Thesis: Towards Open-world Accessibility

Nahyun Kwon

October 2024

Figure 1: Open-world accessibility

1 Introduction

Recognizing accessibility challenges in real-world environments is essential for creating inclusive and adaptive technologies. Accessibility is complex because it involves understanding both physical and contextual barriers, which vary widely across different scenarios. These contextual barriers include situational and temporary conditions, such as carrying objects, fatigue, or having a child. While object detection and recognition are well-researched in computer vision, detecting human interactions, which are more abstract and dynamic, remains highly challenging. Addressing these accessibility issues requires systems that can recognize objects and interpret human interactions. This task becomes even more difficult in unpredictable, real-world settings.

Open-world accessibility (OWA) is to develop AI systems that can adapt to diverse, unseen accessibility needs and contexts without relying on pre-defined categories or fixed vocabularies. Traditionally, automated accessibility solutions rely on fixed vocabularies (e.g., specific classes of objects), predefined taxonomies (e.g., function-based or improvement-based categories), and set target groups (e.g., individuals with permanent disabilities). This traditional approach, which we call "closed-world accessibility", assumes that all accessibility needs are already known, limiting solutions to a fixed set of categories and scenarios. These systems are typically effective for specific problem spaces, such as indoor personal spaces [7, 1, 5] and public outdoor areas [6]. While effective in such targeted problems, closed-world solutions struggle to address contextual disabilities, because these conditions are highly variable and unpredictable, making it difficult to predefine all possible needs or situations.

OWA emphasizes adaptability, enabling AI systems to adjust to diverse user inputs (images) by leveraging open vocabulary that is capable of handling unfamiliar objects and contexts not encountered during training. This adaptability should support personalized, context-aware analysis that evolves with individual user needs. OWA aims to offer scalable solutions that can continuously adapt to everyday life's ever-changing environments and conditions.

Computer Vision has explored similar open-world problems in many ways. For example, zero-shot recognition refers to the ability of models to correctly classify objects or actions they have never encountered before by leveraging semantic information or knowledge transfer. Few-shot recognition enables models to learn from a limited number of examples, allowing them to adapt to new categories with minimal supervision. These techniques are used to build systems that can generalize beyond pre-defined datasets.

The system must dynamically recognize new situations and adapt without requiring extensive retraining. My research sets a new benchmark to extend closed-world accessibility solutions into open-world scenarios, enabling them to operate in dynamic and unfamiliar conditions with minimal manual intervention.

[Nahyun: specify contributions here 1. framework 2. transferring closed world to open-world (method) 3. long-term research vision]

2 Recognizing Accessibility Issues in Open-world Settings

2.1 Objective

The primary objective is to develop AI-powered tools that detect contextual disabilities in open-world environments. To achieve this, we leverage foundation models to build an AI system that analyzes user-uploaded photos and automatically detects inaccessible items based on the user's specific contextual disabilities.

2.2 Method

[Nahyun: How does each method can contribute this problem? and what's my unique contribution over just using existing methods?] To tackle open-world accessibility, I propose an automated pipeline utilizing deep learning techniques to extract meaningful image features and contextualize foundation models. Here I present possible key components that can contribute to the pipeline.

- **Feature Encoder:** I apply a convolutional neural network (CNN) architecture to process user-uploaded images, generating embeddings that capture significant contextual clues relevant to accessibility.
- **Contrastive Learning:** Self-supervised contrastive learning is applied to enable the encoder to distinguish between relevant and irrelevant contexts, thereby enhancing the model's ability to interpret varying accessibility challenges of users with different contexts presented in the images.
- **Test-Time Adaptation:** We employ test-time adaptation techniques to dynamically adjust the model's responses to the specific contextual needs of users, enhancing its ability to recognize and respond to new accessibility challenges.

2.2.1 Feature Encoders

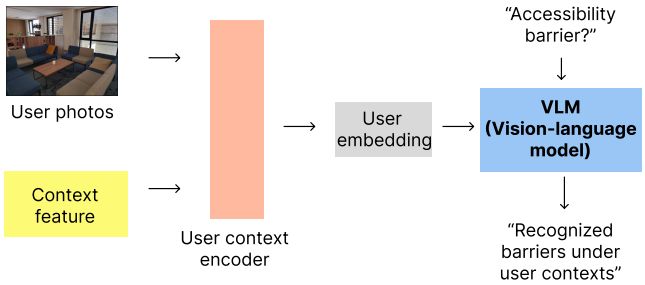

Figure 2: Applying feature encoder in OWA

The approach from the paper USER-LLM [4] can contribute to the thesis by incorporating user-specific embeddings, enabling the system to adapt dynamically based on unique user contexts. Specifically, this approach can support the goals of open-world accessibility as follows:

- **User-Specific Embedding for Contextual Adaptation**: By generating user embeddings based on individual interactions, USER-LLM enables the model to capture user-specific behaviors, preferences, and interactions. This customization aligns with the project's objective of dynamically adapting accessibility recommendations for users, as the model learns from each user's unique contexts in real-time.

- **Enhanced Computational Efficiency**: USER-LLM's embedding-based approach significantly reduces the computational load by representing user history as embeddings instead of relying on long, sequential text prompts. This allows the model to process extensive user interaction histories efficiently, making it more feasible to deploy real-world accessibility tools that can analyze user behaviors over time.
- **Cross-Attention Mechanism for Embedding Integration**: USER-LLM integrates user embeddings with language models through a cross-attention mechanism, which improves the system's ability to draw context-specific information. In the context of accessibility, this approach can help the model better recognize relevant accessibility barriers in user-specific scenarios by effectively focusing on relevant aspects of the user's behavior and preferences.
- **Real-Time User Feedback and Model Adaptation**: USER-LLM allows for a feedback loop where user responses help refine the embeddings, supporting ongoing model adaptation. This aligns with the thesis's goal of evaluating model outputs with real participants who will assess the model's performance and provide feedback, thereby enabling iterative improvements based on actual user needs.

Integrating USER-LLM's approach enhances the system's ability to provide adaptive, personalized accessibility solutions in an open-world setting, leveraging the efficiency and personalization capabilities of user embeddings.

2.2.2 Retrieval-Augmented Customization

[2] introduces a retrieval-augmented customization (REACT) method that enhances pre-trained visual models by leveraging external image-text pairs relevant to the target domain[1]. This approach can contribute to open-world accessibility in the following ways:

- **Domain-Specific Adaptation**: The REACT approach provides a method to augment pre-trained models with external image-text pairs, focusing on specific target domains. In this thesis, REACT can be utilized to gather accessibility-relevant knowledge from online sources, targeting diverse daily scenarios (e.g., cooking, commuting), thereby enhancing the model's contextual understanding of situations where accessibility barriers may occur.
- **Handling Limited Data through Retrieval-Augmentation**: Since it is challenging to collect extensive labeled data for every accessibility scenario, REACT's ability to retrieve and use web-based data without extensive manual labeling aligns with the goals of open-world accessibility. This allows the model to adapt to new, unseen contexts, such as identifying accessibility challenges in varied environments, by leveraging relevant image-text pairs.
- **Customizable and Efficient Model Update**: REACT introduces modular learning blocks, which enable efficient adaptation by training only essential parameters and freezing the rest of the model. This customization aligns with the thesis's goal of achieving user-specific adaptability, allowing the model to efficiently adapt to each participant's unique environment without re-training the entire model.
- **Evaluation of Personalized Accessibility**: By training with REACT's retrieval-augmented data, the model can produce outputs specific to each user scenario, focusing on identifying accessibility barriers relevant to individuals. These results can then be evaluated by participants to test how well the system identifies relevant accessibility issues, validating the model's applicability in real-world settings.

Integrating the REACT approach in this thesis supports the development of adaptive, personalized visual models for accessibility in dynamic environments, enabling efficient customization without the need for exhaustive annotation. This aligns with the thesis's objective to create an open-world accessibility framework.

[1] In this thesis, 'domain' refers to the specific environments, activities, and contexts that comprise users' daily interactions, such as home, campus, or work environments, as well as various tasks like cooking, commuting, or attending meetings. Each domain has distinct characteristics, challenges, and potential accessibility barriers unique to its setting. This contextual understanding allows the model to adapt its approach based on the varying demands of each environment, ensuring that accessibility solutions are both contextually relevant and responsive to the diverse, open-world scenarios encountered by users.

2.2.3 Test-time Adaptation

The SwapPrompt [3] approach provides a dynamic framework for adapting prompts to new, unseen scenarios, which aligns well with the open-world accessibility challenges of this research. Specifically, this approach contributes to the thesis in the following ways:

- **Test-Time Adaptation for Real-World Data**: As this project involves evaluating raw, unsegmented real-world videos from participants, SwapPrompt's unsupervised test-time prompt adaptation method allows the model to handle distribution shifts in user-generated content without requiring labeled data. This feature supports the open-world accessibility requirement by enhancing adaptability in diverse, unpredictable environments.
- **Handling Diverse Contexts**: SwapPrompt's self-supervised contrastive learning and dual-prompt setup (online and target prompts) enable flexible adaptation across different contexts. In this project, this translates to the model dynamically adjusting to various real-world scenarios (e.g., washing hands, cooking, attending meetings) and identifying context-specific accessibility barriers.
- **Scene Understanding and Contextual Segmentation**: The swapped prediction mechanism in SwapPrompt leverages multiple augmentations of an image to improve the model's predictions. This can enhance the model's ability to understand scenes and perform contextual segmentation without requiring user-provided labels, allowing the system to autonomously identify which segments in participants' videos contain accessibility barriers relevant to individual needs.
- **User-Specific Adaptability**: SwapPrompt's exponential moving average (EMA) of prompts helps maintain historical knowledge, supporting consistent personalization by allowing the model to learn from diverse contexts without forgetting previously learned information. This is essential for delivering relevant and acceptable results to participants, as they evaluate the model's outputs based on their specific needs.

Integrating SwapPrompt's approach enhances the model's adaptability, making it well-suited for open-world scenarios characterized by limited labeled data and significant variation in contexts across individuals.

3 Evaluation

To evaluate the performance of my Open-World Accessibility (OWA) system, I will establish a benchmark that measures the system's effectiveness in detecting accessibility challenges and interpreting dynamic contexts from user-uploaded images. The evaluation will focus on the system's ability to handle diverse and unseen objects and situations, relying on both detection and natural language metrics.

3.1 Benchmark Creation

To evaluate the Open-World Accessibility (OWA) system, I will establish a benchmark using labeled video frames from datasets such as **EpicKitchen** and **Ego4D**. Accessibility experts will manually label data points in controlled experiments, focusing on key factors like scene similarity, selected frames with important actions, and varying environmental conditions (e.g., lighting, occlusion). This controlled labeling will accurately evaluate the system's capabilities under specific conditions.

Manual labeling becomes less scalable for real-world evaluations, where variability and unpredictability are more prominent. Instead, the system will be tested on its ability to generalize to unseen objects and contexts with minimal reliance on new labels. To maintain scalability and flexibility, I will use a semi-structured, open-vocabulary taxonomy during labeling, emphasizing broader and abstract categories that allow the system to learn from both labeled and unlabeled data. Techniques such as contrastive learning, self-supervised learning, and test-time adaptation will enable the system to handle diverse, real-world scenarios without the need for extensive manual labeling.

If additional labeling is necessary for complex or ambiguous cases, I will consider selective expert input to ensure accuracy while minimizing costs. This benchmark will provide a balanced approach to evaluating both controlled and real-world performance, ensuring that the system is capable of adapting to dynamic, open-world environments.

3.2 Metrics

I will use a combination of detection metrics and natural language metrics to comprehensively evaluate the system's performance. These are example metrics that can include:

- **Mean Average Precision (mAP):** This detection metric will be used to measure the accuracy of object and interaction detection in the video frames. mAP evaluates the precision-recall tradeoff, making it an ideal metric for assessing the system's ability to identify relevant objects and actions across various contexts.
- **BLEU (Bilingual Evaluation Understudy):** To measure the quality and reliability of the system's natural language outputs, such as context-aware descriptions or accessibility-related insights, I will use BLEU, a standard metric for evaluating the accuracy and fluency of machine-generated text. This metric will ensure that the system's explanations and contextual responses are coherent and align with user needs.
- **Additional Question-Answering Metrics:** I will explore other NLP metrics such as **ROUGE** (Recall-Oriented Understudy for Gisting Evaluation) and **METEOR** (Metric for Evaluation of Translation with Explicit ORdering) to assess the completeness and relevance of the system's responses given specific questions, especially when addressing context-specific accessibility questions (e.g., how this object can be inaccessible when the users have limb injury?).
- **Test-Time Adaptation Performance:** I will also evaluate the effectiveness of test-time adaptation techniques by measuring how well the system adjusts to new and unfamiliar contexts during inference. This will involve comparing the system's performance before and after adaptation, focusing on improvements in detection accuracy and the quality of generated outputs.

3.3 Evaluation Process

The evaluation will proceed in two phases:

- **Phase 1 - Controlled Experiments:** I will first test the system using controlled video frames from the labeled benchmark to ensure it can handle a range of accessibility challenges and contexts. These experiments will evaluate the system's base performance using mAP and natural language metrics.
- **Phase 2 - Real-World Scenarios:** Next, I will test the system in dynamic, real-world environments, analyzing how well it adapts to unseen objects and contexts through test-time adaptation techniques. The results from this phase will highlight the system's ability to generalize and its real-time adaptability to unpredictable accessibility needs.

3.3.1 Controlled Testing

'Controlled' refers to carefully curated and sampled data points from the benchmark dataset, which are selected based on specific criteria. These criteria include:

- **Scene Similarity:** Ensuring that frames or data points are selected from similar scenes with the training data to test the system's consistency in detecting accessibility challenges under controlled environmental factors.
- **Key Actions:** Selecting frames that contain key actions or interactions that are critical for assessing the system's ability to interpret dynamic human behaviors and interactions, such as lifting objects, opening doors, or navigating obstacles.

These controlled experiments will allow for a more precise evaluation of the system's core functionalities by isolating key variables and systematically testing them.

3.3.2 Real-world Testing

In this stage, real-world scenario testing will be conducted using video data from 10-12 participants, capturing diverse daily life activities. These video clips will vary in length and encompass various scenarios, such as washing hands, cooking, going to campus, and attending group meetings. Each participant's data will provide multiple contexts without prior segmentation or labeling, meaning the data will serve as raw input, offering an authentic slice of daily life.

Given this unstructured format, the system will process these inputs, handling scene understanding and segmentation, and identifying potential accessibility barriers for each individual. The model's outputs will then be returned to participants, who will assess the relevance and acceptability of the identified accessibility barriers in their unique contexts. This feedback will be crucial in evaluating the model's real-world applicability and effectiveness.

The combination of a robust benchmark, multiple evaluation metrics, and a two-phase evaluation process will ensure that the OWA system is rigorously tested and capable of addressing real-world accessibility challenges in a scalable, open-world setting.

4 Milestones

I have set the following milestones, which include exploring additional methods to enhance the model's adaptability:

Task Name	Q4 2024	Q1 2025	Q2 2025	Q3 2025	Q4 2025
Literature Review & Initial Model Design					
Development of Initial Model					
Evaluation with Ego4D Dataset					
Submission to UIST 2025					
Model Improvements Based on Initial Results					
Real-world Data Collection & Evaluation					
Submission to CHI 2025					
Defense Writing					

Table 1: Thesis Project Timeline (Quarterly)

- **Literature Review & Initial Model Design (Q4 2024)**: The project begins with an in-depth literature review to build a strong foundation for the proposed model. This stage includes designing the initial version of the model based on the findings.
- **Development of Initial Model (Q4 2024 - Q1 2025)**: Following the initial design, the model's development will continue through the first quarter of 2025, incorporating key findings and preparing for evaluation.
- **Evaluation with Ego4D Dataset (Q1 2025)**: The initial model will undergo its first evaluation phase using the Ego4D dataset, providing valuable insights into its performance and areas for improvement.
- **Submission to UIST 2025 (Q2 2025)**: By the second quarter of 2025, the project will be ready for a formal paper submission to UIST 2025, showcasing the initial model and results from the evaluation.
- **Model Improvements Based on Initial Results (Q2 2025)**: Model improvements will be made in response to initial evaluation results, refining the approach based on feedback and analysis.
- **Real-world Data Collection & Evaluation (Q2 2025)**: This stage will focus on gathering real-world data to test the model in practical settings, enabling further evaluation and refinement.
- **Submission to CHI 2025 (Q3 2025)**: In the third quarter, another paper will be submitted to CHI 2025, presenting the model's advancements and real-world evaluation results.

- **Defense Writing (Q4 2025)**: The final quarter of 2025 is reserved for drafting the thesis defense, consolidating all project findings, and preparing for the defense presentation.

References

[1] Nahyun Kwon, Qian Lu, Muhammad Hasham Qazi, Joanne Liu, Changhoon Oh, Shu Kong, and Jeeeun Kim. Accesslens: Auto-detecting inaccessibility of everyday objects. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2024.

[2] Haotian Liu, Kilho Son, Jianwei Yang, Ce Liu, Jianfeng Gao, Yong Jae Lee, and Chunyuan Li. Learning customized visual models with retrieval-augmented knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15148–15158, 2023.

[3] Xiaosong Ma, Jie Zhang, Song Guo, and Wenchao Xu. Swapprompt: Test-time prompt adaptation for vision-language models. *Advances in Neural Information Processing Systems*, 36, 2024.

[4] Lin Ning, Luyang Liu, Jiaxing Wu, Neo Wu, Devora Berlowitz, Sushant Prakash, Bradley Green, Shawn O'Banion, and Jun Xie. User-llm: Efficient llm contextualization with user embeddings. *arXiv preprint arXiv:2402.13598*, 2024.

[5] American Association of Retired Persons (AARP). Homefit ar. `https://apps.apple.com/us/app/homefit-ar/id1513619492?platform=iphone`, 2020. Accessed: 4/3/2023.

[6] Manaswi Saha, Michael Saugstad, Hanuma Teja Maddali, Aileen Zeng, Ryan Holland, Steven Bower, Aditya Dash, Sage Chen, Anthony Li, Kotaro Hara, et al. Project sidewalk: A web-based crowdsourcing tool for collecting sidewalk accessibility data at scale. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2019.

[7] Xia Su, Han Zhang, Kaiming Cheng, Jaewook Lee, Qiaochu Liu, Wyatt Olson, and Jon E Froehlich. Rassar: Room accessibility and safety scanning in augmented reality. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2024.