and terminologically inconsistent, leading to a substantial gap between free-form expressions and structured task representations. To address this challenge, we propose a hierarchy-aware framework for intent recognition and task-oriented text generation from non-expert satellite instructions. Specifically, we design a hierarchical annotation scheme that models intent levels, parameter structures, inter-element relations, and execution complexity, and we further construct a hierarchical sequence representation for learning. We then introduce a boundary-aware sample organization method based on semantic similarity and structural proximity, together with a retrieval-augmented multi-type negative-sample reorganization strategy to enhance robustness. Finally, we adopt Qwen3-8B with LoRA for parameter-efficient domain adaptation and unified generation of top-level intents and task-oriented outputs. Experiments on a manually curated dataset of 4,025 non-expert satellite instructions show that the proposed method consistently outperforms multiple baselines on both intent classification and task-oriented generation, demonstrating a resource-efficient and scalable solution for natural-language satellite task interfaces.

1. Introduction

1.1 Research Background and Problem Statement

With the wide application of remote sensing satellite systems in land monitoring, disaster response, ecological protection, and urban governance, remote sensing services are gradually evolving from expert-dominated closed processing pipelines toward more intelligent, service-oriented, and interactive paradigms (Kuckreja et al., 2024). In particular, with the rapid development of large language models (LLMs), using natural language to interact with complex remote sensing systems has become an important technical path for reducing system usage barriers and improving service accessibility (Mansourian and Oucheikh, 2024, Zhang et al., 2024b, Manvi et al., 2023).

However, in real-world scenarios, many satellite instructions are issued by non-expert users who lack professional knowledge of remote sensing. Such instructions are often conversational, non-standardized, and highly implicit in parameter expression. As a result, users are often unable to express observation requirements, processing demands, or consultation needs in the form of standardized operational parameters. Therefore, how to map free-form natural language instructions into task representations with explicit operational semantics has become a key fundamental problem in intelligent remote sensing interaction systems.

Unlike general text classification or open-domain question answering, understanding non-expert satellite instructions requires the model not only to identify the user intent but also to organize relevant task information to support downstream satellite applications such as task planning, data recommendation, and human–machine collaborative decision-making (Powell et al., 2023). This problem therefore combines both domain semantic modeling and task-oriented generation, and requires the model to align natural language understanding with operational task structures.

1.2 Key Challenges and Research Idea

Although LLMs provide new opportunities for natural language task understanding, several challenges remain in the scenario of non-expert satellite instructions. First, non-expert users often use vague and colloquial expressions instead of formal operational terms, which creates a clear mismatch between natural language expressions and remote sensing task semantics. As a result, it is difficult for a model to achieve accurate mapping by relying only on general linguistic knowledge (Manvi et al., 2023). Second, strong semantic confusion often exists among neighboring intent categories. Some instructions are highly similar in surface expressions but belong to different operational functions, which makes classification boundaries difficult to distinguish (Wang et al., 2022). Third, high-quality annotated data in this domain are limited, and public resources are scarce. Therefore, zero-shot, few-shot, or conventional supervised learning methods often fail to achieve stable performance (Soudani et al., 2024, Jiang et al., 2024). Finally, this task requires the model not only to output a top-level intent category but also to generate task-oriented results for downstream processing (Wu et al., 2020a), which places higher demands on unified modeling ability.

To address these issues, this paper builds a unified framework from three aspects: hierarchical semantic modeling,

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLIX-B4-2026-643-2026

boundary-aware sample strategy, and parameter-efficient fine-tuning. First, according to the operational semantic characteristics of non-expert satellite instructions, we construct a hierarchical annotation scheme covering intent hierarchy, parameter hierarchy, element relations, and execution complexity. We further design a hierarchical sequence representation so that non-standard natural language instructions can be mapped into a unified semantic space with operational interpretability (Wang et al., 2022, Zhou et al., 2020). Second, to address neighboring intent confusion and low-resource constraints, we propose a hierarchy-aware semantic boundary modeling method. This method jointly models textual semantic similarity and structural proximity to construct a semantic–structural neighborhood score, and further introduces a retrieval-augmented multi-type negative sample reorganization strategy over the existing training set to enhance the model's discrimination ability on complex boundary cases (Robinson et al., 2020, Suresh and Ong, 2021). Finally, we adopt Qwen3-8B as the backbone model and use LoRA for parameter-efficient fine-tuning (Hu et al., 2022). With a unified input–output template, the model jointly performs top-level intent recognition and task-oriented list generation, forming a unified framework from natural language satellite instructions to task-oriented outputs (Valipour et al., 2023, Zi et al., 2023).

1.3 Main Contributions

This paper focuses on the understanding of non-expert satellite instructions. The main contributions are as follows. (1) We propose a hierarchical annotation scheme and a hierarchical sequence representation for non-expert satellite instructions, which provide a unified semantic foundation for the structured representation and modeling of free-form user requests. (2) We propose a hierarchy-aware semantic boundary modeling method together with a retrieval-augmented multi-type negative sample reorganization strategy, which improves the model's ability to distinguish neighboring intent categories, ambiguous expressions, and complex boundary samples. (3) We build a lightweight domain adaptation framework based on Qwen3-8B and LoRA, which enables joint modeling of top-level intent recognition and task-oriented list generation, and improves practicality and stability under low-resource conditions. (4) We conduct experiments on a self-constructed dataset of 4,025 non-expert satellite instructions. The results show that the proposed method outperforms multiple baselines in both intent classification performance and task-oriented output quality.

2. Related Work

2.1 Remote Sensing Satellite Task Understanding and Intelligent Operation

As remote sensing satellite applications continue to develop toward intelligence, autonomy, and service-oriented operation, increasing attention has been paid to satellite task planning, task interpretation, and human–machine collaborative decision-making (Kuckreja et al., 2024, Muhtar et al., 2024, Zhang et al., 2024a). Existing studies mainly focus on two directions. On the one hand, some work investigates remote sensing question answering, image content understanding, and geospatial semantic parsing, with the goal of accessing remote sensing products and geographic information resources through natural language interfaces (Mansourian and Oucheikh, 2024, Zhang et al., 2024b, Dubey et al., 2025). On the other hand, some studies emphasize autonomous satellite planning and intelligent payload control, and attempt to map high-level mission objectives into executable operations (Deleglise et al., 2022).

However, most existing studies are designed for professional operational workflows, structured task inputs, or standardized parameter sets. Their inputs usually have relatively clear semantic boundaries. In contrast, research on understanding natural language satellite instructions from non-expert users remains limited. Such instructions are often conversational, non-standardized, and implicit in parameter expression. They differ from both general remote sensing question answering and standardized satellite task planning inputs. Therefore, this problem requires a combination of natural language understanding, remote sensing task semantic modeling, and task result organization. Based on this observation, this paper focuses on a unified understanding path from non-expert natural language satellite instructions to task-semantic outputs.

2.2 Hierarchical Text Classification and Intent Recognition

Hierarchical text classification aims to assign texts to categories by exploiting the hierarchical structure of the label system. Its core idea is to explicitly model hierarchical dependencies and semantic relations among labels (Wang et al., 2022, Zhou et al., 2020). Related research has evolved from early step-wise classification strategies to modeling frameworks based on deep neural networks, hierarchical label representations, and structure-aware training (Wang et al., 2022, Zhou et al., 2020). More recently, the field has also expanded to methods that combine large language models with hierarchy-aware fine-tuning (Zi et al., 2023). In general, the main focus of this line of work is how to effectively incorporate hierarchical label information into text representation learning and classification decisions, so as to improve discrimination under complex label systems (Wang et al., 2022, Zhou et al., 2020).

Intent recognition is closely related to this topic and has been widely studied in task-oriented dialogue, human–computer interaction, and domain question answering (Wu et al., 2020a, Louvan and Magnini, 2020). Its goal is to identify the core intent from user input and support subsequent task execution. Existing methods usually focus on intent category discrimination, contextual semantic modeling, and domain adaptation, and have achieved strong performance in scenarios where the intent space is relatively stable and expression patterns are relatively concentrated (Louvan and Magnini, 2020, Jbene et al., 2025). However, non-expert satellite instructions are more domain-dependent and task-oriented. They often contain implicit operational elements such as target regions, temporal requirements, resolution preferences, and processing demands, which makes them different from general intent recognition problems (Manvi et al., 2023).

Furthermore, most existing methods for hierarchical text classification and intent recognition are developed for general text scenarios, where the input expressions, label systems, and output goals are relatively fixed (Powell et al., 2023, Wu et al., 2020a). As a result, they cannot directly cover the operational workflow attributes, parameter hierarchy information, complexity variation, and task-oriented output requirements involved in non-expert satellite instructions (Powell et al., 2023). Therefore, based on the idea of hierarchical information modeling, this paper further constructs a hierarchical annotation scheme and a unified representation method for remote sensing operational semantics, so as to enhance the structured understanding of such instructions.

2.3 Similarity-Based Data Augmentation and Sample Mining

In low-resource text classification tasks, similarity-based sample organization and sample mining have been widely used to improve model discrimination ability (Robinson et al., 2020, Suresh and Ong, 2021). The basic idea is to exploit similarity relations among samples in the representation space and construct more informative training pairs or groups, thereby enhancing intra-class compactness and inter-class separability (Robinson et al., 2020). Typical studies usually organize samples of different difficulty levels, such as easy negatives, hard negatives, and semi-hard negatives, to strengthen the model's ability to learn classification boundaries (Suresh and Ong, 2021).

In hierarchical classification scenarios, however, sample relations are determined not only by surface-level text semantics, but also by factors such as label hierarchy, parameter patterns, and structural characteristics (Wang et al., 2022, Suresh and Ong, 2021). Therefore, recent studies have begun to introduce hierarchical labels, class priors, and structural constraints into similarity modeling, in order to improve the accuracy of sample selection and boundary learning (Wang et al., 2022, Suresh and Ong, 2021). Nevertheless, most existing work is conducted on general text classification tasks or public benchmark datasets, and cannot directly characterize the complex task boundaries found in non-expert natural language satellite instructions (Powell et al., 2023). Based on this observation, this paper jointly models textual semantic similarity and structural proximity, and further performs retrieval-augmented multi-type negative sample reorganization within the existing training set, so as to organize boundary-sensitive samples more accurately.

In addition to sample organization, parameter-efficient fine-tuning of large language models has also provided important support for domain text understanding (Hu et al., 2022, Valipour et al., 2023, Xu et al., 2023). Although general LLMs have shown strong ability in natural language understanding and generation, there is often a clear distribution gap between general corpora and domain-specific texts. As a result, zero-shot or few-shot inference alone is often insufficient for stable adaptation to specialized scenarios, while full-parameter fine-tuning requires high computational cost and may introduce overfitting risks on small domain datasets (Valipour et al., 2023, Xu et al., 2023). Parameter-efficient fine-tuning methods achieve domain transfer by updating only a small number of additional parameters, and are therefore highly practical under low-resource conditions (Hu et al., 2022, Valipour et al., 2023). For the task of understanding non-expert satellite instructions, such methods can preserve the general semantic capability of large models while learning domain terminology, expression patterns, and output formats. This makes them an important technical basis for the model adaptation strategy adopted in this paper (Hu et al., 2022, Valipour et al., 2023).

3. Method

3.1 Task Definition and Overall Framework

3.1.1 Task Definition This paper studies the problem of understanding non-expert satellite instructions. Given a satellite-related request expressed in natural language by a general user, the model is required to accomplish two objectives simultaneously. The first is to identify the top-level operational intent category of the request. The second is to generate a task-oriented list that can support downstream operational processing. Unlike conventional single-label text classification, this problem combines intent classification and lightly structured task representation generation, and can therefore be formulated as a joint learning task.

Let the input request text be

$$x = \{w_1, w_2, \dots, w_n\}, \tag{1}$$

where w_i denotes the i-th word or subword token and n is the input sequence length. The first output of the model is the top-level intent label

$$y \in Y = \{\mathrm{ACQ}, \mathrm{CTRL}, \mathrm{MON}, \mathrm{DATA}, \mathrm{INFO}\}, \tag{2}$$

where ACQ, CTRL, MON, DATA, and INFO denote acquisition requests, on-orbit control requests, operational monitoring requests, data processing and delivery requests, and consultation and rule question-answering requests, respectively. The second output is a task-oriented list

$$T = \{t_1, t_2, \dots, t_m\}, \tag{3}$$

where t_j denotes the j-th component in the task-oriented output, such as task type, target region, time requirement, resolution preference, sensor mode, product form, or additional notes. Accordingly, the task can be formulated as a mapping from natural language instructions to a joint output space:

$$f : x \mapsto (y, T). \tag{4}$$

Compared with general task-oriented intent recognition, the problem studied in this paper has two notable characteristics. On the one hand, the input text is conversational and contains implicit semantics. On the other hand, the output must not only identify the correct task category, but also organise structured information for remote sensing task understanding as completely as possible. Therefore, instead of adopting a purely flat classification framework, this paper builds a unified approach through hierarchical annotation, similarity-driven sample enhancement, and parameter-efficient fine-tuning.

3.1.2 Overall Framework The proposed framework mainly consists of four stages: preprocessing and parameter extraction, hierarchical annotation and hierarchical sequence representation, hierarchy-aware retrieval-augmented multi-type negative sample reorganisation, and parameter-efficient fine-tuning. The overall workflow is as follows.

1. First, the original non-expert satellite requests are preprocessed and their basic operational parameters and intent-related parameters are extracted. Specifically, we compile a comprehensive dictionary enumerating all possible keywords, synonyms, and regular expressions for common satellite task parameters (e.g., geographic locations, temporal expressions, and sensor types). A rule-based matching algorithm is then utilized for initial element extraction, providing structural inputs for subsequent hierarchical modelling.

2. Second, samples are structurally annotated from four aspects, namely intent hierarchy, parameter hierarchy, element relations, and execution complexity, and are further organised into a unified hierarchical sequence representation.

https://doi.org/10.5194/isprs-archives-XLIX-B4-2026-643-2026 | © Author(s) 2026. CC BY 4.0 License.

3. Third, based on the unified representation, semantic similarity, hierarchical proximity, parameter patterns, and complexity information are jointly considered to construct boundary-aware neighbourhood relations between samples. Positive samples and multiple types of negative samples are then organised from the existing training set.

4. Finally, Qwen3-8B is adopted as the backbone large language model, and LoRA is used for parameter-efficient fine-tuning to realise top-level intent recognition and task-oriented list generation.

As shown in Fig. 1, the proposed framework combines hierarchical representation, boundary-aware sample strategy, and lightweight domain adaptation to achieve unified modelling from natural language satellite instructions to task-semantic outputs.

3.2 Hierarchical Annotation and Hierarchical Sequence Representation

3.2.1 Intent Hierarchy Annotation This paper divides the top-level intents of non-expert natural language satellite instructions into five categories: ACQ, CTRL, MON, DATA, and INFO. On this basis, mid-level operational subcategories and fine-grained low-level action semantic types are further defined, forming a three-level intent hierarchy:

$$L_i = \left(l_i^{(1)}, l_i^{(2)}, l_i^{(3)}\right), \tag{5}$$

where $l_i^{(1)}$ denotes the top-level intent category of sample x_i, $l_i^{(2)}$ denotes its operational subcategory, and $l_i^{(3)}$ denotes a finer-grained action-level semantic type. This hierarchical structure organises originally flat labels into a semantic framework with operational workflow attributes, providing more fine-grained structural priors for subsequent boundary modelling.

3.2.2 Parameter Hierarchy, Relation, and Complexity Annotation In addition to the intent hierarchy, the parameter information in each request is also organised hierarchically. According to their roles in task understanding, parameters are divided into three categories: core parameters, constraint parameters, and optional parameters:

$$P_i = \left(P_i^{\text{core}}, P_i^{\text{cons}}, P_i^{\text{opt}}\right), \tag{6}$$

where P_i^{core}, P_i^{cons}, and P_i^{opt} denote the sets of core parameters, constraint parameters, and optional auxiliary parameters of sample x_i, respectively. This design is used to characterise semantic differences in parameter organisation among different satellite instructions.

Since different parameter elements do not exist independently, this paper further introduces element relation modelling. For sample x_i, its relation structure can be expressed as

$$R_i = \left(R_i^{\text{parent}}, R_i^{\text{sibling}}, R_i^{\text{cross}}\right), \tag{7}$$

where R_i^{parent} denotes the dependency relations between the intent category and parameter elements, R_i^{sibling} denotes the co-occurrence relations among different parameter elements, and R_i^{cross} denotes cross-intent composite semantic relations. Although these relations are not predicted independently in this paper, they provide supplementary information for subsequent structural proximity modelling.

In addition, execution complexity annotation is introduced to describe the overall complexity of a satellite instruction from three dimensions: linguistic organisation, task semantics, and technical processing. The complexity representation of sample x_i is defined as

$$C_i = \left(c_i^{\text{syn}}, c_i^{\text{sem}}, c_i^{\text{tech}}\right), \tag{8}$$

where c_i^{syn}, c_i^{sem}, and c_i^{tech} denote syntactic complexity, semantic complexity, and technical complexity, respectively. Based on the overall score, each sample is further mapped to a discrete complexity level, which is used for selecting and organising samples with similar complexity in subsequent modelling.

3.2.3 Hierarchical Sequence Representation To incorporate the above multi-dimensional hierarchical annotations into a unified modelling framework, this paper further designs a hierarchical sequence representation. For sample x_i, its intent hierarchy, parameter hierarchy, complexity information, and relation information are encoded into a structured sequence:

$$H_i = \text{Serialize}(L_i, P_i, C_i, R_i), \tag{9}$$

where $\text{Serialize}(\cdot)$ denotes the sequence serialisation function. To improve the operability of the representation, the hierarchical sequence is further explicitly expanded as

$$H_i = [l_i^{(1)}, l_i^{(2)}, l_i^{(3)}, P_i^{\text{core}}, P_i^{\text{cons}}, P_i^{\text{opt}}, C_i^{\text{level}}, R_i^{\text{tag}}], \tag{10}$$

where C_i^{level} denotes the discrete complexity level and R_i^{tag} denotes the discretised relation structure tag. This representation transforms hierarchical annotation from a static label collection into a unified structural interface for subsequent sample neighbourhood modelling and training sample organisation.

3.3 Sample Neighbourhood Modelling and Multi-Type Negative Sample Reorganisation

3.3.1 Sample Neighbourhood Modelling In low-resource scenarios, the organisation of training samples has a strong influence on model performance. To construct neighbourhood relations that better match task semantic boundaries, this paper first models the proximity relations among samples in the unified representation space.

Given samples x_i and x_j, a text encoder is used to obtain their semantic representations:

$$\mathbf{h}_i = E(x_i), \quad \mathbf{h}_j = E(x_j), \tag{11}$$

where $E(\cdot)$ denotes the text encoding function. Based on cosine similarity, the textual semantic similarity between the two samples is defined as

$$S_{\text{sem}}(i, j) = \frac{\mathbf{h}_i \cdot \mathbf{h}_j}{\|\mathbf{h}_i\| \, \|\mathbf{h}_j\|}. \tag{12}$$

However, relying only on textual semantic similarity is insufficient for accurately characterising the true boundaries among satellite instruction samples. Therefore, this paper further introduces structural proximity modelling. The structural proximity

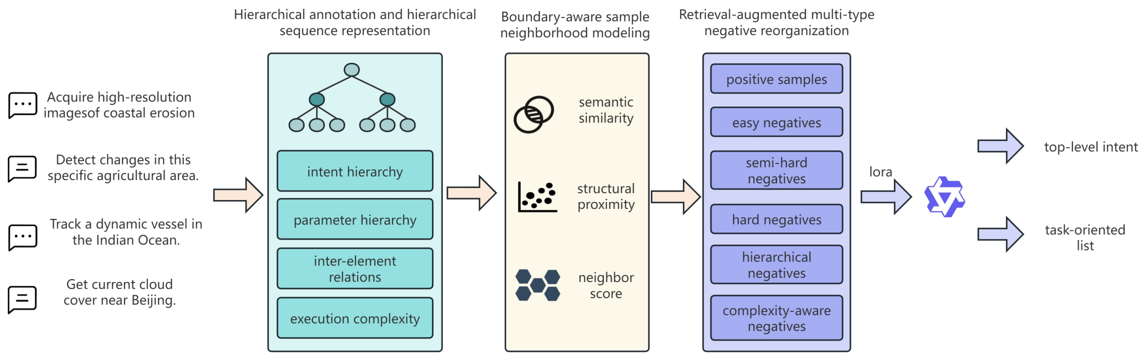

Figure 1. Overall framework of the proposed method.

between samples x_i and x_j is defined as

$$S_{\text{str}}(i,j) = \beta_1 S_{\text{hier}}(i,j) + \beta_2 S_{\text{param}}(i,j) + \beta_3 S_{\text{comp}}(i,j) + \beta_4 S_{\text{rel}}(i,j), \tag{13}$$

where $\beta_1, \beta_2, \beta_3, \beta_4 \geq 0$ and

$$\beta_1 + \beta_2 + \beta_3 + \beta_4 = 1. \tag{14}$$

Each component is defined as follows.

(1) Hierarchical similarity Hierarchical similarity $S_{\text{hier}}(i,j)$ measures the closeness of two samples in the intent hierarchy. Suppose the three-level intent labels of x_i and x_j are $\left(l_i^{(1)}, l_i^{(2)}, l_i^{(3)}\right)$ and $\left(l_j^{(1)}, l_j^{(2)}, l_j^{(3)}\right)$, respectively. Then

$$S_{\text{hier}}(i,j) = \gamma_1 \mathbf{1}\left(l_i^{(1)} = l_j^{(1)}\right) + \gamma_2 \mathbf{1}\left(l_i^{(2)} = l_j^{(2)}\right) + \gamma_3 \mathbf{1}\left(l_i^{(3)} = l_j^{(3)}\right), \tag{15}$$

where $1 > \gamma_1 > \gamma_2 > \gamma_3 \geq 0$.

(2) Parameter pattern similarity Parameter pattern similarity $S_{\text{param}}(i,j)$ measures the similarity of parameter role organisation between two samples. Let P_i^{core}, P_i^{cons}, and P_i^{opt} denote the core, constraint, and optional parameter sets of sample x_i, respectively. Then

$$S_{\text{param}}(i,j) = \eta_1 J(P_i^{\text{core}}, P_j^{\text{core}}) + \eta_2 J(P_i^{\text{cons}}, P_j^{\text{cons}}) + \eta_3 J(P_i^{\text{opt}}, P_j^{\text{opt}}), \tag{16}$$

where

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}, \tag{17}$$

denotes the Jaccard similarity, and $\eta_1 + \eta_2 + \eta_3 = 1$.

(3) Complexity similarity Complexity similarity $S_{\text{comp}}(i,j)$ measures the closeness of two samples in terms of overall complexity. If the complexity levels of x_i and x_j are c_i and c_j, respectively, then

$$S_{\text{comp}}(i,j) = 1 - \frac{|c_i - c_j|}{C_{\max}}, \tag{18}$$

where $C_{\max}$ denotes the maximum possible difference in complexity levels.

(4) Relation structure similarity Relation structure similarity $S_{\text{rel}}(i,j)$ measures the overlap between the relation structure tags of two samples, and is defined as

$$S_{\text{rel}}(i,j) = J(R_i^{\text{tag}}, R_j^{\text{tag}}). \tag{19}$$

Finally, by jointly combining semantic similarity and structural proximity, the boundary-aware neighbourhood score is defined as

$$S(i,j) = \alpha S_{\text{sem}}(i,j) + (1-\alpha) S_{\text{str}}(i,j), \tag{20}$$

where $\alpha \in [0,1]$ is a balancing coefficient. For each anchor sample x_i, candidate samples are sorted in descending order according to $S(i,j)$, yielding its boundary-aware neighbourhood set.

3.3.2 Positive and Negative Sample Organisation and Retrieval-Augmented Mechanism

Based on the sample neighbourhood modelling described above, this paper further organises positive samples and multiple types of negative samples from the existing training set to improve the model's discrimination ability across different boundary scales. For anchor sample x_i, the positive set is preferentially selected from same-class samples with higher hierarchical consistency and greater structural similarity. When such candidates are insufficient, neighbouring samples with the same top-level intent and similar complexity are added.

For heterogeneous samples, this paper divides negative samples into five types according to the boundary-aware neighbourhood score and hierarchical difference information.

1. **Easy negatives**: samples with different top-level intents from the anchor and low overall similarity. They help the model quickly establish coarse inter-class separation boundaries.

2. **Semi-hard negatives**: samples with different classes but some similarity in local semantics or parameter patterns. They help improve training stability.

3. **Hard negatives**: samples that are highly similar to the anchor in semantics but belong to different classes, especially neighbouring confusing samples. They are used to strengthen fine-grained boundary learning.

4. **Hierarchical negatives**: samples that share the same top-level intent as the anchor but differ at the middle or lower semantic levels, or heterogeneous samples located near hierarchical boundaries. They are used to improve the model's sensitivity to hierarchical boundary changes.

5. **Complexity-aware negatives**: samples with complexity levels similar to that of the anchor but different operational semantics. They are used to prevent the model from making decisions based only on shallow complexity cues such as sentence length or parameter quantity.

Based on this design, the retrieval-augmented sample set of anchor sample x_i is constructed as

$$\mathcal{A}_i = \{x_i, P_i, N_i^{\text{easy}}, N_i^{\text{semi}}, N_i^{\text{hard}}, N_i^{\text{hier}}, N_i^{\text{comp}}\}, \quad (21)$$

where P_i denotes the positive sample set, and N_i^{easy}, N_i^{semi}, N_i^{hard}, N_i^{hier}, and N_i^{comp} denote the five negative sample sets, respectively. In this way, the model can explicitly perceive intra-class aggregation relations, inter-class boundary relations, and neighbouring confusion caused by hierarchy and complexity perturbations during training, thereby improving its ability to learn complex semantic boundaries.

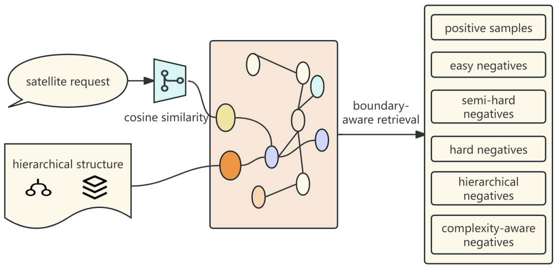

Figure 2. Boundary-aware sample strategy and retrieval-augmented multi-type negative sample reorganisation process.

As shown in Fig. 2, the sample neighbourhood modelling and retrieval-augmented mechanism together form the boundary-aware sample strategy process.

3.4 Parameter-Efficient Fine-Tuning Framework Based on Qwen3-8B

3.4.1 Backbone Model and Lightweight Adaptation Strategy To achieve joint understanding and output generation for non-expert natural language satellite instructions, this paper adopts Qwen3-8B as the backbone large language model. This model provides strong Chinese semantic understanding, instruction-following ability, and structured output capability, while maintaining a reasonable balance between performance and training cost. Considering the limited size of the dataset used in this work, full-parameter fine-tuning would not only require high computational cost but also introduce a higher risk of overfitting. Therefore, LoRA is adopted for parameter-efficient fine-tuning.

Let W denote the original weight matrix of a linear layer. Its update under LoRA can be written as

$$W' = W + BA, \quad (22)$$

where B and A are trainable low-rank matrices. In this way, the model can learn domain terminology, expression patterns, and output structures in the satellite instruction scenario with relatively low computational overhead.

3.4.2 Input–Output Design During training and inference, this paper adopts a unified instruction-style input–output format to jointly model top-level intent recognition and task-oriented list generation. Let the input satellite instruction text be x_i, and let the model be denoted by M_θ, where θ represents the LoRA-adapted model parameters. The output of the model is written as

$$M_\theta(x_i) = (\hat{y}_i, \hat{T}_i), \quad (23)$$

where $\hat{y}_i$ denotes the predicted top-level intent category, and $\hat{T}_i$ denotes the generated task-oriented list.

To improve training consistency, the following unified template is adopted:

Input Template: Task: Understand the following non-expert satellite request and output its top-level intent and task-oriented list. To help distinguish complex semantic boundaries, please refer to the following reference samples:

[Reference Context]
Positive Reference: {Text of a sampled positive instance} → Intent: {intent label}
Negative Reference: {Text of a sampled negative instance} → Intent: {intent label}
Request: {user request}

Output Template: Top-level intent: {intent label} Task-oriented list:

- Task Type: {task type}
- Target Region: {target region}
- Time Requirement: {time requirement}
- Resolution Preference: {resolution preference}
- Sensor / Product Preference: {sensor or product preference}
- Notes: {notes}

This design preserves the flexibility of generative models while providing a relatively stable output interface for subsequent human review and operational execution.

3.4.3 Joint Training Objective This paper mainly adopts supervised fine-tuning to learn the mapping from request text to target output based on the enhanced training set. Let y_i^* denote the ground-truth top-level intent label of sample x_i, and let T_i^* denote its ground-truth task-oriented list. The overall training objective is defined as

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda \mathcal{L}_{\text{gen}}, \quad (24)$$

where $\mathcal{L}_{\text{cls}}$ denotes the intent classification loss, $\mathcal{L}_{\text{gen}}$ denotes the task-oriented generation loss, and λ is a balancing coefficient. By jointly optimising these two objectives, the model can simultaneously learn category discrimination ability and task element organisation ability within a unified training framework.

It should be noted that this paper adopts a strategy of unified training and separate evaluation. During training, top-level intent recognition and task-oriented list generation are jointly learned through a unified output template. During evaluation, intent classification metrics and task-oriented output metrics are calculated separately. This strategy preserves the integrity of satellite instruction understanding while also facilitating independent analysis of classification performance and output quality.

4. Experiments

4.1 Dataset Construction and Experimental Settings

4.1.1 Self-Constructed Dataset To evaluate the effectiveness of the proposed method in realistic scenarios, this paper constructs a dataset containing 4,025 non-expert natural language satellite instructions. The data are collected from anonymised natural language satellite request samples, with a focus on common input forms from general users in scenarios such as remote sensing data acquisition, task consultation, task expression, and result feedback. Unlike public general-domain text classification datasets, this dataset exhibits clear domain-specific characteristics. The instruction texts are often conversational, contain non-standard terminology, have incomplete parameter expressions, and involve implicit intents, which makes the dataset more consistent with real-world application scenarios.

According to the hierarchical annotation scheme proposed in this paper, all samples are manually annotated from the following aspects: (1) top-level intent category; (2) hierarchical intent labels; (3) parameter hierarchy information; (4) element relation labels; (5) execution complexity labels; and (6) task-oriented outputs.

The top-level intent categories are divided into five classes: ACQ, CTRL, MON, DATA, and INFO. To ensure annotation quality, we adopt a double-annotation and review procedure. Samples with inconsistent annotations are discussed and revised, and a third-party reviewer is introduced when necessary. This process yields a structured dataset that can be used for model training and testing.

4.1.2 Dataset Split and Experimental Settings The dataset is divided into training and test sets, while the distributions of intent categories, expression styles, and complexity levels are kept as consistent as possible across subsets. The category distribution is shown in Table 1.

Table 1. Category distribution of the self-constructed non-expert satellite instruction dataset.

Intent Class	Description	Train	Test	Total
ACQ	Acquisition requests	651	156	807
CTRL	On-orbit control requests	639	164	803
MON	Operational monitoring requests	646	159	805
DATA	Data processing and delivery requests	641	163	804
INFO	Consultation and rule Q&A requests	648	158	806
Total		**3225**	**800**	**4025**

All experiments are conducted under a unified training and evaluation framework. Qwen3-8B is used as the backbone model and is fine-tuned with LoRA. In terms of input organisation, a unified instruction template is adopted so that top-level intent recognition and task-oriented list generation are learned within the same supervised fine-tuning framework. The main hyperparameters include learning rate, batch size, number of training epochs, LoRA rank, maximum sequence length, and the weighting parameters used in sample neighbourhood modelling.

4.2 Baselines and Evaluation Metrics

4.2.1 Baselines To evaluate the effectiveness of the proposed method, baseline methods are introduced from both the intent classification and task-oriented output perspectives, covering general pre-trained models, large language models, and representative hierarchical text classification methods.

For general pre-trained models, BERT-base and MacBERT-base are used as classification baselines. For large language models, Qwen3-8B and Llama-3-8B are selected as open-source models of comparable scale. For each model, zero-shot, few-shot, and LoRA fine-tuning settings are considered, so as to compare the effects of different adaptation strategies and analyse generalisation across backbone models. For hierarchical text classification methods, HiAGM(Wu et al., 2020b)and HPT(Wang et al., 2022) are selected as representative baselines, corresponding to hierarchy dependency modelling and hierarchy-aware prompt learning, respectively. These public methods are adapted to the task setting of the self-constructed satellite instruction dataset and are evaluated under the same data split and metric criteria.

It should be noted that the intent classification experiment focuses on classification performance and therefore includes all baselines. In contrast, the task-oriented output experiment only includes methods that can produce structured outputs. For BERT-base and MacBERT-base, task results are generated using a classifier-plus-template-filling strategy. **Ours** denotes the complete method proposed in this paper, which integrates hierarchical sequence representation, boundary-aware sample modelling, and multi-type negative sample reorganisation within the Qwen3-8B + LoRA framework.

4.2.2 Evaluation Metrics This paper evaluates model performance from two aspects: intent classification performance and task-oriented output quality. A strategy of unified training and separate evaluation is adopted.

(1) Intent classification metrics

- **Accuracy**: overall classification accuracy.
- **Macro-F1**: the average F1 score over all classes, which is used to measure overall performance under potential class imbalance.

(2) Task-oriented output metrics

- **Task Intent Accuracy**: the accuracy of intent consistency in the task-oriented output.
- **Structured Exact Match (EM)**: a prediction is considered correct only when all key fields and their values are completely correct.
- **Field-level F1**: the F1 score calculated based on the matching of key fields and their values.

https://doi.org/10.5194/isprs-archives-XLIX-B4-2026-643-2026

- **Human Evaluation Score**: a manual assessment of output quality from the aspects of correctness, readability, and practical usability.

For human evaluation, let A_i, R_i, and U_i denote the scores of correctness, readability, and usability for the i-th sample, respectively. The sample-level human score is defined as

$$H_i = \frac{A_i + R_i + U_i}{3}. \tag{25}$$

The final human evaluation score over a test set with N samples is computed as

$$\text{Human Score} = \frac{1}{N} \sum_{i=1}^{N} H_i. \tag{26}$$

All three dimensions are rated on a 1–5 scale by domain-informed annotators following a unified guideline.

4.3 Main Experimental Results

Table 2. Intent classification results of different methods.

Method	Accuracy (%)	Macro-F1 (%)
BERT-base	91.08	89.36
MacBERT-base	91.87	90.21
HiAGM	92.46	90.84
HPT	93.14	91.63
Zero-shot Llama-3-8B	86.52	83.27
Few-shot Llama-3-8B	89.43	87.12
Zero-shot Qwen3-8B	89.82	87.54
Few-shot Qwen3-8B	92.81	91.17
Qwen3-8B + LoRA	94.67	93.38
Ours	**96.52**	**95.81**

4.3.1 Main Results on Intent Classification Table 2 presents the results of different methods on the intent classification task. Overall, general pre-trained models can perform basic intent recognition, but their ability to distinguish fine-grained category boundaries remains limited. By introducing structural information, hierarchical text classification methods such as HiAGM and HPT further improve performance over BERT-style models, which indicates that hierarchical modelling is beneficial for this task. In comparison, large language models show a certain level of generalisation under zero-shot and few-shot settings, but their performance remains constrained when domain adaptation is insufficient. After LoRA fine-tuning is introduced, classification performance improves substantially. The proposed method achieves the best results on both Accuracy and Macro-F1, which shows that hierarchical sequence representation, boundary-aware sample modelling, and multi-type negative sample reorganisation effectively improve intent discrimination for complex satellite request scenarios.

4.3.2 Results on Task-Oriented Output Generation In addition to intent classification, this paper further compares the performance of different methods on task-oriented output generation. The results are shown in Table 3.

Table 3 reports the results of different methods on task-oriented output generation. Overall, template-based pipeline methods can produce relatively stable structured results, but they tend to suffer from insufficient field coverage and incomplete semantic recovery in complex request scenarios, which limits their performance. In contrast, zero-shot and few-shot large language models have stronger natural language generation ability, but their outputs still fluctuate in structural stability and parameter accuracy due to limited domain constraints.

After LoRA fine-tuning is introduced, both Qwen3-8B and Llama-3-8B show clear improvements across all metrics, indicating that parameter-efficient fine-tuning effectively enhances the model’s ability to learn terminology, field organisation, and output patterns in the satellite task domain. Furthermore, the proposed method achieves the best results on Slot-F1, ROUGE-L, Exact Match, and Human Score. This demonstrates that the proposed hierarchical representation, boundary-aware modelling, and multi-type negative sample reorganisation improve both task semantic understanding and output generation quality, resulting in task-oriented outputs that are more complete, more accurate, and more usable in practice.

4.4 Ablation Study

The following ablation variants are considered in this paper:

- **Full model**: the complete proposed model;
- **w/o hierarchical annotation**: removing the hierarchical annotation scheme and using only top-level intent labels;
- **w/o hierarchical sequence**: removing the hierarchical sequence representation and keeping only the original text input;
- **w/o boundary-aware retrieval**: removing the semantic–structural joint neighbourhood modelling and using only random or flat sample organisation for training;
- **w/o negative augmentation**: removing the retrieval-augmented multi-type negative sample reorganisation;
- **w/o LoRA adaptation**: removing parameter-efficient fine-tuning or using only the base output setting.

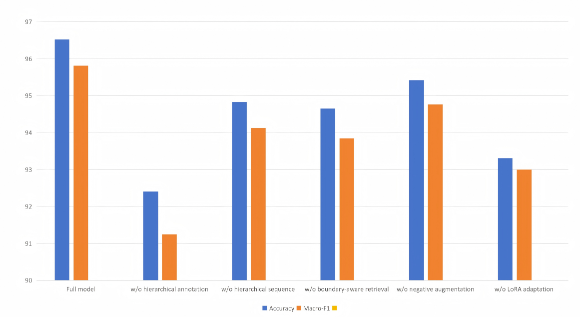

Figure 3. Ablation results of different model variants on the intent classification task.

The quantitative results of the ablation study are reported in Table 4.The results show that the full model achieves the best overall performance, which indicates that the different modules work well together. The largest performance drop is observed when hierarchical annotation is removed, showing that hierarchical annotation is the foundation of the whole framework and directly affects subsequent structural representation and sample organisation. Further performance degradation is observed when hierarchical sequence representation is removed,

Table 3. Task-oriented output results of different methods.

Method	Slot-F1 (%)	ROUGE-L (%)	Exact Match (%)	Human Score
BERT-base + Template	79.12	76.28	63.54	3.61
MacBERT-base + Template	81.46	78.37	66.12	3.74
Zero-shot Llama-3-8B	70.95	73.84	46.32	3.28
Few-shot Llama-3-8B	81.72	82.96	63.28	3.86
Zero-shot Qwen3-8B	72.43	75.26	48.61	3.41
Few-shot Qwen3-8B	83.88	84.69	65.74	3.97
Llama-3-8B + LoRA	87.94	89.15	78.63	4.18
Qwen3-8B + LoRA	89.76	90.68	81.84	4.29
Ours	**93.86**	**94.18**	**88.43**	**4.56**

Table 4. Results of the ablation study.

Variant	Accuracy	Macro-F1
Full model	**96.52**	**95.81**
w/o hierarchical annotation	93.68	92.75
w/o hierarchical sequence	94.83	94.12
w/o boundary-aware retrieval	94.65	93.84
w/o negative augmentation	95.42	94.76
w/o LoRA adaptation	93.31	93.00

which confirms that a unified hierarchical sequence is necessary for fully exploiting structural semantic information.Fig. 3 provides a more intuitive visual comparison of the performance differences among the ablation variants.

When boundary-aware retrieval and negative augmentation are removed, the model performance also decreases to different degrees. This indicates that boundary-aware sample strategy and multi-type negative sample reorganisation effectively improve the model’s ability to distinguish neighbouring intent classes and complex boundary samples. Compared with other components, the gain brought by negative sample reorganisation is slightly smaller, but it still provides a stable contribution to overall performance.

In addition, performance drops clearly when LoRA adaptation is removed, which shows that parameter-efficient fine-tuning is also critical for adapting large language models to this domain. Overall, hierarchical representation, boundary-aware sample modelling, negative sample reorganisation, and LoRA adaptation together constitute the main sources of performance improvement in the proposed method.

5. Conclusion

5.1 Summary and Main Findings

This paper studies the understanding of satellite requests issued by non-expert users and proposes a hierarchy-aware framework that maps free-form natural language inputs to unified “top-level intent + task-oriented list” outputs.

Methodologically, the proposed approach combines hierarchical annotation and sequence representation, boundary-aware sample strategy with retrieval-augmented multi-type negative reorganisation, and parameter-efficient adaptation of Qwen3-8B via LoRA for joint intent recognition and task-oriented generation.

Experiments on a self-constructed dataset of 4,025 non-expert satellite instructions show that the proposed method consistently outperforms strong baselines in both intent classification and task-oriented generation. Ablation results further confirm the effectiveness of the main components of the framework.

5.2 Limitations and Future Work

Despite the encouraging results, several limitations remain. First, the current study is validated on a self-constructed dataset, and further expansion is needed in both data scale and expression diversity, especially for multi-turn interactions and cross-task requests. Second, the current task-oriented output is still represented in a lightly structured textual form, which is convenient for unified modelling but still insufficient for seamless integration with downstream satellite operation systems.

Future work will focus on expanding the dataset to cover more realistic and diverse non-expert requests, and on mapping task-oriented textual outputs into more standardised and executable task representations for downstream operational workflows.

In summary, this work provides a unified hierarchy-aware framework for non-expert satellite request understanding and offers practical support for building more accessible intelligent satellite task interfaces.

References

Deleglise, H., Interdonato, R., Bégué, A., d’Hôtel, E. M., Roche, M., Teisseire, M., 2022. Natural language processing (nlp) methods to support satellite data interpretation in early warning systems in west africa. *Living Planet Symposium 2022 (LPS2022)*, ESA, 2–p.

Dubey, G., Sharma, S., Mishra, V., Rishikesh, 2025. Visual question answering with satellite images. *International Conference on Data Analytics & Management*, Springer, 455–465.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W. et al., 2022. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2), 3.

Jbene, M., Chehri, A., Saadane, R., Tigani, S., Jeon, G., 2025. Intent detection for task-oriented conversational agents: a comparative study of recurrent neural networks and transformer models. *Expert Systems*, 42(2), e13712.

Jiang, Z., Ma, X., Chen, W., 2024. Longrag: Enhancing retrieval-augmented generation with long-context llms. *arXiv preprint arXiv:2406.15319*.

Kuckreja, K., Danish, M. S., Naseer, M., Das, A., Khan, S., Khan, F. S., 2024. Geochat: Grounded large vision-language model for remote sensing. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 27831–27840.

Louvan, S., Magnini, B., 2020. Recent neural methods on slot filling and intent classification for task-oriented dialogue systems: A survey. *Proceedings of the 28th international conference on computational linguistics*, 480–496.

Mansourian, A., Oucheikh, R., 2024. ChatGeoAI: Enabling geospatial analysis for public through natural language, with large language models. *ISPRS International Journal of Geo-Information*, 13(10), 348.

Manvi, R., Khanna, S., Mai, G., Burke, M., Lobell, D., Ermon, S., 2023. Geollm: extracting geospatial knowledge from large language models. arXiv. *arXiv preprint arXiv:2310.06213*.

Muhtar, D., Li, Z., Gu, F., Zhang, X., Xiao, P., 2024. Lhrs-bot: Empowering remote sensing with vgi-enhanced large multimodal language model. *European Conference on Computer Vision*, Springer, 440–457.

Powell, C., Berquand, A., Riccardi, A., 2023. Natural language processing for explainable satellite scheduling. *SPACEOPS 2023*, 349.

Robinson, J., Chuang, C.-Y., Sra, S., Jegelka, S., 2020. Contrastive learning with hard negative samples. *arXiv preprint arXiv:2010.04592*.

Soudani, H., Kanoulas, E., Hasibi, F., 2024. Fine tuning vs. retrieval augmented generation for less popular knowledge. *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, 12–22.

Suresh, V., Ong, D., 2021. Not all negatives are equal: Label-aware contrastive loss for fine-grained text classification. *Proceedings of the 2021 conference on empirical methods in natural language processing*, 4381–4394.

Valipour, M., Rezagholizadeh, M., Kobyzev, I., Ghodsi, A., 2023. Dylora: Parameter-efficient tuning of pre-trained models using dynamic search-free low-rank adaptation. *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 3274–3287.

Wang, Z., Wang, P., Huang, L., Sun, X., Wang, H., 2022. Incorporating hierarchy into text encoder: a contrastive learning approach for hierarchical text classification. *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, 7109–7119.

Wu, C.-S., Hoi, S. C., Socher, R., Xiong, C., 2020a. Tod-bert: Pre-trained natural language understanding for task-oriented dialogue. *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, 917–929.

Wu, Z., Wang, S., Gu, J., Khabsa, M., Sun, F., Ma, H., 2020b. CLEAR: Contrastive Learning for Sentence Representation. *arXiv preprint arXiv:2012.15466*.

Xu, Y., Xie, L., Gu, X., Chen, X., Chang, H., Zhang, H., Chen, Z., Zhang, X., Tian, Q., 2023. Qa-lora: Quantization-aware low-rank adaptation of large language models. *arXiv preprint arXiv:2309.14717*.

Zhang, W., Cai, M., Zhang, T., Zhuang, Y., Li, J., Mao, X., 2024a. EarthMarker: A visual prompting multimodal large language model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 63, 1–19.

Zhang, W., Cai, M., Zhang, T., Zhuang, Y., Mao, X., 2024b. EarthGPT: A universal multimodal large language model for multisensor image comprehension in remote sensing domain. *IEEE Transactions on Geoscience and Remote Sensing*, 62, 1–20.

Zhou, J., Ma, C., Long, D., Xu, G., Ding, N., Zhang, H., Xie, P., Liu, G., 2020. Hierarchy-aware global model for hierarchical text classification. *Proceedings of the 58th annual meeting of the association for computational linguistics*, 1106–1117.

Zi, B., Qi, X., Wang, L., Wang, J., Wong, K.-F., Zhang, L., 2023. Delta-lora: Fine-tuning high-rank parameters with the delta of low-rank matrices. *arXiv preprint arXiv:2309.02411*.

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLIX-B4-2026-643-2026