Generative AI-Based Application for Producing Tourism Video Blogs with Proximity and Direction to Points of Interest

Hayate Eguchi [1], Iori Sasaki [2], Min Lu [2], Tomihiro Utsumi [2], Ryo Sato [2], Masatoshi Arikawa[2]

[1] Graduate School of Engineering Science, Akita University, Akita, Japan - m8025502@s.akita-u.ac.jp
[2] Faculty of Informatics and Data Science, Akita University, Akita, Japan – sasaki_iori@gipc.akita-u.ac.jp

Keywords: AI-Enabled Urban Tourism, Context-Aware Captioning, Tourism Video Blogs, Geofencing, Generative AI.

Abstract

Taking and sharing videos of tourist attractions has become a common activity among tourists. When accompanied by captions and audio, these videos serve as an effective way of conveying impressions and information about the places visited through social media. Such content not only enriches the post-travel experience of individuals but also contributes to promoting local tourism and stimulating inbound demand. However, producing highly informative video content poses challenges in that it requires editing skills and reliability of information. This study establishes a method for automatically overlaying captions onto videos by (1) estimating appropriate time periods during which points of interest (POIs) are captured within the camera's field of view, and (2) generating explanatory comments with a suitable word count for the corresponding durations. This method was implemented in the author's video blog application to enable users to easily share the appeal of a region. In a field experiment simulating tourist movement and POI filming within a predefined guide area, the average error between the time a POI appeared in the video and the calculated caption display duration was approximately 1.8 seconds, with a maximum error of 4.0 seconds. This level of accuracy is considered sufficient for viewers to associate each caption with the corresponding POI as it appears in the video. Furthermore, the text length of the generated captions was also reasonable for the display duration, and their content was confirmed to be factually accurate through qualitative evaluation. Future improvements should incorporate the users' personal experiences into the caption generation.

1. Introduction

With the widespread adoption of smartphones, recording and sharing videos at tourist destinations has become a common practice among tourists. These videos, often enriched with captions and audio, serve as a means of conveying impressions and information about visited places and are widely shared via social media. Such content not only enhances personal post-travel experiences but also contributes to the promotion of regional tourism and the stimulation of inbound demand—key objectives in smart city development.

Creating informative video content requires editing skills and the credibility of information. Determining the appropriate timing and content of explanatory captions based on tourist behaviour and camera orientation presents a significant challenge. Location-based services aiming to enhance the tourism experience through user behaviour tracking based on geofencing technology have been proposed in the past (Garg et al., 2017; Sasaki et al., 2024). However, existing geofencing approaches are limited to roughly detecting a user's proximity to landmarks and have not yet reached the capability of recognizing the user's attention or interest within the geofenced area.

Therefore, this study proposes a novel context-aware framework for caption generation in tourism video blogs, which detects tourist's focus on points of interest (POIs) by utilizing mobile sensors such as GPS and orientation sensors. By enabling dynamic content generation based on location and user behaviour, this framework encourages citizen participation, enhances regional appeal, and improves the digital experience through intelligent media generation based on sensor technology.

A video blogging application developed by the authors consists of a recording mode and a viewing mode, as shown in Figure 1. After recording video, walking trajectory, and orientation data, the caption generation process is executed in two stages: the first detects proximity and direction toward POIs and estimates appropriate caption display durations (discussed in Section 2); the second employs large language models to generate captions with appropriate length and sufficient background content based on the durations (discussed in Section 3). In addition, this study conducts an experimental evaluation in real-world environments to verify the usefulness of the developed application, as described in Section 4.

Figure 1. The video blog application developed by the authors. The left shows the recording screen, and the right displays the viewing screen of the generated video blog.

2. Geospatial Context Extraction for Determining Timing of Captioning

2.1 Definition of Guide Area

POI refers to a specific location on a map and is commonly used in navigation systems and map applications. In this application,

it refers to tourist spots that users visit during sightseeing, with each POI defined by geographic coordinates.

To enable AI-generated explanatory captions (as described in Section 3), a unique guide document is associated with each POI. These documents consist of highly reliable textual data authored by experts. As an example, the POI for "Meitoku Library" located in Akita City, Akita Prefecture, is associated with document data based on descriptions of "Meitoku Library" found on Wikipedia and the official website of Akita City.

Geofencing is a technology that sets a virtual boundary (geofence) around a specific location and triggers certain actions based on user events related to that boundary. By using device location data obtained through GPS and other methods, it is possible to detect entry into and exit from a defined geofence. Detecting entry into a geofence is an effective method for easily identifying user proximity to a specific location (Garzon and Deva, 2014). In this study, we define the guide area using this technology. For each major POI within the guide area, a circular geofence is placed centred on its location, enabling the system to recognize when the user is staying near a POI. Figure 2 shows an example configuration using Akita University in Akita City, Akita Prefecture as the guide area.

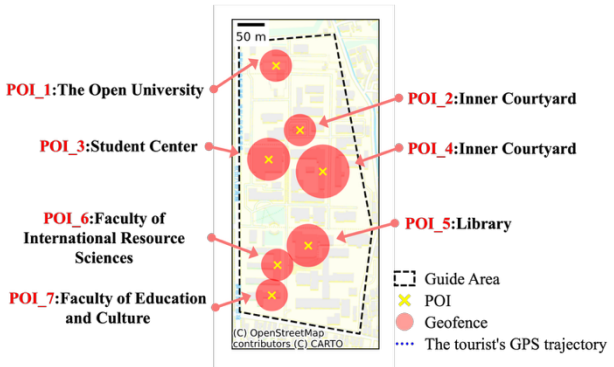

Figure 2. Example of guide area with geofences on the campus of Akita University, Japan.

2.2 Detection of Approach via Geofencing

To detect a user's approach to a POI using geofencing, the system determines whether the user is inside or outside the geofence. Figure 3 illustrates the recorded trajectory data, with coordinate point (reference points of the trajectory) color-coded based on whether they were determined to be inside or outside the geofence. For each recorded coordinate point, if the coordinate falls within the geofence area defined on the geographic space, the user is considered to have approached the corresponding POI at that location.

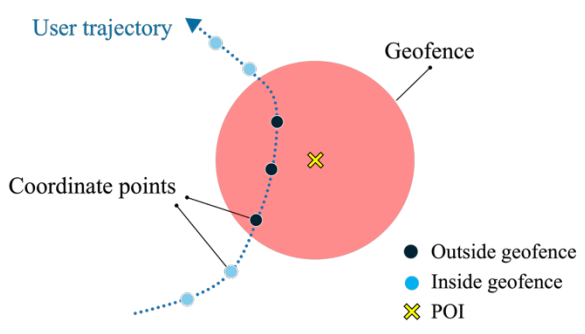

Figure 3. User movement trajectory and geofence inside/outside determination based on coordinate points.

2.3 Detection of Camera Viewing Angle Using Azimuth Data

To detect the time intervals during which the user is pointing the camera at a POI, the system determines whether the POI lies within the camera's viewing angle. Let $heading\_angle$ represent the camera's orientation data obtained from the azimuth sensor, POI_angle the directional data indicating the location of the POI relative to the user's coordinate point. and $view$ the camera's viewing angle. Using Equation 1, it is determined whether the POI lies within the camera's viewing range $(-\frac{view}{2} \sim \frac{view}{2})$ centred on $heading\_angle$. Figure 4 illustrates the $view$ centred on the recorded $heading\_angle$, and the POI_angle indicating the direction to the POI.

$$heading\_angle - view/2 < POI\_angle < heading\_angle + view/2, \quad (1)$$

where

- $heading\_angle$ = azimuth value from the orientation (direction) sensor
- POI_angle = azimuth from the user to the POI
- $view$ = field of view of the smartphone camera

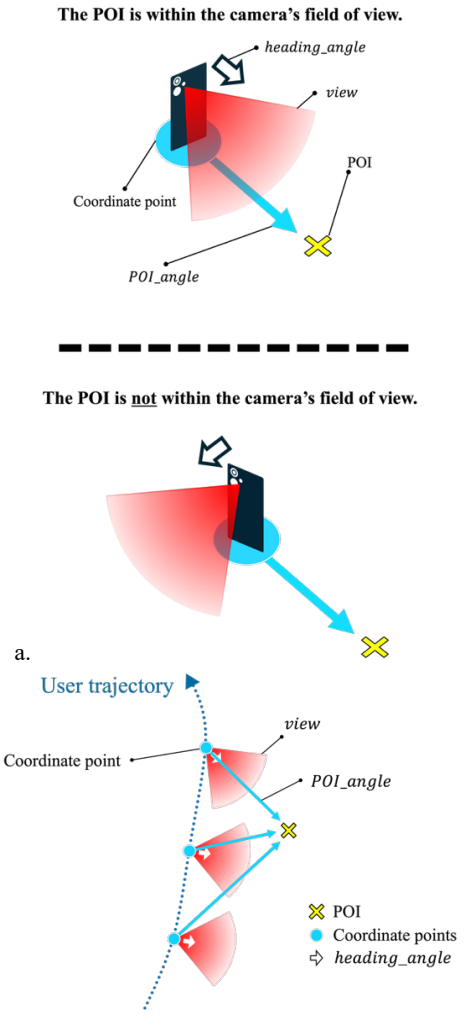

Figure 4. Determining whether the POI is captured within the camera's viewing angle (a) and overview of the trajectory (b).

2.4 Calculation of Caption Display Duration Based on Acquired Data

To overlay explanatory captions on recorded video, it is necessary to determine the caption display duration that provides appropriate timing for their appearance. In this study, the period during which the user directs their attention toward a POI and points the camera at it is defined as the optimal caption display duration. The method for determining this duration is described below.

First, the user's proximity to a POI is detected. Let $coordinates$ be the location data recorded by the GPS sensor, $heading\_angle$ the directional data recorded by the orientation sensor, and $time$ the timestamp at the moment of data recording. Then, the user's trajectory data is defined as $Tr = \{p_1, \dots, p_n\}, p_i = (coordinates_i, heading\_angle_i, time_i)(0 < i \leq n)$. The trajectory data and video data are time-synchronized based on $time$. By determining whether the location is inside or outside the geofence for each reference point $p_i(0 < i \leq n)$ in the trajectory data, a geofence dwell list $S = \{s_1, \dots, s_m\}, s_j = (poi\_id,\ p_{enter}, \dots, p_{leave})(0 < j \leq m)$ is extracted. The points p_{enter} and p_{leave} represent the reference point immediately after entering and just before exiting the geofence corresponding to the poi_id.

Next, the time during which the user points the camera at the POI is estimated. From the list of user location coordinates $\{p_{enter}.coordinates, \dots, p_{leave}.coordinates\}$ contained in each geofence dwell data s_j, and the location coordinates $POI_{poi\_id}.coordinates$ of the POI corresponding to poi_id, a direction list representing the direction from the user to the POI $\{POI\_angle_{enter}, \dots, POI\_angle_{leave}\}$ is calculated. Figure 5 illustrates this list of directions.

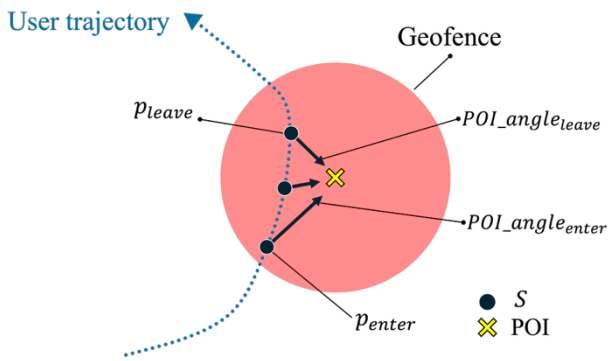

Figure 5. Direction list from the user to POIs and the user movement trajectory.

The direction list is then compared with the list of orientation data $\{p_{enter}.heading\_angle, \dots, p_{leave}.heading\_angle\}$ obtained from the azimuth sensor to perform the camera's viewing angle determination described in Section 2.3. In this process, the field of view of the smartphone camera—denoted as $view$—is used for the judgment. Since the field of view varies depending on the smartphone model, it is necessary to configure an appropriate value for each device. For example, in the case of an iPhone 12, the viewing angle is 80° (Apple Inc, 2020).

Based on the above processes, a list of valid dwell intervals during which the user is pointing the camera at the POI $S' = \{s'_1, \dots, s'_m\}, s'_j = (poi\_id, p_{camera\_in}, p_{camera\_out})(0 < j \leq m)$ is extracted. The difference between the timestamps of these time intervals, denoted as $p_{camera\_out}.time - p_{camera\_in}.time$, represents the caption display duration.

3. Dynamic Generation of Explanatory Captions

3.1 Caption Length Based on Caption Display Duration

If the amount of text in an explanatory caption exceeds the available display time, viewers may not be able to finish reading it before it disappears. Conversely, if the text is too short for the display duration, the caption may feel unnecessarily prolonged, potentially causing viewers to lose focus. For these reasons, it is necessary to generate captions with an appropriate number of characters suited to the display duration. Assuming Japanese users and letting wpm represent the number of characters a person can read per minute (Words Per Minute), the ideal number of characters in a caption, w, is defined by Equation 2 below.

$$w = \frac{wpm}{60}\left(p_{camera_{out}} - p_{camera_{in}}\right), \qquad (2)$$

3.2 Generation of Explanatory Captions Using Generative AI

Next, explanatory captions are generated using GPT-4, a large language model provided by OpenAI (OpenAI, 2023). By leveraging GPT-4's summarization capabilities, explanatory captions are generated from guide documents associated to each poi_id. Let l denote the number of characters displayed per caption. The prompt sent to GPT-4 consists of the guide document, preceded by an instruction such as "Please summarize the following content in approximately l characters." As a result, GPT-4 is queried k times, where k is the largest integer satisfying $k \leq w/l$, and the generated captions are sequentially displayed on the screen. The process of generating the explanatory captions is shown in Figure 6.

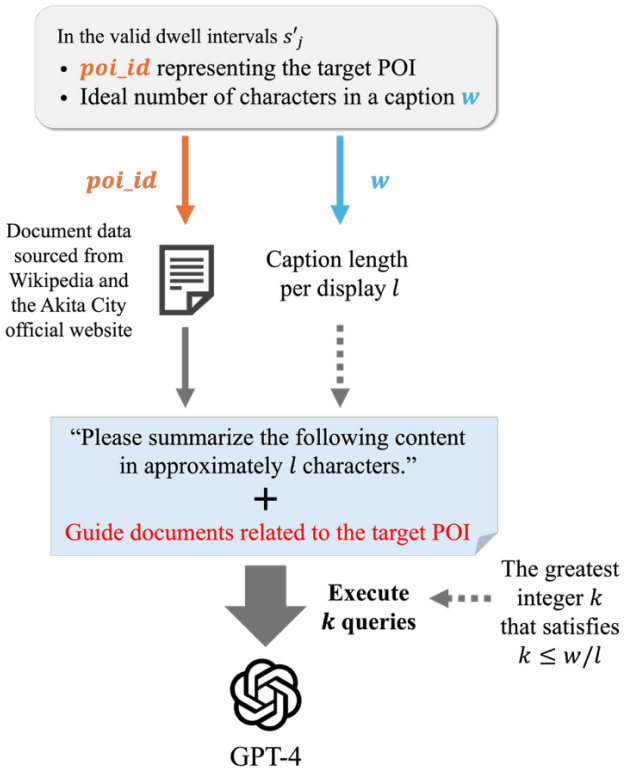

Figure 6. The process of generating the explanatory captions.

4. Field Evaluation of Video Blog Generation in Urban Parks

4.1 Experimental Settings

This section presents an evaluation of the usefulness of the developed video blog creation application. After defining guide areas in a real-world environment, a simulated tourist moves within the area and captures footage of POIs. The appropriateness of the calculated caption display times and the generated caption content is then assessed.

The experimental environment was set in Senshu Park in Akita City, Akita Prefecture, which served as the guide area. Senshu Park is a tourist attraction introduced on the official website of Akita City (Akita City, 2025). Within its grounds are features such as statues, ponds, and cultural facilities, making it a popular strolling area for both residents and tourists. Based on these characteristics, the park was considered suitable as an experimental environment for the application. The POIs designated within the guide area for the experiment were: Kogetsu Pond (POI_1), Statue of Lord Satake Yoshitaka (POI_2), Meitoku Library (POI_3), Akita City Cultural Creation Center (POI_4), and Akita Arts Theatre “Mille Has” (POI_5). These POIs were selected because each differs in terms of spatial size and shape, factors believed to affect the accuracy of calculating caption display duration. Guide documents related to these locations were cited from the official website of Akita City and Wikipedia.

In the experiment, the developed application was implemented on an Apple iPhone 12. A single participant (a student from the Faculty of Science and Engineering), assuming the role of a tourist, walked through the defined guide area while recording video and collecting data. Figure 7 shows the defined guide area and the obtained GPS trajectory. Next, the geofence dwell time and the timing of caption appearance and disappearance were measured from the collected data. The timing of objects entering and exiting the video frame was measured visually and used to evaluate the accuracy of the calculated caption display duration. For the dynamic generation of explanatory captions, parameters were set to wpm=360 and l=60, assuming a Japanese-speaking user. The generated captions were then evaluated based on two criteria: (1) whether the amount of text was appropriate, and (2) whether the content was sufficient for contextual understanding.

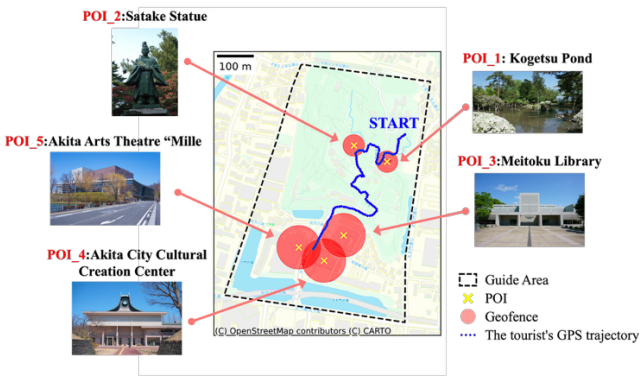

Figure 7. Map of the guide area defined in Senshu Park, and the tourist movement trajectory obtained from experiment.

4.2 Experimental Results

Figure 8 show the timelines for each POI’s geofence dwell time (top, blue), caption display time (middle, orange), and object appearance time (bottom, green). The timeline times represent elapsed time from the start, with each timeline displaying approximately 10.0 seconds before and after the geofence dwell period.

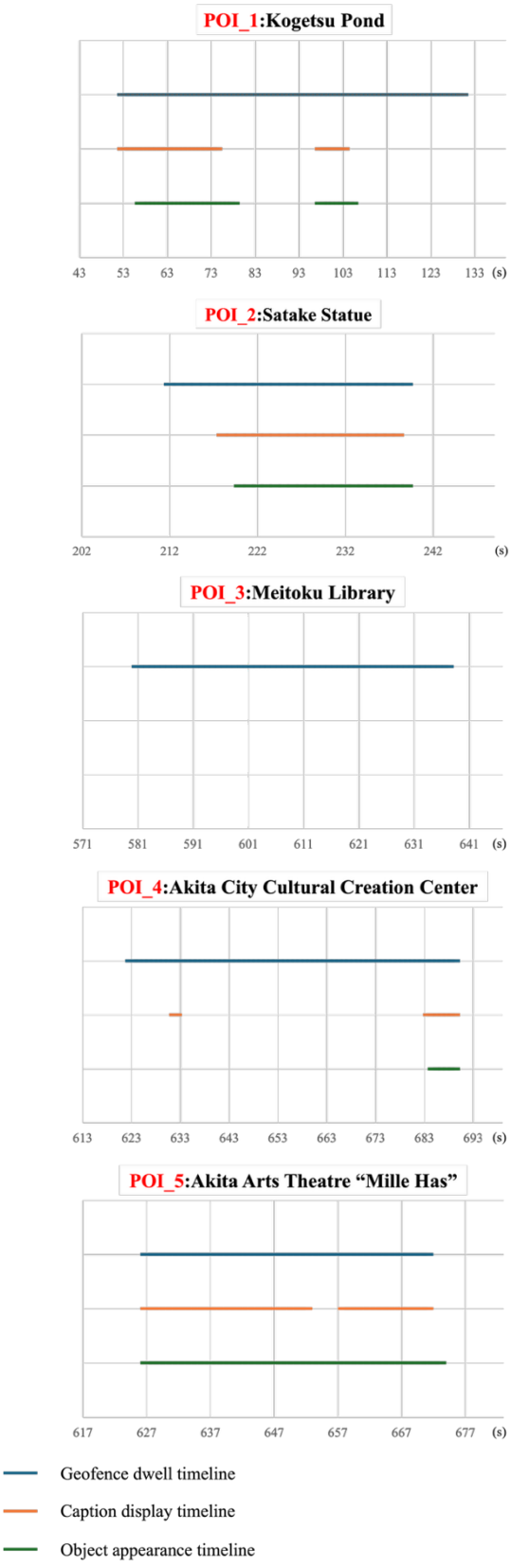

Figure 8. The resulting timelines for each POI. The x-axis values represent the number of seconds elapsed from the start time.

Overall measurement results indicated that the average error between caption display duration and object appearance time was about 1.8 seconds, with a maximum error of approximately 4.0 seconds. Additionally, for POI_3, although the user remained within the geofence, the POI did not appear in the video; therefore, only the geofence dwell time is shown.

Figure 9 shows the caption generated for the POI_1. The number of characters in the generated captions were: POI_1 (57 characters), POI_2 (61 characters), POI_4 (36 characters), and POI_5 (58 characters). Overall, no incorrect information that was not included in the guide documents was observed to be generated. However, infrequently, the generated text mixed plain form and polite form styles.

Figure 9. Japanese guide captions generated for POI_1: Kogetsu Pond. The captions states that in 2002, six lotus rhizomes were donated to Kogetsu Pond by Mr. Yozo Innami, the president of the Lotus Culture Research Association, and efforts have since been made to preserve the pure quality of the Ooga Lotus.

5. Discussion

5.1 Caption Timing Accuracy

In our evaluation (Section 4), the error between the caption display time and the object's appearance time was measured. For all POIs, the displayed captions fell within an error range that did not hinder the understanding of the correspondence between the captions and their respective subjects, demonstrating sufficient accuracy in the calculation of caption display timing.

For POI_4 and POI_5, errors occurred such as captions disappearing while the object was still appearing, or captions appearing when the object was not visible at all. This is believed to be influenced by the spatial size of the objects. For relatively small spaces like POI_1 and POI_2, the timing at which the corresponding POI coordinates were judged to be within the camera's viewing angle closely matched the actual timing of their frame-in. On the other hand, when capturing large facilities such as POI_4 and POI_5, even when these objects were visually within the frame, the field of view judgment based on POI coordinates did not consider them to be inside the camera's viewing angle. This behavior is thought to have caused unexpected errors. As a countermeasure, a camera viewing angle adjustment function can be defined based on the spatial size of the POI.

In the measurement results for POI_4, a significant time discrepancy was observed between the object's actual frame-in timing and the caption display duration. The timeline (Figure 8) indicates that during the period when the incorrect caption was displayed, the adjacent POI, POI_5, was being recorded. This suggests that while recording POI_5 and POI_4 was mistakenly detected. As a countermeasure to this issue, potential solutions include shaping the geofence more appropriately to match the target object (e.g., using a polygon aligned with the building) or allowing users to pre-select the POI they are recording to help prevent misdetections.

5.2 Caption Content Quality

Regarding the generated captions, all POIs except for POI_4 yielded character counts close to the predefined target length. The reason for the shorter caption at POI_4 is likely due to the shorter appearance time of the object compared to other POIs, which was around 6 seconds, resulting in a reduced amount of generated text. As a countermeasure, if a caption is likely to be generated with insufficient content for adequate background understanding, possible approaches include canceling the generation itself or slowing down the playback speed to extend the on-screen duration of the object.

In addition, the mixing of plain and polite forms in the generated text is believed to be influenced by inconsistencies in the writing style of the source guide documents, which vary between POIs. As a countermeasure, the writing style of the guide documents can be standardized, or specific instructions can be added to the generation prompt to ensure consistency in style—either fully plain or polite.

Across all cases, no captions significantly exceeded 60 characters, and no false information outside the scope of the guide documents was observed. This indicates that the generated captions are not excessively verbose and provide sufficient explanation for background understanding.

6. Conclusion

In this study, we developed a tourist video blog creation application aimed at enabling anyone to easily share the appeal of their local area. The application utilizes mobile sensors built into smartphones and the text summarization capabilities of generative AI. Results from validation experiments conducted in real-world environments revealed that combining geofencing-based proximity detection with azimuth data–based frame-in judgment effectively reduces the gap between the actual appearance of an object in the video and the caption display time. Furthermore, the AI-driven adjustment of text length functioned appropriately, enabling the dynamic generation of captions that support background understanding.

The experimental results presented in this paper suggest that the spatial extent of the target spot influences the accuracy of caption display timing. Moving forward, we will define guide areas not only within the region used in this study but also in other tourist locations and verify through implementation whether the system can adapt to various environments. In addition, the current validation involves only a small number of participants and is limited in its assumed environment. In this regard, further evaluations will consider issues such as magnetic interference in urban settings that may affect sensor accuracy, challenges arising from users holding their smartphones at various angles, and battery consumption caused by real-time processing or continuous sensor monitoring.

Regarding the generation of explanatory captions, we aim to build a mechanism that incorporates not only prepared guide documents but also users' impressions and newly recorded text information. Based on these, we aim to develop communication tools that more effectively express individual experiences.

Acknowledgements

This research was supported partly by JSPS KAKENHI Grant Numbers JP24K15631, JP24K02981, JP25K00559, and JP23K11362.

References

Akita City, 2025. Senshu Park – Living Information [in Japanese], https://www.city.akita.lg.jp/kurashi/doro-koen/1003685/1007159/index.html (27 June 2025).

Apple Inc, 2020. iPhone 12 - Technical Specifications. https://support.apple.com/en-asia/111876 (27 June 2025).

Garzon, S.R., Deva, B., 2014. Geofencing 2.0: taking location-based notifications to the next level. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UBICOMP'14)*, 921–932. doi.org/10.1145/2632048.2636093.

Garg, A., Choudhary, S., Bajaj, P., Agrawal, S, 2017. Smart geo-fencing with location sensitive product affinity. In *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, Redondo Beach, CA, USA, 7–10 November 2017; 1–10. doi.org/10. 1145/3139958.3140059.

OpenAI, 2023. GPT-4 Technical Report. arXiv (Version 6). doi.org/10.48550/arXiv.2303.08774.

Sasaki, I., Arikawa, M., Lu, M., Utsumi, T., Sato, R. 2024. Data-Driven Geofencing Design for Point-of-Interest Notifiers Utilizing Genetic Algorithm. *ISPRS International Journal of Geo-Information* 13(6), 174. doi.org/10.3390/ijgi13060174.