Towards Scalable Web Accessibility Audit with MLLMs as Copilots

Ming Gu[1,2], **Ziwei Wang**[1,2], **Sicen Lai**[1,3], **Zirui Gao**[1,2], **Sheng Zhou**[1,3*], **Jiajun Bu**[1,2]

[1]Zhejiang Key Laboratory of Accessible Perception and Intelligent Systems, Zhejiang University
[2]College of Computer Science and Technology, Zhejiang University
[3]School of Software Technology, Zhejiang University
{gmwork, wangziwei98, laisicen, gaozirui.zju, zhousheng_zju, bjj}@zju.edu.cn

Abstract

Ensuring web accessibility is crucial for advancing social welfare, justice, and equality in digital spaces, yet the vast majority of website user interfaces remain non-compliant, due in part to the resource-intensive and unscalable nature of current auditing practices. While WCAG-EM offers a structured methodology for site-wise conformance evaluation, it involves great human efforts and lacks practical support for execution at scale. In this work, we present an auditing framework, AAA, which operationalizes WCAG-EM through a human-AI partnership model. AAA is anchored by two key innovations: GRASP, a graph-based multimodal sampling method that ensures representative page coverage via learned embeddings of visual, textual, and relational cues; and MaC, a multimodal large language model-based copilot strategy that supports auditors through cross-modal reasoning and intelligent assistance in high-effort tasks. Together, these components enable scalable, end-to-end web accessibility auditing, empowering human auditors with AI-enhanced assistance for real-world impact. We further contribute four novel datasets designed for benchmarking core stages of the audit pipeline. Extensive experiments demonstrate the effectiveness of our methods, providing insights that small-scale language models can serve as capable experts when fine-tuned.

Code & Datasets — https://github.com/eaglelab-zju/AAA
Extended Version — https://arxiv.org/pdf/2511.03471

1 Introduction

Web accessibility is a foundational principle in the pursuit of an inclusive digital environment, ensuring that all users including those with disabilities can perceive, navigate, and interact with online content (Web Accessibility Initiative (WAI) 2024; Sharif et al. 2022). Despite the widespread adoption of standards such as the Web Content Accessibility Guidelines (WCAG) (Consortium 2024b), and substantial efforts devoted to web accessibility evaluation, the state of web accessibility remains alarmingly poor. A recent study reported that 94.8% of homepages across one million websites contained accessibility violations (WebAIM 2025). Emerging research suggests that this stagnation stems not from a lack of education or tooling, but from the intrinsic complexity of web accessibility as a resource management problem (Abramovich and Patitsas 2024; Elglaly et al. 2024). This means that *the time-consuming and labor-intensive nature of Web Accessibility Audits (WAA) increasingly misaligned with the growing scale and maintenance cost of modern websites* (SolarWinds Worldwide 2025).

To address WAA, the World Wide Web Consortium (W3C) introduced the Website Accessibility Conformance Evaluation Methodology (WCAG-EM) (Group 2014), a five-step protocol designed to standardize evaluation procedures. However, it lacks a corresponding technical framework that supports scalable execution in practice. In this context, **scalability** refers to two critical capabilities: (1) accelerating audit processes via automation, and (2) minimizing unavoidable manual effort through intelligent human–AI collaboration. Yet, most existing tools operate only at the page or element level, *covering only fragments of the WCAG-EM pipeline* (Huang et al. 2024). This narrow scope hinders scalability with bottlenecks in both time and labor.

To overcome the limitations, we propose a comprehensive framework anchored in three pillars: Automation, AI, and Auditor (AAA). *AAA operationalizes five procedures aligned with WCAG-EM's five steps*, including web crawling, automated checks, page sampling, manual evaluation, and reporting/remediation, with the goal of enabling scalability across the full audit lifecycle. Despite advances in automating tasks such as crawling and hard-coded checks, *two fundamental challenges remain*. **First**, existing page sampling methods fail to satisfy WCAG-EM's representativeness requirements. Recent clustering-based approaches rely primarily on textual similarity (Hambley et al. 2023), overlooking the rich multimodal semantics of web pages including visual layout, textual content and hyperlink relationships, which are essential for capturing diversity and representativeness. **Second**, intelligent assistance for manual auditing tasks remains underexplored. Current methods offer minimal assistance in high-effort tasks such as identifying accessibility-critical components, which often demand sophisticated multimodal reasoning, making them particularly burdensome for human auditors to collect.

To tackle these challenges, we first introduce Graph-based Representative PAge Clustering for SamPling (GRASP), a novel multimodal approach that generates WCAG-EM-

*Corresponding Author

compliant representative page subsets. GRASP defines representativeness across three complementary dimensions: textual semantic, visual layout, and linkage relationships, and employs graph neural networks (GNNs) to learn a unified embedding space for representative clustering. A dedicated structure learning module further improves sampling quality by mutually enhancing representativeness and clustering. In parallel, we explore the emerging potential of multimodal large language models (MLLMs) in accessibility workflows (Ara and Sik-Lanyi 2025). We present MLLMs as Copilot Assistant, Auditor and Consultant (MaC), a holistic AI companion designed to support multiple stages of WAA. By enabling cross-modal reasoning, MaC assists in identifying audit-critical elements and pages, thereby accelerating both sampling and manual evaluation. Furthermore, it broadens audit coverage by facilitating the evaluation of underrepresented accessibility issues, particularly those affecting less commonly addressed disabilities like cognitive impairment. To support future research, we also release four new datasets tailored for distinct stages of the WAA pipeline. These datasets address the current lack of accessibility-specific benchmarks. Our contributions are summarized as:

- **Scalable WAA Framework**: We propose a full-lifecycle audit framework AAA aligned with WCAG-EM, advancing scalability across web accessibility audit lifecycle.
- **Multimodal Sampling Method**: We introduce GRASP, a novel graph-based multimodal page sampling technique satisfying WCAG-EM representativeness criteria.
- **MLLM as Copilot Strategy**: We introduce MAC, a versatile MLLM-powered strategy augmenting multiple labor-intensive procedures via multimodal reasoning.
- **Benchmark Datasets**: We release four new datasets tailored to different stages of the AAA pipeline, facilitating comprehensive evaluation and comparison.
- **Empirical Insights**: Through extensive experiments, we demonstrate the effectiveness of our methods and uncover the potential of small MLLMs as domain experts.

2 Related Work

Web Accessibility Audit (WAA). Traditional tools like WAVE (WebAIM, Utah State University 2025) and Axe (Deque Systems, Inc 2025) rely on hard-coded checks that detect syntactic issues (e.g., missing alt text), but often missing contextual and semantic aspects (López-Gil and Pereira 2024; Ara and Sik-Lanyi 2025). Recent advances leverage large language models (LLMs) to provide intelligent evaluation and repair, aiming to reduce manual effort (Huang et al. 2024; Othman, Dhouib, and Nasser Al Jabor 2023). Nevertheless, accessibility gaps persist: a 2025 audit of one million websites revealed WCAG violations on 94.8% of home pages (WebAIM 2025). Studies suggest this stems less from tool limitations and more from resource constraints (Abramovich and Patitsas 2024). Auditing remains time- and labor-intensive, especially as website scale increases nowadays (SolarWinds Worldwide 2025). *Existing approaches predominantly focus on individual elements or pages, lacking a framework to address labor and resource challenges across the full site-wise audit lifecycle.*

Page Sampling for WAA. Full-site evaluation is infeasible for large sites, making page sampling essential for producing representative results. WCAG-EM outlines two dimensions: (1) *the individual level*, which targets pages with critical accessibility relevance (e.g., essential functionality, accessibility statements, or home-linked common pages), and (2) *the collective level*, which ensures diversity and representativeness across the site. However, existing methods often address narrow aspects, such as URL patterns (Zhang et al. 2015b), Web Accessibility Quantitative Metric (Zhang et al. 2015a), or structure-based active learning (Yu et al. 2020), falling short of multi-level requirements. A recent method, Web Structure Derived Clustering (SDC) (Hambley et al. 2023), attempts to align with collective-level sampling by clustering. Yet, *it excludes individual-level sampling, potentially omitting accessibility-critical pages, and relies solely on shallow statistical textual features, lacking semantic depth and multimodal integration*.

LLM Applications in WAA. Several studies have explored the use of LLMs for web accessibility evaluation (Huang et al. 2024; Othman, Dhouib, and Nasser Al Jabor 2023; He, Huq, and Malek 2025), demonstrating impressive automation in addressing element- and page-level issues. However, they mainly focus on text semantic alignment (Zhong et al. 2025), such as text or code generation, and *rarely involve more modularities and complex reasoning about accessibility knowledge which MLLMs are capable of*. **Moreover**, most researches are confined to evaluation and remediation, *overlooking the broader applicability across the entire lifecycle addressing labor-intensive steps*.

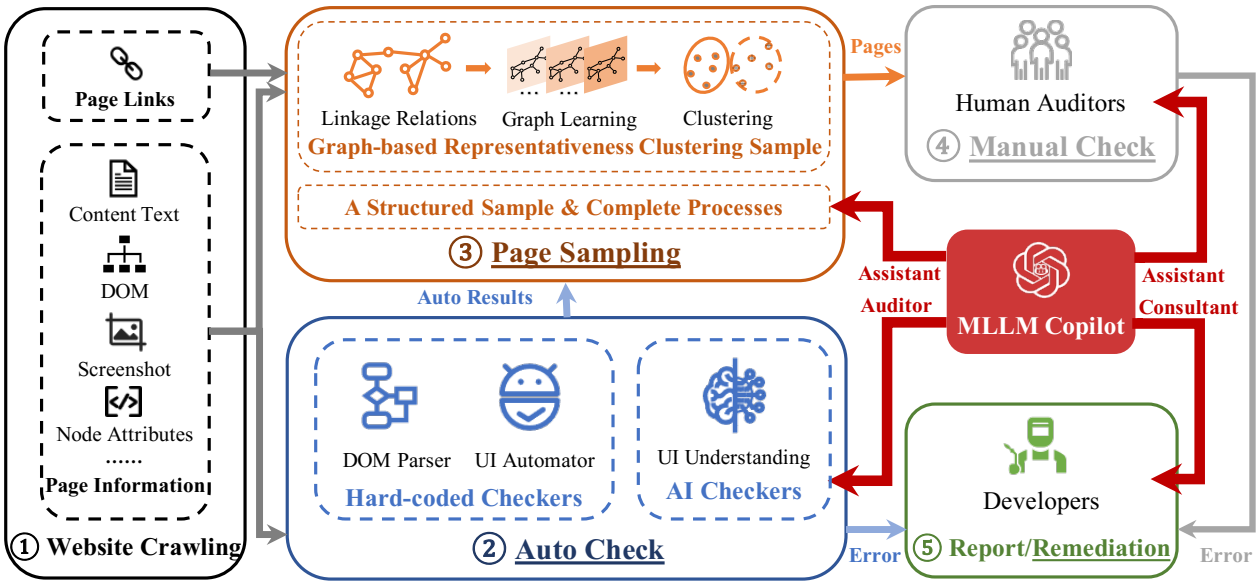
3 AAA: Scalable WAA Framework

3.1 Pipeline of the Proposed Framework AAA

We propose a scalable WAA framework for large or multiple sites centered on Automation, AI, and Auditor (AAA). Here, *AI* denotes ***artificial intelligence*** technologies such as computer vision and natural language processing, which can understand abstract concepts beyond rule-based programmatic ***automation***. Moreover, since no single tool can independently determine whether a website meets accessibility standards (L. Holliday 2020), and given the reliability and application challenges associated with LLMs, such as hallucinations (Kaddour et al. 2023), knowledgeable evaluation by human ***auditors*** remains essential. Inspired by the five-step guidance of WCAG-EM (Group 2014), we reinterpret it from a technical perspective and organize it into five structured procedures. An overview of AAA along with a comparison to WCAG-EM is in Figure 1.

Website Crawling. For large-scale or multi-site evaluations, manually defining the evaluation scope, as required by WCAG-EM, is impractical. To address this, automated website crawling is introduced to systematically explore and extract site structures and content at scale.

Auto Check. Automated checks in AAA are performed using two types of checkers: ❶ *Hard-coded checkers*, which include tools based on static DOM parsing (e.g., Axe (Deque Systems, Inc 2025)) and dynamic UI testing frameworks (e.g., Selenium (Project 2025)). These tools provide deter-



WCAG-EM	(1) Define the Evaluation Scope, (2) Explore the Target Website	(3) Select a Representative Sample	(4) Audit the Selected Sample	(5) Report the Evaluation Findings
AAA	① Website Crawling	③ Page Sampling	② Auto Check & ④ Manual Check	⑤Report/ Remediation

Figure 1: Overview of AAA.

ministic accuracy but are limited in assessing semantic-level issues. ❷ *AI-powered checkers*, which leverage intelligent technologies to perform visual and textual semantic analysis, enabling the detection of accessibility violations that require deeper contextual understanding.

Page Sampling. WCAG-EM prescribes 3 steps to construct a representative sample: (i) Include a structured sample, (ii) Include a randomly selected sample, (iii) Include complete processes. Except for random sampling, these steps require sophisticated semantic understanding at multiple levels. To alleviate the labor burden, we introduce: ❶ A novel deep-learning method integrating visual, textual, and relational information to optimize the *selection of representative diverse pages*. ❷ The use of MLLMs to *recognize structured samples and complete processes*, leveraging their strong semantic understanding capabilities (Li et al. 2024).

Manual Check. While manual evaluation follows WCAG 2.2 (Consortium 2024b,a) success criteria, its scalability and coverage remains a challenge. Two key optimizations are introduced by MLLM-powered assistance: ❶ *Pre-extraction of accessibility-critical items*: Some critical accessibility issues occur infrequently across a website, potentially reducing audit coverage in a sampling-based pipeline. MLLMs assist in identifying these elements in advance to ensure a faster and comprehensive evaluation. ❷ *Automation of evaluating underrepresented accessibility issues*: Leveraging the powerful multi-modal reasoning capabilities of MLLMs (Kil et al. 2024), we can automate the evaluation of accessibility issues that typically receive less attention.

Report/Remediation. The remediation process is integrated into the reporting step for rich and actionable feedback with automated remediation potentials. First, reports should contain detailed violation descriptions and developer-friendly repair suggestions. Second, recent advances in AI-driven code generation provide promising avenues for automated accessibility fixes.

3.2 Challenges in Implementing AAA

Representative Page Sampling for Scalable Audits. ***First***, existing approaches to page sampling in WAA are predominantly based on statistical analysis of DOM text (Hambley et al. 2023), which lack a deeper understanding of multimodal semantics like *visual styles and layouts*, *textual topics, and functional diversity*. ***Second***, with the advancement of MLLMs in comprehending complex semantics across modalities, it is now feasible to automatically identify many key web page types previously reliant on manual inspection, like common web pages and essential functionality in WCAG-EM Step 2.a "*Identify common web pages*" and Step 2.b "*Identify essential functionality of the website*". MLLMs enables the construction of page samples that are not only statistically representative but also semantically aligned with accessibility requirements, largely saving labor.

Comprehensive MLLMs Integration in WAA. ***First***, most existing works on LLMs-powered WAA only explore checks that have been well covered by existing automated tools and limited within textual information (Zhong et al. 2025; He, Huq, and Malek 2025), leaving out evaluation that relies on complex modalities or semantics . ***Second***, the existing work on the application of MLLMs in WAA is mostly

Sub-step		Type	Description	Methods
3.a	Structured Sample	Common Web Pages and States	Linked Directly from the Main Entry Point (Home Page)	MaC
		Relevant Web Page and States	Relevant for People with Disabilities and the Accessibility of the Website	
		Additional Web Pages and States	Essential Functionality	
			Web Technologies Relied Upon	
			Variety of Web Page Types	GRASP
3.b	Randomly Selected Sample		Number: 10% of the 3.a sample	—
3.c	Complete Processes		Pages belonging to a series presenting a complete process	MaC

Table 1: Three Sub-steps (3.a, 3.b, 3.c) of WCAG-EM Step 3 Achieved by Our Proposed Methods MaC and GRASP.

limited to evaluation or rededication (Suh et al. 2025), and has not fully explored its potential in the entire lifecycle.

Datasets for WAA Benchamarking. Standardized evaluation datasets are still lacking, which include not only cases of accessibility issues that are not detectable by automated tools, but also an evaluation of the application of MLLMs in other steps of the WAA lifecycle.

4 GRASP: Graph-based Page Sampling

WCAG-EM Step 3 (S3, *"Select a Representative Sample"*) has three sub-steps as shown in Table 1, in which the randomly selected sample in 3.b can be trivially realized as a random sample with certain tools (Consortium 2024b). Therefore, we focus on the other two sub-steps, which are achieved by a two-fold method. From the ***individual perspective***, when the selection of each page is based on the individual characteristic of itself, MLLMs are used as an alternative to the labor-intensive human recognition of (1) common and relevant web pages and states, (2) two kinds of additional web pages and states and (3) all pages of complete processes. From the ***collective perspective***, as tasks not solvable by scale is a well-known challenge faced by MLLMs (Kaddour et al. 2023), which calls for other UI understanding models to deal with this task where a whole picture of different pages across the website is needed to be captured for sampling representative pages among them. Individual sampling will be introduced in next section, and for collectivitive sampling, we propose Graph-based Representative pAge cluStering for samPling (GRASP). Figure 2 presents the overview of GRASP.

4.1 Triple Representativeness for Page Variety

S3 defines the variety of web page types as *varying styles, layouts, structures, and functionality with varying support for accessibility* (Group 2014). GRASP addresses this by taking into account three types of page representativeness from the perspectives of text, layouts and linkages.

Textual Semantic Representativeness. *Traditional approaches based on token frequency or lexical analysis fall short, as they tend to reflect only surface-level textual features, neglecting deeper semantic structures and functional intentions.* To address this, we leverage BERT (Devlin 2018), a contextualized language model that captures the nuanced meaning of words and phrases within their broader linguistic and structural context. For instance, it distinguishes between the use of "submit" in a login form versus in a feedback module by attending to nearby content and structural patterns. This capability makes BERT particularly well-suited for extracting semantically representativeness.

Layout Visual Representativeness. With the advancement of web technologies, *the textual structures of DOMs have become increasingly inadequate for reflecting the rendered visual layout of modern web pages, especially in dynamically generated single-page applications.* To overcome this, we employ Vision Transformers (ViT) (Dosovitskiy 2020) to learn visual representations directly from screenshots. This enables robust extraction of layout-level representativeness, capturing the spatial and visual organization of web pages beyond what is available via DOM analysis.

Linkage Relational Representativeness. Moreover, websites inherently contain rich hyperlink structures that are often overlooked in existing approaches. These linkages naturally form a graph structure that encodes functional relatedness and semantic proximity across multiple pages. *For instance, pages belonging to the same functional module frequently exhibit clustering behavior within the hyperlink network, while pages with similar layouts tend to share common linking patterns or structural relationships.* To model this, we adopt Graph Neural Networks (GNNs), which are well-suited for learning structured data (Hamilton, Ying, and Leskovec 2017). GNNs support the integration of node attributes with topological context, making them ideal for capturing and fusing this third modality of representativeness.

4.2 Representativeness-enhanced Page Sampling

GNN-based Graph Representativeness Clustering. Recent clustering-based page sampling approaches (Hambley et al. 2023) leverages shallow statistical representations derived from textual content, which fail to capture the deeper semantic nuances of text, and more critically, neglect both the visual structure and inter-page relational context. To address this, we introduce a GNN-based graph clustering approach that explicitly integrates multiple modalities into the clustering process. Our method consists of two key stages. (1) Modality-specific Representation Learning, (2) Semantic Fusion via GNN Message Passing. The first stage is:

$$\mathbf{H}_t = \text{BERT}(\text{text}_{\texttt{DOM}}),\ \mathbf{H}_v = \text{ViT}(\text{image}_{\texttt{screen}}), \quad (1)$$

$$\mathbf{X} = \mathbf{H}_t||\mathbf{H}_v, \mathbf{H}_g = \text{GNN}(\mathbf{X}, \mathbf{A}), \mathbf{C} = \mathcal{C}(\mathbf{H}_g), \quad (2)$$

where $||$ is concatenation, and $\mathbf{A}$ is the adjacent matrix of the hyperlinks. $\mathcal{C}$ is a clustering method like k-means, and $\mathbf{C}$ is the clustering assignments of web pages. *This allows*

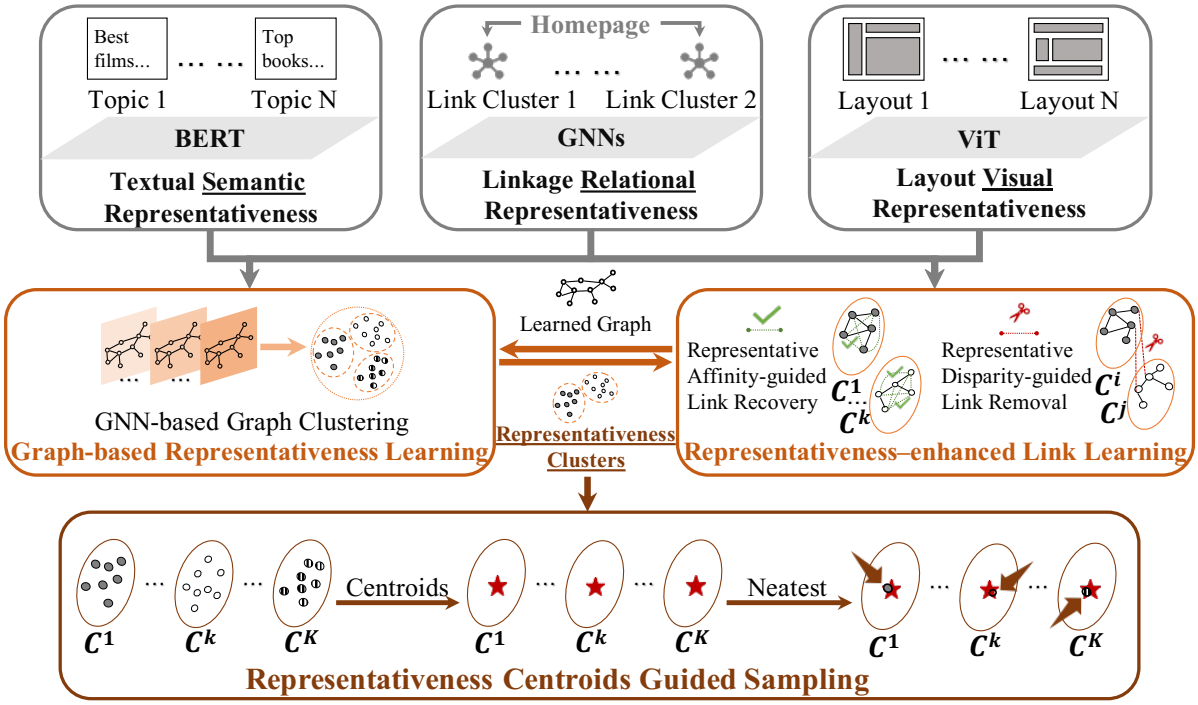


Figure 2: Overview of GRASP.

us to learn fused embeddings that encapsulate the combined representativeness across all modalities.

Representativeness-enhanced Graph Learning. While hyperlink structures offer a natural foundation for graph-based modeling of websites, they often suffer from noise and sparsity. We propose a representativeness-enhanced graph learning approach that refines the raw hyperlink structure using *representativeness affinity and disparity* inspired by graph structure learning (Gu et al. 2023; Liu et al. 2022). Specifically, we leverage the representativeness clustering results derived from textual and visual representation to guide the reconstruction of the linkage graph. This refinement involves both the removal of noisy links and the recovery of semantically meaningful missing ones. Formally, we define the hyperlink edge removal and recovery sets as:

$$\mathcal{E}_{\text{rm}} = \mathcal{S}_{\text{sim}}(\mathbf{C}, \mathbf{H}_g, \gamma), \mathcal{E}_{\text{rc}} = \mathcal{S}_{\text{dis}}(\mathbf{C}, \mathbf{H}_g, \beta), \quad (3)$$

$$\mathcal{E}_{\text{new}} = (\mathcal{E}_{\mathbf{A}} \cup \mathcal{E}_{\text{rc}}) \setminus \mathcal{E}_{\text{rm}}, \quad (4)$$

where γ and β control the thresholds for the least and most semantically similar node pairs, respectively, as determined by similarity functions $\mathcal{S}_{\text{sim}}$ and $\mathcal{S}_{\text{dis}}$ operating over cluster assignments $\mathbf{C}$ and node embeddings $\mathbf{H}_g$. *The set $\mathcal{E}_{rm}$ identifies representativeness-disparity guided redundant or misleading links to be pruned from the original edge set $\mathcal{E}_{\mathbf{A}}$, while $\mathcal{E}_{rc}$ identifies representativeness-affinity guided strong candidate connections.* The refined set $\mathcal{E}_{\text{new}}$ is obtained by adding promising edges and removing low-quality ones.

Representative Centroids-based Sample Selection. We perform representative page sampling by selecting exemplar nodes from each cluster. For each cluster c_i of $\mathbf{C}$, we first compute its centroid μ_i and then select the representative node v_i closest to the centroid. The final sampled page set is obtained by aggregating all selected nodes across clusters:

$$v_i^* = \arg\min_{v \in c_i} \|\mathbf{H}_g(v) - \mu_i\|_2, \mathcal{P}_{\text{sample}} = \bigcup_i \{v_i^*\}. \quad (5)$$

This strategy ensures that the sampled pages are deemed the most representative of the semantic, visual, and relational characteristics captured by their cluster.

5 Proposed MLLMs Strategies and Datasets

5.1 MaC: MLLMs as Various Copilots

Current applications of LLMs in accessibility face two limitations. (1) ***Limited Exploration of Evaluation.*** Most existing applications focus on a narrow range of rules, often overlapping with those addressed by traditional tools (e.g., missing alt text) (He, Huq, and Malek 2025) . We argue that the true potential of MLLMs lies in enabling fairer and more comprehensive audits by filling evaluation gaps for ***underrepresented disabilities*** *through expert-informed multimodal reasoning*. (2) ***Restricted Scope of Applications.*** The application of MLLMs in web accessibility remains underexplored beyond evaluation and remediation, particularly in resource-intensive stages like page sampling and manual auditing. We argue that integrating MLLMs across the full audit pipeline can alleviate these bottlenecks by supporting human-in-the-loop workflows, enhancing scalability and enabling more holistic accessibility.

To address them, an integrated strategy is proposed to explore the competence of MLLMs as Copilots (MaC).

Assistant: *Automating Labor-Intensive Tasks through Multimodal Reasoning*. We explore two key use cases: ❶*Individuality-based Page Sampling*: The WCAG-EM relies on manually identifying structured sample pages based on individual factors like functional role and structural position. By combining web crawling with MLLMs' multimodal reasoning, we automate this process, enabling informed sampling that *captures accessibility-critical Individuality beyond what is achievable through collective data-driven methods* like GRASP. ❷*Pre-audit Element Localization*: MLLMs can be used to preprocess pages, automatically identifying and labeling candidate elements for manual review. This transforms manual auditing into a more efficient process of *test without search*, where human experts validate elements without needing to navigate the page exhaustively.

Auditor: *Identifying Underrepresented Accessibility Barriers*. Recent studies highlight an overemphasis on visual and auditory disabilities in accessibility guidelines and tools, often *overlooking cognitive and situational disabilities that resist rule-based detection* (Abramovich and Patitsas 2024). MLLMs, with their contextual and inferential capabilities, offer a promising alternative. We focus on WCAG 2.2 success criteria 3.3.8 and 3.3.9, which address accessible authentication and require reasoning about cognitive demands in user verification. MLLMs can help identify such mechanisms and assess potential barriers by interpreting page semantics at a higher level of abstraction than existing tools.

Consultant: *Providing Informed Remediation Suggestions*. The consultant role envisions MLLMs as intelligent agents for recommending fixes to accessibility issues . While promising, we focus on the assistant and auditor roles to address foundational scalability challenges, leaving the consultant role for future research.

5.2 AWA: Datasets of AI for Web Accessibility

In addition, we contend that *a major barrier to evaluating WAA is the lack of comprehensive datasets tailored to accessibility*. To overcome this, we propose four datasets to advance the application of AI for Web Accessibility (AWA).

Triple-representativeness Page Sampling (TPS). We have developed a dataset consisting of 495 publicly accessible websites, categorized into 117 distinct classes (please refer to Appendix). These websites were crawled using an automated web crawler, ensuring that no private or sensitive data was involved. On average, each website contains 196 webpages (ranging from a minimum of 104 to a maximum of 200), totaling 97,246 pages. The dataset includes the following data for each page: (1) the page's DOM, (2) a screenshot of the page, (3) auto-check results covering 131 rules, where each rule corresponds to the number of violations detected (using Axe-core (Deque Systems, Inc 2025)), and (4) an adjacency matrix representing the website's overall linkage graph. The inclusion of this novel adjacency matrix allows for more granular web accessibility sampling.

Accessibility-relevant Page Recognition (APR). To evaluate the ability of MLLMs Assistants in individuality-based page sampling, we constructed a manually annotated dataset consisting of 968 pages from five websites of different classes (entertainment, job search, e-commerce, government & organizations, and social media), covering *four category labels defined by WCAG-EM* for a structured sample: (1) Common Web Pages and States, (2) Relevant Web Pages and States, (3) Pages of Essential Functionality, and (4) Pages of Web Technologies Relied Upon for Conformance. This dataset contains 951 human labels of four types with an equal distribution of positive and negative labels. Fifty cases are selected as a few-shot fine-tuning set.

CAPTCHA of Cognitive Tests (CCT). Completely Automated Public Turing test to tell Computers and Humans Apart (CAPTCHA) is a widely utilized mechanism for distinguishing human users from automated bots, typically as part of login verification processes. *It presents various cognitive tasks designed to challenge and differentiate human recognition abilities from those of machines, making it an ideal test scenario for evaluating cognitive accessibility.* In this study, we have collected a dataset of 1,985 CAPTCHA images from the internet, spanning 17 distinct categories of authentication requirements (see Appendix). Among these categories, three meet the criteria outlined in WCAG 2.2 Success Criteria 3.3.9 for cognitive disabilities. We also provide a 50% train split drawn from each class for fine-tuning.

Complete Process Extraction (CPE). To assess MLLMs in Pre-audit Element Localization, we annotated 1,199 pages from the APR dataset, marking pages that contain *five key components relevant to complete processes or accessibility defined by WCAG-EM*: (1) search bar, (2) select/filter panel, (3) input form, (4) CAPTCHA, and (5) contact information. 598 positive labels and 601 negative labels are annotated. Given the low occurrence frequency of some elements, we constructed 50 representative cases for few-shot fine-tuning, with an equal split between positive and negative examples.

6 Experiments

6.1 Baselines and Experimental Settings

Baselines and Experimental Settings: For the *collective page sampling*, we utilize the TPS dataset and compare GRASP with *five statistical representations* proposed by a recent study on Web Structure Derived Clustering for Accessibility Page Sampling (SDC)(Hambley et al. 2023) and its dimension reduction variants with t-SNE (Van der Maaten and Hinton 2008). We also assess two variants of GRASP: one leveraging GCN(Kipf and Welling 2016) tailored for homophilic graphs, where nodes with similar labels tend to be interconnected, and another utilizing IGNN (Gu et al. 2025) for heterophilic graphs, where nodes of differing labels are more likely to be adjacent. For the *MaC*, we adopt the APR, CCT, and CPE datasets, with various models including GPT-4o (200B), GPT-4o-mini (8B), Qwen2.5-VL-72B (Qwen2.5), Intern2-VL-8B (Chen et al. 2023), and MiniCPM-V 8B (CPM) (Yao et al. 2024). See more details in the extended version. **Metrics**: SDC evaluates page sampling using the mean internal cosine similarity of clusters, which reflects cluster cohesiveness but overlooks the quality of the **sampled** results. We propose two enhanced metrics: (i) **The mean inter-cluster cosine similarity S_{sampled} of sampled nodes** in the layout and textual embedding spaces derived from BERT and ViT. *A lower value indicates greater*

Method	Layout Space		Textual Space	
	S_{sampled}	$D_{\text{intra-inter}}$	S_{sampled}	$D_{\text{intra-inter}}$
SDC_content	56.66	9.96	89.29	2.73
+TSNE	55.14	6.46	88.32	1.66
SDC_struc_cont	55.61	11.53	89.59	1.93
+TSNE	55.89	8.91	88.89	1.60
SDC_structure	56.11	11.07	89.77	1.39
+TSNE	55.93	10.07	89.16	1.51
SDC_tags	54.18	10.76	88.76	2.12
+TSNE	55.80	9.02	89.05	1.51
SDC_tree	54.17	10.55	88.79	2.09
+TSNE	55.86	8.81	88.85	1.63
GRASP_GCN	51.54	13.05	86.99	1.59
GRASP_IGNN	**44.31**	**14.94**	**80.45**	**7.40**

Table 2: Mean Performance across 495 Websites of GRASP on TPS. The smallest S_{sampled} and largest $D_{\text{intra-inter}}$ are highlighted in **bold**, while the second are underlined.

	Category	GPT-4o	4o-mini	Qwen	CPM
Element	Form	77.95	**87.06**	34.36	61.54
	Contact	50.38	24.04	46.03	**59.15**
	Select/Filter	87.60	**89.91**	87.50	62.69
	Search	**98.01**	77.29	88.31	73.21
	CAPTCHA	**95.33**	93.20	94.85	87.62
Page	Com.	68.65	**72.41**	47.90	59.26
	Ess.	77.46	**79.71**	65.91	64.33
	Rel.	35.44	12.81	**80.21**	44.57
	Tech.	**87.32**	9.40	64.29	60.12

Table 3: F1 on APR and CPE Datasets.

diversity in sample nodes, suggesting more distinct representativeness. (ii) **the difference $D_{\text{intra-inter}}$ between the mean intra-cluster cosine similarity of all nodes and the mean inter-cluster cosine similarity of sampled nodes**. *A larger difference indicates that not only the clusters are internally cohesive but also sample nodes are distinct from one another.* (2) For MaC tasks, we use accuracy for recognition and extraction on APR and CPE, and use precision, recall and macro F1-score for classification on CCT.

6.2 Performance Analysis

GRASP Performance. The results is presented in Table 2. Several key observations can be made: **First**, *GRASP variants consistently yield more representative samples*, evidenced by lower inter-cluster layout and textual semantic similarities S_{sampled}, as well as higher differences $D_{\text{intra-inter}}$. **Second**, *GRASP demonstrates superior performance when utilizing the heterophilic IGNN*, compared to the homophilic GCN. This suggests that the linkage relationships within websites are more likely to exhibit diverse connections across distinct semantic clusters. **Finally**,*the five SDC representations show comparable performance to each other, with inclusion of t-SNE mostly improving results*. GRASP_IGNN demonstrates significantly better results across both textual and visual spaces, while SDC only performs relatively better in the textual space compared to GRASP_GCN.

Model	T.	Exist.	P.	R.	F1	Vio.
MiniCPM-sft 8B	✓	85.64	15.00	10.88	11.72	38.20
Intern2-VL 8B	✓	**100**	**49.54**	**43.85**	**45.58**	**99.88**
GPT-4o-mini 8B		91.06	27.19	16.66	19.33	93.33
GPT-4o 200B		96.30	34.24	27.34	29.16	97.47
Qwen2.5-VL 72B		90.75	39.79	32.47	34.55	84.96

Table 4: Results on Dataset CCT. *T.* is short for training, while *Exist.* means recall of existence of CAPTCHAs. *P., R., F1* are precision, recall and macro F1-score, respectively. *Vio.* is the accuracy of violation judgement of cogniton test.

MaC Assistant Performance. MaC performance in individuality-based page sampling and pre-audit element localization is presented in Table 3 and in the extended version. **First**, *the MLLM assistant demonstrates high accuracy*, exceeding 50% for most high-level multimodal semantic understanding and recognition tasks, with several reaching over 90%, indicating promising capabilities. **Second**, *larger MLLMs mostly outperform smaller MLLMs with the first or second ranks, although both large and small MLLMs exhibit varying preferences and strengths*. GPT-4o excels in extracting smaller elements such as contact forms and search boxes, whereas GPT-4o-mini performs better with larger components like forms and CAPTCHAs. This suggests that smaller MLLMs can also find effective use, *highlighting the critical role of selecting an appropriate model or integrating multiple models to improve performance*.

MaC Auditor Performance. The results of the MaC on CAPTCHA cognition are presented in Table 4 with several key observations as follows. (1) *While the recognition of the existence of CAPTCHAs approaches 100%, the classification results still leave considerable room for optimization.* This may be partly attributed to the semantic similarities between the types. However, even when CAPTCHAs are similar, the distinct cognitive tests and operational requirements can introduce different barriers. For example, there are several initial recognition tasks, each followed by different subsequent tasks such as matching, segmentation, or recognition. These variations may introduce different cognitive challenges. (2) Nevertheless, *MLLMs demonstrate high accuracy in determining whether CAPTCHAs might impede users with cognitive impairments.* This suggests their reasoning capabilities regarding functionality and barriers remain strong, compensating for classification shortcomings.

7 Conclusion

In response to the scalability challenges in web accessibility audits, we present a full-lifecycle WAA framework that operationalizes WCAG-EM through integration of Automation, AI, and Auditor (AAA). Our contributions address critical resource bottlenecks in both sampling and evaluation, including GRASP for representative multimodal page sampling, MAC strategies for MLLMs as Copilots for scalable WAA, and a suite of benchmark datasets. They advance the state of scalable, WCAG-EM-aligned accessibility auditing and lay the groundwork for more scalable WAA practices.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No.62372408).

References

Abramovich, S.; and Patitsas, E. 2024. " Slipping through the cracks": A Duoethnography of Web Accessibility. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility*, 1–6.

Ara, J.; and Sik-Lanyi, C. 2025. Automated evaluation of accessibility issues of webpage content: tool and evaluation. *Scientific Reports*, 15(1): 9516.

Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; Li, B.; Luo, P.; Lu, T.; Qiao, Y.; and Dai, J. 2023. InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks. *arXiv preprint arXiv:2312.14238*.

Consortium, W. W. W. 2024a. How to Meet WCAG (Quick Reference). https://www.w3.org/WAI/WCAG22/quickref/. Accessed: 2025-3-26.

Consortium, W. W. W. 2024b. Web Content Accessibility Guidelines (WCAG) 2.2. https://www.w3.org/TR/WCAG22/. Accessed: 2025-3-26.

Deque Systems, Inc. 2025. AXE-CORE. https://github.com/dequelabs/axe-core. Accessed: 2025-3-26.

Devlin, J. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dosovitskiy, A. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Elglaly, Y. N.; Baker, C. M.; Ross, A. S.; and Shinohara, K. 2024. Beyond HCI: The need for accessibility across the CS curriculum. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1*, 324–330.

Group, W. W. 2014. Website Accessibility Conformance Evaluation Methodology (WCAG-EM) 1.0. https://www.w3.org/TR/WCAG-EM. Accessed: 2025-3-26.

Gu, M.; Yang, G.; Zhou, S.; Ma, N.; Chen, J.; Tan, Q.; Liu, M.; and Bu, J. 2023. Homophily-enhanced structure learning for graph clustering. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 577–586.

Gu, M.; Zheng, Z.; Zhou, S.; Liu, M.; Chen, J.; Tan, Q.; Li, L.; and Bu, J. 2025. Making Classic GNNs Strong Baselines Across Varying Homophily: A Smoothness–Generalization Perspective. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Hambley, A.; Yesilada, Y.; Vigo, M.; and Harper, S. 2023. Web structure derived clustering for optimised web accessibility evaluation. In *Proceedings of the ACM Web Conference 2023*, 1345–1354.

Hamilton, W.; Ying, Z.; and Leskovec, J. 2017. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30.

He, Z.; Huq, S. F.; and Malek, S. 2025. Enhancing Web Accessibility: Automated Detection of Issues with Generative AI. *Proceedings of the ACM on Software Engineering*, 2(FSE): 2264–2287.

Huang, C.; Ma, A.; Vyasamudri, S.; Puype, E.; Kamal, S.; Cheema, S.; and Lutz, M. 2024. Deep-Learning Approaches for Optimized Web Accessibility: Correcting Violations and Enhancing User Experience.

Kaddour, J.; Harris, J.; Mozes, M.; Bradley, H.; Raileanu, R.; and McHardy, R. 2023. Challenges and Applications of Large Language Models. arXiv:2307.10169.

Kil, J.; Mai, Z.; Lee, J.; Chowdhury, A.; Wang, Z.; Cheng, K.; Wang, L.; Liu, Y.; and Chao, W.-L. H. 2024. Mllm-compbench: A comparative reasoning benchmark for multimodal llms. *Advances in Neural Information Processing Systems*, 37: 28798–28827.

Kipf, T. N.; and Welling, M. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

L. Holliday, E. 2020. The Compliance Mindset: Exploring Accessibility Adoption in Client-Based Settings. In *Proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, 1–3.

Li, W.; Fan, H.; Wong, Y.; Yang, Y.; and Kankanhalli, M. 2024. Improving context understanding in multimodal large language models via multimodal composition learning. In *Forty-first International Conference on Machine Learning*.

Liu, N.; Wang, X.; Wu, L.; Chen, Y.; Guo, X.; and Shi, C. 2022. Compact Graph Structure Learning via Mutual Information Compression. In *Proceedings of the ACM Web Conference 2022*, 1601–1610.

López-Gil, J.-M.; and Pereira, J. 2024. Turning manual web accessibility success criteria into automatic: an LLM-based approach. *Universal Access in the Information Society*, 1–16.

Othman, A.; Dhouib, A.; and Nasser Al Jabor, A. 2023. Fostering websites accessibility: A case study on the use of the Large Language Models ChatGPT for automatic remediation. In *Proceedings of the 16th International Conference on PErvasive Technologies Related to Assistive Environments*, 707–713.

Project, T. S. 2025. Selenium. https://www.selenium.dev/. Accessed: 2025-3-26.

Sharif, A.; Pruekcharoen, P.; Ramesh, T.; Shang, R.; Williams, S.; and Hsieh, G. 2022. "What's going on in Accessibility Research?" Frequencies and Trends of Disability Categories and Research Domains in Publications at ASSETS. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility*, 1–5.

SolarWinds Worldwide, L. 2025. Webpages Are Getting Larger Every Year, and Here's Why it Matters. https://www.pingdom.com/blog/webpages-are-getting-larger-every-year-and-heres-why-it-matters/. Accessed: 2025-4-9.

Suh, H.; Tafreshipour, M.; Malek, S.; and Ahmed, I. 2025. Human or LLM? A Comparative Study on Accessible Code Generation Capability. *arXiv preprint arXiv:2503.15885*.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Web Accessibility Initiative (WAI). 2024. Introduction to Web Accessibility. https://www.w3.org/WAI/fundamentals/accessibility-intro/. Accessed: 2024-11-26. First published: February 2005. Last updated: 7 March 2024.

WebAIM. 2025. WebAIM: The WebAIM Million - The 2025 report on the accessibility of the top 1,000,000 home pages. https://webaim.org/projects/million/. Accessed: 2025-4-9.

WebAIM, Utah State University. 2025. WAVE. https://wave.webaim.org/. Accessed: 2025-3-26.

Yao, Y.; Yu, T.; Zhang, A.; Wang, C.; Cui, J.; Zhu, H.; Cai, T.; Li, H.; Zhao, W.; He, Z.; et al. 2024. MiniCPM-V: A GPT-4V Level MLLM on Your Phone. *arXiv preprint arXiv:2408.01800*.

Yu, Z.; Bu, J.; Shen, C.; Wang, W.; Dai, L.; Zhou, Q.; and Zhao, C. 2020. A multi-site collaborative sampling for web accessibility evaluation. In *Computers Helping People with Special Needs: 17th International Conference, ICCHP 2020, Lecco, Italy, September 9–11, 2020, Proceedings, Part I 17*, 329–335. Springer.

Zhang, M.; Wang, C.; Bu, J.; Yu, Z.; Lu, Y.; Zhang, R.; and Chen, C. 2015a. An optimal sampling method for web accessibility quantitative metric. In *Proceedings of the 12th International Web for All Conference*, 1–4.

Zhang, M.-n.; Wang, C.; Bu, J.-j.; Yu, Z.; Zhou, Y.; and Chen, C. 2015b. A sampling method based on URL clustering for fast web accessibility evaluation. *Frontiers of Information Technology & Electronic Engineering*, 16(6): 449–456.

Zhong, M.; Chen, R.; Chen, X.; Fogarty, J.; and Wobbrock, J. O. 2025. ScreenAudit: Detecting Screen Reader Accessibility Errors in Mobile Apps Using Large Language Models. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–19.