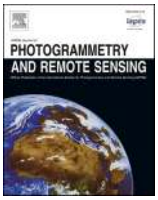
Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

journal homepage: www.elsevier.com/locate/isprsjprs



Estimating urban flooding depth by integrating multimodal image-text data: A segment-level direct preference optimization-based multimodal large language model

Tianyou Chu[a], Yumin Chen[a,d,*], Rui Zhu[b,*], Fei Zeng[c]

[a] School of Resource and Environmental Sciences, Wuhan University, Hubei, Wuhan 430079, China
[b] Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A*STAR), 1 Fusionopolis Way, Singapore 138632, Republic of Singapore
[c] 32004 Reserve Group of the Information Support Force of the People's Liberation Army, Hubei 430000, China
[d] Hubei Luojia Laboratory, Hubei, Wuhan 430079, China

ARTICLE INFO

Keywords:
Flooding depth estimation
Multimodal Large Language Model
Direct Preference Optimization
Social media

ABSTRACT

Urban flood mapping Massive and multi-dimensional social media data provide precious opportunities for the rapid collection and assessment of urban flooding depth. However, effectively and robustly estimating water depth from these multimodal data remains a significant challenge. Although previous studies integrated several existing models, they increase model complexity and hinder joint optimization across different modalities. This paper proposes a Segment-level Direct Preference Optimization-based Multimodal Large Language Model (SDPO-MLLM) for estimating flood depth by integrating image-text data. Our contributions include the design of a hybrid training strategy combining Supervised Fine-Tuning (SFT) and SDPO to reduce inaccurate responses. Additionally, a novel structured workflow is designed, including: (1) dataset preprocessing and construction; (2) event-based extraction of flood location and depth descriptions from text; (3) generation of water depth descriptions from images and videos; (4) classification of water depth descriptions based on multiple reference objects; and (5) quantification of depth categories into numerical values. Empirical experiments are conducted on a dataset containing 2843 text records and 1563 images. The evaluation results show that SDPO-MLLM outperforms other unimodal methods, generating structured and organized results from text, and identifying flooding depth from images based on reference objects. As a case study in Wuhan, Shenzhen and Beijing, the multimodal water depth extracted from social media data is quantified and fused to map and analyze waterlogging-prone areas, demonstrating satisfactory generalization and adaptability of the developed model under various flood scenarios. Our research offers valuable insights for rapid mapping and analysis of urban waterlogging severity.

1. Introduction

In recent decades, flooding has become more frequent and severe due to the increasing occurrence of extreme weather events driven by global climate change, coupled with the growth of populations and the expansion of impervious surfaces. Flooding caused by rainfall can range from minor inconveniences, such as disruption to transport systems and daily activities, to more severe consequences, such as damage to infrastructure, resulting in significant economic losses and presenting risks to public security (Moftakhari et al., 2018; Wang et al., 2021; Ouyang et al., 2022b). Collecting and extracting water depth information during flood events can provide insight into the severity of flooding in different areas, which is crucial for implementing emergency response plans. Moreover, data-driven monitoring also supports urban planning efforts, such as the design of flood-proof architecture, to mitigate the impacts of urban flooding (Feng et al., 2022; Mustafa, 2023).

Various methods have been developed to collect flooding depth information (Huang et al., 2018; Liao et al., 2023; Saleh et al., 2024). Traditional methods, such as field surveys and stream gauges, though accurate in results, often require high labor or equipment costs and have limited detection ranges, making them unsuitable for large-scale monitoring. Flood simulation methods based on complex physical

[*] Corresponding authors.
E-mail addresses: ymchen@whu.edu.cn (Y. Chen), zhur@a-star.edu.sg (R. Zhu).

https://doi.org/10.1016/j.isprsjprs.2025.10.013
Received 6 April 2025; Received in revised form 10 September 2025; Accepted 14 October 2025

mechanisms require substantial computational resources, rendering them ineffective for real-time surveys. The method of monitoring using optical or Synthetic Aperture Radar (SAR) imagery is limited by the satellite's revisit cycle, making it difficult to obtain timely images during flood events, particularly in cases of short-term heavy rainfall.

Compared to the above methods, massive social media data provide an opportunity to collect flood information in real time (Li et al., 2023b; Wang et al., 2024c). During extreme disasters or emergencies, the public often spontaneously uses social media platforms like X (formerly Twitter), TikTok, and Sina Weibo, to request assistance or disseminate disaster information (Hou et al., 2024). However, the primary challenge lies in the effective and robust extraction of flood depth and location data from multimodal social media content, including text, images, and videos. Current research primarily focuses on developing water depth estimation methods based on a single data type or modality (Wan et al., 2024), potentially overlooking valuable information from other sources and introducing biases into flood mapping and analysis. An alternative approach involves integrating these models into a pipeline (Yan et al., 2023), which not only increases model complexity and computational requirements but also hampers the collaborative optimization of multitasking and the modeling of interrelations between different modalities.

For extraction of flood location and water depth from text, single-task models struggle to establish the correspondence relationships between these elements, making it difficult to accurately identify water depths at multiple locations. For estimation of water depth from images and videos, fixed-category image classification or detection models lack semantic associations and reasoning capabilities, limiting their scalability and flexibility. Moreover, with respect to flood information localization, most data lack geotags (Lamsal et al., 2022) and location information is typically conveyed through textual descriptions. The variety and flexibility of location descriptions, especially for non-contiguous geographic entities or complex spatial relationships (Stock et al., 2022), further hinder the application and analysis of water depth information (Sathianarayanan et al., 2024).

To tackle these challenges, Multimodal Large Language Models (MLLMs) (Bubeck et al., 2023; Wu et al., 2024b) are employed to construct a unified flooding depth estimation model for text, images and videos. These models possess cross-modal comprehension and semantic association capabilities and have been explored in several domains such as smart cities (Duan et al., 2024), social media analysis (Chu et al., 2025), and cybersecurity (Ali and Ghanem, 2025). However, in practical applications, the results generated by MLLMs still exhibit hallucination (Huang et al., 2025). For example, although MLLMs can identify objects in images relatively accurately, they can generate missing or incorrect water depth descriptions. The behavior of MLLMs, which does not align well with human preferences, makes it challenging when applied to flood disaster scenarios.

Supervised fine-tuning (SFT) is a basic optimization method, but it may introduce or amplify hallucinations due to a mismatched learning objective (Ouyang et al., 2022a). An alternative approach is Reinforcement Learning with Human Feedback (RLHF) (Kaufmann et al., 2024), which builds on an SFT model. RLHF involves training an additional reward model using a large set of preference-labeled data to evaluate and guide the MLLMs. However, RLHF typically requires running three models simultaneously—the reference model, the reward model, and the policy model—which leads to increased computational complexity and memory usage. Direct Preference Optimization (DPO) (Rafailov et al., 2023) offers another alternative by directly parameterizing the reward model, thereby eliminating the need for separate reward model training and scoring during fine-tuning. DPO reduces computational overhead and often achieves comparable or better performance with fewer preference-labeled samples, improving sample efficiency. Despite these advancements, MLLMs often generate short, inaccurate fragments in the responses regarding flooding depth, particularly when describing the water depth of reference objects in images. These errors are typically overlooked by evaluation mechanisms that focus on overall response quality (Wang et al., 2024d). Therefore, developing a targeted fine-tuning method is essential to correct these segment-level errors and enhance the accuracy of the generate responses.

This paper proposes a Segment-level Direct Preference Optimization-based Multimodal Large Language Model (SDPO-MLLM) focused on the extraction of urban flooding depth from multimodal social media flood data. The multimodal water depth extraction task is first decomposed into three subtasks and a multimodal dataset is annotated for fine-tuning and evaluation of the model. Then, a SDPO loss combined with SFT loss is proposed to train the MLLM, while the Low-Rank Adaptation (LoRA) (Hu et al., 2022) method is used to decrease the computational cost during training. Finally, the performance on the multimodal tasks is comprehensively evaluated, while its effectiveness is further validated through water depth quantification as well as water depth mapping and analysis. Overall, the primary content and contributions of this paper are summarized as follows:

- A novel SDPO-MLLM is proposed, which integrates a training strategy combining supervised fine-tuning with Segment-level Direct Preference Optimization, effectively facilitating model alignment for water depth information extraction by integrating image-text data.
- The efficient and accurate extraction methods for multimodal data are constructed in SDPO-MLLM to adapt the generative response of MLLM, including text water depth extraction, image water depth extraction, and water depth level classification.
- An event-based extraction method is proposed to enhance waterlogged area localization, which improves the structured and organized extraction of discontinuous and overlapping locations in the text.
- A water-level classification and quantification method is developed that establishes classification levels and estimates the depth range for each category based on multiple reference objects, enhancing the overall water depth estimation for waterlogged areas containing multiple objects.
- The innovative approach integrating multimodal image-text data is used to assess waterlogged areas, further validating the model's effectiveness and contributing to waterlogging risk management and response strategies in a large-scale urban environment.

2. Related work

The estimation of urban flooding depth from social media data relies primarily on three data modalities: text, images, and videos. For text data, two main approaches are used to extract water depth information. The first approach (Khan et al., 2022) classifies the text into different water depth levels and assesses the overall water depth of the entire text. However, the method cannot handle multiple water depth descriptions presented simultaneously within the same text, such as several numerical values for water depth. The second approach (Aarthy et al., 2022) first extracts water depth descriptions and then estimates water depth using keyword matching or classification methods. While this approach can accurately identify multiple water depth descriptions, it fails to establish spatial associations between water depths and location descriptions because the extraction process of each is performed in isolation. To address these limitations, this study models the task as event-based water depth extraction, where each flood event contains both a water depth description and a location argument, allowing effective extraction of water depth information from complex textual content.

For image data, current methods attempt to estimate water depth information by analyzing visual cues within the images. One approach (Chaudhary et al., 2019; Feng et al., 2020; Liu et al., 2024) uses object detection models to classify reference objects at various water levels (e.g., people, cars, bicycles) into different categories, thereby estimating water depth. While this approach can stably identify multiple objects and their corresponding water levels when sufficient high-quality

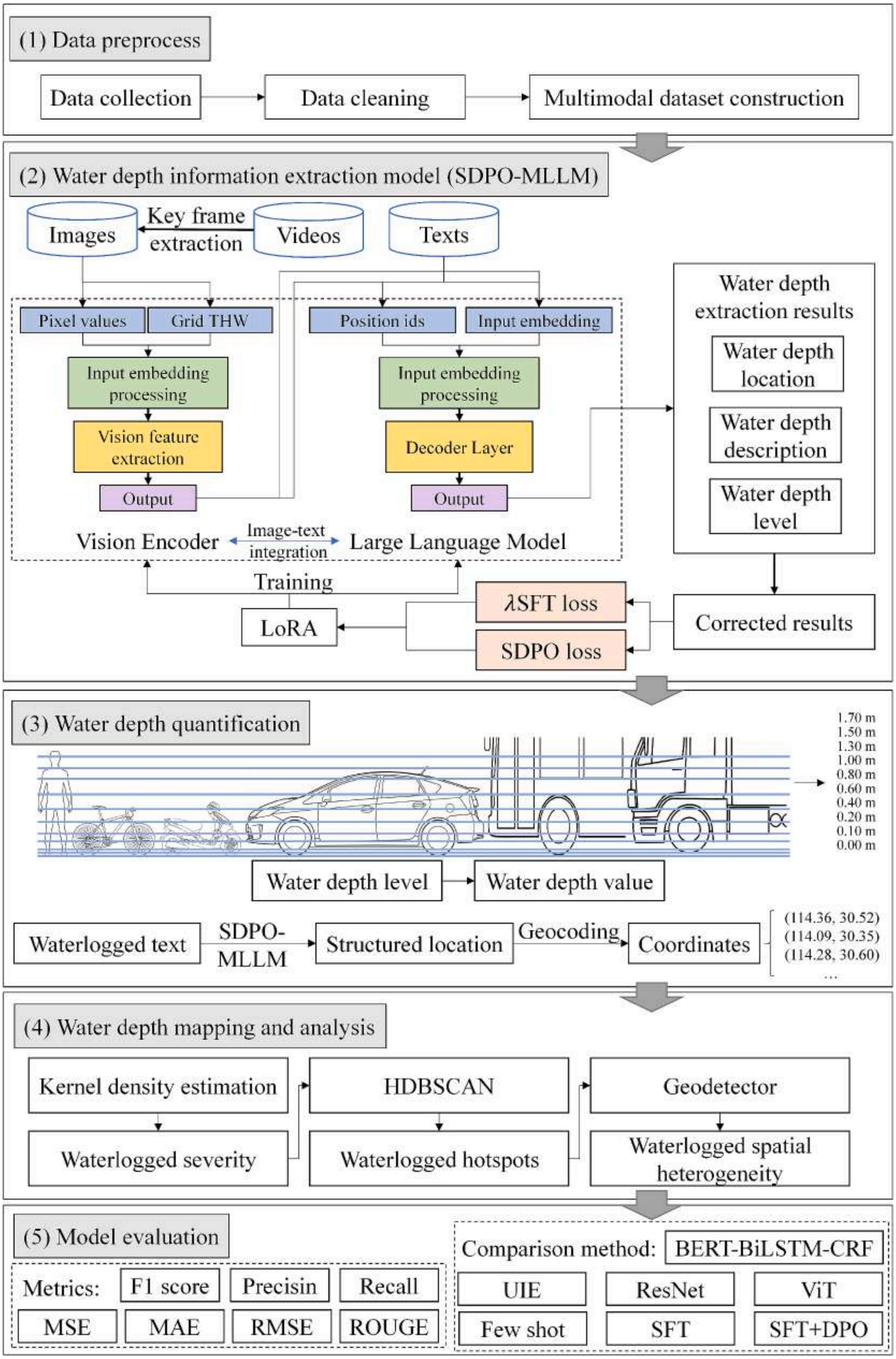


Fig. 1. The SDPO-MLLM framework.

training data is available, its implementation remains challenging and resource-intensive for data annotation. In addition, the method's reliance on predefined reference object types limits the scalability of the model. Another approach (Wu et al., 2024a; Yan et al., 2023) estimates the overall water depth of an image using image classification or regression. For example, Chaudhary et al. (2020) uses a multi-task ranking method to regress the water depth. However, this method is susceptible to interference from the image background, and the water depth at which each reference object is submerged may differ. In addition, some studies attempt to estimate water depth numerically by utilizing fixed-size reference objects or assuming ideal reference objects, such as road signs (Alizadeh Kharazi and Behzadan, 2021), traffic cones (Jiang et al., 2020), and pedestrians (Li et al., 2023a). For example, Qin and Shen (2025) estimate water depth by analyzing the refraction-induced displacement of road markings in images captured by traffic cameras. While these methods are reliable in specific environments, their implementation is challenging with diverse social media data. For the video data, research on water depth estimation is limited (Zhu et al., 2024b). In most cases, key frames are obtained from videos and treated as separate frames for analysis. For example, Hao et al. (2022) extracts the time periods when vehicles appear in surveillance videos and detects their water levels using object detection. In this paper, the MLLMs that integrate image-text data are utilized to enhance the robustness and flexibility of water depth estimation.

For the localization of flooding-related data, one approach (Feng et al., 2020) is to use geotags from social media data, which provide the coordinates of the poster's location. However, only 1–2 % of posts contain this information, and the location provided does not necessarily correspond to the actual flooding site. Another approach extracts location descriptions from text. The informal and irregular nature of location descriptions in social media data, along with the infrequent use of structured address formats, presents a challenge to accurately extracting this information. Previous methods (Berragan et al., 2023) mainly depend on Named Entity Recognition (NER) models to detect specific toponymies, like administrative divisions, streets, and points of interest (POIs). However, these models often struggle to model non-continuous or hierarchically nested locations (Chen et al., 2022). This study presents a generative extraction method based on MLLMs to extract and organize complex location expressions, thereby improving the accuracy of subsequent location geocoding and fully exploiting social media data.

By constructing generative tasks, multi-task depth estimation models can be developed based on MLLMs that integrate multimodal data, facilitating collaborative modeling across tasks and improving both performance and robustness. These models have already been preliminarily explored for their potential in disaster management field (Zhang et al., 2024). For example, Hu et al. (2023) used geo-knowledge to guide Generative Pre-trained Transformer (GPT) in location information extraction from disaster messages. Zhu et al. (2024a) developed a flood knowledge-constrained Large Language Models (LLMs) to improve citizens awareness of flood disasters. Additionally, Akinboyewa et al. (2024) applied GPT-4 to estimate flood depth of social media images. However, current methods rely solely on carefully designed prompts to implement specific tasks, limiting the full potential of MLLMs. Efficient fine-tuning is therefore required. Existing approaches primarily focus on fine-tuning LLMs. The study proposes Kahneman-Tversky Optimization (KTO) (Ethayarajh et al., 2024), based on prospect theory, to model human preferences. Another approach introduces Odds Ratio Preference Optimization (ORPO) (Hong et al., 2024), which integrates an Odds Ratio loss into the SFT loss. Additionally, Simple Preference Optimization (SimPO) (Meng et al., 2024) method is developed by omitting the reference model in the DPO loss to reduce the computational cost. However, fine-tuning methods for task-specific MLLMs require further investigation (Yu et al., 2024). This paper presents a Segment-level DPO fine-tuning method to build a robust and generalizable system for extracting water depth and its location from various social media sources, including text, images, and videos.

3. Methodology

3.1. Model framework

The framework of SDPO-MLLM is presented in Fig. 1. The model consists of the following: (1) Data preprocess; (2) Water depth information extraction model (SDPO-MLLM); (3) Water depth quantification; (4) Water depth mapping and analysis; (5) Model evaluation.

Table 1
The keywords used for crawling flood messages.

Category	Keywords
Flooding-related	Ponding, flooding, inundation, water immersion, water rise, water disaster, urban waterlogging, flood, submersion, water depth
Weather-related	Heavy rain, rainstorm, downpour, intense rainfall, heavy precipitation, rainwater, precipitation, storm
Natural environment-related	Low-lying terrain, river, waterway
Infrastructure-related	Drainage system, drainage pipeline, flood resistance, flood prevention facilities, flood embankment

3.2. Data preprocess

3.2.1. Data collection

The flood-related data is collected from Sina Weibo (https://weibo.com/) and the Wuhan City Message Board (https://liuyan.cjn.cn/). Data from Sina Weibo is retrieved using an Application Programming Interface (API) based on predefined keywords and a specified time range. To minimize the risk of missing flood-related posts, an extensive list of relevant keywords is incorporated. Table 1 details the keywords, which are manually refined and expanded by referencing previous studies (Li et al., 2023b). Additionally, to enhance the variety of the dataset, Wuhan City Message Board data is included as an additional data source. Each record from both data sources includes a textual description, with some records also containing multiple images or a video.

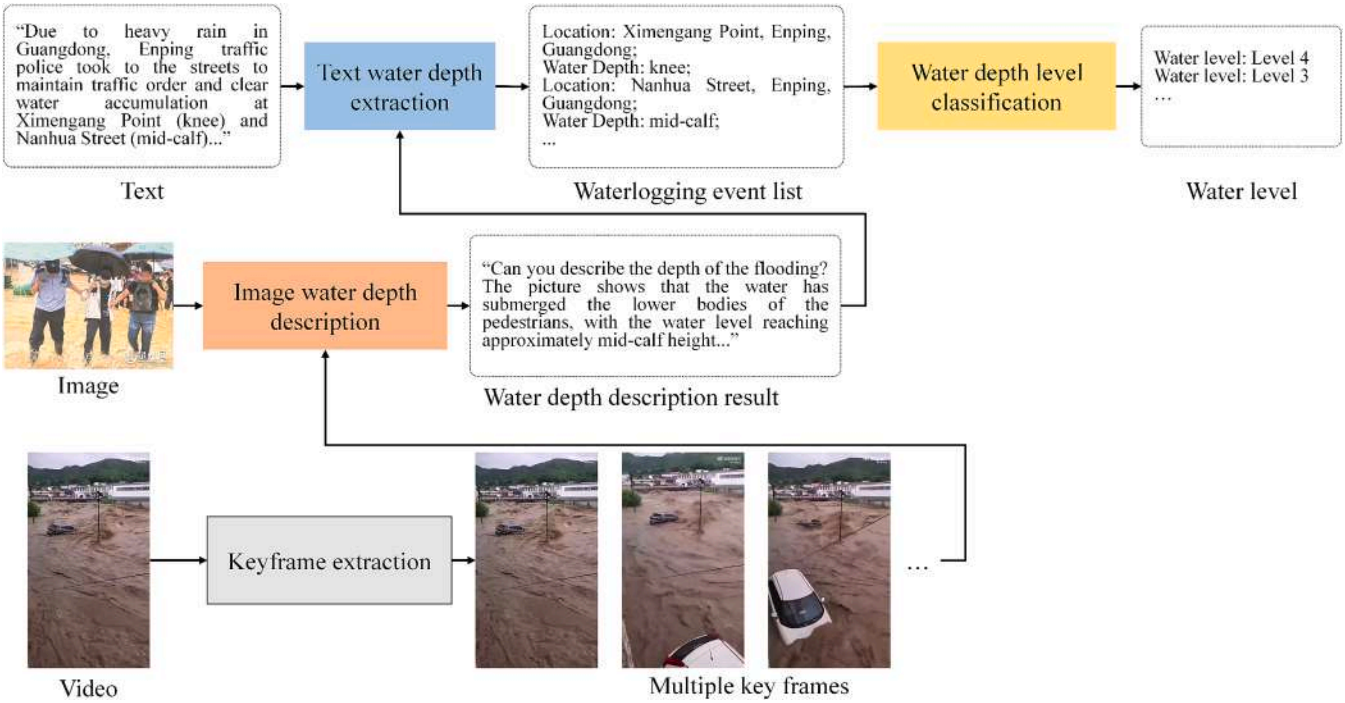
3.2.2. Data cleaning

The dataset collected using keywords contains irrelevant and duplicate posts. Irrelevant posts include advertisements, entertainment content, garbled text, posts that contain the keyword but are semantically unrelated, and those describing other disasters or weather events. Duplicates consist of reposts or repeated posts with identical or highly similar text. Flood-related posts are required to indicate that flooding occurs at a specific location, either through a geo-tag or a textual geographic description to enable spatial localization, while information about the severity of flooding is not mandatory. Data cleaning is performed as follows: regular expressions remove emojis, user information, and URLs; duplicates are detected using the single-pass method with Sentence-BERT (Reimers and Gurevych, 2019) embeddings, discarding texts with cosine similarity above 0.8. Posts are retained only if they contain geo-tags or textual toponyms, the latter extracted using NER tools (He and Choi, 2021). The same tools are also used to identify and exclude posts mentioning celebrities or public figures. Finally, a binary classifier based on Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019), fine-tuned on annotated data, filters the remaining semantically irrelevant samples.

3.2.3. Multimodal dataset construction

As shown in Fig. 2, to effectively extract water depth information, this study decomposes the task into three subtasks based on the characteristics of MLLMs: text water depth extraction, image water depth description, and water depth level classification. The cleaned dataset is then used to construct a multimodal water depth dataset for training and evaluating MLLMs. In addition, the video task is treated as an image water depth description of multiple key frames. Therefore, the following sections primarily focus on modeling water depth extraction from text and images.

For text data, text water depth extraction is treated as an event extraction task, as texts often refer to waterlogging at multiple locations. The water depth description at each location is considered an event comprising two arguments: location description and water depth description. If certain information is missing, it will be left blank. The spatial description includes administrative divisions, points of interest,


Text
Text water depth extraction
Waterlogging event list
Water depth level classification
Water level
Image
Image water depth description
Water depth description result
Video
Keyframe extraction
Multiple key frames
…

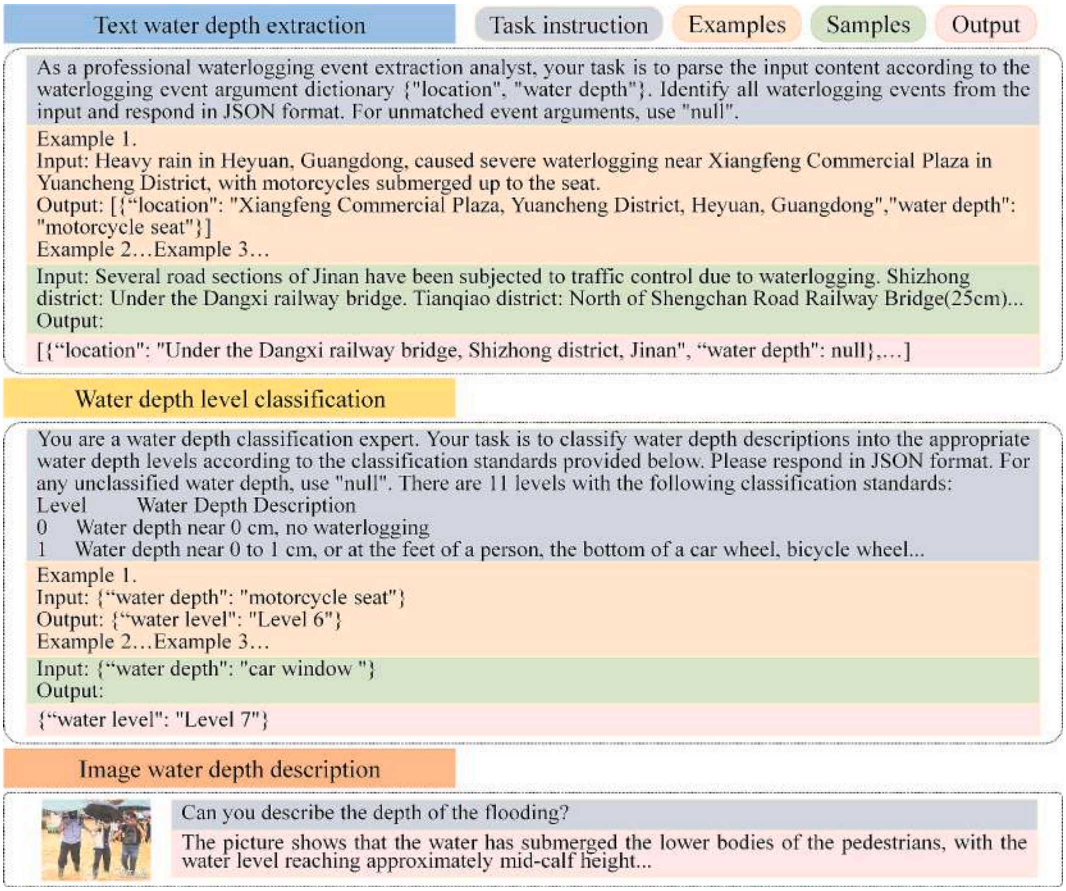
"Due to heavy rain in Guangdong, Enping traffic police took to the streets to maintain traffic order and clear water accumulation at Ximengang Point (knee) and Nanhua Street (mid-calf)..."

Location: Ximengang Point, Enping, Guangdong;
Water Depth: knee;
Location: Nanhua Street, Enping, Guangdong;
Water Depth: mid-calf;
...

Water level: Level 4
Water level: Level 3
…

"Can you describe the depth of the flooding? The picture shows that the water has submerged the lower bodies of pedestrians, with the water level reaching approximately mid-calf height..."

Fig. 2. The multimodal data process.


Text water depth extraction
Task instruction
Examples
Samples
Output
Water depth level classification
Image water depth description

As a professional waterlogging event extraction analyst, your task is to parse the input content according to the waterlogging event argument dictionary {"location", "water depth"}. Identify all waterlogging events from the input and respond in JSON format. For unmatched event arguments, use "null".

Example 1.
Input: Heavy rain in Heyuan, Guangdong, caused severe waterlogging near Xiangfeng Commercial Plaza in Yuancheng District, with motorcycles submerged up to the seat.
Output: [{"location": "Xiangfeng Commercial Plaza, Yuancheng District, Heyuan, Guangdong","water depth": "motorcycle seat"}]
Example 2…Example 3…

Input: Several road sections of Jinan have been subjected to traffic control due to waterlogging. Shizhong district: Under the Dangxi railway bridge. Tianqiao district: North of Shengchan Road Railway Bridge(25cm)...
Output:

[{"location": "Under the Dangxi railway bridge, Shizhong district, Jinan", "water depth": null},…]

You are a water depth classification expert. Your task is to classify water depth descriptions into the appropriate water depth levels according to the classification standards provided below. Please respond in JSON format. For any unclassified water depth, use "null". There are 11 levels with the following classification standards:

Level	Water Depth Description
0	Water depth near 0 cm, no waterlogging
1	Water depth near 0 to 1 cm, or at the feet of a person, the bottom of a car wheel, bicycle wheel...

Example 1.
Input: {"water depth": "motorcycle seat"}
Output: {"water level": "Level 6"}
Example 2…Example 3…

Input: {"water depth": "car window "}
Output:

{"water level": "Level 7"}

Can you describe the depth of the flooding?

The picture shows that the water has submerged the lower bodies of the pedestrians, with the water level reaching approximately mid-calf height...

Fig. 3. Prompt templates for the three water depth extraction subtasks.

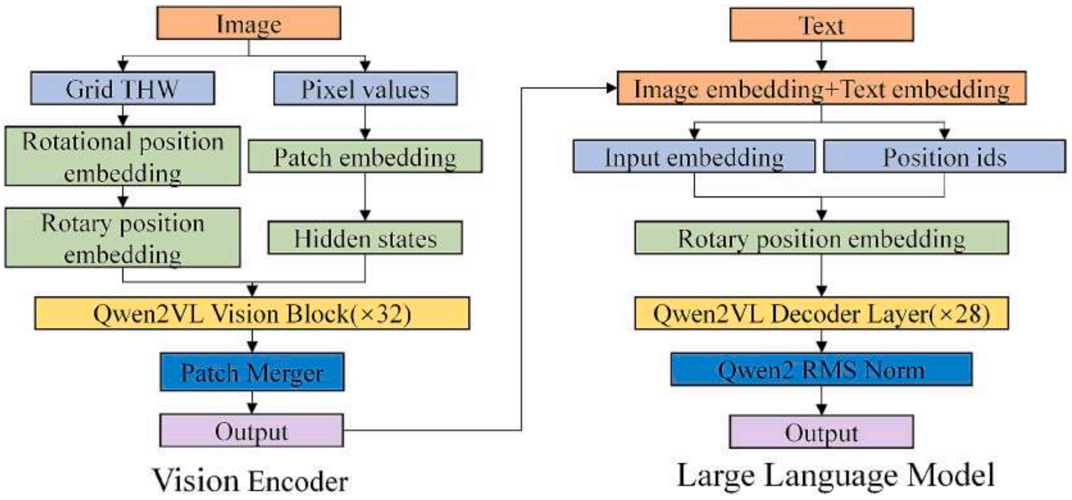
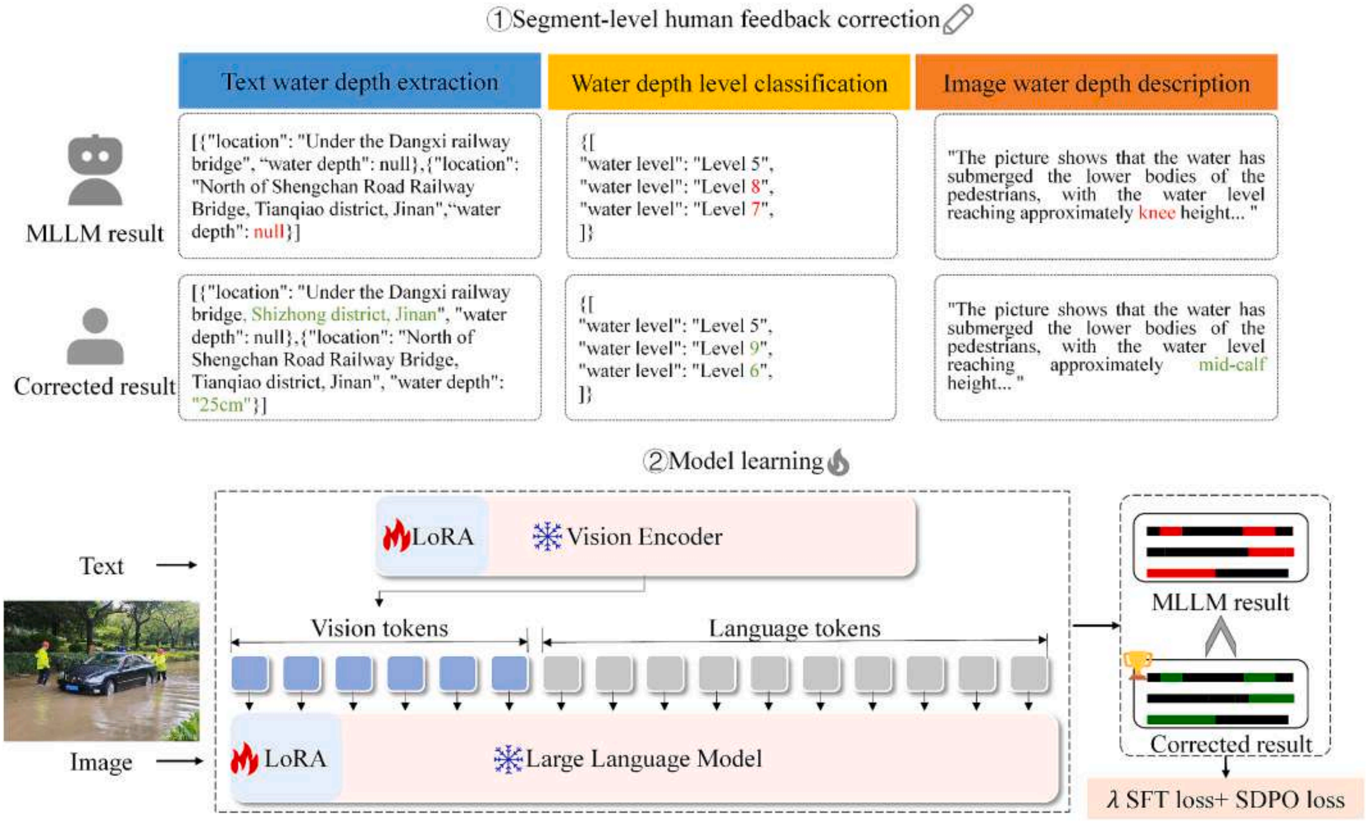


Image
Grid THW
Pixel values
Rotational position embedding
Patch embedding
Rotary position embedding
Hidden states
Qwen2VL Vision Block(×32)
Patch Merger
Output
Vision Encoder
Text
Image embedding+Text embedding
Input embedding
Position ids
Rotary position embedding
Qwen2VL Decoder Layer(×28)
Qwen2 RMS Norm
Output
Large Language Model

Fig. 4. Qwen2-VL-7b model structure.

roads, addresses, etc., while also considering discontinuous location entities and complex spatial relationships. For example, in the sentence of "Due to heavy rain in Guangdong, Enping traffic police took to the streets to maintain traffic order and clear water accumulation at Ximengang Point and Nanhua Street" two discontinuous and overlapping locations are mentioned: "Ximengang Point, Enping, Guangdong" and "Nanhua Street, Enping, Guangdong". Completely extracting both locations, rather than only "Ximengang Point" and "Nanhua Street" can reduce ambiguities caused by similarly named locations during subsequent geocoding and localization processes. The water depth descriptions generally fall into two categories. The first are absolute descriptions, such as '5cm' or '20 cm'. The second are relative descriptions, which refer to the water depth in relation to a reference object, such as 'the water level reached the motorcycle seat'. Subjective terms such as 'very deep' or 'quite deep' are ignored because they are difficult to quantify.

For image data, image water depth description involves instructing the MLLM to identify and describe the flooded parts of reference objects, like 'knee' for a person or 'tire' for a vehicle. The responses are first collected from the MLLMs, then incorrect segments, especially incorrect or missing water depth descriptions, are corrected to construct positive samples of the dataset. The image descriptions are then processed by the text water depth extraction task, which sets the location description to null or replaces it with the flooded object to extract the water depth of multiple targets in the image. This two-step approach allows the model to focus on identifying objects and their corresponding water depths in



MLLM result:

[{"location": "Under the Dangxi railway bridge", "water depth": null},{"location": "North of Shengchan Road Railway Bridge, Tianqiao district, Jinan","water depth": null}]

{[
"water level": "Level 5",
"water level": "Level 8",
"water level": "Level 7",
]}

"The picture shows that the water has submerged the lower bodies of the pedestrians, with the water level reaching approximately knee height... "

Corrected result:

[{"location": "Under the Dangxi railway bridge, Shizhong district, Jinan", "water depth": null},{"location": "North of Shengchan Road Railway Bridge, Tianqiao district, Jinan", "water depth": "25cm"}]

{[
"water level": "Level 5",
"water level": "Level 9",
"water level": "Level 6",
]}

"The picture shows that the water has submerged the lower bodies of the pedestrians, with the water level reaching approximately mid-calf height... "

Fig. 5. Model learning process with SFT and SDPO loss.

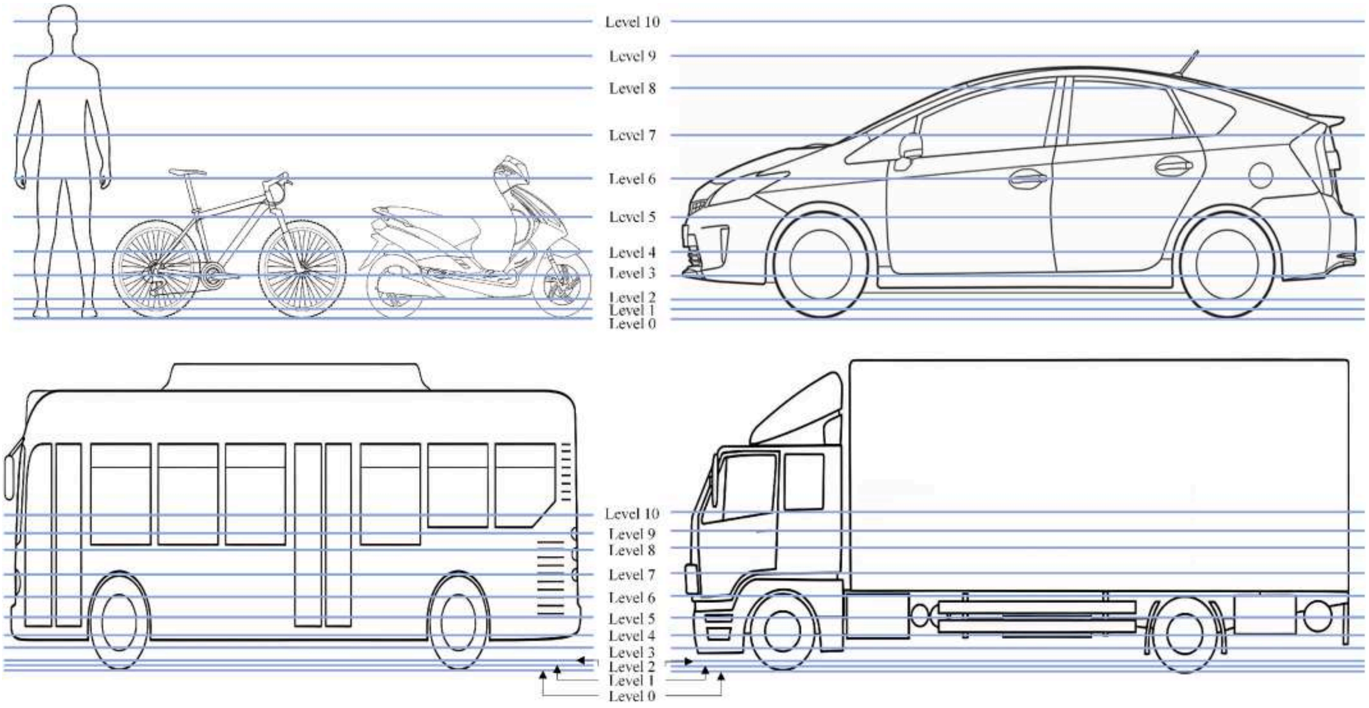


Fig. 6. The water levels corresponding to the reference objects.

the image during the first stage, while ensuring that the description task in MLLMs is more generalizable than directly classifying or detecting water depth levels for objects in the image. Finally, water depth level classification categorizes the extracted water depth descriptions from text and images into predefined levels.

To avoid model overfitting, a set of prompts is designed specifically for each of the three tasks, as shown in Fig. 3. These prompts consist of task instruction, few shot examples, and samples to process. Task instruction specifies the task to be performed and the required output format, such as producing textual results in JSON format. Few shot examples include three input samples along with their corresponding outputs in JSON format. Including more examples could exceed the maximum input length and increase the inference time, so the number of examples is limited to three. In addition, no examples are included for the image water depth description task. A separator is added between the examples and the samples to be processed to reduce the likelihood of the model replicating the examples provided.

3.3. Water depth information extraction model (SDPO-MLLM)

Since the collected social media data is mainly composed of Chinese, Qwen2-VL-7b (Wang et al., 2024b) is selected as the base model to construct the multimodal water depth information extraction model. The training set of this model contains more Chinese corpus, has relatively fewer parameters, so it has relatively better adaptability to the water depth dataset. Fig. 4 shows the two components of the model: A Vision Transformer (ViT) (Dosovitskiy et al., 2021) serving as the visual encoder, and a Qwen-2 (Yang et al., 2024) LLM for generating results.

To enable the model to generate the desired responses based on instructions, especially when describing image content with water depth information, a training method for MLLMs is designed that combines SFT and SDPO losses, which can be represented as Eq. (1).

$$\mathscr{L}_{total} = \lambda \mathscr{L}_{SFT} + \mathscr{L}_{SDPO} \# \quad (1)$$

where λ is weight parameter. The SFT loss is employed to prompt the model to mimic the dataset while adjusting the probability distribution of model to meet the requirements of SDPO, which is given by Eq. (2).

$$\mathscr{L}_{SFT} = -\mathbb{E}_{(x,y)}[log\pi(y|x)]\# \quad (2)$$

where $\pi(y|x)$ is the model's predicted probability for the expected output y given the input x, and $log\pi(y|x)$ is given by Eq. (3).

$$log\pi(y|x) = \sum\nolimits_{y_i \in y} logp(y_i|x,y_i)\# \quad (3)$$

where y_i is the i-th token of the response y. As shown in Fig. 5, considering the relatively long textual responses related to image water depth description task, where the part describing the water depth is relatively short, it is difficult to achieve fine-grained water depth alignment using DPO. Therefore, segment- level DPO is used to align text fragments where the model's judgment of the water depth is incorrect. DPO transforms the reinforcement learning objective in RLHF into a supervised learning objective that expresses the reward function $r(x,y)$ in terms of both its optimal policy model $\pi(y|x)$ and reference model $\pi_{ref}(y|x)$. The reward function can be represented as Eq. (4).

$$r(x,y) = \beta log\frac{\pi(y|x)}{\pi_{ref}(y|x)} + \beta logZ(x)\# \quad (4)$$

where β is a constant and $Z(x)$ is the partition function. Furthermore, the DPD learning objective is defined by Eq. (5).

$$\begin{aligned}\mathscr{L}_{DPO} = &-\mathbb{E}_{(x,y_w,y_l)}[log\sigma(r(x,y_w) \\ &- r(x,y_l))] = -\mathbb{E}_{(x,y_w,y_l)}\left[log\sigma\left(\beta log\frac{\pi(y_w|x)}{\pi_{ref}(y_w|x)} - \beta log\frac{\pi(y_l|x)}{\pi_{ref}(y_l|x)}\right)\right]\#\end{aligned} \quad (5)$$

where the reference model $\pi_{ref}(y_w|x)$ refers to the initial base model, whose parameters remain fixed, while only the policy model $\pi(y|x)$ is updated i.e. DPO allows to optimize policy models directly from paired preference data without the reward model. To improve the learning ability from the corrected text fragment, especially the water depth text fragment, it is essential to increase its contribution to the overall score of the text. Therefore, the scoring method is refined to a weighted sum of the text fragments, which is given by Eq. (6).

Table 2
The water level and estimation depth subject to a person.

Level	Body parts	Height range (m)	Estimation of water depth (m)	Depth range under ± 10 % variation (m)
0	None	0.00	0.00	0.00 – 0.00
1	Instep	0.00 – 0.01	0.01	0.01 – 0.01
2	Ankle	0.01 – 0.10	0.10	0.09 – 0.11
3	Calf	0.10 – 0.20	0.20	0.18 – 0.22
4	Knee	0.20 – 0.40	0.40	0.36 – 0.44
5	Thigh	0.40 – 0.60	0.60	0.54 – 0.66
6	Hip	0.6 – 0.80	0.80	0.72 – 0.88
7	Waist	0.80 – 1.00	1.00	0.90 – 1.10
8	Chest	1.00 – 1.30	1.30	1.17 – 1.43
9	Neck	1.30 – 1.50	1.50	1.35 – 1.65
10	Temple, eyes	1.50 – 1.70	1.70	1.53 – 1.87

$$log\pi(y|x) = \frac{1}{N}\left[\sum_{y_i \in y_u} logp(y_i|x,y_{<i}) + \gamma\sum_{y_i \in y_c} logp(y_i|x,y_{<i})\right]\# \tag{6}$$

where y_u is the unchanged fragment, y_c denotes the corrected text fragment, and $\gamma > 1$ is a weight parameter. As the value of γ increases, the impact of y_c on the total score also grows. N is used for normalization to prevent longer responses from receiving higher scores, where $N = |y_u| + \gamma|y_c|$.

In addition, LoRA is employed to improve the training efficiency. This method conceptualizes the model training as an incremental process applied to the original parameters. To encode the parameter increments with fewer parameters, a low-rank decomposition is performed for the pre-trained weight matrix $W_0 \in \mathbb{R}^{d\times k}$ to represent the parameter updates ΔW, which is given by Eq. (7).

$$W_0 + \Delta W = W_0 + BA\# \tag{7}$$

where $B \in \mathbb{R}^{d\times r}$, $A \in \mathbb{R}^{r\times k}$, $and\, r \ll \min(d,k)$, During fine tuning, W_0 is fixed, while the parameters of A and B are trained. The forward process $h = W_0x$ can be represented as $h = W_0x + \Delta Wx = W_0x + BAx$.

3.4. Water depth quantification

In the water level classification task, the water levels are divided into 11 categories, extending the criteria from previous studies (Yan et al., 2023), as shown in Fig. 6. A person approximately 170 cm tall is used as a reference, and the height is then mapped to other reference objects, including bicycles, motorcycles, cars, etc. To better utilize the classified depth levels for estimating specific depth values, an estimated depth is constructed for each level, as shown in Table 2. This approach allows the conversion of depth levels into numerical values, which facilitates the calculation of the average depth of multiple targets in images or videos.

To evaluate the impact of reference objects on water depth estimation and to quantify the uncertainty introduced by ambiguous or variable reference objects, a sensitivity analysis was performed. For each level, the corresponding water depth was recalculated by perturbing the original reference heights by ± 10 %, and the results are presented in Table 2. The analysis shows that such perturbations lead to a maximum deviation of 0.01–0.17 m in the estimated water depth, depending on the level. Furthermore, in practical situations, extreme deep-water events are relatively rare, whereas shallow flooding events occur much more frequently. This results in a skewed distribution of water depths in the dataset, under which the overall average absolute deviation across all samples tends toward 0.01 m. Overall, these findings demonstrate that the water level classification method is robust, providing reliable depth estimations even under ambiguous or variable reference conditions.

In addition, to locate waterlogged areas, the structured location extracted by SDPO-MLLM is geocoded using the Gaode API to obtain coordinates, as it provides more reliable and accurate geocoding services within the study areas. However, the API has certain limitations in global applications. It is primarily optimized for Chinese address formats, such as processing abbreviations, misspellings, incomplete, or redundant address information. Moreover, its geographic coverage mainly focuses on mainland China. For studies requiring worldwide coverage, alternative services such as Google Maps or Mapbox may be more appropriate (Geoapify, 2021).

Then, flooding depth extracted from text, images, and videos is integrated. During integration, records containing location information are first retained. The next step is to determine if the record specifies a water depth. If an exact depth value is provided, it is used as the water depth for that location, disregarding estimates from other modalities. If the depth is described relative to a reference object or is not explicitly mentioned, images and videos are used to estimate water depth. The final water depth for each location is calculated by averaging the depths derived from all three modalities.

3.5. Water depth mapping and analysis

To identify the spatio-temporal distribution and trends of waterlogging hotspots in social media data, Kernel Density Estimation (KDE) is applied to analyze the multimodal water depth information at the city scale. The KDE can be used to identify urban waterlogging hotspot areas, which typically correspond to regions prone to waterlogging. These areas are strongly linked to factors such as urban geography, development, and rainfall intensity. For a spatial location (x,y), the KDE function $f(x,y)$ can be represented as the probability density of flood points $\{(x_1,y_1), (x_2,y_2), \cdots, (x_n,y_n)\}$ over the spatial area, as given by the Eq. (8).

$$f(x,y) = \frac{1}{r^2}\sum_{i=1}^{n}\left[\frac{3}{\pi}\cdot k_i\left(1-\left(\frac{d_r}{r}\right)^2\right)^2\right](d_r < r)\# \tag{8}$$

where r is the search radius, the weight of water depth at the flood point (x_i,y_i) is denoted by k_i, while d_r is the length between the flood point and the location (x,y). Only flood points within the radius r contribute to the calculation.

Additionally, clustering based on the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBCSAN) algorithm is applied to identify regions with concentrated flood points, and by comparing the kernel density values of these clusters, the severity of flooding in these areas is assessed. HDBSCAN (Campello et al., 2013) is a clustering algorithm that extends DBSCAN by incorporating a hierarchical approach, enabling it to handle datasets with varying densities effectively. It introduces the concept of mutual reachability distance to account for density differences. This distance is calculated as the maximum of the core distances of two points and the actual distance between them. HDBSCAN then uses these distances to construct a minimum spanning tree (MST) that represents the connectivity of points based on density. The MST is then condensed into a hierarchy of clusters by varying the density threshold. Clusters form and merge dynamically as the density level changes. HDBSCAN has an average-case time complexity of approximately O(nlogn), making it efficient and scalable for large datasets.

To further explore the impact and variation of various geographical factors on waterlogging hotspot regions, Geodetector (Wang et al., 2024a) is employed to assess the spatial heterogeneity of waterlogging hotspots, which divides the geographic space into different regions and uses the q-value to quantify how effectively X explains Y. A higher q-value signifies a greater explanatory influence of the factor on Y, whereas the p-value indicates the statistical significance of the factor. The formula for calculating the q-value is calculated using Eq. (9).

$$q(\mathrm{Y}|\mathrm{X}) = 1 - \frac{\sum_{i=1}^{n} \mathrm{N}_i \bullet \sigma_i^2}{\mathrm{N}\cdot\sigma^2} \in [0,1]\# \tag{9}$$

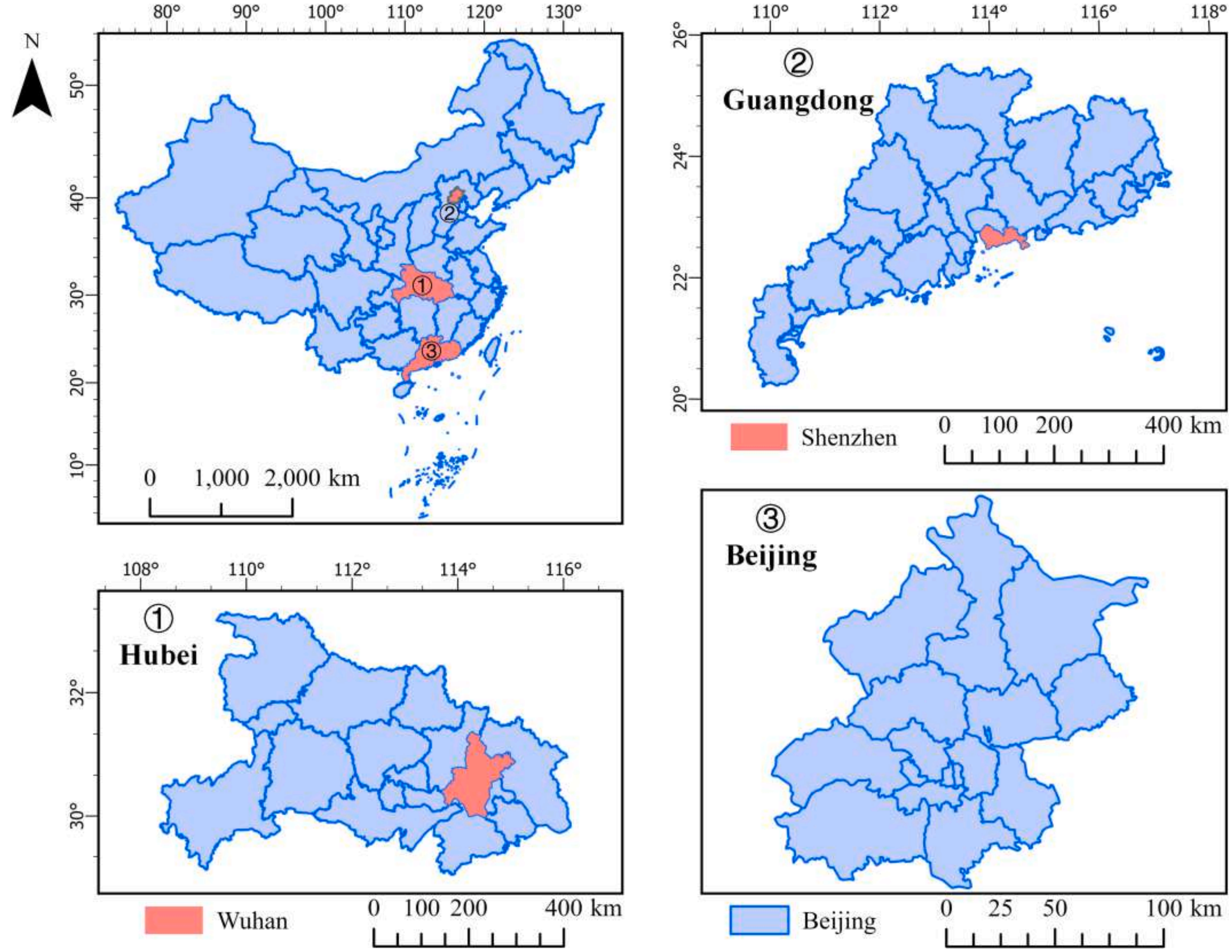


Fig. 7. Study area for urban waterlogging assessment.

where i refers to the category or division of X, N is the overall count of regional units, N_i is the count of regional units in category i, σ_i^2 is the variance of Y within category i, and σ^2 is the total variance of the dependent variable across all regions.

3.6. Model evaluation

For the text water depth extraction task, precision, recall and F1-score are used to evaluate the results of location and water depth description, which is given by Eqs. (10)–(12).

$$F1 = \frac{2 \times P^*R}{P + R}\# \tag{10}$$

$$P = \frac{TP}{TP + FP}\# \tag{11}$$

$$R = \frac{TP}{TP + FN}\# \tag{12}$$

Since LLMs often add connective words to improve the flow of responses, especially for discontinuous entities, this can prevent the extracted results from exactly matching the true labels. Therefore, extraction results with a sequence match similarity greater than 80 % to the true labels are considered correct (Han et al., 2024). In addition, the extracted location results are further verified by converting them to coordinates through geocoding, with a distance threshold of 500 m used to determine correct localization. MSE, MAE, and RMSE are used to further evaluate the water depth extraction results after classification and quantification. These metrics provide an overall estimation of the disparity between the true and estimated depth, which are given by Eqs. (13)–(15).

$$MSE = \frac{1}{n}\sum\nolimits_{i=1}^{n}(y_i - x_i)^2\# \tag{13}$$

$$MAE = \frac{1}{n}\sum\nolimits_{i=1}^{n}|y_i - x_i|\# \tag{14}$$

$$RMSE = \sqrt{\mathrm{MSE}}\# \tag{15}$$

where y_i is predicted water depth and x_i is true value. For the image water depth description task, Recall-Oriented Understudy for Gisting Evaluation (ROUGE) is utilized to examine the consistency of the output water depth text, which evaluates the degree of n-grams matching between the output text and the target text. The F1-ROUGE score is calculated using Eq. (16).

$$\mathrm{F1}_{ROUGE} = \frac{2 \times P_{ROUGE}{}^*R_{ROUGE}}{P_{ROUGE} + R_{ROUGE}}\# \tag{16}$$

where P_{ROUGE} is the fraction of the count of matching texts to the output

Table 3
Location extraction and localization result on the text water depth dataset.

Method	Extraction F1-score	Extraction Precision	Extraction Recall	Localization F1-score	Localization Precision	Localization Recall
BERT-BiLSTM-CRF	42.40 %	36.84 %	49.93 %	56.79 %	50.77 %	64.44 %
UIE	50.19 %	**82.45 %**	36.08 %	54.56 %	**90.35 %**	39.08 %
SDPO-MLLM (Few shot)	63.24 %	64.98 %	61.59 %	70.44 %	72.85 %	68.19 %
SDPO-MLLM (SFT)	78.83 %	77.99 %	79.70 %	86.35 %	88.95 %	83.89 %
SDPO-MLLM (SFT + DPO)	81.82 %	80.94 %	82.72 %	86.98 %	89.20 %	84.86 %
SDPO-MLLM (SFT + SDPO)	**83.19 %**	81.91 %	**84.50 %**	**87.51 %**	88.95 %	**86.11 %**

text, and R_{ROUGE} is the ratio of the count of overlapping texts to the target text. ROUGE-1, ROUGE-2 and ROUGE-L are used to measure concordance at different levels. These metrics calculate overlap based on 1-gram, 2-gram and longest common subsequence, respectively. Similarly, MSE, MAE and RMSE are used to assess the error after the image water depth descriptions are quantified. Finally, the F1-score is employed to assess the water level classification task, where words such as 'very deep' and 'fairly deep' are classified as a separate category and excluded from the subsequent water depth quantification.

4. Experiments and results

4.1. Experiments

The multimodal water depth dataset contains 2843 text records and 1563 images derived from collected, cleaned and annotated social media data. A test set is constructed from three cities—Wuhan, Shenzhen, and Beijing—with each city contributing 100 texts and 50 images. The remaining data are divided into training and validation sets at an 8:2 ratio. The text annotations include 4491 location descriptions and 1267 water depth descriptions. Each image contains one or more water depth description texts, with each description corresponding to the water depth extraction results and water level classification labels.

The fine-tuning process is performed for 10 epochs on 4 RTX A6000 GPUs, with the SFT loss weight parameter λ of 1, the SDPO loss parameter β of 0.1, and the corrected text fragment weight of 5. The LoRA parameters are configured to rank 8, alpha 32, and dropout probability 0.05. During model inference, repetition penalty is set to 1, temperature to 0.01, top p to 0.001, and top k to 1 to minimize randomness in the model outputs. Additionally, regular expressions are used to extract results in JSON format from the model's responses. If the model does not return the correct format, the inference process is repeated until the output could be correctly recognized, with a maximum of 10 attempts.

As shown in Fig. 7, Wuhan, Shenzhen, and Beijing in China are selected as study areas to further evaluate the applicability and effectiveness of SDPO-MLLM in different urban environments. Wuhan, characterized by low-lying terrain, frequent extreme rainfall events, and rapid urban expansion, faces a heightened risk of waterlogging. The case study focuses on urban waterlogging in Wuhan in 2022, where SDPO-MLLM is employed to extract water depth information from flood-related multimodal social media data. The integrated multimodal water depth data is then used to map and analyze inundated areas.

Shenzhen and Beijing are chosen to further test the model's applicability under different flooding conditions. Shenzhen, a coastal city in southern Guangdong Province adjacent to Hong Kong, experienced extreme rainfall from September 7 to 8, 2023, when Typhoon Haikui made landfall, leading to severe waterlogging risks. Flood-related messages were collected from September 7 to 16 to extract and analyze water depth information. Beijing, an inland city in northern China bordering Tianjin and Hebei Province, suffered extraordinary rainfall from July 28 to August 2, 2023, due to Typhoon Doksuri. The event caused flooding in rivers such as the Yongding and Juma within the municipality. Posts were collected from July 28 to August 28 for experimentation.

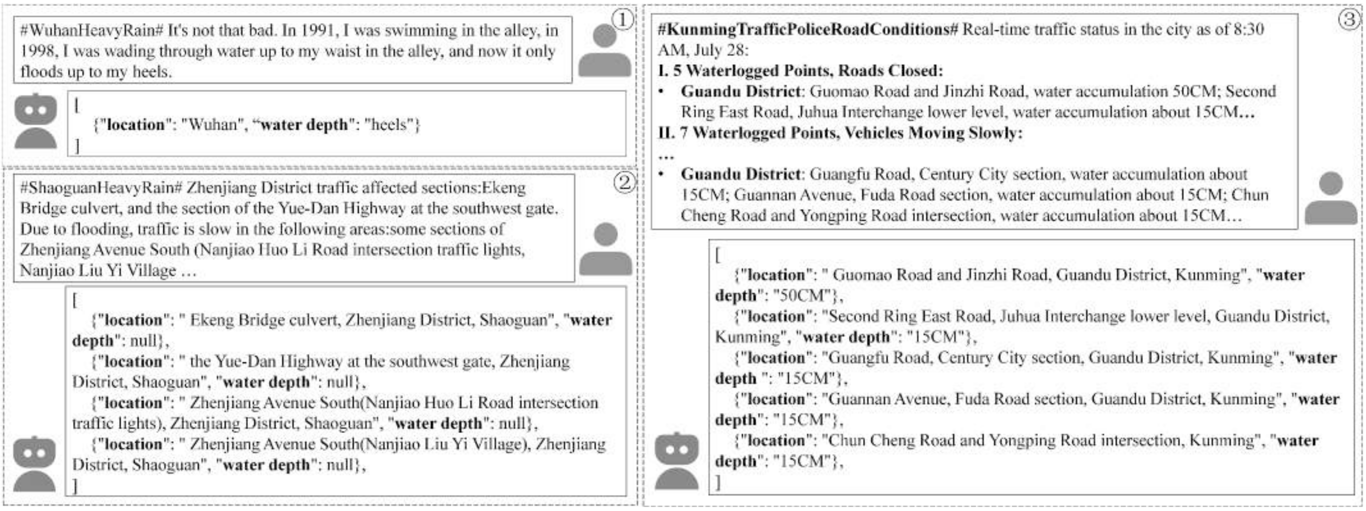
4.2. Results

The performance of SDPO-MLLM on three subtasks related to multimodal water depth information extraction is assessed and contrasted with other baseline models. For the text water depth dataset, the performance on location and water depth extraction tasks is evaluated and compared with BERT-BiLSTM-CRF, UIE, and SDPO-MLLM under different configurations. BERT-BiLSTM-CRF (Yan et al., 2023) combines the strengths of three components: BERT, which serves as an encoder layer that provides context-aware embeddings of input text; BiLSTM (Bi-directional Long Short-Term Memory), which models sequential dependencies bidirectionally; and CRF (Conditional Random Field), which acts as a decoder to produce globally optimal tag sequences. During training, water depth data is converted from dialog format to BIO format. UIE (Universal Information Extraction) (Lou et al., 2023) is a unified framework for information extraction tasks that treats these tasks as generative problems and uses the Structured Extraction Language for unified representation, which improves the adaptability and effectiveness across different extraction tasks. In addition, SDPO-MLLM is evaluated under different configurations: Few-shot, SFT, SFT + DPO, and SFT + SDPO. The Few-shot SDPO-MLLM performs the water depth information extraction task using a prompt and a few examples, without fine tuning. The SFT, SFT + DPO, and SFT + SDPO variants of SDPO-MLLM correspond to models trained with their respective loss functions.

Table 3 presents the location extraction and localization results. The SDPO-MLLM with SFT + SDPO achieved the best performance, with an improvement of 1.37–40.79 % on location extraction F1-score and 0.53–32.95 % on localization F1-score compared to other models. The accuracy of location extraction and localization implemented by the MLLMs generally outperformed BERT-BiLSTM-CRF and UIE because these two models lack the capability to handle longer and non-contiguous location descriptions. For instance, the UIE model tends to show higher precision but lower recall, indicating that it fails to extract some location entities. On the other hand, due to the limited sample size,

Table 4
Water depth extraction and quantification result on the text water depth dataset.

Method	Extraction F1-score	Extraction Precision	Extraction Recall	Quantification MSE	Quantification MAE	Quantification RMSE
BERT-BiLSTM-CRF	49.12 %	58.33 %	42.42 %	0.1048	0.1498	0.3237
UIE	57.42 %	77.92 %	45.45 %	0.0855	0.1284	0.2924
SDPO-MLLM (Few shot)	77.47 %	80.99 %	74.24 %	0.3431	0.2904	0.5857
SDPO-MLLM (SFT)	83.52 %	84.50 %	82.58 %	0.0517	0.0576	0.2274
SDPO-MLLM (SFT + DPO)	83.85 %	85.16 %	82.58 %	0.0276	0.0545	0.1660
SDPO-MLLM (SFT + SDPO)	**86.26 %**	**86.92 %**	**85.61 %**	**0.0210**	**0.0324**	**0.1449**



#WuhanHeavyRain# It's not that bad. In 1991, I was swimming in the alley, in 1998, I was wading through water up to my waist in the alley, and now it only floods up to my heels.

```
[
    {"location": "Wuhan", "water depth": "heels"}
]
```

#ShaoguanHeavyRain# Zhenjiang District traffic affected sections:Ekeng Bridge culvert, and the section of the Yue-Dan Highway at the southwest gate. Due to flooding, traffic is slow in the following areas:some sections of Zhenjiang Avenue South (Nanjiao Huo Li Road intersection traffic lights, Nanjiao Liu Yi Village …

```
[
    {"location": " Ekeng Bridge culvert, Zhenjiang District, Shaoguan", "water depth": null},
    {"location": " the Yue-Dan Highway at the southwest gate, Zhenjiang District, Shaoguan", "water depth": null},
    {"location": " Zhenjiang Avenue South(Nanjiao Huo Li Road intersection traffic lights), Zhenjiang District, Shaoguan", "water depth": null},
    {"location": " Zhenjiang Avenue South(Nanjiao Liu Yi Village), Zhenjiang District, Shaoguan", "water depth": null},
]
```

#KunmingTrafficPoliceRoadConditions# Real-time traffic status in the city as of 8:30 AM, July 28:

I. 5 Waterlogged Points, Roads Closed:

- **Guandu District**: Guomao Road and Jinzhi Road, water accumulation 50CM; Second Ring East Road, Juhua Interchange lower level, water accumulation about 15CM**…**

II. 7 Waterlogged Points, Vehicles Moving Slowly:

…

- **Guandu District**: Guangfu Road, Century City section, water accumulation about 15CM; Guannan Avenue, Fuda Road section, water accumulation about 15CM; Chun Cheng Road and Yongping Road intersection, water accumulation about 15CM…

```
[
    {"location": " Guomao Road and Jinzhi Road, Guandu District, Kunming", "water depth": "50CM"},
    {"location": "Second Ring East Road, Juhua Interchange lower level, Guandu District, Kunming", "water depth": "15CM"},
    {"location": "Guangfu Road, Century City section, Guandu District, Kunming", "water depth ": "15CM"},
    {"location": "Guannan Avenue, Fuda Road section, Guandu District, Kunming", "water depth": "15CM"},
    {"location": "Chun Cheng Road and Yongping Road intersection, Kunming", "water depth": "15CM"},
]
```

Fig. 8. Text water depth extraction examples.

these two models struggle to align with the requirements of this task. In contrast, the MLLMs, even in a Few-shot scenario using only three examples, performed better than both BERT and UIE. Furthermore, the Few-shot model achieves an F1-score of 63.24 %. After fine-tuning with SFT, the performance improves by 15.59 %. The performance is further enhanced when trained with SFT + DPO, while the SFT + SDPO configuration provides the best results. The results show that segment-based weighting effectively scores text and facilitates the model to learn from specific segments. Finally, the localization accuracy after location extraction is further evaluated, given the challenges with generative models in extracting locations with precision. The experiments indicate that the performance of localization is generally better than that of location extraction, indicating that while MLLMs may not extract locations with complete precision, they still perform well in terms of localization. For example, the F1-score difference between the model trained with SFT and the model trained with SFT + DPO is 2.99 % for location extraction but only 0.63 % for localization. This suggests that the inability of the model to accurately extract locations has minimal impact on the final analysis of the geographic distribution of waterlogging.

Table 4 shows the performance of the models on the water depth extraction and quantification tasks. The SDPO-MLLM with SFT + SDPO achieves the best performance, with an F1-score improvement ranging from 2.41 % to 37.14 % in the water depth extraction tasks. For water depth quantification, the model also shows improvements in MAE, MSE, and RMSE. Compared to the Few-shot model, the performance of the trained model shows a noticeable improvement in water depth extraction, especially after training with SFT + SDPO. This improvement can be attributed to the conciseness of the water depth expressions, allowing SDPO to focus on learning from these short fragments, resulting in significant gains.

As shown in Fig. 8, the examples of water depth extraction from text illustrate that SDPO-MLLM demonstrates a deep understanding of water depth-related information. The first example is related to the heavy rainfall in Wuhan, where three descriptions of water depth appear in the text: those of 1991, 1998, and the current situation. The model correctly extracted the current water depth. The second example describes a heavy rain event in Shaoguan. The results indicate that the model not only extracts waterlogging locations from complex textual descriptions of places but also organizes these place names systematically, facilitating a clearer understanding of the flood situation and more accurate location localization. The third example is a road condition warning issued by the Kunming traffic police. The results show that both the locations of waterlogging and the corresponding water depths are correctly identified. However, certain deficiencies remain in the location extraction. For example, the phrase ‘Guomao Road and Jinzhi Road’ in the original text is ambiguous, as it could refer to either two separate locations or a single combined area. The model interprets it as the latter, and such ambiguity can lead to inaccuracies in the extraction. Furthermore, in Section II, some extracted locations related to Guandu District lack the prefix ‘Guandu District’, such as ‘Chuncheng Road and Yongping Road Intersection, Kunming’. This problem may be due to the considerable textual distance between these locations and the mention of ‘Guandu District’, which weakens their contextual association.

Table 5
F1-score for image water depth description result.

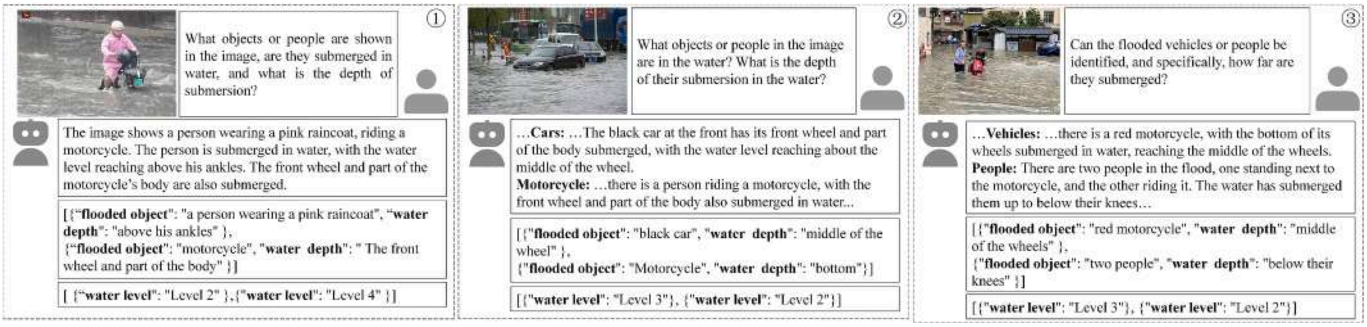
Method	Rouge-1	Rouge-2	Rouge-L
SDPO-MLLM (Few shot)	76.45 %	66.59 %	72.13 %
SDPO-MLLM (SFT)	87.29 %	80.74 %	83.96 %
SDPO-MLLM (SFT + DPO)	86.72 %	80.15 %	83.37 %
SDPO-MLLM (SFT + SDPO)	**87.51 %**	**80.94 %**	**83.99 %**

Table 6
Image-based water depth quantification result.

Method	MSE	MAE	RMSE
ResNet	0.0524	0.1338	0.2288
VIT	0.0297	0.0925	0.1723
SDPO-MLLM (Few shot)	0.0293	0.0553	0.1713
SDPO-MLLM (SFT)	0.0104	0.0283	0.1020
SDPO-MLLM (SFT + DPO)	0.0110	0.0288	0.1047
SDPO-MLLM (SFT + SDPO)	**0.0076**	**0.0249**	**0.0870**

Table 5 shows the evaluation results for the image water depth description task using SDPO-MLLM with different configurations. The model trained with SFT + SDPO performs best on three metrics. In addition, the models trained with SFT, SFT + DPO, and SFT + SDPO all outperform the Few-Shot model; however, the performance differences among these three training methods are minimal. This is because the training data is derived from minor modifications of erroneous segments in the MLLM output, resulting in high similarity between the generated and reference texts and minimal differences in accuracy.

To further evaluate the quantified water depth of the image, the generated descriptions are processed through a workflow for extraction, classification, and quantification, and then compared with ResNet (He et al., 2016) and ViT. In the extraction process, considering the descriptions do not include location information, the location argument is replaced with flooded object to better align with the content of the



What objects or people are shown in the image, are they submerged in water, and what is the depth of submersion?

The image shows a person wearing a pink raincoat, riding a motorcycle. The person is submerged in water, with the water level reaching above his ankles. The front wheel and part of the motorcycle's body are also submerged.

[{"**flooded object**": "a person wearing a pink raincoat", "**water depth**": "above his ankles" },
{"**flooded object**": "motorcycle", "**water depth**": " The front wheel and part of the body" }]

[{"**water level**": "Level 2" },{"**water level**": "Level 4" }]

What objects or people in the image are in the water? What is the depth of their submersion in the water?

…**Cars:** …The black car at the front has its front wheel and part of the body submerged, with the water level reaching about the middle of the wheel.
Motorcycle: …there is a person riding a motorcycle, with the front wheel and part of the body also submerged in water...

[{"**flooded object**": "black car", "**water depth**": "middle of the wheel" },
{"**flooded object**": "Motorcycle", "**water depth**": "bottom"}]

[{"**water level**": "Level 3"}, {"**water level**": "Level 2"}]

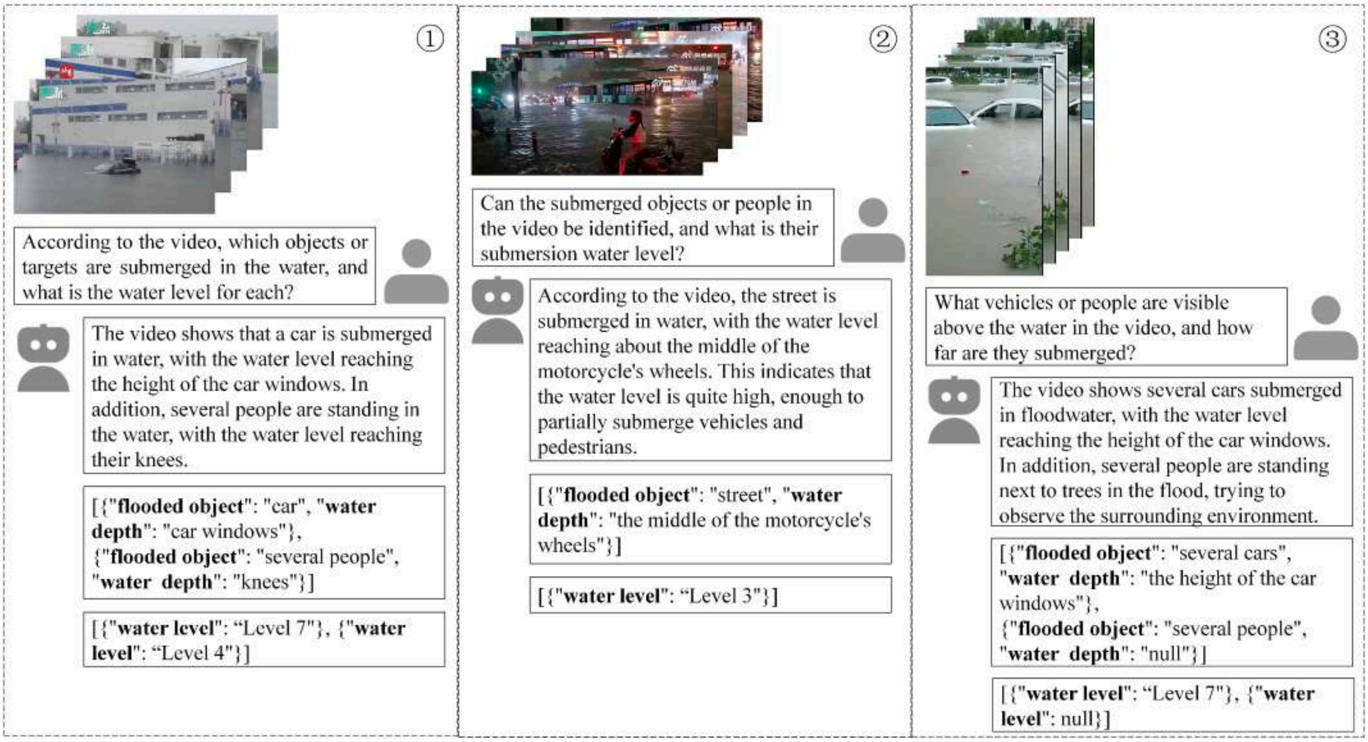
Can the flooded vehicles or people be identified, and specifically, how far are they submerged?

…**Vehicles:** …there is a red motorcycle, with the bottom of its wheels submerged in water, reaching the middle of the wheels.
People: There are two people in the flood, one standing next to the motorcycle, and the other riding it. The water has submerged them up to below their knees…

[{"**flooded object**": "red motorcycle", "**water depth**": "middle of the wheels" },
{"**flooded object**": "two people", "**water depth**": "below their knees" }]

[{"**water level**": "Level 3"}, {"**water level**": "Level 2"}]

Fig. 9. Image water depth extraction examples.



According to the video, which objects or targets are submerged in the water, and what is the water level for each?

The video shows that a car is submerged in water, with the water level reaching the height of the car windows. In addition, several people are standing in the water, with the water level reaching their knees.

[{"**flooded object**": "car", "**water depth**": "car windows"},
{"**flooded object**": "several people", "**water depth**": "knees"}]

[{"**water level**": “Level 7"}, {"**water level**": “Level 4"}]

Can the submerged objects or people in the video be identified, and what is their submersion water level?

According to the video, the street is submerged in water, with the water level reaching about the middle of the motorcycle's wheels. This indicates that the water level is quite high, enough to partially submerge vehicles and pedestrians.

[{"**flooded object**": "street", "**water depth**": "the middle of the motorcycle's wheels"}]

[{"**water level**": “Level 3"}]

What vehicles or people are visible above the water in the video, and how far are they submerged?

The video shows several cars submerged in floodwater, with the water level reaching the height of the car windows. In addition, several people are standing next to trees in the flood, trying to observe the surrounding environment.

[{"**flooded object**": "several cars", "**water depth**": "the height of the car windows"},
{"**flooded object**": "several people", "**water depth**": "null"}]

[{"**water level**": “Level 7"}, {"**water level**": null}]

Fig. 10. Video water depth extraction examples.

descriptions. ResNet is a convolutional neural network (CNN) that incorporates residual connections. During training, the ResNet-101 architecture is fine-tuned for image water level classification. ViT divides images into multiple tokens using a grid, serializes them, and leverages the Transformer to capture global dependencies within the image. The ViT model is trained using a 12-layer Transformer network for classification. Table 6 shows the quantification results for the image water depth dataset, where SDPO-MLLM with SFT + SDPO outperforms other methods in MSE, MAE and RMSE, indicating that the proposed approach improves the ability of MLLM to detect image water depth.

In addition, Fig. 9 presents examples of water depth extraction from images using the SDPO-MLLM with SFT + SDPO. In the first example, the image depicts a person wearing a pink raincoat riding a motorcycle through the water. The model successfully identifies both the person and the motorcycle, and infers the water level. In the second example, the model detects and estimates the flood depth of the car in the front and the motorcycle in the behind. The third example shows a police officer helping a resident cross a flooded area, with the model estimating the water level at the person's location. These examples also illustrate certain limitations of the model. In example 1, the model does not account for the person sitting on the motorcycle, resulting in an overestimation of the water level compared to when the person is standing. Additionally, in example 3, the model incorrectly describes the resident as sitting on a motorcycle, whereas the person is actually standing beside it. This misinterpretation may be due to visual occlusion or ambiguity. Despite these errors, the model generally provides accurate descriptions of water depth in the images.

Table 7
Water level classification result.

Method	F1-score	MSE	MAE	RMSE
BERT	78.18 %	0.0826	0.0943	0.2875
SDPO-MLLM (Few shot)	45.65 %	0.2022	0.2052	0.4496
SDPO-MLLM (SFT)	85.67 %	0.0746	0.0629	0.2731
SDPO-MLLM (SFT + DPO)	88.09 %	0.0592	0.0532	0.2434
SDPO-MLLM (SFT + SDPO)	**90.17 %**	**0.0176**	**0.0330**	**0.1327**

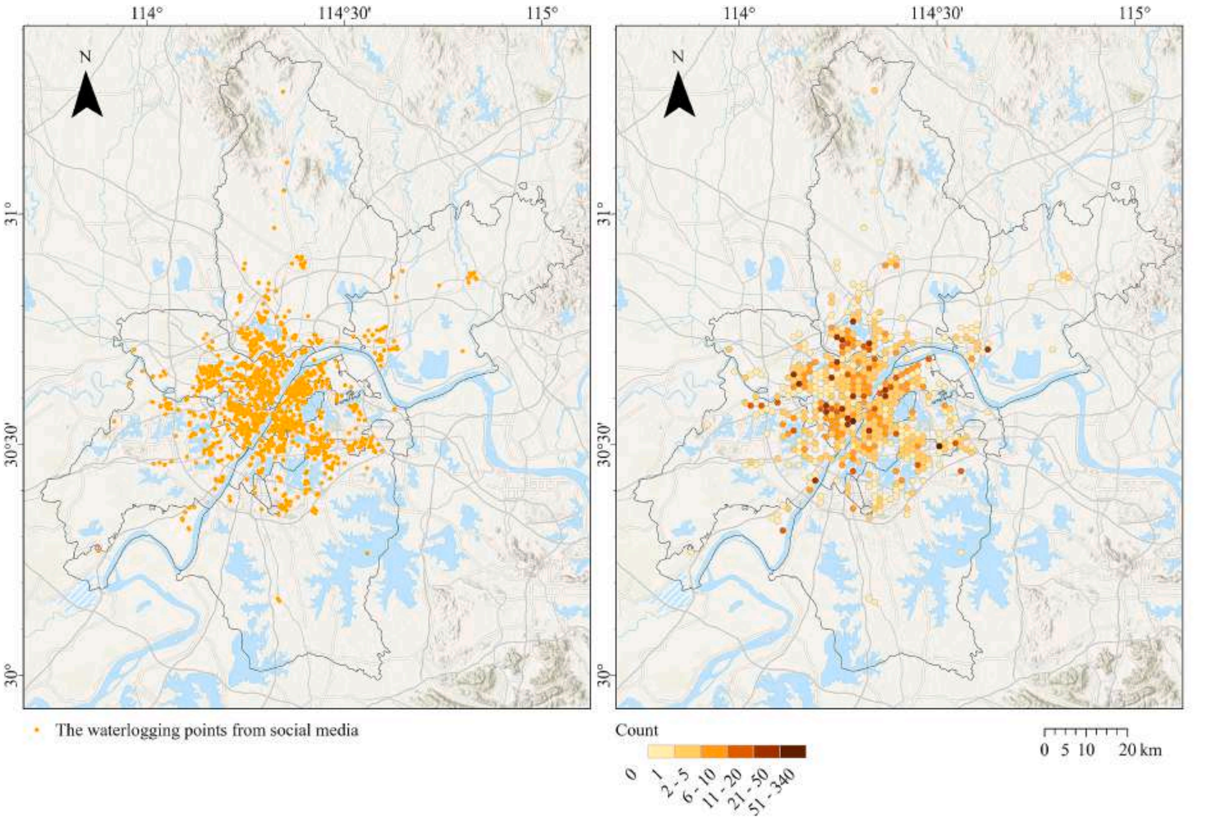
Furthermore, Fig. 10 presents examples of water depth information extraction from videos. Since the base model Qwen-2-VL supports video understanding, the keyframe extraction process prior to video input is unnecessary. However, the model basically treats the video as a sequence of frames, similar to processing multiple images. To avoid exceeding GPU memory limits, the maximum pixel value for the Qwen-

Table 8
Text evaluation results in the three regions.

Method	Wuhan Location Extraction F1-score	Wuhan Water Extraction F1-score	Shenzhen Location Extraction F1-score	Shenzhen Water Extraction F1-score	Beijing Location Extraction F1-score	Beijing Water Extraction F1-score
BERT-BiLSTM-CRF	45.74 %	57.47 %	40.29 %	56.57 %	27.54 %	53.19 %
UIE	47.62 %	59.04 %	56.34 %	77.97 %	34.30 %	61.05 %
SDPO-MLLM (Few shot)	83.37 %	91.92 %	57.98 %	66.17 %	68.50 %	68.25 %
SDPO-MLLM (SFT)	91.28 %	86.96 %	86.51 %	79.43 %	79.65 %	80.00 %
SDPO-MLLM (SFT + DPO)	88.48 %	93.04 %	78.89 %	84.93 %	75.76 %	77.17 %
SDPO-MLLM (SFT + SDPO)	**94.58 %**	**96.58 %**	**92.57 %**	**87.84 %**	**92.18 %**	**81.16 %**

Table 9
Image water depth quantification results in the three regions.

Method	Wuhan MAE	Wuhan RMSE	Shenzhen MAE	Shenzhen RMSE	Beijing MAE	Beijing RMSE
ResNet	0.0086	0.0926	0.3029	0.4459	0.1017	0.1810
VIT	0.0247	0.1573	0.2427	0.3417	0.1645	0.2426
SDPO-MLLM (Few shot)	0.0781	0.2794	0.1116	0.3677	0.0908	0.3427
SDPO-MLLM (SFT)	0.0028	0.0528	0.0429	0.1441	0.0226	0.0941
SDPO-MLLM (SFT + DPO)	0.0090	0.0949	0.0345	0.1174	0.0188	0.0725
SDPO-MLLM (SFT + SDPO)	**0.0014**	**0.0375**	**0.0246**	**0.0923**	**0.0120**	**0.0546**



(a) The distribution of waterlogging points
(b) The distribution of the number of waterlogging points

Fig. 11. The waterlogging spatial distribution in Wuhan in 2022.

2-VL hyperparameters is set to 50,176 pixels (224 × 224) per frame, with a total of 40 frames. The hyperparameter setting limits the model's ability to process longer duration content. However, the responses indicate that SDPO-MLLM is still able to detect the water depth of objects in the video. For instance, in the first example, it correctly recognizes that the water level reaches the character's knees. However, the description of the vehicle's water level should state that it reaches the door handles rather than the windows. While the answer is not entirely accurate, it is fairly close, likely due to the down sampling of the video, which caused the targets to blur and resulted in recognition errors.

Finally, Table 7 show the water level classification result, where the SDPO-MLLM based on SFT + SDPO achieves the best performance,

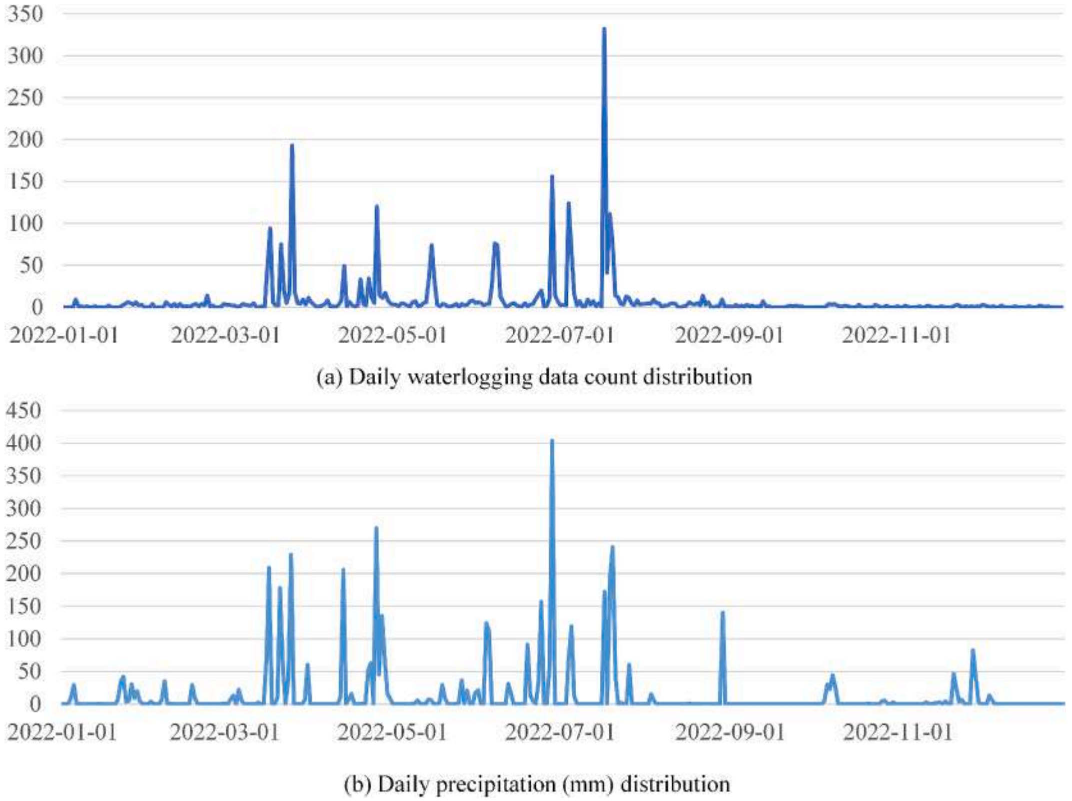


Fig. 12. Daily precipitation and waterlogging data distribution in Wuhan.

improving the F1-score from 2.08 % to 44.52 %. In addition, BERT, as an encoder-based model, excels in text embedding and classification, outperforming the Few-Shot MLLM. However, the fine-tuned MLLM still demonstrates superior performance. Furthermore, the SDPO-MLLM, based on SFT + SDPO, outperforms other training strategies, further improving performance.

4.3. Water depth mapping and analysis result

The adaptability of the model is evaluated using test sets from three regions: Wuhan, Shenzhen, and Beijing. The evaluation results of text extraction are presented in Table 8, while the image-based water depth quantification results are reported in Table 9. As shown in Table 8, SDPO-MLLM achieved the best overall performance in both location and water depth extraction tasks across the three regions. Compared with other models, the improvements are substantial, demonstrating the effectiveness of the SDPO strategy in cross-regional generalization. Table 9 presents the proposed SDPO-MLLM (SFT + SDPO) consistently achieved the lowest errors in Wuhan, Shenzhen and Beijing, further confirming the robustness of the model across diverse flood characteristics.

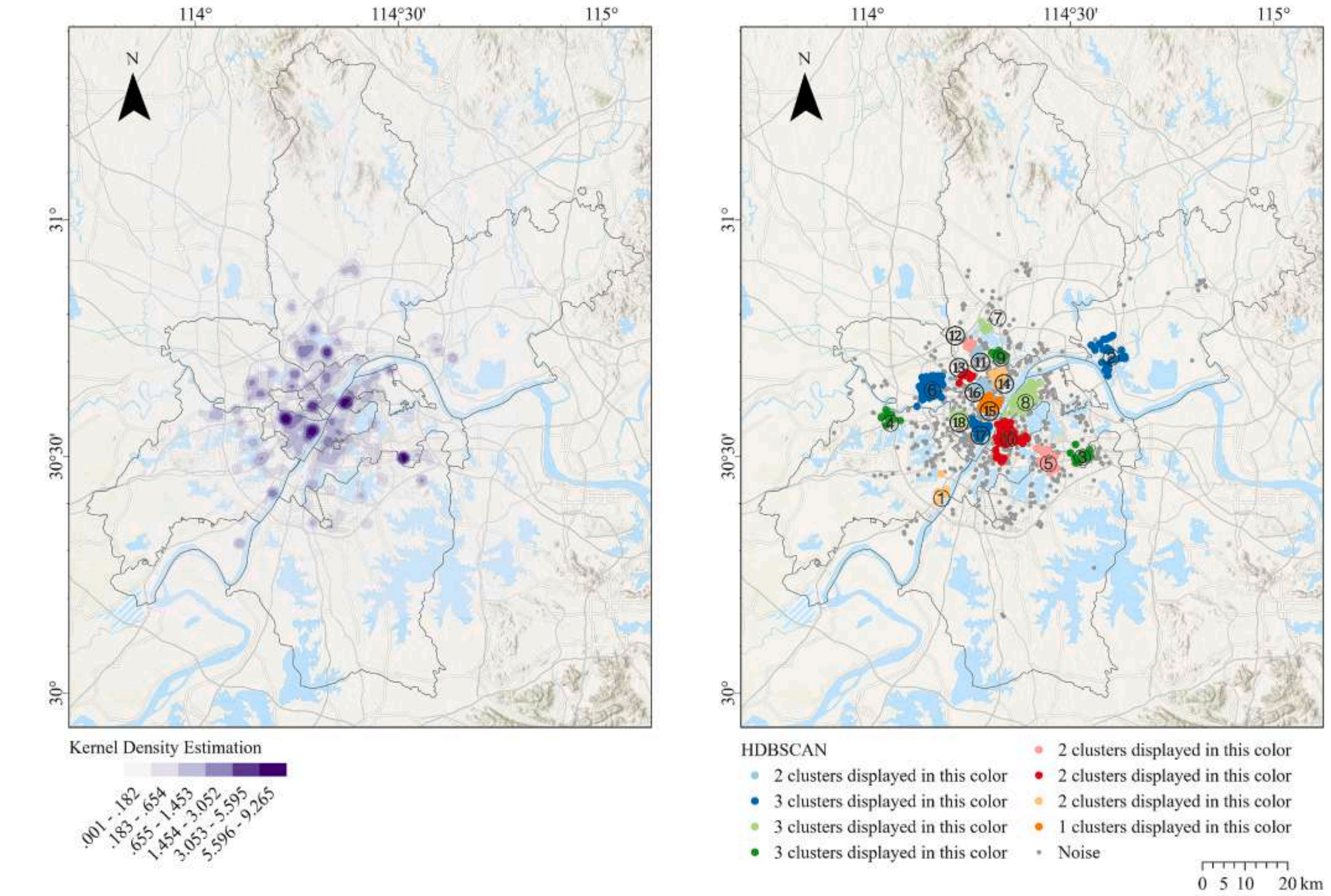
In addition, urban flooding in these regions is analyzed and mapped. As shown in Fig. 11, the waterlogging points in Wuhan in 2022 are mainly concentrated in the central urban area and extended along both banks of the Yangtze River. Specifically, Fig. 11(a) presents the spatial distribution of 2809 waterlogging points with depth information, while Fig. 11(b) shows their counts aggregated within a 1.5 km hexagonal grid. This area corresponds to the city's densely populated core, where certain locations exhibit a higher density of waterlogging reports, suggesting more severe flooding conditions.

Fig. 12 shows the daily waterlogging record reports in Wuhan in 2022, along with the corresponding precipitation. Normality tests (Shapiro-Wilk) indicated that neither variable follows a normal distribution. Therefore, a Spearman correlation analysis was conducted, resulting in a correlation coefficient of 0.37 ($p < 0.001$), suggesting a moderate positive relationship between the two variables over time. Most waterlogging records are concentrated between late March and May, and again in July, peaking around July 17.

The KDE method is then performed to assess the detailed waterlogging situation. The water depth of each record is used as the kernel weight, with a search radius of 2000 m. The KDE results for waterlogging in 2022, shown in Fig. 13(a), indicate the presence of several density centers with relatively severe waterlogging conditions. To assess the conditions of these density centers, the HDBSCAN algorithm is used to cluster the waterlogging points, with the minimum number of points per cluster set to 30. Fig. 13(b) presents the clustering results from HDBSCAN, suggesting that some of the clustered regions spatially correspond to areas with higher density values in KDE, such as clusters 3, 8, 15, 17, and 18.

In addition, Table 10 shows the average KDE values of the points within each cluster and their corresponding locations in major street blocks, which presents the severity of waterlogging in each cluster, by ranking the KDE values. Among them, cluster 3 has the highest KDE value, which is mainly concentrated in Jiufeng. Close behind are cluster 18, which is distributed across Yongfeng, and cluster 17, which is spread over Wuli Dun, Cuiwei Jie, and Jiangdi Jie. Meanwhile, three clusters (9, 11, and 12) are located in the same region, Panlong Cheng. Although these clusters are not highly ranked, they still highlight the widespread and significant waterlogging problem in the area. The results show that the application of KDE and HDBSCAN to analyze waterlogging points extracted from SDPO-MLLM effectively identifies areas prone to waterlogging, thereby also validating the performance of the model.

Furthermore, the case study shows that, although both information extraction and situational analysis are important, accurate information extraction is more critical. Inaccuracies in the extraction of key elements can distort downstream spatial analyses and lead to incorrect interpretations. In contrast, situational analysis typically involves more subjective reasoning and judgment, allowing for interpretive flexibility,


114°
114°30'
115°
31°
30°30'
30°
N
Kernel Density Estimation
.001 - .182
.183 - .654
.655 - 1.453
1.454 - 3.052
3.053 - 5.595
5.596 - 9.265
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
HDBSCAN
2 clusters displayed in this color
3 clusters displayed in this color
3 clusters displayed in this color
3 clusters displayed in this color
2 clusters displayed in this color
2 clusters displayed in this color
2 clusters displayed in this color
1 clusters displayed in this color
Noise
0 5 10 20 km

(a) The kernel density estimation map of waterlogging points (b) The spatial clustering map of waterlogging points

Fig. 13. The spatial clustering of waterlogging points in Wuhan.

Table 10
The KDE values of spatial clusters and their corresponding street blocks.

Cluster ID	KDE Value	Street block
3	7.8321	Jiufeng
18	6.4682	Yongfeng Jie
17	5.1545	Wuli Dun, Cuiwei Jie, Jiangdi Jie
8	4.6805	Yangyuan, Xujia Peng, Heping Jie, Hongwei Lu
9	3.5894	Panlong Cheng
15	3.2077	Wansong Jie
11	2.4049	Panlong Cheng
13	2.1906	Jinyin Hu
14	1.7385	Houhu, Tazi Hu
7	1.5574	Hengdian
2	1.4346	Junshan
10	1.3635	Zhongnan Lu, Luonan, Shouyi Lu
6	1.1956	Jinhe, Wujia Shan, Changqing Jie
12	1.1507	Panlong Cheng
4	0.9383	Caidian
1	0.9278	Yangluo
16	0.9098	Hanxing Jie, Changqing Huayuan
5	0.2551	Guandong, Guanshan

and minor inaccuracies have a limited impact on the overall outcome. Therefore, accurate and robust information extraction is essential for trustworthy situational analyses.

Finally, a rainfall event occurring between July 17 and July 20, 2022, along with the corresponding waterlogging KDE results, is analyzed. Spatial distribution maps of daily waterlogging severity are generated by applying KDE to the waterlogging points extracted from SDPO-MLLM. Daily precipitation data is sourced from the GPM (Global Precipitation Measurement) Level 3 Final data with a 10-kilometer resolution. Table 11 presents the results obtained using Geodetector's factor analysis to assess the effect of precipitation on waterlogging severity. The findings indicate that daily precipitation explained the driving factors of waterlogging severity to varying degrees. On July 20, 2022, the precipitation factor accounted for approximately 48.01 % of the spatial heterogeneity in water depth, reaching its maximum value. This suggests that precipitation significantly influences the spatial stratification heterogeneity of waterlogging severity.

Table 11
Results of spatial heterogeneity detection of waterlogging severity and precipitation.

Date	q-value	p-value
July 17, 2022	0.3688	<0.01
July 18, 2022	0.3972	<0.01
July 19, 2022	0.179	<0.01
July 20, 2022	0.4801	<0.01

Additionally, Fig. 14 presents the spatiotemporal distribution of precipitation and waterlogging severity between July 17 and July 20, 2022, which shows a significant spatiotemporal consistency between waterlogging severity and precipitation. From July 17 to July 20, the precipitation area shows a trend of moving from south to north. On the 17th, the precipitation is relatively low, but a significant number of waterlogging reports are still generated. This is due to the data processing method, which gives priority to retaining earlier records in case

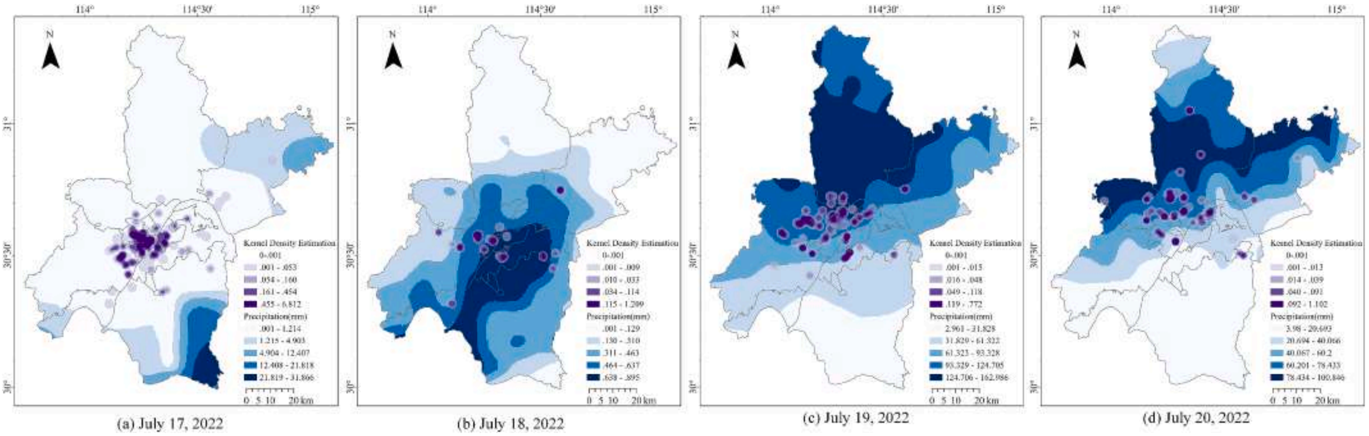


Fig. 14. The spatial distribution of precipitation and KDE analysis in Wuhan from July 17 to 20, 2022.

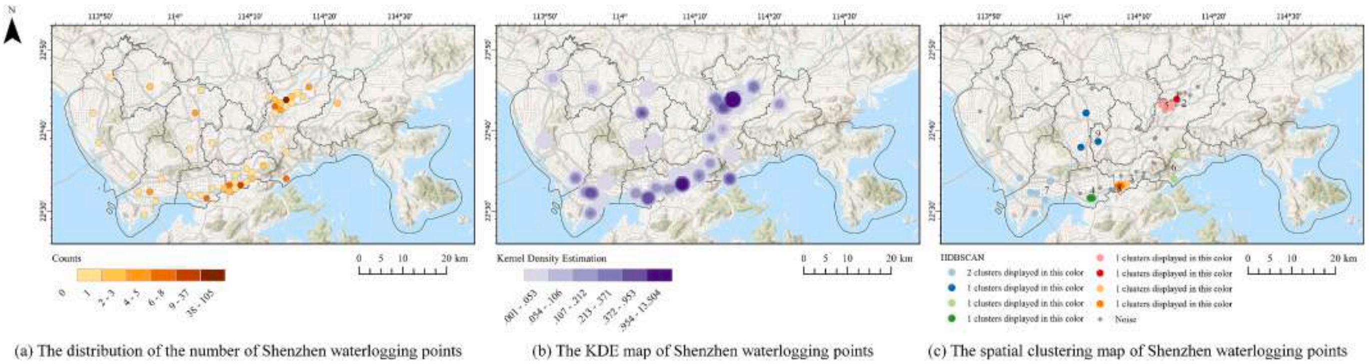

N
113°50'
114°
114°10'
114°20'
114°30'
22°50'
22°40'
22°30'
Counts
0
1
2-3
4-5
6-8
9-37
38-105
0 5 10 20 km
Kernel Density Estimation
.001 - .053
.054 - .106
.107 - .212
.213 - .371
.372 - .953
.954 - 13.504
HDBSCAN
2 clusters displayed in this color
1 clusters displayed in this color
1 clusters displayed in this color
1 clusters displayed in this color
1 clusters displayed in this color
1 clusters displayed in this color
1 clusters displayed in this color
1 clusters displayed in this color
Noise
(a) The distribution of the number of Shenzhen waterlogging points
(b) The KDE map of Shenzhen waterlogging points
(c) The spatial clustering map of Shenzhen waterlogging points

Fig. 15. The results of urban waterlogging analysis from waterlogging points in Shenzhen.

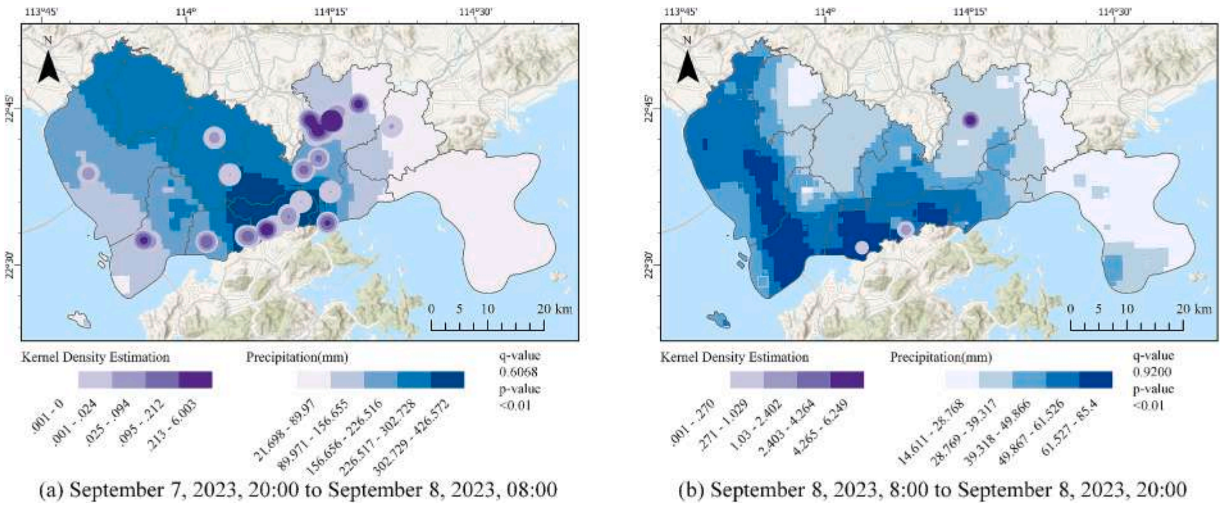
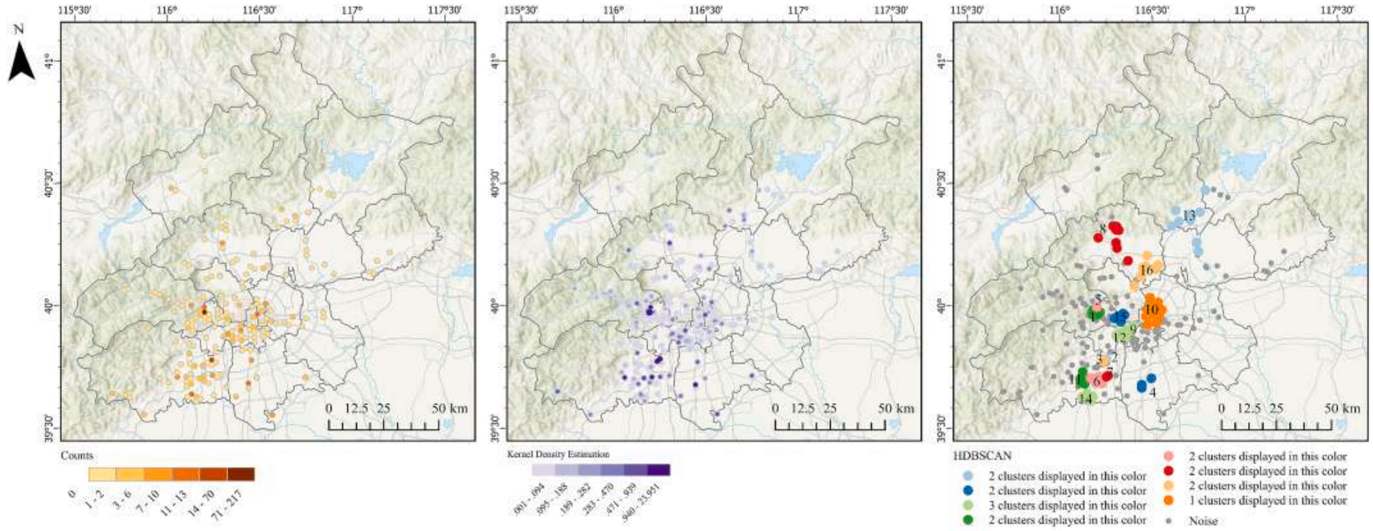


Fig. 16. The spatial distribution of precipitation and KDE analysis in Shenzhen, September 7–8, 2023.

of duplicates. As a result, a higher number of reports from the early stages of rainfall are retained, reducing the efficiency loss caused by processing duplicate reports. On the 18th, as the rain decreased, the number of waterlogging reports also decreased. On the 19th and 20th, as the rainfall moved northward, the spatial distribution of waterlogging also gradually shifted northward, consistent with the movement of the rainfall. Overall, waterlogging reports are closely related to the spatial pattern of daily rainfall.

Furthermore, Fig. 15 presents the results of urban waterlogging analysis in Shenzhen. Fig. 15(a) presents the distribution of waterlogging points aggregated within a 1.5 km hexagonal grid, with a total of 356 valid data points collected in the city. These points are mainly concentrated in the northern Longgang District and the southern Futian and Luohu Districts. Fig. 15(b) illustrates the KDE results based on extracted water depth information, which indicate that the severely affected areas extend inland from the southern coastline. Fig. 15(c) shows the spatial clustering results, where the numbers assigned to each cluster represent their ranking based on KDE values in descending order.


115°30'
116°
116°30'
117°
117°30'
41°
40°30'
40°
39°30'
N
0 12.5 25 50 km
Counts
0
1 - 2
3 - 6
7 - 10
11 - 13
14 - 70
71 - 217
Kernel Density Estimation
.001 - .094
.095 - .188
.189 - .282
.283 - .470
.471 - .939
.940 - 23.951
HDBSCAN
2 clusters displayed in this color
2 clusters displayed in this color
3 clusters displayed in this color
2 clusters displayed in this color
2 clusters displayed in this color
2 clusters displayed in this color
2 clusters displayed in this color
1 clusters displayed in this color
Noise
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16

(a) The distribution of the number of Beijing waterlogging points

(b) The KDE map of Beijing waterlogging points

(c) The spatial clustering map of Beijing waterlogging points

Fig. 17. The results of urban waterlogging analysis from waterlogging points in Beijing.

The most severe clusters are located in Longcheng Subdistrict of Longgang District (1, 2, 5), Dongmeng Subdistrict of Luohu District (3, 8), and Fubao Subdistrict of Futian District (4).

Fig. 16 illustrates the spatial distribution of precipitation and flood severity from 20:00 on September 7 to 20:00 on September 8, 2023. The results indicate that precipitation was relatively high during the first 12 h, accompanied by widespread waterlogging reports. In the subsequent 12 h, precipitation decreased markedly, and waterlogging reports also declined, occurring only in a few rainfall centers, demonstrating a clear spatiotemporal consistency. Moreover, Fig. 16 presents the results of the Geodetector factor analysis. During the two periods, precipitation factors explained 60.68 % and 92.00 % of the spatial heterogeneity in water depth, respectively, indicating that precipitation played a dominant role in flood occurrence.

Fig. 17 presents the results of urban waterlogging analysis in Beijing. Fig. 17(a) shows the distribution of waterlogging points aggregated within a 3 km hexagonal grid, with a total of 710 valid data points collected. These points are primarily concentrated in the southern and central parts of the city, including Fangshan, Mentougou, and Fengtai Districts. Fig. 17(b) illustrates the KDE results based on extracted water depth information, which reveal multiple flooding centers, indicating that the flooding event had a wide spatial impact. Fig. 17(c) displays several clustered areas, where the numbers next to each cluster represent their ranking based on KDE values in descending order. The most severe clusters are located in Liangxiang Town of Fangshan District (1, 5), Longquan Town of Mentougou District (2, 3), and Dingfuzhuang Township of Daxing District (4).

Fig. 18 illustrates the spatial distribution of precipitation and KDE-based waterlogging reports from July 29 to August 3, 2023. Overall, precipitation was mainly concentrated in the central and southern parts of Beijing, showing a temporal pattern of first increasing and then decreasing, with the peak occurring on July 30. Waterlogging reports exhibited a similar trend. Fig. 18 also presents the results of the Geodetector factor analysis, which indicate that precipitation explained the spatial differentiation of waterlogging reports to varying degrees on different dates. These visualization results further demonstrate the model's ability to accurately capture the spatial distribution of flood risk across regions with varying flood characteristics.

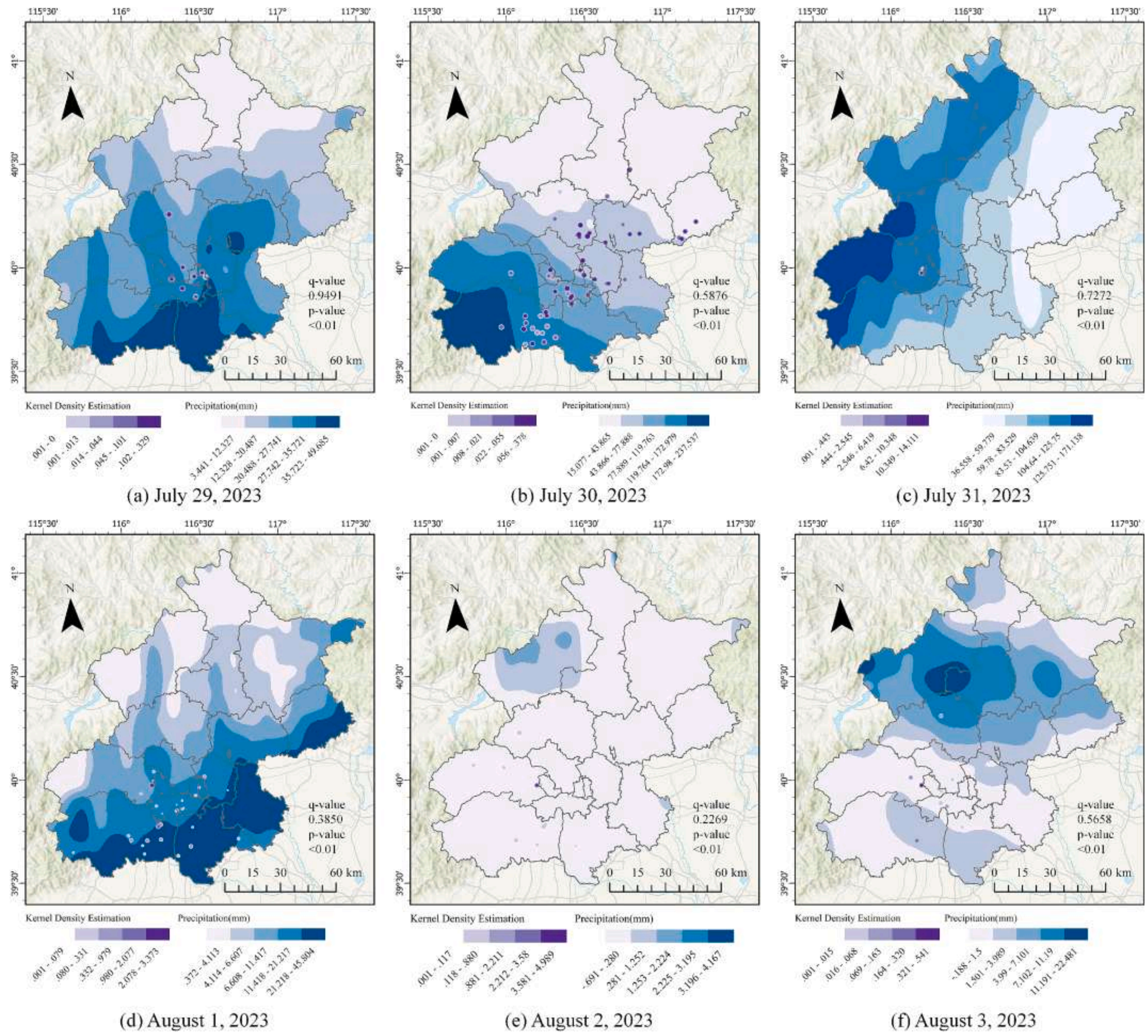
5. Discussion

5.1. Accuracy and efficiency after model quantization

Model quantization can reduce resource consumption and increase the speed of model inference, making it more applicable to large-scale social media data. In this section, the efficiency and performance of the quantization-applied SDPO-MLLM in extracting water depth information is further evaluated. Gradient-based Post-training Quantization (GPTQ) (Frantar et al., 2023) is used to quantize the model into INT3, INT4, and INT8 precision, which compresses the weights after training and adjusts the parameter errors through gradient optimization.

Fig. 19 compares the performance of different quantized models. Specifically, Fig. 19(a) demonstrates the inference time of each model on the test dataset, indicating that the time taken decreases linearly with the level of quantization. Fig. 19(b) illustrates the quantified performance of the model on three water depth information extraction tasks. It is evident that the location extraction task experiences minimal performance loss for the INT4 and INT8 models, while the loss is more significant for the INT3 model. In addition, the water depth extraction task shows little performance degradation across all quantized models, while the water level classification task shows a gradual performance degradation. In the image water depth description task, all models except the INT3 model exhibit relatively small Rouge-1, Rouge-2, and Rouge-L losses. Fig. 19(c) shows the results of the quantization of the water depth text extracted by the model. It becomes evident that the error gradually increases with the level of quantization, with the INT8 model exhibiting relatively smaller errors. Overall, the INT8 model improves inference speed while minimizing accuracy loss, thereby enhancing the efficiency of water depth information extraction.

In practical scenarios, the proposed model can be extended to enable real-time or near real-time detection of floodwater depth. During the data collection phase, scheduled tasks and asynchronous requests can be employed to achieve high concurrency and rapid acquisition of relevant information. In the data cleaning phase, vector databases can be utilized to store and retrieve text embeddings of streaming data, followed by NER filtering and BERT-based binary classification filtering. This process substantially reduces the number of samples to be processed, thereby allowing subsequent model inference to efficiently handle peak traffic. Furthermore, the SDPO-MLLM model can be deployed locally using techniques such as vLLM and DeepSeed to optimize GPU



(a) July 29, 2023

(b) July 30, 2023

(c) July 31, 2023

(d) August 1, 2023

(e) August 2, 2023

(f) August 3, 2023

Fig. 18. The spatial distribution of precipitation and KDE analysis in Beijing, July 29-August 3, 2023.

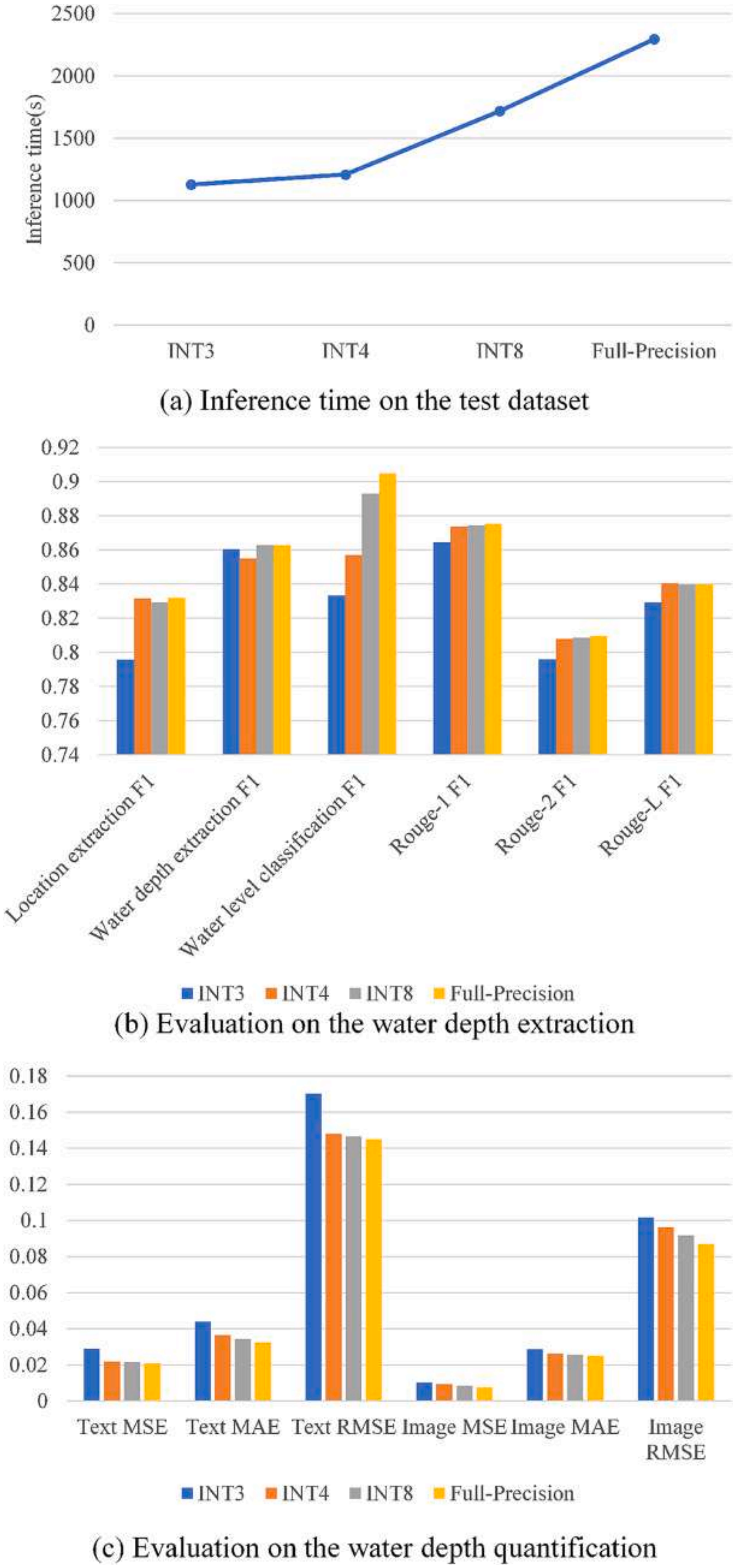
inference. Quantization techniques can also be adopted to accelerate inference speed. The multimodal workflow supports various combinations of data inputs, thereby enabling flexible and scalable real-time water depth detection. In summary, the proposed approach features a modular architecture that is well-suited for real-time applications, with strong scalability that facilitates adaptation to diverse scenarios.

5.2. Model generalization performance on different water depth estimation tasks

To further validate the model's performance, it is evaluated on a different water depth dataset. The dataset (Wan et al., 2024) consists of 2000 images of vehicles submerged in water, including sedans, SUVs, and trucks. The dataset annotations are provided in YOLO format, where each vehicle is assigned a water depth category and a bounding box for object detection. Water depth is classified into five levels, corresponding to levels 0, 3, 5, 6, and 8, based on the water level standards outlined in Section 3.4. During the evaluation, the detection results are quantified using the estimated water depth shown in Table 2, and the MSE, MAE, and RMSE are calculated to evaluate the average water depth for each image.

SDPO-MLLM is compared to YOLOv8 (Wan et al., 2024), DINO (Zhang et al., 2023), and Qwen2-VL-7b. YOLOv8 (You Only Look Once version 8) is an object detection method that balances accuracy and speed. DINO (Detection Transformer with Improved denoising anchor boxes) is an object detection model based on the Detection Transformer framework. It introduces improved denoising anchor boxes and enhances detection capabilities through multi-scale feature fusion and efficient training strategies. Qwen2-VL-7b is the base model parameters without fine-tuning using the proposed method, while SDPO-MLLM is the model parameters fine-tuned on the multimodal water depth dataset using SFT + DDPO. Qwen2-VL-7b and SDPO-MLLM are further fine-tuned using a vehicle dataset with SFT. The prompt includes instruct model output detection box of vehicles and classification criteria for each water level. The response format follows the structure:

"<|object_ref_start|>L3 vehicle<|object_ref_end|><|box_start|>(x1,



(a) Inference time on the test dataset

(b) Evaluation on the water depth extraction

(c) Evaluation on the water depth quantification

Fig. 19. Performance comparison on different quantized models.

Table 12
The evaluation results on the vehicle water depth dataset.

Method	MSE	MAE	RMSE
YOLOv8	0.0616	0.1913	0.2483
DINO	0.0475	0.1627	0.2178
Qwen2-VL-7b	0.0389	0.1062	0.1972
SDPO-MLLM (SFT + SDPO)	**0.0378**	**0.1051**	**0.1944**

y1),(x2,y2)<|box_end|>”,

where multiple detection results are concatenated sequentially.

As shown in Table 12, the results indicate that the fine-tuned SDPO-MLLM achieved the best performance because its ability to estimate water depth is improved during prior training. Fig. 20 shows the detection results of four different methods alongside the ground truth. The results show that all methods have high accuracy in vehicle detection, with only DINO missing a few vehicles. For water level estimation, MLLM-based methods, especially SDPO-MLLM, show superior accuracy. This improvement can be attributed to the language module's ability to better understand the water level classification criteria. In addition, the integration of image and text effectively improves the ability to discriminate between different water depths.

5.3. Limitations and future enhancements

This study demonstrates that the proposed SDPO-MLLM achieves strong performance in extracting water depth information from multimodal social media data. However, the results of the experiments and the subsequent analyses reveal several challenges and opportunities for future research.

Although the SFT + SDPO fine-tuning strategy effectively mitigates hallucination risks, potential risks remain, particularly in safety–critical contexts such as emergency response and urban flood management. Erroneous or fabricated information could mislead decision-makers, resulting in delayed or inappropriate actions. Several factors contribute to hallucinations. For instance, in extreme flood scenarios where reference objects are fully submerged or absent, the lack of reliable visual anchors can lead to uncertain or biased predictions. This study focuses on cases where reference features are present. To address situations with very limited references, future work may incorporate additional information sources (Wieland et al., 2025), such as hydrodynamic models, remote sensing imagery, and elevation data, to improve robustness.

The characteristics of input imagery also influence estimation accuracy. Shooting angle can affect the perceived size, clarity, and visibility of reference objects. Images captured from oblique or elevated viewpoints may distort spatial relationships. By contrast, ground-level

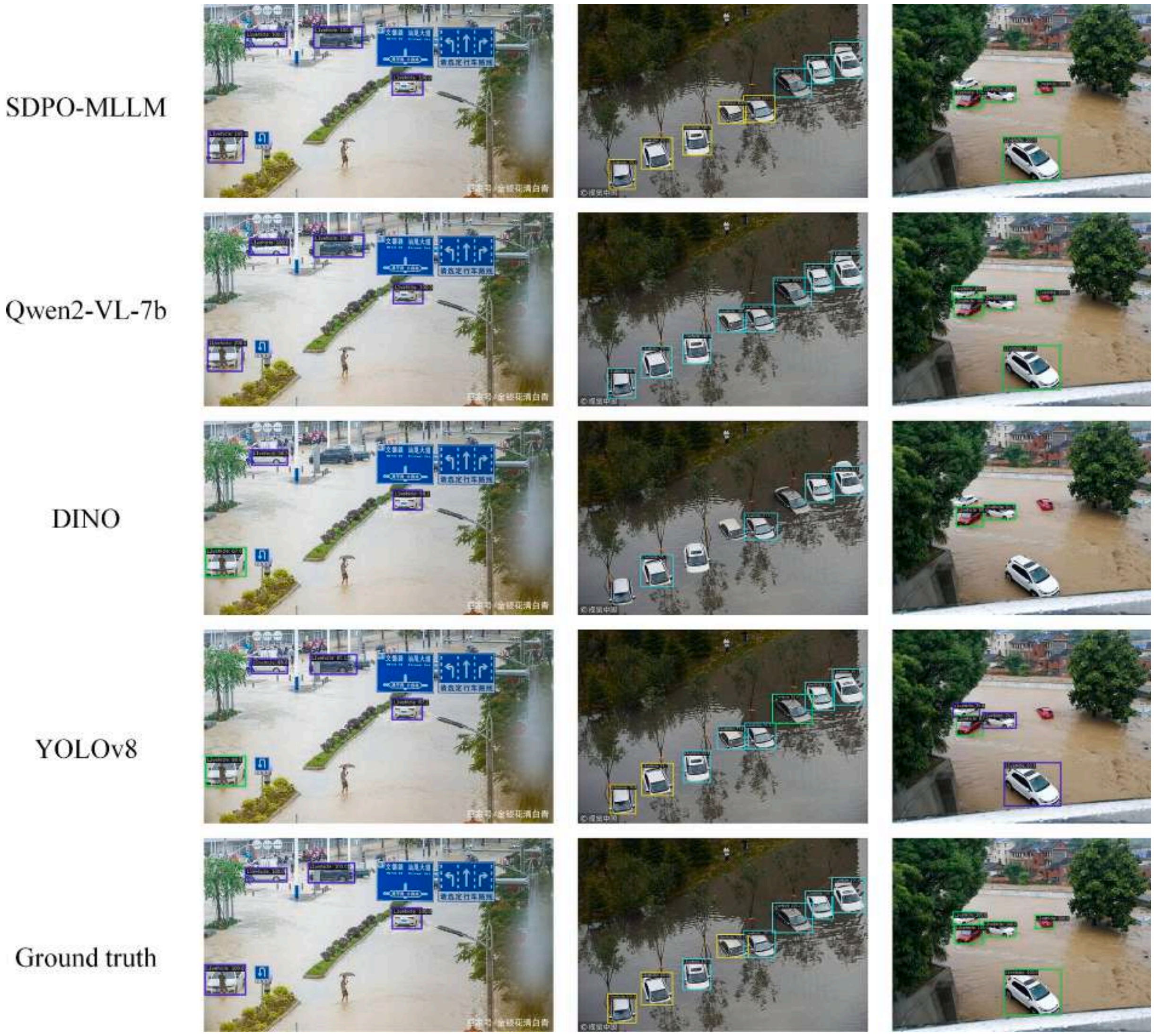


Fig. 20. The detection examples of the vehicle water depth dataset.

images in which the water surface and reference targets are clearly visible and unobstructed tend to yield more reliable results. In addition, incorporating images with fixed-height reference structures (e.g., lampposts) can reduce the effect of viewing angles, provided that such structures contain distinct and easily recognizable markers distributed along their height to enable accurate identification of submergence levels. Future studies may extend this approach by exploring other types of reference objects with similar properties.

There are several strategies that could reduce hallucinations and errors in complex or occluded environments. First, integrating reasoning-capable LLMs, such as DeepSeek-R1 (DeepSeek-AI, 2025), with vision encoders could improve reasoning and detection in challenging flood scenarios. Second, incorporating human pose estimation models enables more accurate interpretation of human references (Feng et al., 2020). Third, applying confidence thresholds to flag uncertain predictions coupled with human-in-the-loop (HITL) systems would allow automated pipelines to operate while maintaining expert oversight (Wilchek et al., 2023). Although SDPO-MLLM provides a fully automated framework that integrates multimodal inputs and minimizes human involvement in structuring water depth information from unstructured social media data, limited verification remains essential when dealing with low-quality or ambiguous imagery, or with misleading textual content such as sarcasm or irony. Furthermore, incorporating sensor data from IoT networks represents a promising direction for improving estimation accuracy and reliability (Kamel Boulos et al., 2011). Data from water level sensors, rain gauges, and traffic cameras not only provide precise timestamps and georeferenced readings to calibrate and validate predictions but also enable correction mechanisms to identify noise or anomalous reports in social media data, as well as to address spatial ambiguity or temporal lag.

Beyond hallucination risks, potential biases in data sources also merit attention. The dataset is primarily derived from Sina Weibo and citizen message boards, which may not accurately reflect the diversity of all populations. Linguistic styles and expression patterns influence the training process and extraction performance. User-generated content is inherently heterogeneous in reliability, tone, and spatial precision, which in turn may affect model outputs. The current model relies on pretrained MLLMs primarily trained on high-resource languages and mainstream platforms. In regions with sparse social media activity or high linguistic diversity, performance may be limited by insufficient training samples, restricting both coverage and comprehensiveness. Future research should explore domain adaptation, cross-lingual transfer learning, and the model's transferability to less structured platforms (e.g., TikTok videos) to improve robustness in multilingual and low-resource environments (Hong et al., 2025). Integrating complementary data sources, such as official reports or sensor networks, could also counterbalance social media biases and enhance monitoring coverage.

In terms of data privacy, this study only extracted task-relevant information from social media content, specifically water depth and location descriptions related to flood events. All data were collected from publicly accessible posts, and no private or restricted content was accessed. Personally identifiable information (e.g., names, contact details, user handles) was removed during preprocessing. Geographic references were limited to public descriptions (e.g., streets or neighborhoods) and were not linked to individual users. Water depth values represent environmental observations rather than personal information. Data usage complied with the platforms' terms of service, and all results were analyzed and presented at an aggregated level using statistics and mapping. Future work will explore stronger privacy-preserving techniques, such as differential privacy, to further enhance protection (Boulemtafes et al., 2020).

6. Conclusions

A novel model (SDPO-MLLM) is proposed to effectively extract multimodal water depth information by integrating image-text data, demonstrating advanced performance and robustness. To accommodate different social media data types and the generative responses of MLLMs, three water depth extraction subtasks are designed, and corresponding datasets are constructed: textual water depth extraction, image water depth description, and water depth level classification. These tasks are systematically integrated into a structured workflow to facilitate multimodal water depth extraction from text, image, and video data.

To optimize the model parameters, the fine-tuning strategy SFT + SDPO is introduced, which effectively reduces hallucinations in the model's responses to water depth information. In addition, LoRA is incorporated to reduce computational resource consumption and improve training efficiency. Evaluation results demonstrate that SDPO-MLLM outperforms single-modal water depth extraction methods. In text water depth extraction, the proposed method generates structured and organized results by understanding textual content. In image water depth description, the model analyzes image content and specifies the water depth of observed objects. Furthermore, multimodal water depth extracted from multi-source social media data is quantified and fused to map and assess waterlogging severity. In the case study, waterlogging prone areas in Wuhan, Shenzhen and Beijing are delineated using KDE and HDBSCAN, providing valuable insights for urban flood management. Factor analysis using Geodetector reveals that precipitation significantly explains waterlogging severity, further validating the model's capability.

Finally, the quantization and generalization performance are assessed. Results from post-training quantization using GPTQ (INT3, INT4, INT8) show that while lower-bit quantization significantly reduces inference time, it may lead to varying degrees of accuracy loss. Notably, the INT8 quantized model achieves a favorable trade-off: it maintains high accuracy across the three water depth tasks while substantially improving inference efficiency. These findings suggest that SDPO-MLLM can be effectively deployed in resource-constrained environments or real-time scenarios, balancing performance and efficiency. Additional experiments on other water depth datasets and in regions with different flood characteristics, such as Shenzhen and Beijing, confirm that after fine-tuning, the proposed model demonstrates strong generalization and applicability across diverse and complex scenarios.

CRediT authorship contribution statement

Tianyou Chu: Writing – original draft, Software, Methodology, Investigation, Conceptualization. **Yumin Chen:** Writing – review & editing, Supervision, Methodology, Investigation, Funding acquisition, Conceptualization. **Rui Zhu:** Writing – review & editing, Supervision, Investigation. **Fei Zeng:** Writing – review & editing, Validation, Data curation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 42471456, the National Key R&D Program of China (project No. 2024YFC3811000) and the Fundamental Research Funds for the Central Universities of China (Grant No. 2042022dx0001).

References

Aarthy, K.P., Ramesh, K., Prasanna Venkatesh, V.P., Chitrakala, S., Bhatt, S.C.M., 2022. Social Media Analysis for Flood Nuggets Extraction using Relevant Post Filtration. In: Chaki, N., Devarakonda, N., Cortesi, A., Seetha, H. (Eds.), Proceedings of

International Conference on Computational Intelligence and Data Engineering. Springer Nature, pp. 201–212. https://doi.org/10.1007/978-981-16-7182-1_17.

Akinboyewa, T., Ning, H., Lessani, M.N., Li, Z., 2024. Automated floodwater depth estimation using large multimodal model for rapid flood mapping. Comput. Urban Sci. 4, 12. https://doi.org/10.1007/s43762-024-00123-3.

Ali, A., Ghanem, M.C., 2025. Beyond Detection: Large Language Models and Next-Generation Cybersecurity. SHIFRA 2025, 81–97. 10.70470/SHIFRA/2025/005.

Alizadeh Kharazi, B., Behzadan, A.H., 2021. Flood depth mapping in street photos with image processing and deep neural networks. Comput. Environ. Urban Syst. 88, 101628. https://doi.org/10.1016/j.compenvurbsys.2021.101628.

Berragan, C., Singleton, A., Calafiore, A., Morley, J., 2023. Transformer based named entity recognition for place name extraction from unstructured text. Int. J. Geogr. Inf. Sci. 37, 747–766. https://doi.org/10.1080/13658816.2022.2133125.

Boulemtafes, A., Derhab, A., Challal, Y., 2020. A review of privacy-preserving techniques for deep learning. Neurocomputing 384, 21–45. https://doi.org/10.1016/j.neucom.2019.11.041.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y.T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M.T., Zhang, Y., 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. 10.48550/arXiv.2303.12712.

Campello, R.J.G.B., Moulavi, D., Sander, J., 2013. Density-based Clustering based on Hierarchical Density estimates. In: Pei, J., Tseng, V.S., Cao, L., Motoda, H., Xu, G. (Eds.), Advances in Knowledge Discovery and Data Mining. Springer, Berlin, Heidelberg, pp. 160–172. https://doi.org/10.1007/978-3-642-37456-2_14.

Chaudhary, P., D'Aronco, S., Leitão, J.P., Schindler, K., Wegner, J.D., 2020. Water level prediction from social media images with a multi-task ranking approach. ISPRS J. Photogramm. Remote Sens. 167, 252–262. https://doi.org/10.1016/j.isprsjprs.2020.07.003.

Chaudhary, P., D'Aronco, S., Moy de Vitry, M., Leitão, J.P., Wegner, J.D., 2019. FLOOD-WATER LEVEL ESTIMATION FROM SOCIAL MEDIA IMAGES. ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci. IV-2-W5, 5–12. https://doi.org/10.5194/isprs-annals-IV-2-W5-5-2019.

Chen, P., Xu, H., Zhang, C., Huang, R., 2022. Crossroads, Buildings and Neighborhoods: A Dataset for Fine-grained Location Recognition, in: Carpuat, M., de Marneffe, M.-C., Meza Ruiz, I.V. (Eds.), Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Presented at the NAACL-HLT 2022, Association for Computational Linguistics, Seattle, United States, pp. 3329–3339. 10.18653/v1/2022.naacl-main.243.

Chu, T., Chen, Y., Wilson, J.P., Liu, L., 2025. A large language model-enhanced argument extraction and clustering model for urban hotspot event detection using crowdsourced data. Expert Syst. Appl. 293, 128760. https://doi.org/10.1016/j.eswa.2025.128760.

DeepSeek-AI, 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. 10.48550/arXiv.2501.12948.

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Burstein, J., Doran, C., Solorio, T. (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (long and Short Papers). Presented at the NAACL-HLT 2019. Association for Computational Linguistics, Minneapolis, Minnesota, pp. 4171–4186. https://doi.org/10.18653/v1/N19-1423.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, in: International Conference on Learning Representations.

Duan, Z., Cheng, H., Xu, D., Wu, X., Zhang, X., Ye, X., Xie, Z., 2024. CityLLaVA: Efficient Fine-Tuning for VLMs in City Scenario. Presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7180–7189.

Ethayarajh, K., Xu, W., Muennighoff, N., Jurafsky, D., Kiela, D., 2024. Model Alignment as Prospect Theoretic Optimization, in: Proceedings of the 41st International Conference on Machine Learning. Presented at the International Conference on Machine Learning, PMLR, pp. 12634–12651.

Feng, Y., Brenner, C., Sester, M., 2020. Flood severity mapping from Volunteered Geographic Information by interpreting water level from images containing people: a case study of Hurricane Harvey. ISPRS J. Photogramm. Remote Sens. 169, 301–319. https://doi.org/10.1016/j.isprsjprs.2020.09.011.

Feng, Y., Huang, X., Sester, M., 2022. Extraction and analysis of natural disaster-related VGI from social media: review, opportunities and challenges. Int. J. Geogr. Inf. Sci. 36, 1275–1316. https://doi.org/10.1080/13658816.2022.2048835.

Frantar, E., Ashkboos, S., Hoefler, T., Alistarh, D., 2023. OPTQ: Accurate Quantization for Generative Pre-trained Transformers, in: The Eleventh International Conference on Learning Representations.

Geoapify, 2021. Geocoding Services Comparison: Which One is the Best?, https://www.geoapify.com/top-geocoding-services-comparison/. (Accessed 28 July, 2025).

Han, R., Yang, C., Peng, T., Tiwari, P., Wan, X., Liu, L., Wang, B., 2024. An Empirical Study on Information Extraction using Large Language Models. 10.48550/arXiv.2305.14450.

Hao, X., Lyu, H., Wang, Z., Fu, S., Zhang, C., 2022. Estimating the spatial-temporal distribution of urban street ponding levels from surveillance videos based on computer vision. Water Resour. Manag. 36, 1799–1812. https://doi.org/10.1007/s11269-022-03107-2.

He, H., Choi, J.D., 2021. The Stem Cell Hypothesis: Dilemma behind Multi-Task Learning with Transformer Encoders. In: Moens, M.-F., Huang, X., Specia, L., Yih, S.W. (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Presented at the EMNLP 2021. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, pp. 5555–5577. https://doi.org/10.18653/v1/2021.emnlp-main.451.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep Residual Learning for Image Recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778. 10.1109/CVPR.2016.90.

Hong, J., Lee, N., Thorne, J., 2024. ORPO: Monolithic Preference Optimization without Reference Model. In: Al-Onaizan, Y., Bansal, M., Chen, Y.-.-N. (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Presented at the EMNLP 2024. Association for Computational Linguistics, Miami, Florida, USA, pp. 11170–11189. https://doi.org/10.18653/v1/2024.emnlp-main.626.

Hong, S., Lee, S., Moon, H., Lim, H., 2025. Cross-Lingual Adaptation of Domain-specific LLMs through Code-Switching and Embedding transfer. In: Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Eugenio, B.D., Schockaert, S. (Eds.), Proceedings of the 31st International Conference on Computational Linguistics. Presented at the COLING 2025. Association for Computational Linguistics, Abu Dhabi, UAE, pp. 9184–9193.

Hou, H., Shen, L., Jia, J., Xu, Z., 2024. An integrated framework for flood disaster information extraction and analysis leveraging social media data: a case study of the Shouguang flood in China. Sci. Total Environ. 949, 174948. https://doi.org/10.1016/j.scitotenv.2024.174948.

Hu, E.J., shen, yelong, Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., 2022. LoRA: Low-Rank Adaptation of Large Language Models, in: International Conference on Learning Representations.

Hu, Y., Mai, G., Cundy, C., Choi, K., Lao, N., Liu, W., Lakhanpal, G., Zhou, R.Z., Joseph, K., 2023. Geo-knowledge-guided GPT models improve the extraction of location descriptions from disaster-related social media messages. Int. J. Geogr. Inf. Sci. 37, 2289–2318. https://doi.org/10.1080/13658816.2023.2266495.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., Liu, T., 2025. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, challenges, and Open questions. ACM Trans. Inf. Syst. 43, 42: 1–42:55. https://doi.org/10.1145/3703155.

Huang, X., Wang, C., Li, Z., 2018. A near real-time flood-mapping approach by integrating social media and post-event satellite imagery. Ann. GIS 24, 113–123. https://doi.org/10.1080/19475683.2018.1450787.

Jiang, J., Qin, C.-Z., Yu, J., Cheng, C., Liu, J., Huang, J., 2020. Obtaining Urban Waterlogging Depths from Video Images using Synthetic image Data. Remote Sens. (Basel) 12, 1014. https://doi.org/10.3390/rs12061014.

Kamel Boulos, M.N., Resch, B., Crowley, D.N., Breslin, J.G., Sohn, G., Burtner, R., Pike, W.A., Jezierski, E., Chuang, K.-Y.-S., 2011. Crowdsourcing, citizen sensing and sensor web technologies for public and environmental health surveillance and crisis management: trends, OGC standards and application examples. Int. J. Health Geogr. 10, 67. https://doi.org/10.1186/1476-072X-10-67.

Kaufmann, T., Weng, P., Bengs, V., Hüllermeier, E., 2024. A Survey of Reinforcement Learning from Human Feedback. 10.48550/arXiv.2312.14925.

Khan, Q., Kalbus, E., Zaki, N., Mohamed, M.M., 2022. Utilization of social media in floods assessment using data mining techniques. PLoS One 17, e0267079. https://doi.org/10.1371/journal.pone.0267079.

Lamsal, R., Harwood, A., Read, M.R., 2022. Where did you tweet from? Inferring the origin locations of tweets based on contextual information, in: 2022 IEEE International Conference on Big Data (Big Data). Presented at the 2022 IEEE International Conference on Big Data (Big Data), pp. 3935–3944. 10.1109/BigData55660.2022.10020460.

Li, J., Cai, R., Tan, Y., Zhou, H., Sadick, A.-M., Shou, W., Wang, X., 2023a. Automatic detection of actual water depth of urban floods from social media images. Measurement 216, 112891. https://doi.org/10.1016/j.measurement.2023.112891.

Li, Y., Osei, F.B., Hu, T., Stein, A., 2023b. Urban flood susceptibility mapping based on social media data in Chengdu city, China. Sustainable Cities Soc. 88, 104307. https://doi.org/10.1016/j.scs.2022.104307.

Liao, Y., Wang, Z., Chen, X., Lai, C., 2023. Fast simulation and prediction of urban pluvial floods using a deep convolutional neural network model. J. Hydrol. 624, 129945. https://doi.org/10.1016/j.jhydrol.2023.129945.

Liu, B., Li, Y., Feng, X., Lian, P., 2024. BEW-YOLOv8: a deep learning model for multi-scene and multi-scale flood depth estimation. J. Hydrol. 132139. https://doi.org/10.1016/j.jhydrol.2024.132139.

Lou, J., Lu, Y., Dai, D., Jia, W., Lin, H., Han, X., Sun, L., Wu, H., 2023. Universal information extraction as unified semantic matching, in: Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence, AAAI'23/IAAI'23/EAAI'23. AAAI Press, pp. 13318–13326. 10.1609/aaai.v37i11.26563.

Meng, Y., Xia, M., Chen, D., 2024. SimPO: Simple Preference Optimization with a Reference-Free Reward, in: Advances in Neural Information Processing Systems (NeurIPS).

Moftakhari, H.R., AghaKouchak, A., Sanders, B.F., Allaire, M., Matthew, R.A., 2018. What is Nuisance Flooding? defining and monitoring an Emerging Challenge. Water Resour. Res. 54, 4218–4227. https://doi.org/10.1029/2018WR022828.

Mustafa, D.F.Y., 2023. Flood-proof Architectural Solutions for Urban Environments: a Review. KHWARIZMIA 2023, 155–164, 10.70470/KHWARIZMIA/2023/016.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., Lowe, R., 2022a. Training language models to follow instructions with human feedback, in: Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22. Curran Associates Inc., Red Hook, NY, USA, pp. 27730–27744.

Ouyang, M., Kotsuki, S., Ito, Y., Tokunaga, T., 2022b. Employment of hydraulic model and social media data for flood hazard assessment in an urban city. J. Hydrol.: Reg. Stud. 44, 101261. https://doi.org/10.1016/j.ejrh.2022.101261.

Qin, J., Shen, P., 2025. Refraction-based waterlogging depth measurement using solely traffic cameras for transparent flood monitoring. J. Hydrol. 655, 132917. https://doi.org/10.1016/j.jhydrol.2025.132917.

Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C.D., Finn, C., 2023. Direct preference optimization: your language model is secretly a reward model, in: Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23. Curran Associates Inc., Red Hook, NY, USA, pp. 53728–53741.

Reimers, N., Gurevych, I., 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In: Inui, K., Jiang, J., Ng, V., Wan, X. (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Association for Computational Linguistics, Hong Kong, China, pp. 3982–3992, 10.18653/v1/D19-1410.

Saleh, T., Weng, X., Holail, S., Hao, C., Xia, G.-S., 2024. DAM-Net: Flood detection from SAR imagery using differential attention metric-based vision transformers. ISPRS J. Photogramm. Remote Sens. 212, 440–453. https://doi.org/10.1016/j.isprsjprs.2024.05.018.

Sathianarayanan, M., Hsu, P.-H., Chang, C.-C., 2024. Extracting disaster location identification from social media images using deep learning. Int. J. Disaster Risk Reduct. 104, 104352. https://doi.org/10.1016/j.ijdrr.2024.104352.

Stock, K., Jones, C.B., Tenbrink, T., 2022. Speaking of location: a review of spatial language research. Sp. Cogn. Comput. 22, 185–224. https://doi.org/10.1080/13875868.2022.2095275.

Wan, J., Qin, Y., Shen, Y., Yang, T., Yan, X., Zhang, S., Yang, G., Xue, F., Wang, Q.J., 2024. Automatic detection of urban flood level with YOLOv8 using flooded vehicle dataset. J. Hydrol. 639, 131625. https://doi.org/10.1016/j.jhydrol.2024.131625.

Wang, H., Zhou, J., Tang, Y., Liu, Z., Kang, A., Chen, B., 2021. Flood economic assessment of structural measure based on integrated flood risk management: a case study in Beijing. J. Environ. Manage. 280, 111701. https://doi.org/10.1016/j.jenvman.2020.111701.

Wang, J., Haining, R., Zhang, T., Xu, C., Hu, M., Yin, Q., Li, L., Zhou, C., Li, G., Chen, H., 2024a. Statistical Modeling of Spatially Stratified Heterogeneous Data. Ann. Am. Assoc. Geogr. 114, 499–519. https://doi.org/10.1080/24694452.2023.2289982.

Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R., Liu, D., Zhou, C., Zhou, J., Lin, J., 2024b. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. 10.48550/arXiv.2409.12191.

Wang, W., Zhu, X., Lu, P., Zhao, Y., Chen, Y., Zhang, S., 2024c. Spatio-temporal evolution of public opinion on urban flooding: Case study of the 7.20 Henan extreme flood event. Int. J. Disaster Risk Reduct. 100, 104175. https://doi.org/10.1016/j.ijdrr.2023.104175.

Wang, Z., Bi, B., Pentyala, S.K., Ramnath, K., Chaudhuri, S., Mehrotra, S., Zixu, Zhu, Mao, X.-B., Asur, S., Na, Cheng, 2024d. A Comprehensive Survey of LLM Alignment Techniques: RLHF, RLAIF, PPO, DPO and More. 10.48550/arXiv.2407.16216.

Wieland, M., Schmidt, S., Resch, B., Abecker, A., Martinis, S., 2025. Fusion of geospatial information from remote sensing and social media to prioritise rapid response actions in case of floods. Nat. Hazards 121, 8061–8088. https://doi.org/10.1007/s11069-025-07120-7.

Wilchek, M., Hanley, W., Lim, J., Luther, K., Batarseh, F.A., 2023. Human-in-the-loop for computer vision assurance: a survey. Eng. Appl. Artif. Intel. 123, 106376. https://doi.org/10.1016/j.engappai.2023.106376.

Wu, L., Liu, Y., Zhang, J., Zhang, B., Wang, Z., Tong, J., Li, M., Zhang, A., 2024a. Identification of flood depth levels in urban waterlogging disaster caused by rainstorm using a CBAM-improved ResNet50. Expert Syst. Appl. 255, 124382. https://doi.org/10.1016/j.eswa.2024.124382.

Wu, S., Fei, H., Qu, L., Ji, W., Chua, T.-S., 2024b. NExT-GPT: any-to-any multimodal LLM, in: Proceedings of the 41st International Conference on Machine Learning, ICML'24. JMLR.org, Vienna, Austria, pp. 53366–53397.

Yan, Z., Guo, X., Zhao, Z., Tang, L., 2023. Achieving fine-grained urban flood perception and spatio-temporal evolution analysis based on social media. Sustain. Cities Soc. 105077. https://doi.org/10.1016/j.scs.2023.105077.

Yang, A., Yang, B., Hui, B., Zheng, B., Yu, B., Zhou, C., Li, Chengpeng, Li, Chengyuan, Liu, D., Huang, F., Dong, G., Wei, H., Lin, H., Tang, J., Wang, J., Yang, Jian, Tu, J., Zhang, J., Ma, J., Yang, Jianxin, Xu, J., Zhou, J., Bai, J., He, J., Lin, J., Dang, K., Lu, K., Chen, K., Yang, K., Li, M., Xue, M., Ni, N., Zhang, P., Wang, P., Peng, R., Men, R., Gao, R., Lin, R., Wang, S., Bai, S., Tan, S., Zhu, T., Li, T., Liu, T., Ge, W., Deng, X., Zhou, X., Ren, Xingzhang, Zhang, X., Wei, X., Ren, Xuancheng, Liu, X., Fan, Y., Yao, Y., Zhang, Y., Wan, Y., Chu, Y., Liu, Y., Cui, Z., Zhang, Z., Guo, Z., Fan, Z., 2024. Qwen2 Technical Report. 10.48550/arXiv.2407.10671.

Yu, T., Yao, Y., Zhang, H., He, T., Han, Y., Cui, G., Hu, J., Liu, Z., Zheng, H.-T., Sun, M., 2024. RLHF-V: Towards Trustworthy MLLMs via Behavior Alignment from Fine-Grained Correctional Human Feedback, in: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Presented at the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 13807–13816. 10.1109/CVPR52733.2024.01310.

Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L., Shum, H.-Y., 2023. DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection, in: The Eleventh International Conference on Learning Representations.

Zhang, Y., Wei, C., He, Z., Yu, W., 2024. GeoGPT: an assistant for understanding and processing geospatial tasks. Int. J. Appl. Earth Obs. Geoinf. 131, 103976. https://doi.org/10.1016/j.jag.2024.103976.

Zhu, J., Dang, P., Cao, Y., Lai, J., Guo, Y., Wang, P., Li, W., 2024a. A flood knowledge-constrained large language model interactable with GIS: enhancing public risk perception of floods. Int. J. Geogr. Inf. Sci. 38, 603–625. https://doi.org/10.1080/13658816.2024.2306167.

Zhu, X., Guo, H., Huang, J.J., 2024b. Urban flood susceptibility mapping using remote sensing, social sensing and an ensemble machine learning model. Sustain. Cities Soc. 105508. https://doi.org/10.1016/j.scs.2024.105508.