https://doi.org/10.1038/s44387-025-00006-w

AI system facilitates people with blindness and low vision in interpreting and experiencing unfamiliar environments

Haozhe Lin[1,6], Jiangtao Gong[2,6], Yu Wang[1,6], Jinsong Zhang[3,6], Bing Bai[1], Yan Zhang[2], Luyao Wang[2], Chenyu Wei[1], Yancheng Cao[2], Kun Li[3], Ruqi Huang[4] ✉ & Guyue Zhou[2,5] ✉

Engaging with nature significantly enhances well-being, yet millions of individuals with blindness and low vision (BLV) are often excluded from these benefits due to constrained environmental perception. Here, we introduce VIPTour, an AI-driven system powered by the FocusFormer algorithm, which transforms complex scenes into structured, personalized graphs using tailored attention mechanisms and a BLV-in-the-Loop Adapter. Through intuitive, user-centered interaction, VIPTour facilitates active exploration and in-depth comprehension during dynamic sightseeing, while enabling accurate, long-lasting recollection and effective communication among BLV individuals post-journey. Extensive experiments demonstrate that VIPTour significantly enhances positive emotions and memory retention, with a 67.9% increase in positive emotional response, a 94.7% rise in arousal, a 772.73% improvement in cognitive mapping accuracy, and a 200% enhancement in long-term memory accuracy. These results underscore VIPTour's ability to deliver an unparalleled, enjoyable, and memorable experience, promising profound benefits for the BLV community.

Visiting natural environments, such as parks, has been identified as a significant benefit for both physical and mental well-being[1]. Millions of individuals with blindness or low vision (BLV) also express a keen interest in proactively engaging with these unknown beauties[2]. While there have been many attempts to improve the quality of life of BLV individuals, previous studies have assumed that BLV individuals are aware of their goals and have focused on providing functional assistance such as navigation and obstacle avoidance, leaving BLV individuals to passively engage with the world[3–13]. Consequently, during leisure travel in unfamiliar environments, BLV individuals often experience a profound sense of helplessness and are compelled to rely on the assistance of their friends, family, and volunteers. This reliance often impedes their ability to actively explore and comprehend unfamiliar environments during dynamic sightseeing, as well as recollect and communicate with other BLV individuals after a journey[2,4,5]. Therefore, it is crucial to aid BLV individuals in understanding and relishing unfamiliar environments.

Technologies supporting independent mobility and scene comprehension for BLV individuals have been extensively discussed. Existing assistive solutions, such as canes, and artificial intelligence (AI) solutions aid BLV individuals in navigation and obstacle avoidance[6–13]. For example, NavDog[12], a robotic guide dog system, can help BLV individuals navigate to their destination while avoiding environmental obstacles. Additionally, OmniScrib[13] allows BLV individuals to access audio descriptions of the environment from videos. However, these solutions, assuming BLV individuals are aware of their goals and providing functional assistance, fall short in helping them perceive and understand unfamiliar environments, consequently leading to passive engagement. In reality, what BLV individuals urgently need is the ability to actively engage with unfamiliar environments. However, due to the overwhelming amount of visual information in unfamiliar settings surpassing the perceptual capabilities of BLV individuals, this task remains extremely challenging.

Here, we present an AI-driven System, named VIPTour (Fig. 1B, and Supplementary A), which aids BLV individuals in interpreting and experiencing unfamiliar environments. Unlike existing functional assistance technologies, VIPTour empowers BLV individuals with active exploration, in-depth comprehension, accurate and long-lasting recollection, and effective communication with other BLV individuals (Fig. 1A). VIPTour comprises a set of lightweight, portable, consumer-grade devices (Fig. 1B-i) and a brand-new AI framework called FocusFormer (Fig. 1B-iii). FocusFormer considers aesthetics, freshness, and basic needs as main

[1]Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing, 100084, China. [2]Institute for AI Industry Research (AIR), Tsinghua University, Beijing, 100084, China. [3]College of Intelligence and Computing, Tianjin University, Tianjin, 300350, China. [4]Tsinghua Shenzhen International Graduate School, Shenzhen, 518055, China. [5]School of Vehicle and Mobility, Tsinghua University, Beijing, 100084, China. [6]These authors contributed equally: Haozhe Lin, Jiangtao Gong, Yu Wang, Jinsong Zhang. ✉e-mail: ruqihuang@sz.tsinghua.edu.cn; zhouguyue@air.tsinghua.edu.cn

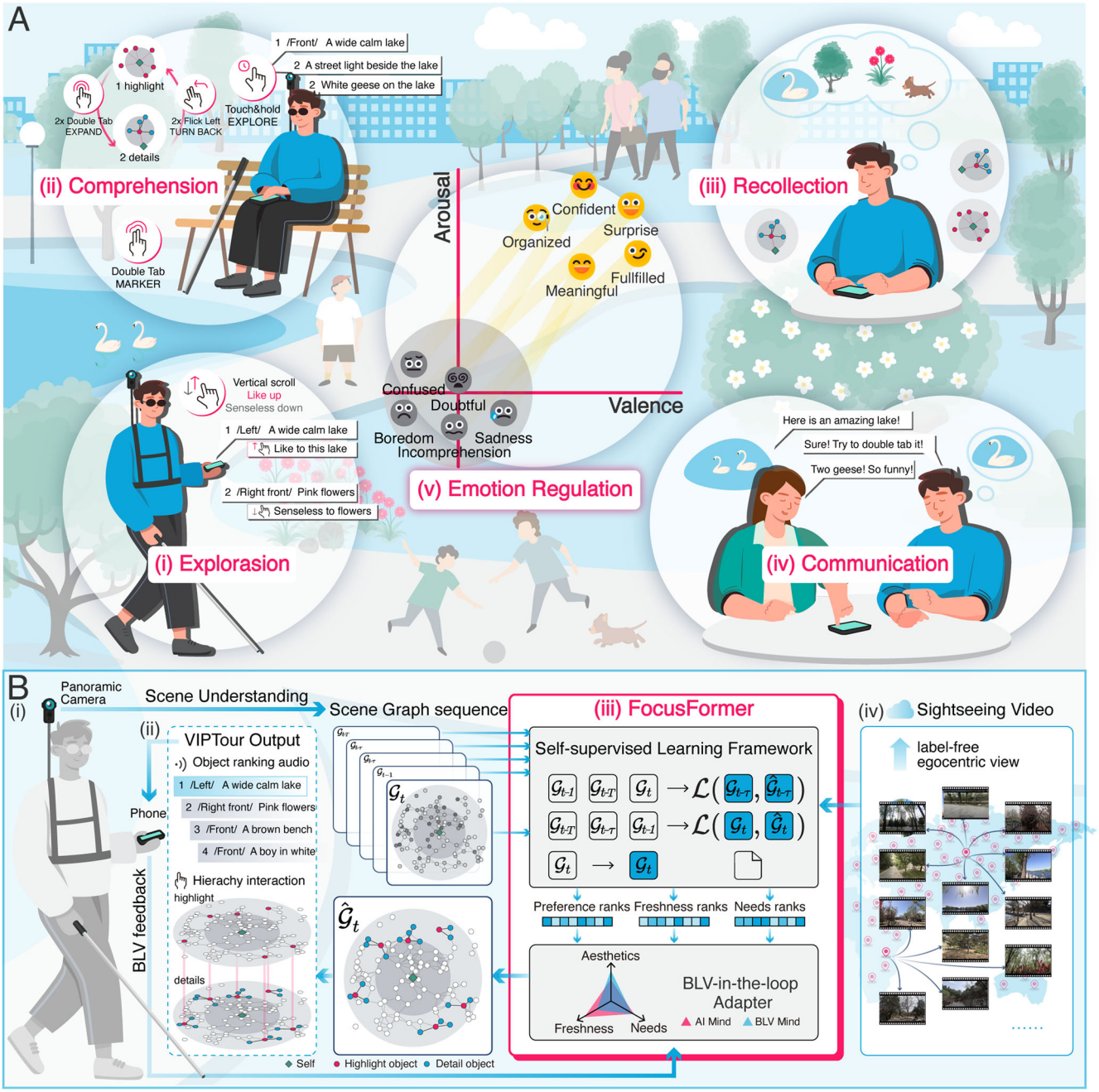

Fig. 1 | The overview of the VIPTour system. A VIPTour enables BLV individuals to understand and relish unfamiliar environments, encompassing exploration, comprehension, recollection, and communication. i) VIPTour gives BLV a heads-up about anything interesting, novel, or necessary nearby. ii) VIPTour exhibits a hierarchical organization of environmental information on a smartphone with a touchscreen interface. iii) VIPTour grants BLV individuals the capability to replay their tour experience. iv) VIPTour facilitates the dissemination of these experiences to other BLV individuals. **B** The VIPTour includes a novel deep learning algorithm framework, named FocusFormer, leveraging easily accessible hardware. i) The hardware components of the VIPTour system comprise lightweight, portable, consumer-grade devices, including a camera and a smartphone. ii) BLV individuals interact with the VIPTour system through efficient multisensory interaction techniques, like audio and hierarchy tactile interaction. iii) FocusFormer processes semantic graph sequences extracted from ego-view panoramic videos as input, and highlights relevant objects as output. Additionally, FocusFormer includes a BLV-in-the-Loop Adapter to interact with BLV individuals, through which FocusFormer can identify the preference of different BLV individuals and provide personalized assistance. iv) Thousands of public tourism videos are collected to train FocusFormer in a self-supervised manner, which is beneficial for effectively reducing aesthetic bias caused by manual annotations and learning semantic cooccurrence relationships among thousands of objects.

factors and develops tailored attention mechanisms to distill overwhelming information in unfamiliar environments, and it restructures vast amounts of information into a sparse and hierarchical graph structure. Based on this well-structured graph, FocusFormer interacts with BLV individuals through a smartphone application, and gradually understands their preferences through a nuanced BLV-in-the-Loop Adapter, and ultimately helps BLV individuals establish comprehensive and personalized cognition maps (Fig. 1B-ii). FocusFormer is trained through a graph masking self-supervised learning scheme using thousands of tourism videos from sighted tourists (Fig. 1B-iv, and Supplementary B.1), which effectively reduces the potential aesthetic bias arising from manual annotations and learns semantic cooccurrence relationship among thousands of objects.

The VIPTour system was thoroughly evaluated by over 30 BLV individuals by supporting them to actively engage with unfamiliar environments (Fig. 1A). Extensive analysis of their momentary emotion self-reports, performance on externalizing cognitive maps via verbal description, memory retention of the cognitive map, and their physiological measures reveals the vital role of VIPTour in bringing positive emotion and accurate long-lasting memory to BLV individuals (Fig. 1A-v). The compared experiment (a condition with or without the FocusFormer) further proved that the AI-driven VIPTour system offers well-structured and meaningful information, which aligns with cognitive principles of fluency and leads to an unprecedentedly enjoyable and memorable experience.

Results

VIPTour: AI system for BLVs understanding unfamiliar environments

Research has established that, akin to their sighted counterparts, BLV individuals harbor a strong desire for active engagement, which can result in an enhanced sense of control and improved psychological well-being[14–16]. However, the information about an unfamiliar environment can be overwhelming for BLV individuals, especially during dynamic sightseeing. Furthermore, unlike sighted individuals who can easily recall and share their experiences through photographs, BLV individuals encounter barriers in recollection and communication. Based on extensive literature and BLV individuals' feedback collected over several rounds of co-design sessions, we have developed VIPTour with the following considerations specifically tailored for BLV individuals.

Distilling overwhelming information for BLV individuals. When assisting BLV individuals in unfamiliar environments, it is important to provide compact but effective information while avoiding cognitive overload. Our research suggests that BLV individuals place considerable emphasis on factors such as aesthetics, freshness, and basic needs to explore and comprehend the unfamiliar environments. However, discerning aesthetics and freshness is subjective due to the diverse preferences among BLV individuals. Therefore, we propose FocusFormer algorithm with several attention mechanism to distill effective information for BLV individuals, which is trained through a graph masking self-supervised learning scheme to mitigate subjective annotation bias.

Catering to personalized interests of each BLV individual. Engagement requirements of BLV individuals differ significantly from those of sighted individuals, such as obstacle avoidance and wayfinding in unknown spaces[4,17]. Further, individual preferences and interests vary in the context of engagement. The VIPTour system, therefore, gathers specific needs of BLV individuals from survey data (see Supplementary B.2) including 118 participants for pre-tuning our algorithms. Additionally, we develop a BLV-in-the-Loop Adapter to incorporate feedback from BLV individuals during sightseeing for providing personalized assistance.

Enhancing environmental cognition through hierarchical Graph. Given the visual channel's superior bandwidth compared to tactile and auditory channels—estimated to be 0.01% and 1%, respectively[18]—presenting complex visual information to BLV individuals poses a significant challenge. The VIPTour system employs a dual-layered Hierarchical Interaction (Fig. 1B-i), incorporating highlighted spots from FocusFormer and scene mapping to facilitate a simplified and sensible information presentation. This approach has been proven effective for BLV individuals in previous research[19,20]. Participants can opt to zoom into the scene graph according to the learned hierarchy, enhancing their exploration of node relationships reflected in the scene.

Promoting recollection and communication among the BLV community. Prior research suggests that the BLV community exhibits a strong desire to share experiences and foster a strong sense of community[4]. To enable information sharing and emotional communication among BLV individuals, the VIPTour system features options for recording, storing, and sharing experiences (Fig. 1A-iii, iv). With these features, VIPTour allows BLV individuals to document and share scenes and moments encountered during their journeys, promoting a sense of connection and facilitating the exchange of knowledge and experiences within their social networks.

FocusFormer: the core of VIPTour system

FocusFormer aims to distill overloaded contextual information and align it with the personalized interest of each BLV individual, so that VIPTour can efficiently and effectively interact with BLV individuals given their limited bandwidth. FocusFormer mimics how humans would select views when assisting and accompanying blind people during sightseeing, and then dissect the motivation of BLV into three sources: the aesthetics of the unfamiliar scenes, the freshness (or novelty) of scenes arising from the incoming temporal dynamics, and the warning signals regarding daily needs (e.g., drinking, toilet usage, obstacle avoidance). Moreover, to learn these aspects for guidance, FocusFormer is trained in a self-supervised masked semantic graph scheme by only exploiting the rich structure within the unlabeled training data itself[21–25].

FocusFormer architecture

As outlined in Fig. 2A, the FocusFormer comprises four subnetworks. Each incoming video frame is initially translated into semantic graph sequence[26] using the scene graph generation technique[27–30] before being fed into FocusFormer for a training forward pass. The predicted output of FocusFormer is a reordered list of object instances corresponding to the nodes of the input scene graph. These nodes are ranked according to the predicted "BLV interest scores" associated with each object instance/node in the list, which FocusFormer uses for guiding decisions and interactions with BLV individuals.

The **background subnetwork** learns to extract the background objects enormously distributed in the training data, which may include objects such as "sky", "sun", "trees" and other common objects that are ubiquitously present in the parks but are potentially of least interests of tourists. These likely background objects are predicted with higher background scores, denoted as background score S_b. The **attraction subnetwork** infers the objects that sighted people may have focused on. This subnetwork reflects situations where sighted people might have steered the camera toward specific objects out of interest, akin to how one might consistently observe an appealing landscape. Objects captured by the attraction subnetwork are considered potentially interesting to sighted people and are assigned high attraction scores S_a by FocusFormer. However, high attraction scores might also include background nodes since these are likely to be constantly present during the tour. To compensate for the effect of background noise, we further define the pruned version of the attraction scores as aesthetics score S_A, computed as $S_A = S_a - \alpha \cdot S_b$, where α is a predefined hyperparameter defined as background weight. Meanwhile, the **freshness subnetwork** detects the novel objects in dynamic scenes, reflecting the emergence of objects not observed in the recent past. We assign a freshness score S_F to each object, measuring its novelty. FocusFormer also incorporates the obstacle avoidance and needs information through the **BLV needs subnetwork**, which draws on surveys collected from the BLV individuals (see Supplementary B.2). We assign S_N to represent the Need score, which is normalized according to the scale of S_A and S_F. After training procedure is complete, FocusFormer eventually computes and associates each class object present in the training data with three predicted scores S_A, S_N, and S_F. We define $S_{c,a}$, $S_{c,b}$, $S_{c,A}$, $S_{c,N}$, and $S_{c,F}$, to indicate each specific scores are assigned to specific class c. However, for general illustrating purpose (as in Fig. 2), we omit the c subscript to avoid cluttering.

FocusFormer SSL training philosophy

Distilling useful information for BLV individuals is challenging to achieve through supervised learning, as people have different interpretations of

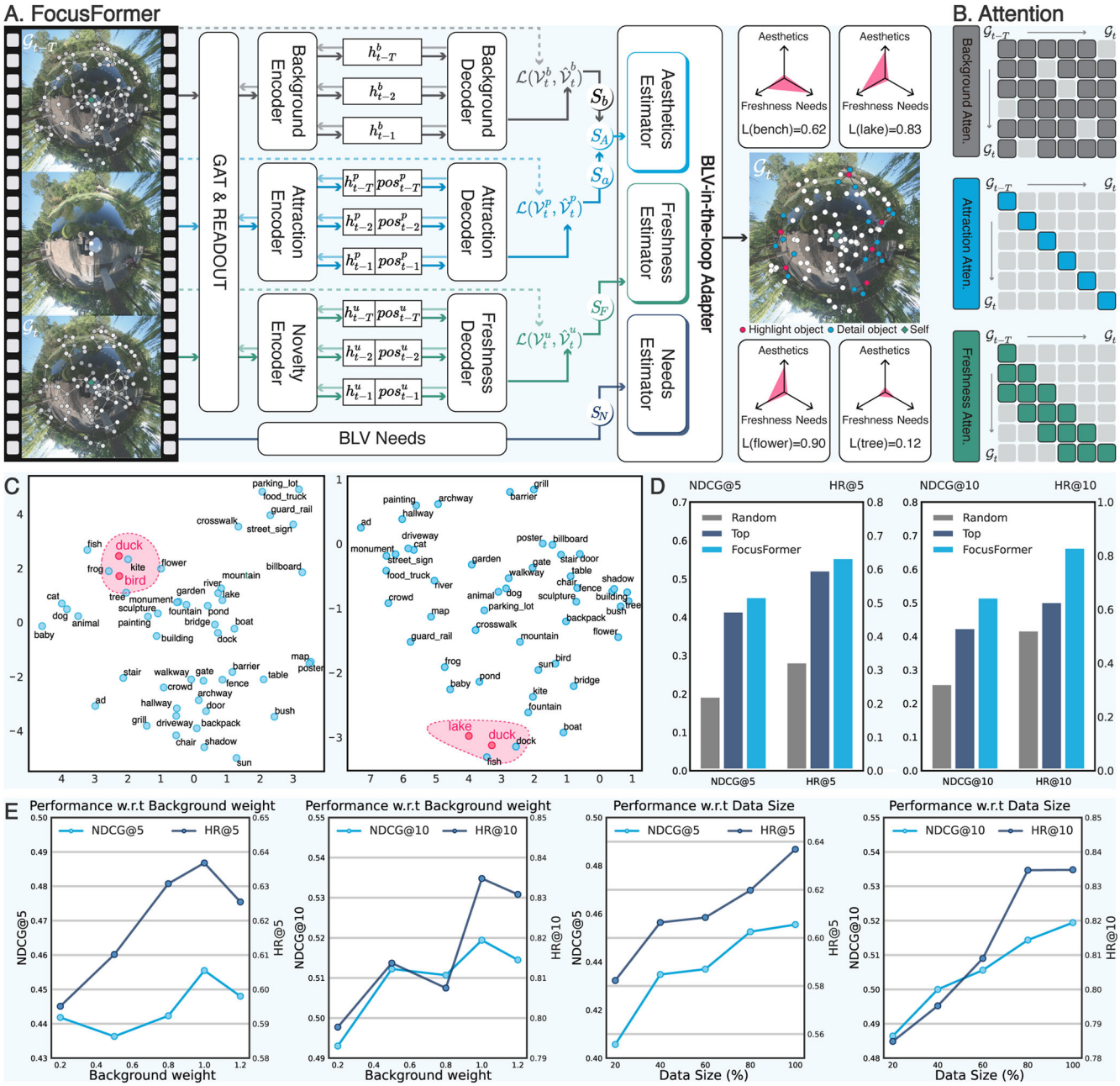

Fig. 2 | FocusFormer distills valuable contextual information for BLV individuals. A The Architecture of FocusFormer is composed of four subnetworks (background, attraction, freshness, and BLV needs. The background subnetwork infers the common background across various parks covered by the training data. The attraction subnetwork calculates the scores of potentially interesting objects during brief, local sessions. The freshness subnetwork identifies and captures unique and unexpected scenes during the sightseeing. The BLV needs subnetwork recognize the basic needs for BLV individuals. This architecture is trained under different masking strategies. **B** Different attention mechanism for the FocusFormer network. These units instill diverse inductive biases, allowing each subnetwork to detect distinct patterns that correspond to the background, attraction, and freshness characteristics. **C** The t-SNE visualization of embedding representing each class after training via (left) BERT and FocusFormer (right). FocusFormer captures semantically meaningful contextual information based on tourism training data. **D** Evaluation of the FocusFormer architecture. The inferred highlight signals are also compared with the constructed baselines including "Top" and "Random," demonstrating FocusFormer's accuracy in capturing sighted individuals' preferences. **E** Ablation study against the change of hyperparameter α, and the size of training data. The performance of FocusFormer keeps improving as training size increases, with an optimal point in α values.

beauty, making it difficult to annotate accurately. For instance, some individuals may prefer splendid landscapes, while others favor human history, and some enjoy the sounds of cicadas and birds. Therefore, we propose a graph masking self-supervised learning scheme to train FocusFormer, leveraging a vast collection of tourism videos from sighted tourists, which can effectively reduce the potential aesthetic bias arising from manual annotations and learn semantic cooccurrence relationship among thousands of objects.

Particularly, we use the semantic graph sequence corresponding to the tour videos as input for FocusFormer. For different subnetworks, we design different graph mask strategies and attention mechanism. FocusFormer is trained to recover these masked graphs as the objective. Through effective masked self-supervised training, FocusFormer can distill useful information to select the most relevant touring highlights from the three proposed touring incentives, potentially interesting to BLV people, without overwhelming them. The training mechanism of each subnetwork is illustrated

as in Fig. 2B. FocusFormer masks certain percentage of the nodes in the scene graphs according to the predefined masked patterns, and predict those masked nodes according to the specific training reconstruction objective in different subnetworks. The masked patterns are implemented through special form of attention mechanisms. The Background attention utilizes a positional-unaware design, capable of indiscriminately attending to randomly selected frames within a video segment. This stochastic approach ensures the extracted background is uncorrelated to any specific scene within the segment, thereby enabling the identification of common background elements throughout various scenes. The Attraction attention is architecturally constructed to attend to a sequential selection of frames. The design ensures the attention function operates across each pair of image embeddings, confined within a short consecutive local session. Consequently, entities with persistent presence in the scene, distinct from background objects, are associated with elevated pruned attraction scores. Finally, the Freshness attention predicts consistent entities that may linger in the scene, utilizing a positional-aware attention mechanism across all previous instances. This aids in the identification of novel events and helps spot the presence of fresh objects within the scene.

BLV-in-the-Loop Adapter. During inference time on a BLV sightseeing, FocusFormer gradually learns to comprehend the preference of each BLV individuals and provide the personalized assistance through interaction. Specifically, with the inference time functioning unit named BLV-in-the-Loop Adapter, the weights associated with S_A, S_N, and S_F will be iteratively updated during the tour, adjusting the final guiding list in real-time. The overall BLV interest score, considering all of the guiding information, is computed as $L = \lambda \cdot S_A + \beta \cdot S_N + \gamma \cdot S_F$. Specifically, the weights λ, β and γ are the weights respectively learned through the Aesthetics Estimator, Needs Estimator, and Freshness Estimator, respectively, using the maximum likelihood estimation method (MLE) learning procedure. According to VIPTour guiding philosophy, the objects with the highest overall score L (i.e, higher Aesthetics Score, Freshness Score, Needs score, and lower Background Score) will be eventually assigned a higher overall score L and will thus receive higher priority and rank in the guiding list.

During the tour guidance, VIPTour will prioritize highlighting the top attractions to BLV individuals based on the list ranked according to their overall scores L. The final presentation layout and hierarchy of the view will also be updated according to these scores. Through MLE, the BLV-in-the-Loop adaptor receives BLV feedbacks in the form of "likes" and "dislikes" for each object. The BLV-in-the-Loop Adapter then learns to adjust the weights between the three sources S_A, S_N, and S_F. (See Methods for more details). This MLE iteration leads to updated test time tour guiding plans as the λ, β and γ that reflect the tailored personal BLV interests are updated. This process can be viewed as the participant implicitly selecting important parameters that defined how the scene graph prioritizes the nodes, where FocusFormer tries to estimate the true value of these weights.

FocusFormer evaluations. Figure 2C presents the resultant t-SNE[31] embedding representing each object class, revealing interesting insights from the training. The plot at the left demonstrates the embedding obtained via conventional BERT[32] pre-training. Conversely, the right t-SNE plot compares the changes in the relative t-SNE embedding distribution after FocusFormer training. Intriguingly, unlike the BERT which clusters semantically similar objects together (i.e., duck and bird are in closer positions), FocusFormer training uncovers the unique underlying contextual information during sighted individuals' tourism. For example, the "duck" embedding is located in positions closer to "dock", "pond", "fish", "bridge" and other "pond" related classes for FocusFormer, showcasing the extracted contextual relationship between "duck" and "pond" under the particular occasion of tour in park.

Figure 2D quantifies the accuracy of how well FocusFormer predicted objects with the highest predicted aesthetics scores S_A align with the actual preferences of the tourist who captured the video. We compare FocusFormer with two considered baseline models here. FocusFormer chooses the top five objects with the highest S_A scores. The "Random" baseline randomly selects the top five classes in the ranking list, while "Top" calculates the frequencies of each object class (each instance will count as 1) out of the total number of nodes across all training scene graphs, and chooses the top five objects with highest frequencies for each view. We assess how these returned top five objects from each algorithm correspond to the true preference (annotated) of the tourist according to the evaluation metrics Normalized Discounted Cumulative Gain (NDCG@5) and Hit Ratio (HR@5)[33]. FocusFormer demonstrated a significant advantage over the two baseline models, justifying the effectiveness of using self-supervised learning to extract the statistical patterns in terms of aesthetics (S_A), reflecting a relatively accurate prediction on landscape background (S_b) and touring attraction (S_a).

Figure 2E ablates the effect of the "Background Weight" hyperparameter α and illustrates an optimal point of the α value in achieving the best "Preference" prediction accuracy. Notably, the prediction accuracy of FocusFormer continually improves as the size of the training data increases, justifying the critical role of training data scales in terms of FocusFormer's self-supervised training efficiency. We also report the influence of other hyper-parameters in Supplementary C.2.

Interpret and experience unfamiliar environments with VIPTour

VIPTour system aids the BLV individuals in interpreting and experiencing the unfamiliar environments across four key aspects, including exploration, comprehension, recollection and communication. We collected emotion regulation performance data from these scenarios with the participation of 33 individuals (See Supplementary D.1 for participant recruitment details), thereby demonstrating the efficacy of VIPTour.

Exploration. While walking in an unfamiliar environment, BLV individuals can gain an essential understanding of the surroundings through VIPTour's voice broadcast (Fig. 3a). In this scenario, BLV individuals can receive information from VIPTour and provide feedback through single-handed interaction using their smartphone while navigating with their regular aid (e.g., a white cane). A BLV-in-the-Loop Adapter (Fig. 3b) is utilized to adjust the weights between the three sources, leading to updated tour guide plans displayed in the subsequent capture frame. Therefore, the VIPTour system primarily offers a quick "glance" at the surroundings, including the name, direction and attributes of the selected items, as determined by the FocusFormer framework's ranking score. With VIPTour, the valence of the post-test was significantly increased by 50% (66.67%) (3.00 ± 0.67 $Md \pm SD$ versus 2.00 ± 0.87 $Md \pm SD$, $p = 0.04$, effectsize $= 0.16$, paired Wilcoxon signed-rank test) and the arousal of the post-test's was significantly increased by 200% (120%) (1.00 ± 1.05 $Md \pm SD$ versus 3.00 ± 0.73 $Md \pm SD$, $p = 0.02$, effectsize $= 0.18$, paired Wilcoxon signed-rank test) in exploration (Fig. 3g). The detailed experimental procedure can be referred to Supplementary D.2.

Comprehension. Upon encountering picturesque locations, BLV individuals can interact with VIPTour system with a tactile interface, enabling them to gain a profound understanding of their surroundings (Fig. 3c). Given that typical scenes contain numerous objects, we opted for a hierarchical structure to visualize objects information, as its efficiency has been validated for BLV individuals in previous studies[19,20]. BLV individuals can the surroundings by slowly moving their touch on the smartphone screen (Fig. 3d). Additionally, they can zoom in or out of the detail layer with familiar daily smartphone gestures. This allows BLV individuals to carefully explore the current scene and reconstruct their mental map independently. With VIPTour, the valence of the post-test was significantly increased by 75% (2.00 ± 1.30 $Md \pm SD$ versus 3.00 ± 0.60 $Md \pm SD$, $p = 0.048$, effectsize $= 0.16$, paired Wilcoxon signed-rank test) and the arousal of the post-test was significantly increased by 50% (2.00 ± 0.97 $Md \pm SD$ versus 3.00 ± 0.87 $Md \pm SD$,

Fig. 3 | Typical application scenarios. a Exploration: BLV individuals acquire essential information about their surroundings and provide feedback through single-handed smartphone interaction while using their standard navigation aid (such as a white cane). **b** The BLV-in-the-loop unit: The FocusFormer framework learns and adapts to the individual touring preferences of BLV individuals during interactions. The parameterization reflecting the various needs BLV individuals, as recorded in VIPTour, is updated iteratively throughout the tour. **c** Comprehension: BLV individuals can comprehend objects of interest in detail when they encounter picturesque locations. **d** Hierarchical Interaction: This entails a dual-layered scene structure that incorporates recommendations from the FocusFormer and scene mapping to establish a streamlined information hierarchy. The smartphone touchscreen mirrors the scene graph structure of the nodes described in the audio play. BLV individuals can select to zoom into the scene graph based on the learned hierarchy to better understand the relationships between the nodes reflected in the scene. The edges in the scene graph correspond to the direction and position information of the objects. **e** Recollection: BLV individuals can revisit scenes they explored and memorable events they recorded as voice tags during the trip at any time after the tour. The interaction modality is as same as comprehension. **f** Communication: BLV individuals can share scenes and related voice tags with their friends via a smartphone using the VIPTour system. The interaction modality is the same as comprehension and recollection. **g** Emotion Regulation Effect: Momentary emotion self-reports from BLV individuals indicate an increase in both emotional valence and arousal after using the VIPTour system.

$p = 0.04$, effectsize $= 0.23$, paired Wilcoxon signed-rank test) in comprehension (Fig. 3g). The detailed experimental procedure can be referred to Supplementary D.2.

Recollection. After a sightseeing, BLV individuals can recollect the scenes they explored and memorable events they recorded as voice tags during the trip (Fig. 3e). The interaction modality is the same when they comprehend the unfamiliar environments, eliminating the need for additional instruction. With VIPTour, the valence of the post-test was significantly increased by 100% ($1.71 \pm 1.25\,M \pm SD$ versus $0.86 \pm 0.90\,M \pm SD$, $p = 0.02$, effectsize $= 0.78$, paired two-sided t-test, $n = 14$) and the arousal of the post-test was significantly increased by 100% ($2.00 \pm 1.16\,Md \pm SD$ versus $1.00 \pm 0.98\,Md \pm SD$, $p = 0.015$, effectsize $= 0.25$, paired Wilcoxon signed-rank test) in recollection (Fig. 3g). The detailed experimental procedure can be referred to Supplementary D.3.

Communication. With the VIPTour system, BLV individuals can share the scenes and associated voice tags with their friends via a smartphone (Fig. 3f). Sharing experiences is essential for both sighted and BLV individuals. The personal experiences of BLV individuals not only serve as valuable references for others with visual impairments, but they also foster a strong sense of community[4]. Thus, VIPTour enables individuals to share the scenes they have explored with others, either in person or through social media, allowing BLV individuals to learn about scenes recommended and shared by their peers and the community. The interaction modality is the same as comprehension and Recollection. With VIPTour, the valence of the post-test was significantly increased by

100% ($2.00 \pm 0.25\ Md \pm SD$ versus $1.00 \pm 0.30\ Md \pm SD$, $p = 0.08$, effectsize = 0.09, paired Wilcoxon signed-rank test) and the arousal of the post-test was slightly increased by 21.95% ($1.64 \pm 1.08 M \pm SD$ versus $2.00 \pm 1.04\ M \pm SD$, $p = 0.132$, effectsize = 0.34, paired two-sided t-test, $n = 14$) in communication (Fig. 3g). The detailed experimental procedure can be referred to Supplementary D.4.

The comparative experiment of the FocusFormer algorithm

Participants and experimental design. We recruited participants from a local university for individuals with blindness or low vision. All participants were initially invited to a two-hour park tour. Seven days after the park tour, we invited the participants back for a two-hour recollection and communication experiment. For the communication experiment, the participants were asked to invite a friend with similar age and visual impairments. The entire recruitment and study procedure was approved by the institutional review board. Finally, we had 46 participants in total, where 18 participants attended the exploration and comprehension experiment, and 16 of these attended the subsequent recollection and communication experiment. An additional 16 participants, invited by their friends, attended the communication experiment. The detailed experimental procedure can be referred to Supplementary D.3.

We evaluate the effectiveness of the FocusFormer algorithm by setting up a controlled experiment with both quantitative and qualitative measures. We collected the feedback, memory task performance, and physiological data from participants when using the FocusFormer algorithm ("FocusFormer" condition) and compared it with their feedback when using the Random Algorithm ("Baseline" condition). Each participant experienced both conditions on one day, and the order of each condition is counterbalanced through swapping.

Overall usability and emotion regulation effects. The results suggest the VIPTour facilitate independent mobility for visually impaired individuals. The average usability scores[34] given by participants for Exploration and Comprehension, Recollection and Communication were 80.83, 80.18, 80.00 out of 100, respectively. These scores are either higher than or comparable with other assistance tools for BLV individuals[35,36]. With the VIPTour system, the arousal of the post-test's momentary emotion self-reports increased by an average of 1.07 ± 1.12 ($M \pm SD$, $p = 0.001$, effectsize = 1.03, paired two-sided t-test), which corresponded to an average increase of 94.7% when compared with pre-test. At the same time, the valence of the post-test's momentary emotion self-reports increased by an average of 0.95 ± 1.11 ($M \pm SD$, [min, max] = [−4, 4], $p_{valence} = 0.001$, effectsize = 0.91, paired two-sided t-test, $n = 80$), which corresponded to an average increase of 67.9% when compared with pre-test.

Momentary emotion self-reports. The arousal and valence of the pre-test's momentary emotion self-reports has no significant difference between FocusFormer condition ($n = 9$) and Baseline condition ($n = 9$). After experiment, the arousal of post-test's momentary emotion self-reported increased by an average of $2.00 + 1.33$ ($Md \pm SD$, $p = 0.02$, effectsize = 0.19, paired Wilcoxon signed-rank test) in FocusFormer condition, while it has no significant difference with the Baseline condition (Fig. 4C). The valence of the post-test's momentary emotion self-reported increased by an average of $1.00 + 1.24$ ($Md \pm SD$, $p_{valence} = 0.02$, effectsize = 0.18, paired Wilcoxon signed-rank test) in FocusFormer condition, which corresponded to an average increase of 50.0% (93.33%) when compared with Baseline Condition (Fig. 4D).

Performance on externalizing cognitive maps via verbal description. Participants were required to describe the scene they just explored as detailed as possible to externalize the cognitive map they reconstructed through the VIPTour system. Two trained raters rated each participant's verbal description based on the rating scheme (see Supplementary Table S6) independently. Inter-rater reliability was conducted on 48.28% of the data, where the inter-rater Kappa was greater than 0.89 ($p < 0.001$). With FocusFormer, the performance on externalizing cognitive maps increased 18.18% when compared to the Baseline Condition ($2.08 \pm 1.39 M \pm SD$ versus $1.76 \pm 1.20\ M \pm SD$, $p = 0.02$, effectsize = 0.25, two-sided t-test, $n = 16$). For the analysis, the accuracy of cognitive map verbal description was extracted for three parts: the accuracy of object name, object direction and object attribute. With FocusFormer, the accuracy of object name was significantly increased 52.35% (32.26%) ($2.59 \pm 0.86\ Md \pm SD$ versus $1.70 \pm 0.85\ Md \pm SD$, $p = 0.047$, effectsize = 0.755, Mann-Whitney U test, between-subject, $n = 16$) in exploration, and 55.94% ($3.15 \pm 0.79 M \pm SD$ versus $2.02 \pm 0.57\ M \pm SD$, $p = 0.001$, effectsize = 0.688, two-sided t-test, $n = 14$) in comprehension. At the same time, the accuracy of object direction was significantly increased 772.73% (173.92%) ($2.88 \pm 1.33\ Md \pm SD$ versus $0.33 \pm 1.30\ Md \pm SD$, $p = 0.007$, effectsize = 1.166, Mann-Whitney U test, between-subject, $n = 14$) in comprehension with FocusFormer (Fig. 4E).

Memory retention of the cognitive map. We conducted both the short-term memory (STM) test and long-term memory (LTM) test of comprehension by requiring the participants ($n = 14$) to give answers to a same question list (Supplementary Table S7). The STM test was conducted after experiment directly and the LTM test was conducted about seven days ($d = 7.13 \pm 0.96\ M \pm SD$) after experiment. Besides, we invited the participants into memory experiment and collected their STM performance data after that. Two trained raters rated each participant's answers based on the rating scheme (see Supplementary Table S8) independently. Inter-rater reliability was conducted on 31.82% of the data, where the inter-rater Kappa was 1.00 ($p < 0.001$). Without FocusFormer, the LTM score (Fig. 4F) significantly reduced by 2.00 ± 1.73 ($Md \pm SD$, $p = 0.017$, effectsize = 0.12, paired Wilcoxon signed-rank test, $n = 14$) compared to their STM score in Baseline condition. However, with FocusFormer, the participants' LTM performance increased 200% (85.71%) ($3.00 \pm 1.37\ Md \pm SD$ versus $1.00 \pm 1.09\ Md \pm SD$, $p = 0.001$, effectsize = 1, 48, Mann-Whitney U test, between-subject, $n = 14$), when compared with Baseline condition. Besides, the participants also got a significantly better STM after memory than Baseline ($1.77 \pm 1.34\ Md \pm SD$ versus $1.38 \pm 1.17\ Md \pm SD$, $p = 0.003$, effectsize = 1.35, Mann-Whitney U test, between-subject, $n = 14$).

Physiological measures. The Electrodermal Activity (EDA) and Heart Rate Variability (HRV) of participants ($n = 12$) were recorded by an E4 wristband[https://www.empatica.com/en-int/research/e4/] during the experiment. For analysis, we conducted time-normalization for different phases of each participant's experiment. With the FocusFormer, the EDA data (Fig. 4G) was significantly increased by 9.17% ($2.38 \pm 0.37\ M \pm SD$ versus $2.18 \pm 0.11\ M \pm SD$, $p < 0.001$, effectsize = 1.45, two-sided z-test, $n = 12$) in exploration and by 23.77% ($2.76 \pm 0.30\ M \pm SD$ versus $2.23 \pm 0.08\ M \pm SD$, $p < 0.001$, effectsize = 2.35, two-sided z-test, $n = 12$) in comprehension when compared with baseline condition. Simultaneously, the HRV data (Fig. 4H) was significantly decreased by 3.46% ($327.75 \pm 109.52\ M \pm SD$ versus $339.09 \pm 82.35\ M \pm SD$, $p = 0.004$, effectsize = 0.73, two-sided z-test, $n = 12$) in exploration and by 21.25% ($248.84 \pm 49.68\ M \pm SD$ versus $301.48 \pm 34.95\ M \pm SD$, $p < 0.001$, effectsize = 1.23, two-sided z-test, $n = 12$) in comprehension.

Memorable tourism experience scales and Social virtual reality photo sharing experience questionnaires. We collected the subjective experience of participants during the sightseeing and sharing experiment after the trip through tourism memorable tourism experience scales[37] and Social virtual reality photo sharing experience questionnaires[38]. With FocusFormer, the participants ($n = 18$) reported significant higher meaningfulness ($5 \pm 0.68\ Md \pm SD$ versus $4 \pm 1.04\ Md \pm SD$, $p = 0.04$, effectsize = 1.09, Mann-Whitney U test, between-subject, Fig. 4I) and higher quality of interaction ($26 \pm 1.89\ Md \pm SD$ versus $21 \pm 4.96\ Md \pm SD$, $p = 0.026$, effectsize = 1.49, Mann-Whitney U test, between-subject, $n = 14$, Fig. 4J).

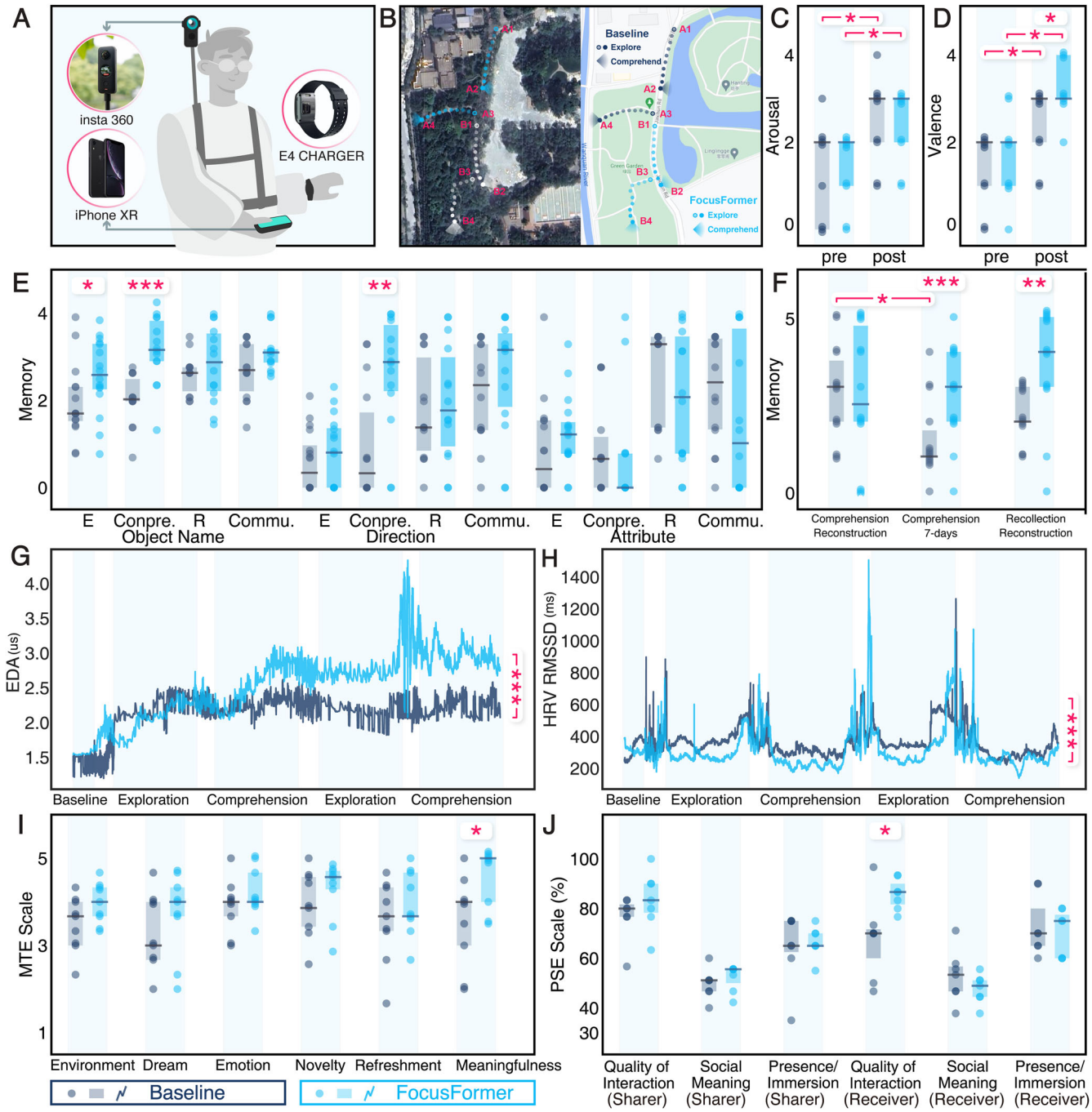

Fig. 4 | Compared experiment to evaluate the effectiveness of FocusFormer algorithm. A, B The experimental site and experimental condition design. The between-within-subject design paradigm is utilized in this experiment. **C, D** The momentary emotion self-reported by BLV individuals in two conditions. Both the valence and arousal of BLV's emotion become significantly higher after the trip in both conditions. The post valence of FocusFormer group is significantly higher than baseline condition. **E** The performance on externalizing cognitive maps via verbal description task, the accuracy of object name, object direction and object attribute show higher in FocusFormer than Baseline condition. **F** The long-term memory (the memory effect lasted seven days) is also significantly more accurate in FocusFormer condition than the Baseline condition. **G, H** On a time-normalized scale, it becomes apparent that the increase in both Electrodermal Activity (EDA) and Heart Rate Variability (HRV) in FocusFormer condition. Solid lines denote mean values. The grey background area denotes exploration and comprehension. **I, J** The BLV participants involving in our experiment reported significant higher meaningfulness and higher quality of interaction based on memory tourism experience (MTE) scale and photo sharing experience (PSE) questionnaire.

Discussion

Millions of people worldwide suffer from vision impairment, often experiencing feelings of isolation and frustration as they struggle to establish a connection with unfamiliar environments. While numerous efforts have been made to enhance the quality of life for visually impaired individuals, previous studies have predominantly focused on functional assistance, such as navigation and obstacle avoidance, assuming that individuals are aware of their goals in unfamiliar environments, leaving BLV individuals to passively engage with the world. However, what BLV individuals truly need is a sense of control over their surroundings. Therefore, in this paper, we introduce the VIPTour system, for the first time, it empowering BLV individuals to actively explore unfamiliar environments, establish an in-depth

comprehension, maintain accurate and long-lasting recollection, and communicate with other BLV individuals, which aids BLV individuals in understanding and relishing unfamiliar environments.

Technically, VIPTour encompasses both algorithmic and interactive innovations. At the algorithmic level, we propose FocusFormer, a neural network that integrates two key advancements: (1) multi-attention mechanisms to extract meaningful information from complex scenes—focusing on aesthetics, novelty, and basic needs—thereby reducing the cognitive load on BLV users; and (2) the BLV-in-the-Loop Adapter, which enables real-time, personalized interaction by dynamically adapting to individual preferences. At the interaction level, we developed dedicated software that leverages intuitive auditory and tactile feedback, optimizing the use of BLV users' limited perceptual bandwidth. Together, these innovations address the unique challenges of understanding unfamiliar environments. Compared to previous technologies with the potential to provide environmental information descriptions for BLV users, such as video captioning, VIPTour not only offers a real-time interactive hardware-software system but further advances the field through its FocusFormer algorithm, which filters redundant visual information to recommend meaningful, personalized environmental exploration spaces for BLV users.

Research[39,40] indicates that the presentation of organized and engaging information enhances individuals' pleasure and facilitates deeper memory encoding. Cognitive and psychological theories propose that humans have a natural inclination towards processing well-structured and meaningful information, leading to a more enjoyable and memorable experience. One possible explanation for this phenomenon is the concept of cognitive fluency, which relates to the ease with which information is processed and understood. Clear and organized information presentation reduces cognitive load, enabling individuals to concentrate their mental resources on comprehending and assimilating the content. This improved processing fluency leads to a positive affective response, as individuals perceive the information as more pleasant. Moreover, the interaction between freshness and familiarity plays a role in the effect of organized and interesting information on memory. Freshness stimulates curiosity and attracts attention, while familiarity provides a sense of cognitive comfort and coherence. Information presented in a structured and engaging way balances freshness and familiarity, maintaining individuals' interest and engagement. This optimal stimulation level promotes active memory encoding and consolidation, leading to enhanced retention and recall.

Our empirical evidence supports that a self-supervised learning technique aptly captures the cognitive fluency discussed above, excelling particularly at revealing how different concepts during tourism scenes are statistically related. Without relying on human annotations, self-supervised learning models predict the required statistical patterns using only unlabeled tourism data, leveraging the rich structure of the data itself. This approach removes potential bias in touring preference labeling and encourages the model to learn meaningful deep representations related to tourism that reflect the true inherent contextual structure of landscapes and tourist views. While the adopted self-supervised learning method learns to extract interesting contextual information, its supervised counterpart may be more limited. These tailored design considerations of FocusFormer enable the VIPTour system to successfully model the desired cognitive fluency, thereby improving the tourism experience for BLV individuals. Therefore, the VIPTour system, incorporating FocusFormer, provides organized and engaging information associated with increased pleasure and deeper memory encoding for BLV participants. By aligning with cognitive principles of fluency and leveraging the interplay between freshness and familiarity, this information facilitates more effective information processing and memory formation.

Empirical evidence demonstrates that the VIPTour system, in conjunction with the FocusFormer framework, successfully captures and models these cognitive fluency principles. Participants' overall feedback was quite positive, and their emotional state significantly improved when using the VIPTour system. Particularly, when comparing the VIPTour system with and without the FocusFormer, we observed a significant impact on the participants' positive emotions and accurate long-lasting memory. These results indicate that VIPTour leads to an unprecedentedly enjoyable and memorable experience, which will have a profound impact on BLV individuals. We also hope that this work can raise awareness for the BLV community and inspire further research on improving the quality of their life.

Method

The architecture of VIPTour was carefully tailored to meet the goal of BLV touring guide, i.e., to communicate with BLV via succinct and selective touring guiding information in a way that how human think. To mimic how human select the views when accompanying blind people during sightseeing, we decompose the touring incentive of BLV into 3 sources: the aesthetics of the unknown scenes, the freshness of scenes stemming from the temporal dynamics, and the warning signals regarding daily needs (e.g., drinking, toilet usage) and obstacle avoidance. We propose FocusFormer architecture, which learns to associate each object with three scores, namely "Aesthetics Score $\boldsymbol{S_A}$", "Freshness Score $\boldsymbol{S_F}$", and "Need Score $\boldsymbol{S_N}$", each reflective of one of the three guiding sources from enormous crowd sourced data. Through effective training on FocusFormer, VIPTour only selects and returns touring highlights potentially of interests to BLV people, without overwhelming BLV with the complex scene layout.

Architecture of FocusFormer

FocusFormer is essentially a deep Transformer neural network architecture implemented with three types of attention modules. Each attention module mimics a specific type of view selection strategy. The network is input with sequences of scene graphs extracted from the sightseeing views. FocusFormer applies masked modeling techniques through these attention modules on the scene graph sequence, to predict respectively "Aesthetics Score S_A", "Freshness Score S_F", and "Need Score S_N". The FocusFormer then outputs a ranking list of the graph nodes according to the computed overall score L.

Input. For each training video, we first evenly sample n image frames per minutes from the ego-viewed video sequence. We use these total number of $N_{tot} = n \times minutes$ sampled frames as the training images. Each training image is then transformed into scene graphs by employing the scene graph generation (SGG) technique[27–30]. Specifically, each edge E_{ij} in the generated scene graph represents a particular form of relationship between a connected pair of nodes n_i and n_j, where n_i represents the i^{th} node in the graph. Examples of relationships could be "in", "on", "against" and so on. Each node essentially represents a detected object in the scene, such as "fish", "lake", "boy". A generated scene graph of the k^{th} image is then represented as G_k, which is composed of several $\{n_i\text{-}E_{ij}\text{-}n_j\}$ tuples. We build training sequences in the form of $s_l = \{G_l, \ldots, G_{l+m}, \ldots, G_{l+M}\}$, *where the* l^{th} sequence s_l consists of M total number scene graphs. Each training batch includes N_b total number of input training sequences.

Training objective and the sequence mask modeling. We propose a sequential graph level masked modeling paradigm which is the central idea of FocusFormer. The object selection is based on ranking of the objects, where a ranking score $P_l^x \in R^C$ is the output of the FocusFormer, where C is the total number of classes present in the training data. P_l^x assigns each class of objects in the scene with the i^{th} entry denoting a score $p_{l,i}^x$. Superscript x is a place holder that may replace any scores as its name implies, i.e., x may represent S_a, S_b, S_p, S_f, S_n, and each p_l^x is computed given each sequence graph $s_l = \{G_1, \ldots, G_k, \ldots, G_M\}$, i.e., $P_l^x = g(z_l^x) = g(f(s_l)) \in R^C$. Here, function f_x represents the feature extractor out of the deep neural network, where x indicates which subnetwork (e.g., x = b: background, x = f: freshness, x = a: attraction) the feature is from; function g is the associated classifier of FocusFormer that maps the embedding out of each scene graph sequence into C number of classes. FocusFormer resorts to three subsets of ranking scores that are eventually combined to obtain the final ranking list as explained in main

paper. z_l^x denotes the feature respectively out of the subnetworks, where again x indicates which subnetwork the feature is from. We sometimes associate each score with addition subscript, e.g., $S_{c,a,l}$ indicates the attraction score of the c^{th} class in the l^{th} sequence of graphs. The definition of the subscripts become clear when mentioned in their contexts.

Attention

Attraction score. The first goal of FocusFormer is to transfer the sightseeing preference of sighted people during BLV's tour. An important assumption here is that sighted people would pay locally consistent attention to the highlights, therefore naturally leading to consistent scene graph structure within short local sessions when they stop to watch (focus). If there are no captured attraction objects (relatively consistent objects within a video segment), this implies a quick change of the scene, and the quick leaving of the tourist from the scene without a "focus". We hope to locate these attraction objects that have drawn attention from sighted people (during a relatively long and consecutive period of time), and we call them attraction objects. This assumption on attraction objects helps us to trace what objects are potentially of interests of sighted people. To find the "attraction" pattern from the crowed sourced training data, we introduce FocusFormer attention module as Fig. 2B illustrates. The supporting idea is as follows: the network is input with a sequentially temporally ordered input scene graphs (computed from a locally consecutive touring video) with their sequential frame orders, while the prediction objective is to reconstruct the target nodes present in the right next single scene graphs following the input sequence. We design the attention model such that each time only *a single frame* of scene graph sampled from sequence s_l associated with its position embedding is used to predict the nodes in the target scene graph. As training proceeds, only those nodes that are frequently shared within these sequences would return high reconstruction accuracy on predicting the target (in the following neighboring scene) nodes, because attraction modes are mostly probably shared within these locally consecutive frames. The joint use of graph embedding and position embedding helped to achieve the goal of FocusFormer in achieving attraction score in a position sensitive manner (attention applied only on each sole scene with itself in predicting the target scene).

During each batchwise training, the subnetwork of FocusFormer is input with a consecutive sequence of graph input $s_l = \{G_l, \ldots, G_{l+m}, \ldots, G_{l+M}\}$ which is of length M. For each batch of sequence, we have total number of N sequences. The objective function is defined to be:

$$L_a(\theta) = \frac{1}{N}\sum_{l=1}^{N}\sum_{k=1}^{C} 1\{k \in Y^{(l+M+1)}\} \log \frac{\exp\left[(\theta^{(c)})^T z_l^a\right]}{\sum_{j=1}^{C}\exp\left[(\theta^{(j)})^T z_l^a\right]}$$

Here, $\boldsymbol{\theta}^{(j)}$ is the learnable classifier parameter corresponding to the ground truth class of the j^{th} class. $k \in Y^{(l)}$ indicates that at least one of the nodes in the $(l+M+1)^{th}$ scene graph, i.e., the next following scene graph after the input sequence, belongs to the k^{th} class. Here, z_l^a denotes the feature of l^{th} sequence of graphs, and z_l^a is extracted from the attraction subnetwork.

After training for the optimal $\boldsymbol{\theta}^{(j)}$ for all classes via the above training loss, we obtain the eventual training attraction score of the c^{th} class by the following formula:

$$S_{c,a,l} = \widehat{P_{l,c}^{S_a}} = \frac{\exp\left[(\boldsymbol{\theta}^{(c)})^T z_l^a\right]}{\sum_{j=1}^{C}\exp\left[(\boldsymbol{\theta}^{(j)})^T z_l^a\right]} \text{ for class } c,$$

where the score associated with $S_{c,a,l}$ in the last training iteration is then assigned to the c^{th} class of object, as other batchwise training machine learning method does. Afterwards, the attraction score will be integrated with the background score for representing the final interest level of sighted people regarding the objectives, which will be introduced later. During inference time, we simply associate each object of class c in the new test scene with the converged attraction score learned from the training dataset.

Background score

A higher background score indicates that the object is likely distributed everywhere in the video of the tour, which does not provide additional information on touring interests, or on freshness. To predict the background score for each node n_i given a sequence l, we follow the intuition in masked modeling. Given the input sequence of s_l, the FocusFormer randomly masks out a graph input $G_{[mask]}$ through the attention model. The training objective is to reconstruct all the node classes $c \in [1, \ldots C]$ of the masked graph $G_{[mask]}$ with equal probability. This can be reformulated as a conventional classification cross entropy loss, where the "ground truth" classes are the node classes $c \in [1, \ldots C]$ of the masked graph $G_{[mask]}$. The intuition here is: if the network is input with any sampled input scene graphs (computed from a certain touring video) having shuffled frame orders, while the prediction objective is to reconstruct all of the node classes present in another randomly chosen scene graph from the same touring video, the most easily reconstructed object classes would be background objects. Those nodes would return high reconstruction accuracy on these, because background is mostly probably shared across any arbitrarily chosen frames in random order. This is equivalent to applying attention on each pair of randomly chosen scene graphs across the whole dataset, with the goal to reconstruct nodes in alternative scene graph. The loss function can be formulated as a special case of SoftMax cross entropy:

$$L_b(\theta) = \frac{1}{N}\sum_{l=1}^{N}\sum_{k=1}^{C} 1\{k \in Y^{(l+random)}\} \log \frac{\exp\left[(\boldsymbol{\theta}^{(c)})^T z_l^b\right]}{\sum_{j=1}^{C}\exp\left[(\boldsymbol{\theta}^{(j)})^T z_l^b\right]}$$

Here the z_l^b vector is the output learned embedding of each node in the scene graph after applying the background attention in the background subnetwork. The intuition is that the FocusFormer learns to reconstruct the backgrounds nodes with high probability in each arbitrary graph by simply observing any other random $(l + random)^{\{th\}}$ frames during the same tour. The reason is the extracted background should be agnostic of any specific scene of the segment but are shared across the video segments most likely. Predicting the background nodes with high probability will then help the network to achieve low training loss if the background node indeed is present in randomly chosen frame (in the form of scene graph). In contrast, unique novel nodes in the $(l + random)^{\{th\}}$ frame would be assigned relatively low scores during training, as the network cannot easily predict these unseen nodes given arbitrary frames from the tour. The Background score attention module then learns to extract background from the sequence through the specific sampling of scenes.

After obtaining the optimal $\boldsymbol{\theta}^{(j)}$ for all classes, we obtain the background score by the following formula:

$$S_{c,b,l} = \widehat{P_{l,c}^{S_b}} = \frac{\exp\left[(\boldsymbol{\theta}^{(c)})^T z_l^b\right]}{\sum_{j=1}^{C}\exp\left[(\boldsymbol{\theta}^{(j)})^T z_l^b\right]} \text{ for class } c.$$

where the score is then assigned to the objects classes in the scene following the l^{th} sequence. At inference time, the score associated with $S_{c,b,l}$ in the last training iteration is then assigned to the c^{th} class of object, as other batchwise training machine learning method does. During inference time, we simply associate each object in the new test scene with the background score obtained from the training dataset. Afterwards, the attraction score will be integrated with the background score for representing the final interest level of sighted people. The final attraction score will be reflecting the difference between $P_{l,c}^{S_a}$ and $P_{l,c}^{S_b}$, which means that the objects that sighted people concentrate on during specific time periods but are not uniformly present in

the background information are more worthy of attention, formally defined as:

$$S_{c,p,l} = S_{c,a,l} - \alpha \cdot S_{c,b,l} = \widehat{P_{l,c}^{S_a}} - \alpha \cdot \widehat{P_{l,c}^{S_b}} \text{ for class } c.$$

In the subsequent applications, this score $S_{c,p,l}$ is further whitened based on statistical observations over a period of time to have a zero mean and unit standard deviation. For simplicity, we have omitted the formula here.

Freshness score

For freshness detection, position embedding plays a key role in encoding the temporal ordering of the sequence. This sub-structure of the network aims to predict the objects in the following frame by exploiting the temporal causal structure between the frames. We rely on position embedding to encode the temporal ordering of the frames. The training objective is to predict the objects in the target scene following the input sequence. Objects less likely to be constructed during training will be determined as novel objects. For freshness objects feature extractor, FocusFormer use the attention model to sample a few frames from the video in the sequential order, and makes sure the attention function is applied across each pair of image embeddings constrained in such a short consecutive session under the proper encoding of position embedding. The position embedding makes sure the attention is only applied between each pair of scene graphs within such temporally ordered sequence, and that the feature extraction procedure must respects the sequence temporal order. FocusFormer learns semantically meaningful embeddings to extract freshness during the training. Similar to how background is computed, freshness score is computed as:

$$L_f(\theta) = \frac{1}{N}\sum_{l=1}^{N}\sum_{k=1}^{C} 1\left\{k \in Y^{(l+M+1)}\right\} \log \frac{\exp\left[\left(\boldsymbol{\theta}^{(c)}\right)^T z_l^f\right]}{\sum_{j=1}^{C}\exp\left[\left(\boldsymbol{\theta}^{(j)}\right)^T z_l^f\right]}.$$

Here, vector z_l^f is the learned embedding of each node in the scene graph after applying the freshness attention in the freshness subnetwork. After obtaining the optimal $\boldsymbol{\theta}^{(j)}$ for all classes, at the inference stage, for a new consecutive sequence of graph input $s_l = \{G_l, \ldots, G_{l+m}, \ldots, G_{l+M}\}$, we obtain the freshness score of the scene following the l^{th} test sequence according to formula:

$$S_{c,f,l} = \widehat{P_{l,c}^{S_f}} = -\frac{\exp\left[\left(\boldsymbol{\theta}^{(c)}\right)^T z_l^f\right]}{\sum_{j=1}^{C}\exp\left[\left(\boldsymbol{\theta}^{(j)}\right)^T z_l^f\right]} \text{for class } c.$$

Please note that the more unpredictable an object is, the more freshness it brings during the inference time. Therefore, we have associated a negative sign in the formula to account for this unpredictability. In the subsequent applications, this score $S_{c,f,l}$ is further whitened based on statistical observations over a period of time to have a zero mean and unit standard deviation. For simplicity, we have omitted the formula here.

Needs score. The needs score is computed based on how each class of object is needed by a BLV participant according to the questionary and survey presented in B.2. We calculate the scores by summing up the collected scores for each class, and normalize cross all the classes.

The final inference time score of a specific class c in the scene following sl sequence is computed as the weighted sum as in the main paper:

$$L_{c,l} = \lambda \cdot S_{c,p,l} + \beta \cdot S_{c,n,l} + \gamma \cdot S_{c,f,l},$$

where the unit Blind-in-the-loop learns the λ, β, γ through maximum likelihood estimation (MLE) techniques elaborated as follows.

BLV-in-the-Loop Adapter

VIPTour learns to adapt to the personal touring preference of BLV during the interaction with BLV. With the unit called "BLV-in-the-loop Adapter", the parameterization of the various needs of visually-impaired individuals recorded in VIPTour will iteratively update during the tour. These needs can be broken down into three categories: freshness, touring preference, and needs. Freshness refers to new things encountered during the visit, like a pavilion that comes into view. Preference refers to objects of interest to most people, such as lakes and shady areas. Basic needs are specific to the needs of visually-impaired people and may include considerations of safety, such as pedestrians and bicycles, or the availability of rest areas and garbage facilities. The log reader will record the actions of the blind so that the system remembers and processes the preference of the participant. The obtained data then are used to train the "Blind in the Loop" online so that we obtain the trade-off preference weighting parameters between prioritizing the "touring preference from the sighted people" against "the freshness of the object incoming into the new scene" and "other necessary needs owing to safety issues". BLV participant can choose to click "like", or "dislike" upon any recommended item in the view. The parameterization reweighting the three branches in the FocusFormer then will be updated by reading the "like" and "dislike" signals.

Through maximum likelihood method (MLE), BLV-in-the-Loop Adapter learns to adjust the weights between the three sources of recommendation, leading to updated tour guide plans displayed in the next capture frame. This may be viewed as the participant implicitly selecting important parameters that defined how the scene graph prioritizes in screening the nodes. We also make sure the obstacle avoidance warning is always available which reassure us the participant is informatic of the safety and needs relevant objects. The participant with impaired vision would interact with the display by exploring and zooming into the details in the constructed scene graph of the current view in the meanwhile.

VIPTour continuously reads the feedback from the BLV people and accordingly updates the parameterization of the recommendation. Specifically, for the $i^{\{th\}}$ object in the list, we define a weight vector $\omega^T = [\lambda, \beta, \gamma]^T \in R^3$, and score vector $s_i = \left[S_p^{(i)}, S_f^{(i)}, S_n^{(i)}\right]^T$. We define the feedback from BLV as a binary value $m_i^{H(cl)}$. Assuming the BLV user considers each individual recommendation item/object (e.g., tree) "liked" or "disliked" by either clicking feedbacks "like" or "dislike" upon observation of item i. The "like" corresponds to a binary label value $m_i^{H(cl)} = 1$, while "dislike" corresponds to label value $m_i^{H(cl)} = 0$. The log likelihood function of the user's feedback is modeled as:

$$\log p\left(m_i^{H(cl)} \middle| z, \omega\right) = \log \frac{\exp\left(\tau\omega^T s_i\right)^{m_i^{H(cl)}} \exp\left(\tau\omega^T \underline{s_i}\right)^{1-m_i^{H(cl)}}}{\exp\left(\tau\omega^T s_i\right) + \exp\left(\tau\omega^T \underline{s_i}\right)}$$

$$= m_i^{H(cl)} \log \frac{1}{1+\exp\left(\tau\omega^T(\underline{s_i} - s_i)\right)} + \left(1 - m_i^{H(cl)}\right) \log\left(1 - \frac{1}{1+\exp\left(\tau\omega^T(\underline{s_i} - s_i)\right)}\right).$$

One might recognize that equation above is essentially a logistic regression with learnable parameters $\omega \in R^3$ and input $\underline{s} - s_i$. Here, $\underline{s_i}$ is the mean value of s_i across all the objects in the training data. In other words, feedback $m_i^{H(cl)} = 1$ in favor of the current item (e.g., tree) is more likely to contribute to the parameterization of the "like" label, whereas the $m_i^{H(cl)} = 0$ will trigger the optimization given the "unlike" label. τ is the predefined hyperparameter (temperature of softmax).

Given this logistics regression formulation, we use MLE (maximum likelihood estimation technique) to learn vector ω by maximizing the $\log p(m_i^{H(cl)}, |, z, \omega)$. The blind in the loop unit is updated using gradient

decent w.r.t. the MLE:

$$\omega = \omega + \eta \frac{\partial \log p\left(m_i^{H(cl)} \middle| z, \omega\right)}{\partial \omega},$$

where η is the learning rate.

By observing the returned statistics "like" and "dislike" from the BLV feedback, VIPTour offers interpretable outputs via the change of $\omega^T = [\lambda, \beta, \gamma]^T$. Here the λ, β, γ values respectively showcase the rising importance of tour preference learned from sighted people (λ), interests in novel objects (β) and needs (γ) for each particular BLV tourist.

Hierarchical layout

We have designed a dual-layered scene structure that incorporates recommendations from the FocusFormer and scene mapping to facilitate a reasonable and simplified information hierarchy. The smartphone with touchscreen will correspondingly reflect the scene graph structure of the nodes conveyed in the audio play. The participant can select to zoom into the scene graph according to the learned hierarchy so that she can better explore the relationship between the nodes reflected in the scene. The edges in the scene graph would correspond to the direction and position information of the objects.

The clustering of recommended objects was achieved by the Louvain algorithm and the layout was rendered by NetworkX 3.1[41]. Each object was assigned to a node with its attributes and ranking score. The relations between objects and their estimated distance were represented by edges and weights. Each community was grouped by the Louvain algorithm. The node with the highest score appeared on the top layer. Others were on the second layer, which can be accessed through the father node on the top layer. The initial position of each node was an egocentric bird's eye view of the user's surroundings. However, the possibility of node overlapping was high enough to affect the user's tactile experience. Therefore, the Fruchterman-Reingold force-directed algorithm contributed to the spring layout to render both layers of node graphs on users' phones, which avoided overlapping and preserved the approximate position of the object.

Impact of upstream AI/ML techniques on FocusFormer performance

FocusFormer relies on several upstream AI/ML techniques, including object detection and semantic graph generation (SGG), as inputs to its processing pipeline. As such, the performance of these upstream techniques inevitably influences the overall effectiveness of FocusFormer. In this work, we adopted VinVL[42] for object detection and Graph R-CNN[30] for semantic graph generation, which have been widely validated in prior research. Using these techniques, we successfully demonstrated the effectiveness of FocusFormer in achieving its intended goals. However, the impact of other potential upstream techniques on the system remains an open question. While it is theoretically expected that more accurate detection and SGG methods would further enhance the performance of FocusFormer, comprehensive empirical evidence is needed to quantify this relationship. Investigating the sensitivity of FocusFormer to different upstream techniques and their performance levels is an important direction for future work, which will provide deeper insights into the system's robustness and scalability.

Data availability

Data is provided within the manuscript or supplementary information files.

Received: 9 December 2024; Accepted: 14 April 2025;
Published online: 04 June 2025

References

1. Bell, S. L., Phoenix, C., Lovell, R. & Wheeler, B. W. Seeking everyday wellbeing: The coast as a therapeutic landscape. *Soc. Sci. Med.* **142**, 56–67 (2015).
2. Bandukda, M., Holloway, C., Singh, A. & Berthouze, N. N. PLACES: a framework for supporting blind and partially sighted people in outdoor leisure activities. In *Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility*. 1–13 (2020).
3. Virgili, G. & Rubin, G. G. Orientation and mobility training for adults with low vision. *Cochr. Database Syst. Rev.* CD003925 (2010).
4. Bandukda, M., Singh, A., Berthouze, N. & Holloway, C. Understanding experiences of blind individuals in outdoor nature. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. (2019).
5. Siu, K. W. M. Accessible park environments and facilities for the visually impaired. *Facilities* **31**, 590–609 (2013).
6. Guerreiro, J. J. et al. Cabot: Designing and evaluating an autonomous navigation robot for blind people. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility*. 68–82 (2019).
7. Avila, M., Funk, M. & Henze, N. Dronenavigator: Using drones for navigating visually impaired persons. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility*. 327–328 (2015).
8. Bai, J., Lian, S., Liu, Z., Wang, K. & Liu, D. Smart guiding glasses for visually impaired people in indoor environment. *IEEE Trans. Consum. Electron.* **63**, 258–266 (2017).
9. Chuang, T. K. et al. Deep trail-following robotic guide dog in pedestrian environments for people who are blind and visually impaired-learning from virtual and real worlds. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*. 5849–5855 (2018).
10. Dastider, A., Basak, B., Safayatullah, M., Shahnaz, C. & Fattah, S. A. Cost efficient autonomous navigation system (e-cane) for visually impaired human beings. In *2017 IEEE region 10 humanitarian technology conference (R10-HTC).* 650–653 (2017).
11. Lee, Y. H. & Medioni, G. Wearable RGBD indoor navigation system for the blind. In *European Conference on Computer Vision*. 493–508 (2014).
12. Wang, L., Zhao, J. & Zhang, L. Navdog: robotic navigation guide dog via model predictive control and human-robot modeling. In *Proceedings of the 36th Annual ACM Symposium on Applied Computing*. 815–818 (2021).
13. Chang, R. C. et al. OmniScribe: Authoring Immersive Audio Descriptions for 360 Videos. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–14 (2022).
14. Brewer, R. N. & Kameswaran, V. Understanding the power of control in autonomous vehicles for people with vision impairment. In *Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*. 185–197 (2018).
15. Brinkley, J. et al. Exploring the needs, preferences, and concerns of persons with visual impairments regarding autonomous vehicles. *ACM Trans. Accessible Comput. (TACCESS)* **13**, 1–34 (2020).
16. Zhang, Y. et al. "I am the follower, also the boss": Exploring Different Levels of Autonomy and Machine Forms of Guiding Robots for the Visually Impaired. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–22 (2023).
17. Asakawa, S., Guerreiro, J., Ahmetovic, D., Kitani, K. M. & Asakawa, C. The present and future of museum accessibility for people with visual impairments. In Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility. 382–384. (2018).
18. Kokjer, K. J. The information capacity of the human fingertip. *IEEE Trans. Syst., Man, Cybern.* **17**, 100–102 (1987).
19. Lee, J., Peng, Y. H., Herskovitz, J. & Guo, A. Image Explorer: Multi-Layered Touch Exploration to Make Images Accessible. In *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility*. 1-4 (2021).

20. Ahmetovic, D., Kwon, N., Oh, U., Bernareggi, C. & Mascetti, S. Touch screen exploration of visual artwork for blind people. In *Proceedings of the Web Conference 2021*. 2781–2791 (2021).
21. He, K., Fan, H., Wu, Y., Xie, S. & Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9729–9738 (2020).
22. Caron, M. et al. Unsupervised learning of visual features by contrasting cluster assignments. *Adv. Neural Inf. Process. Syst.* **33**, 9912–9924 (2020).
23. Grill, J. B. et al. Bootstrap your own latent - a new approach to self-supervised learning. *Adv. Neural Inf. Process. Syst.* **33**, 21271–21284 (2020).
24. Zbontar, J., Jing, L., Misra, I., LeCun, Y. & Deny, S. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*. 12310–12320 (2021).
25. Carlucci, F. M., D'Innocente, A., Bucci S., Caputo, B. & Tommasi, T. (2019). Domain generalization by solving jigsaw puzzles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2229–2238 (2019).
26. Zellers, R., Yatskar, M., Thomson, S. & Choi, Y. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5831–5840 (2019).
27. Han, K., Wang, Y., Guo, J., Tang, Y. & Wu, E. Vision gnn: An image is worth graph of nodes. *Adv. Neural Inf. Process. Syst.* **35**, 8291–8303 (2022).
28. Xu, D., Zhu, Y., Choy, C. B. & Fei-Fei, L. Scene graph generation by iterative message passing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5410–5419 (2017).
29. Dong, X. et al. Stacked hybrid-attention and group collaborative learning for unbiased scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19427–19436 (2022).
30. Yang, J., Lu, J., Lee, S., Batra, D. & Parikh, D. Graph r-cnn for scene graph generation. In *Proceedings of the European conference on computer vision (ECCV)*. 670–685 (2018).
31. Maaten Van der, L. & Hinton,G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
32. Devlin, J., Chang, M. W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805, (2018).
33. He, X. et al. Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web*. 173–182 (2017).
34. Brooke, J. SUS-A quick and dirty usability scale. *Usability Evaluation Ind.* **189**, 4–7 (1996).
35. Slade, P., Tambe, A. & Kochenderfer, M. J. Multimodal sensing and intuitive steering assistance improve navigation and mobility for people with impaired vision. *Sci. Robot.* **6**, eabg6594 (2021).
36. Tobita, K., Sagayama, K., Mori, M. & Tabuchi, A. Structure and examination of the guidance robot LIGHBOT for visually impaired and elderly people. *J. Robot. Mechatron.* **30**, 86–92 (2018).
37. Kim, J. H., Ritchie, J. B. & McCormick, B. Development of a scale to measure memorable tourism experiences. *J. Travel Res.* **51**, 12–25 (2012).
38. Li, J. et al. Measuring and understanding photo sharing experiences in social virtual reality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14 (2019).
39. Alter, A. L. & Oppenheimer, D. M. Uniting the tribes of fluency to form a metacognitive nation. *Personal. Soc. Psychol. Rev.* **13**, 219–235 (2009).
40. Madan, C. R. & Singhal, A. Encoding the world around us: Motor-related processing influences verbal memory. *Conscious. Cognition* **21**, 1563–1570 (2012).
41. Hagberg, A., Schult, D. & Swart, P. Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conference*. 11–15 (2008).
42. K. et al. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5579–5588 (2021).

Acknowledgements

This work was supported in part by the Beijing Outstanding Young Scientist Program; in part by funding from XIAOMI FOUNDATION; in part by the Ministry of Science and Technology of China under contract no. 2024YFB2809103; and in part by National Natural Science Foundation Youth Fund 62202267.

Author contributions

Haozhe Lin, Ruqi Huang and Guyue Zhou initiated the project. Haozhe Lin, Jiangtao Gong, Ruqi Huang and Guyue Zhou conceived the original idea. Haozhe Lin, Yu Wang, Jinsong Zhang, Bing Bai, Chunyu Wei, Kun Li and Ruqi Huang proposed and implemented the AI algorithm. Jiangtao Gong, Yan Zhang, Luyao Wang, Yancheng Cao and Guyue Zhou designed and conducted BLV experiments. Haozhe Lin, Jiangtao Gong and Yu Wang analyzed the results and prepared the manuscripts.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s44387-025-00006-w.

Correspondence and requests for materials should be addressed to Ruqi Huang or Guyue Zhou.