Behaviour & Information Technology

ISSN: 0144-929X (Print) 1362-3001 (Online) Journal homepage: www.tandfonline.com/journals/tbit20

Bridging human insight and automation: improving alt text generation with human-curated contextual data

Nikolaos Droutsas, Fotios Spyridonis, Damon Daylamani-Zad , Philipp E. Glass & Gheorghita Ghinea

To cite this article: Nikolaos Droutsas, Fotios Spyridonis, Damon Daylamani-Zad , Philipp E. Glass & Gheorghita Ghinea (05 Jun 2026): Bridging human insight and automation: improving alt text generation with human-curated contextual data, Behaviour & Information Technology, DOI: 10.1080/0144929X.2026.2678381

To link to this article: https://doi.org/10.1080/0144929X.2026.2678381

© 2026 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group

Published online: 05 Jun 2026.

Submit your article to this journal

Article views: 63

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=tbit20

BEHAVIOUR & INFORMATION TECHNOLOGY
https://doi.org/10.1080/0144929X.2026.2678381

Taylor & Francis
by informa

OPEN ACCESS Check for updates

Bridging human insight and automation: improving alt text generation with human-curated contextual data

Nikolaos Droutsas, Fotios Spyridonis, Damon Daylamani-Zad, Philipp E. Glass and Gheorghita Ghinea

Brunel University of London, Uxbridge, UK

ABSTRACT

The rapid growth of image-based multimedia content on the Web has intensified the challenge of generating high-quality alternative (alt) text descriptions, which is an essential requirement for inclusive online experiences for people with visual impairments. Although recent advances in machine learning (ML) have enabled large-scale automated alt text generation, the accessibility value of such outputs remains limited. This is due to the context-agnostic datasets used to train existing models, resulting in generic descriptions that fail to meet users' needs in alt text. In this work, we introduce and utilise a human-curated, context-driven dataset of alt text descriptions to train two proof-of-concept ML models aimed at improving alt text quality. We evaluate these models within a controlled, reproducible pipeline and demonstrate that context-aware training leads to statistically significant improvements in human-perceived alt text quality compared to a model trained without contextual inputs. We further examine the role of context-dependent routing and the integration of contextual cues in shaping generated descriptions, both of which are critical but underexplored aspects of alt text accessibility. The findings highlight the value of structured, human-curated contextual data in advancing ML-supported alt text generation and point towards opportunities for hybrid human-AI approaches to inclusive web design.

ARTICLE HISTORY
Received 10 December 2025
Accepted 16 May 2026

KEYWORDS
Accessibility; human-centred AI; alt text generation; context-aware modelling; user study

1. Introduction

Whilst it is the professional responsibility of web content creators towards people with disabilities to cater to the accessibility of the web, web accessibility barriers are still present in 94.8% of website home pages (WebAIM 2025). Barriers affect and impact people with different impairments differently, with past work having attempted to measure the impact of barriers based on reported affect rates and web activity per disability (Droutsas et al. 2024). Alternative text (alt text), i.e. 'text that is programmatically associated with non-text content or referred to from text that is programmatically associated with non-text content' (W3C 2025), is a barrier with particular impact on blind people and people with low vision, as well as on other impairments (Bi et al. 2022; WebAIM 2024). Alt text is accessed via *screen readers*, i.e. assistive software that reads out loud content displayed on computer screens (Lee and Ashok 2022); therefore, screen reader users are unable to interact or be aware of images when alt text is missing. Whilst there have been notable improvements in the last five years in relation to missing alt text on the Web, the same cannot be said about the **suitability** of alt text, i.e. its *clarity* and *accuracy* in relation to the **context** in which the image it substitutes is used, which has decreased by 4.1% during the same period (WebAIM 2023; WebAIM 2025). In fact, it has been shown that unsuitable alt text can be equally or more problematic than missing alt text (Mack et al. 2021) and several reasons for this have been identified in the literature, including the difficulty of the task, the lack of guidelines on what makes alt text suitable, and the reluctance to train in authoring alt text suitably (Harris 2020; Miranda and Araujo 2022; Muehlbradt and Kane 2022).

Central to most work on suitability is the need to cater alt text to the context in which the image it substitutes is used; however, context in alt text has been loosely defined in past work, making it difficult to meaningfully use this concept to train people in authoring alt text suitably (Hanley et al. 2021). Further, and given the increased quantity of multimedia content on the Web, alt text descriptions substituting such content need to be of sufficient volume (Lee and Ashok 2022). However, the traditional approach of manual alt text authorship has been shown to be prohibitive to scale, especially since expert views on what makes alt text suitable vary

CONTACT Fotios Spyridonis fotios.spyridonis@brunel.ac.uk Brunel University of London, Kingston Lane, Uxbridge, UB8 3PH, UK

© 2026 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

due to the complexity of the task (Lengua, Rubano, and Vitali 2022). In response, automated approaches relying on Artificial Intelligence (AI) and the training of models known as Vision-to-Language (V2L) models to generate alt text descriptions at scale have been proposed, but they have been shown to be lacklustre with regards to alt text suitability (Zong et al. 2022). Past work suggests that the absence of context in the training of V2L models contributes to the limited results with regard to the suitability of automatically generated alt text descriptions (Zong et al. 2022). Furthermore, it has been shown that V2L models also underperform when trained on small datasets, which has resulted in the rise of web scraped datasets, such as LAION-5B (Schuhmann et al. 2022) and Conceptual 12M (Changpinyo et al. 2021) using a traditional image processing approach. It is therefore necessary to investigate the potential of using **context-rich datasets** for the training of AI solutions and further evaluate their ability in automatically generating more suitable alt text descriptions. This way, it will be possible to deepen the discourse with a practical automated approach to improve the generation and quality of alt text on the Web at scale.

Accordingly, this work builds upon our previously produced crowdsourced dataset of context-driven alt text descriptions collected through a Game-With-A-Purpose (GWAP) approach (Anonymous, Under review). Crowdsourcing literature highlights GWAPs as the most promising crowdsourcing approach for complex tasks and when scalability is paramount, as they excel at training users and participation is devoid of monetary incentives, respectively (Chamberlain et al. 2013; Tuite 2014), particularly when it comes to similarly complex tasks, such as linguistic annotation. There are several examples of GWAPs for complex annotation, such as anaphora resolution, lexical relation extraction and protein folding, which have gathered large-scale datasets for training ML models (Curtis 2015; Poesio et al. 2013). The validity of such datasets compared to equivalents gathered via other crowdsourcing approaches has also been noted, not least in relation to cost-effectiveness and scalability, as well as the ability to train players to become pseudo-experts in complex annotation tasks (Madge 2020). However, while past GWAP efforts aimed at similar tasks have been reported in the literature (Apostolopoulos, Folmer, and Bebis 2013; Harris 2020; Nguyen et al. 2019; Nguyen et al. 2020; Steinmayr et al. 2011), the resulting datasets more closely resembled image captions and keywords rather than alt text descriptions. Conversely, the dataset utilised in this work captures diverse human-generated image descriptions enriched with contextual cues, such as webpage type, communicative intent, and functional purpose. While our earlier study focused on the design and deployment of the data collection GWAP solution and explored the characteristics of the gathered descriptions, the present research extends this contribution by leveraging this dataset to train and evaluate machine learning (ML) models for generating contextually suitable alt text. Using this dataset as a foundation allows the current study to examine how contextual metadata influences model performance and to explore strategies for producing alt text that better aligns with human communicative goals. Accordingly, the present research addresses the research questions below:

- RQ1: Is the use of a human-generated dataset to train an ML model (trained model) to generate alt text descriptions an effective approach to approximate human average alt text generation compared to pure image processing (control model)?
- RQ2: How does context-dependent routing influence the suitability of generated alt text?
- RQ3: To what extent does learning from structured context prompts improve context-driven alt text generation?

Building on this foundation, the overarching aim of this study is to advance the generation of contextually suitable alt text descriptions through ML techniques informed by human-authored, context-rich descriptions. While our prior work established the design and collection of a context-rich dataset of alt text descriptions, the present research focuses on modelling and evaluation. The main contributions of this paper are therefore as follows:

1. **Dataset operationalisation**: We extend a previously human-curated dataset of context-driven alt text descriptions for supervised ML tasks, establishing a baseline for automated context-aware alt text generation suitable for model development and reproducibility.
2. **Context-aware model development**: We designed and implemented two proof-of-concept ML models that integrate contextual features during training and assess their impact on shaping model performance, thereby offering empirical evidence for the impact of contextual cues on alt text generation within a reproducible pipeline.
3. **Empirical insights for accessible AI design**: We provide empirical findings on how contextual cues influence model performance, offering actionable insights for researchers and practitioners developing

accessible AI systems that better reflect human communicative intent.

Collectively, these contributions extend prior work from data collection towards computational modelling, bridging the gap between human-centred design of alt text and automated generation approaches. This work therefore highlights the effectiveness of human-centred datasets for the training of such ML models, as well as the extent to which these datasets can improve context-driven alt text generation, based on the superior performance of the trained models compared to pure image processing.

This paper is structured as follows. Section 2 discusses related work on the datasets used for the training of V2L models, as well as approaches for gathering such datasets, the role of context therein, and the performance of the models trained on these datasets. Section 3 dives into the dataset used to train the models proposed in this work, including details about how the dataset was gathered and the definition of context in alt text, and presents an overview of the architecture of the models. Section 4 then discusses the performance of the models compared to image processing and the impact of context therein. Finally, Section 5 presents the concluding discussion of this work, where implications of our findings are discussed, not least in relation to the possibility of approximating average human-level alt text quality whilst automating alt text generation.

2. Background and related work

As discussed in the previous section, scaling alt text generation through automated solutions, i.e. V2L models that process images and translate visual information into text descriptions, is paramount. However, V2L models rely on the datasets they are trained on to learn how to generate suitable alt text, with the lack of context in these datasets being flagged as a key reason for the poor quality of the resulting alt text. Accordingly, this section discusses state-of-the-art (SoTA) V2L models and the datasets they are trained on, with a focus on alt text descriptions and the incorporation of context therein.

2.1. Vision-to-language (V2L) model datasets

V2L models typically use encoder-decoder methods, i.e. an image is inputted to the encoder processing the visual information in the image and passing it to the autoregressive language decoder generating the text description (i.e. caption) (Ramos et al. 2023). These models are pre-trained on datasets of image-alt text pairs, with the alt text being typically authored by humans (Chen et al. 2023); however, the resulting datasets are too small in size, negatively impacting the performance of the models. Hence, newer models capitalised on the availability of a wide range of images on the Web, retrieving web images and their associated alt text descriptions via web scraping (Changpinyo et al. 2021). Alt text is particularly sought after for its shortness, making it cost-effective to retrieve and capture a more accurate snapshot of the visual information in the image (Laurençon et al. 2023). This, however, has no implications for improving web accessibility. It has, in fact, been noted that web-scraped alt text descriptions are of particularly poor quality and that the images are removed from the context in which they are used during web scraping, resulting in context-poor captions generated by the models (Zong et al. 2022). Several efforts have, in fact, attempted to address this; for example, Laurençon *et al.* created OBELICS, a dataset that preserved neighbouring text and contextual information alongside images (Laurençon et al. 2023). Relatedly, W. Chen *et al.* created the seed dataset, which addressed subject awareness through image clusters with automatically generated descriptions (Zong et al. 2022), while Desai *et al.* introduced RedCaps, a dataset leveraging Reddit's human-moderated content for richer and more emotionally rich descriptions (Desai et al. 2021). Taken together, these efforts highlight the importance of **context** and **its lack thereof in existing datasets** used for training V2L models, alongside broader ethical concerns related to the imagery, captions, and neglect of accessibility (Birhane, Prabhu, and Kahembwe 2021).

2.2. Context in alt text: a definition

The previous section highlighted the importance of including context in datasets to train V2L models, but it also highlighted that there is no unified understanding of context in the case of alt text. Several interpretations of context were, in fact, appreciated in the various datasets described above and this is despite context being commonly cited in the literature as pivotal for the suitability of alt text (Bi et al. 2022; Petrie, Höckner, and Rosenberger 2022). In response, in this work, we adopt an alt text context definition **(altC)** that we developed in a separate study (Anonymous, Under review) that aimed to address the lack of such a definition in alt text literature. To achieve this, the definition was designed to capture the semantic and contextual factors that influence descriptive quality in alt text, and it was thus framed within two important elements relating to an image (type, function and intent) (Desai et al.

2021; Zong et al. 2022), and to the webpage (topic and purpose) (Shen et al. 2025). For completeness, we provide a clear and concise version of the definition here to ensure the present work is self-contained, and we further include the criteria used to operationalise it within this study. The alt text context (altC) definition therefore comprises:

altC = (Image Type, Webpage Topic, Webpage Purpose, Image Function, Image Intent)

where:

- Image Type: represents the nature of the image (e.g. photograph, diagram, icon)
- Webpage Topic: defines the subject matter of the webpage (e.g. climate change article, e-commerce product page)
- Webpage Purpose: captures the primary goal of the webpage (e.g. informational, educational, commercial)
- Image Function: describes the role of the image within the webpage (e.g. decorative, illustrative, navigational)
- Image Intent: refers to the communicative goal of the image (e.g. supporting content, guiding interaction, evoking emotion).

The above contextual framework is used in the present study to guide model training, evaluation, and interpretation. Specifically, the components in the framework were mapped to a natural language prompt, i.e. a **context prompt**. Again, for completeness, we provide the way the contextual factors in altC above were assembled into a natural language context prompt:

This [Image Type]

is a/is found[1] [Function]

on [Webpage Topic] webpage

with the goal of [Webpage Purpose].

The intention of the image is to [Image Intent].

Table 1. SoTA ML solutions and corresponding datasets for alt text generation.

ML solution	Dataset	Context	CIDEr COCO ZS	VQAv2 ZS acc.
IDEFICS (Laurençon et al. 2023)	OBELICS	Yes	91.8	60.0%
InternVL-Chat (Chen et al. 2024)	6.03B image-text pairs	No	142.4	71.7%
PaLI-17B (Chen et al. 2023)	WebLI	No	149.1	84.3%
SmallCap (Ramos et al. 2023)	COCO (Lin et al. 2014)	No	119.7	N/A

The above prompt structure was then used both for the collection human-curated alt text dataset (see section 3.1), as well as for the training of the ML models using this dataset (see section 3.5). Therefore, in this work, we introduce and empirically evaluate a structured context definition for accessibility-oriented alt text generation. This approach is model-agnostic and could be applied to general-purpose models as well as task-specific captioning systems.

2.3. State-of-the-art (SoTA) V2L models

The discussion in the previous sections revealed limitations in the datasets that V2L models are trained on as a key issue regarding the quality of the generated text descriptions, not least in relation to the lack of context in datasets. This is crucial in the case of alt text, as it has been shown that unsuitable alt text can be equally, or even more problematic than missing alt text (see Section 1); therefore, the ability of current models to scale alt text generation becomes less relevant regarding accessibility (Mangiatordi and Lazzari 2018). Accordingly, in this section, SoTA solutions that utilise these datasets are also discussed and presented in Table 1 below, which provides an overview of relevant SoTA models, the datasets they were trained on, and whether these datasets incorporated context, as well as their performance on commonly reported benchmarks (CIDEr COCO ZS, VQAv2 ZS acc.). It must be noted that while there are additional ML models pre-trained on these datasets in the literature, those were not included as they were not fine-tuned for V2L captioning and were thus unrelated to the focus of this work.

The above table demonstrates the general lack of context in which images are used in the datasets that recent V2L captioning models are trained. Further sources (Chen et al. 2023; Zong et al. 2022) have in fact pointed out that these models most often generate generic captions lacking in contextual information, consequently which have no implication for improving web accessibility via screen readers. PaLI-17B, for example, suggested a jointly scaled multilingual model, achieving SoTA performance on captioning benchmarks, but the generated captions are not relevant for accessibility due to the noise and lack of context in the training dataset, i.e. WebLI (tens of billions of image-alt text pairs) (Chen et al. 2023). InternVL-Chat is also prone to a lack of context in web-scraped training data, despite pairing a gigantic 6 billion parameter vision encoder with a large language model through a large language middleware and a progressive training strategy (Chen et al. 2024). The suitability of generated alt text is similarly compromised in SMALLCAP, i.e. a lightweight

model using a datastore to retrieve image-alt text pairs to lift the need for fine-tuning across domains, as the poor quality of alt text retrieved from the datastore is reflected in the automatically generated alt text (Ramos et al. 2023). Further, IDEFICS is a V2L model that learns from context-rich data, namely from the OBELICS dataset discussed in the previous section, showing an increase in performance via a simplified architecture and the processing of images at their native resolution and aspect ratios (Laurençon et al. 2023). However, IDEFICS is also limited in its ability to automatically generate alt text that is relevant for accessibility, as it is trained on data extracted from the web-scraped Common Crawl dump.

Whilst IDEFICS incorporates context through extracting images' neighbouring text content and links embedded in the images, it naturally inherits limitations such as incorrect grammar, incomplete or misleading metadata, stereotypical notions, bias, hate speech, noise and racism from the Common Crawl dataset that it is based (Luccioni and Viviano 2021). In response, in this work, context is incorporated based on accessible principles that influence the descriptive quality of alt text (see Section 2.2), and the performance of the ML models in this work will be evaluated on alt text quality rather than benchmark metrics, which are more relevant in Computer Vision (CV) (see Table 1). While such CV-based benchmark metrics are indicative of linguistic (CIDEr) and visual-reasoning (VQAv2) performance, they do not evaluate the suitability of generated descriptions for access via screen readers and are therefore inadequate for evaluating alt text quality in the context of accessibility. As a result, two gaps are identified in the automated generation of alt text via ML models: (1) the lack of context in training datasets and (2) the lack of evaluation of the performance of the models on the suitability of the generated alt text. Accordingly, the ML models in this work are thus comparatively evaluated using the following proposed **metrics**:

- **Human-perceived alt text quality** via player rating scores and statistical tests. This aligns well with the need for evaluation to focus on accessibility and to use participants' rating scores for human perceptions of alt text suitability, agreeing well with recent accessibility studies (Kreiss et al. 2022; Leotta, Mori, and Ribaudo 2022; Risi 2025) on the importance of subjective evaluation of alt text.
- **Training effectiveness** via non-parametric inferential statistics and effect size estimation between trained and control versions of the models. These are measured via non-parametric statistical tests, as the data are not normally distributed (Section 4.1.3).
- **Context presence** via a binary presence evaluation of elements of the altC definition (Section 2.2) in automatically generated alt text. This is assessed due to its discussed central role in the suitability of alt text and its lack thereof in alt text that has been automatically generated.

It is clarified that the use of player rating scores and statistical tests is not presented as a novel metric, but rather, as a direct operationalisation of accessibility-oriented quality, reflecting human judgment of clarity, relevance and suitability. It is also noted that the comparison of human ratings between context-aware and non-context aware outputs (Section 4.2) provides direct empirical evidence of the effect of context on perceived alt text quality rather than a standalone indicator of quality.

2.4. Summary

In this section, V2L captioning models and the datasets on which they are trained were discussed, with a focus on their ability to scale alt text generation on par with modern needs while ensuring the quality of the generated alt text descriptions. It was shown that these models require very large datasets of image-alt text pairs for training to boast quality in generated alt text. Although the scale of such datasets has increased, they seldom incorporate the context in which images are used, resulting in poor-quality alt text from an accessibility standpoint. Key limitations, therefore, include the compromise in scalability with manual alt text authorship and the poor quality of automatically generated alt text with current V2L models. It was further shown that these models are reliant on the datasets they are trained often deriving from web scraping for scalability, which presents significant limitations for accessibility regarding alt text suitability, as well as ethical concerns. A further revealed gap was the lack of evaluation of the performance of V2L captioning models on accessibility (i.e. alt text description suitability); instead, evaluation is focused on computer-vision-related metrics (i.e. CIDEr). These challenges are very current and are addressed in this work by using context-driven training of non-expert alt text authors to gather training data (see Section 3.2) and by evaluating the performance of models in terms of alt text quality and the ability to generate context-driven alt text (see Section 4).

3. Methodology

In response to the gaps in the automated generation of suitable alt text descriptions via ML models discussed in the previous section, the next section discusses the dataset used for the training of the models utilised in this work and the data collection protocol involved.

3.1. The TagALTlong dataset: overview

The proposed ML models were trained on a custom dataset of human-authored, context-driven alt text descriptions collected through an online GWAP named TagALTlong to address the lack of similar GWAPs for these tasks based on the success of the GWAP approach for similarly complex tasks (Aliady and Poesio 2024; Kicikoglu et al. 2020). The game was designed to crowdsource alt text descriptions that are concise, context-aware and semantically meaningful, forming a rich resource for training and evaluating automated alt text generation models. The GWAP and dataset were created as part of a prior research study by the authors, which is currently anonymised for peer review.

3.2. Data collection protocol

Participants accessed TagALTlong via a public webpage hosted on itch.io, with backend services implemented on a secure Oracle Virtual Machine running Apache, PHP and MySQL. Upon launching the game, participants were shown a sequence of images drawn from diverse visual categories (e.g. everyday objects, social scenes, artworks and infographics). For each image, participants were presented with a contextual prompt in natural language (e.g. '*This image is a link on a social media webpage with the goal of informing. The intention of the image is to elicit emotion (e.g. compassion)*') to guide the type and tone of alt text description they should provide. These were randomly generated and presented to participants to consider in tandem with images when authoring alt text. Participants were able to play as 'alt text authors', where they were presented with an image and an associated context prompt, and had the option to either author alt text for that image based on the context prompt, or to mark this image as decorative. They were also able to play as 'alt text raters', where they were presented with previously authored alt text by other players, whose identity they were unaware of, and were tasked with rating the quality of the alt text on a predefined 5-point Likert scale (1-5 stars). This dual contribution process produced both qualitative (textual) and quantitative (rating) data. All participants provided informed consent, and the study received institutional ethics approval. No personally identifiable information was collected, and each entry was stored pseudonymously in the MySQL database.

3.3. Dataset composition and characteristics

The final dataset contains 1208 authored alt text descriptions and 1836 rated alt text descriptions contributed by 125 participants. These numbers are on par with highly successful past GWAP design efforts (von Ahn et al. 2006; von Ahn et al. 2007). Each data record includes an alt text description, rating score, an author username, a rater username, a full context prompt (see Section 2.2), and individual context elements (i.e. image type, image function, image intent, webpage topic and webpage purpose). Across all entries, the average alt text description length was 67 characters, and ranged from merely descriptive (e.g. 'Zielinski & Frozen Hand Cream | Button Operation', 'A paper cup with starbucks logo and christmas design') to contextually inferential (e.g. 'A starbucks cup printed with friendly christmas themed line art, on a ledge with a blurred cityscape in the background. There's an inviting drip of latte on the rim.', 'A man in a wetsuit relaxes on the beach after a surfing session, his double-finned surfboard leaned against his head to block the sun, watching the waves roll in past the crags'), providing a diverse training corpus for model training. The incorporation of alt text context within the dataset presents the added benefit of being an alternative to the use of complex filtering pipelines for quality control, which is a further deficiency in the performance of such models (Chen et al. 2023; Lee et al. 2023).

3.4. Data preprocessing

The dataset was pre-processed first in a data cleaning stage for the rating population of alt text descriptions that received at least two rating scores (N = 430 descriptions and 1538 rating scores), while alt text with less than two independent ratings was excluded to avoid introducing greater noise. Next, raters whose ratings had no variation with one another (SD = 0) were examined to further filter noisy data, with three players' ratings being excluded after using a moderation rating provided by the first researcher. Finally, we sought to identify potential outliers for alt text with three or more rating scores, as a minimum of three data points is required for a non-trivial measure variance to estimate consensus or disagreement between raters. This allowed for minimising anomalous individual rating scores, while maintaining a meaningful subset of the

GWAP-generated dataset to be used in the subsequent model training stages.

3.5. Model architecture and training

This section presents an overview of the pipeline and architecture of our ML models that were fine-tuned and trained on the GWAP-generated data to investigate effectiveness in generating alt text descriptions close to average human-level quality (first model), which was named **HumanALT-O-matic**, and in generating alt text that are context-rich (second model), which was named **ContextALT-O-matic**.

3.5.1. HumanALT-O-matic: approximating average human-level quality

To ensure transparency and reproducibility, our model architecture builds on **publicly available** and **established pre-trained models** with the intention of building a deployable modelling pipeline, as opposed to using proprietary models that are closed systems with undocumented architectures, training data and optimisation procedures, which limit reproducibility and makes it difficult to attribute performance differences to specific modelling or contextual factors. We also note that the chosen models were the best available at the time of this research. For the classification component, we adopted Google's BERT-Base (Devlin et al. 2019), which we fine-tuned on our domain-specific dataset, while for alt text generation we employed the T5-Small model (Raffel et al. 2020) with the aim of approximate average human-level quality alt text while automating the process using the GWAP-generated output (images, context prompts, alt text descriptions and rating scores) at different stages of the pipeline (see Figure 1 below). We selected these models due to their strong baseline performance in natural-language understanding and generation, respectively, and their computational feasibility for fine-tuning on our dataset, whilst maintaining a transparent, reproducible and deployable modelling pipeline.

The above figure shows how the dataset images were initially processed by an image analyser, which extracted image information (i.e. image class, scene objects and themes, image description, text in the image) that were then used in tandem with contextual information to generate alt text. Within the image analyser, OpenAI CLIP (ViT-B/16) was used for assigning a class to the image, while Detectron2 with the COCO-PanopticSegmentation/panoptic_fpn_R_101_3x configuration was used for the scene segmentation and analysis complemented by the COCO Panoptic Segmentation dataset to identify thematic elements in the scene, as well as a BLIP-2 and a Python-Tesseract OCR were used for generating an image caption and identifying text in the image, respectively. While Tesseract does not represent the most recent OCR technology, it was selected for reproducibility, open-source availability, and local deployability. All components were applied consistently across experimental conditions to ensure that observed differences are attributable to contextual modelling rather than variations in upstream processing. Accordingly, this image information and their respective context prompts were then used to train the T5-Small model. Two versions of this model were used, i.e. a **control version** initialised from the off-the-shelf model and used without fine-tuning, and a second **trained version** fine-tuned on the GWAP-generated dataset. Specifically, the data deriving from the GWAP gameplay (player-authored alt text descriptions and players' amalgamated rating scores for image-context-alt text tuples) were used in tandem with the image information generated by the image analyser and the context prompts to train the T5-Small model. It was thus made possible to compare the output of the control and trained versions of the model to assess whether the use of our GWAP-generated dataset (see section 3.3) to train the model is an effective approach to approximate average human-level quality alt text. This was assessed in the evaluation of the performance of both models (Section 4). To support the automated generation of context-driven alt text, this two-model approach used a modular two-stage architecture (Figure 2). In the first stage, the BERT-based model is used to categorise each image as decorative ('eye candy') or non-decorative, enabling routing to a different alt text generation pathway based on this categorisation. This categorisation functions as a design mechanism to operationalise context-dependent handling of images within the pipeline. Non-decorative images proceed to the second stage, where the T5-Small model generates alt text using a distinct approach for each version of the model: the control version uses only image information and context prompts, while the trained version also uses player-authored alt text and associated ratings.

The above figure shows how the BERT-based model is used to distinguish between decorative and non-decorative images, as part of the pipeline's routing mechanism. While prior work (Noreskal, Feuilloley, and Charbel 2024) has proposed the classification of images into informative and decorative categories, this distinction has not been explicitly operationalised and integrated into an automated, context-driven alt text generation pipeline and evaluated in terms of its effect

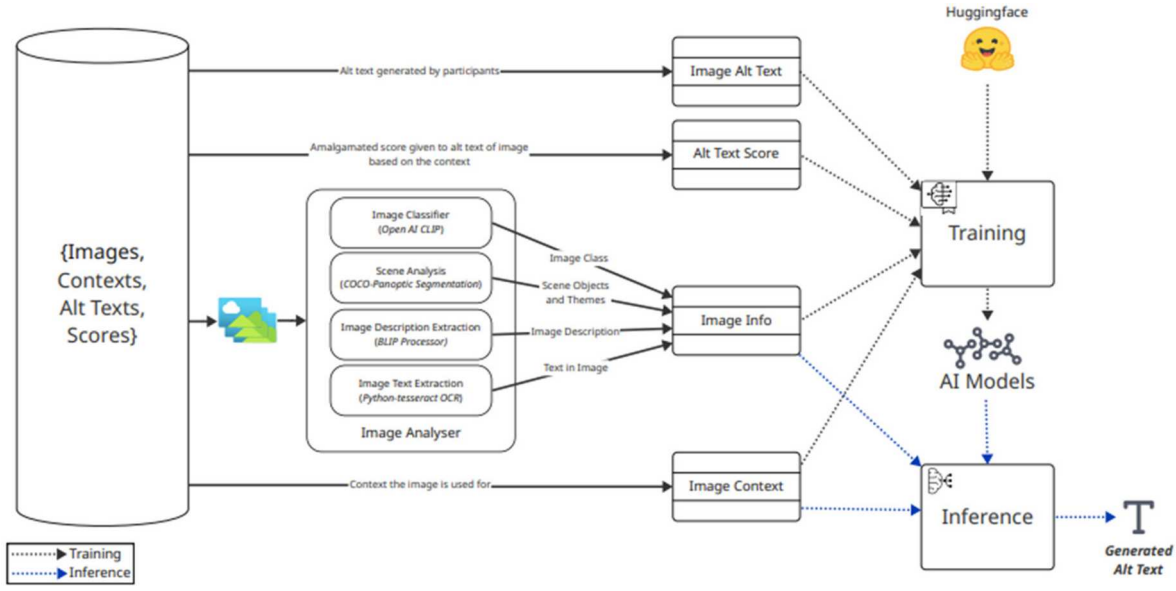

Figure 1. HumanALT-O-matic model pipeline overview.

on alt text suitability. In contrast, the role of our classifier is to operationalise context-dependent routing by identifying images that are treated as decorative ('eye candy') and handled using an alternative descriptive strategy, rather than being passed to the generative model. This design reflects common accessibility guidance, where purely decorative images may not require descriptive alt text. The classifier is therefore not presented as a standalone contribution, but as a functional component within the pipeline that enables context-dependent handling of different image types. Accordingly, the evaluation focuses on the effect of this routing strategy on downstream alt text suitability, rather than on the predictive accuracy of the classifier itself. The predictive performance of this component is not evaluated as a standalone contribution and is

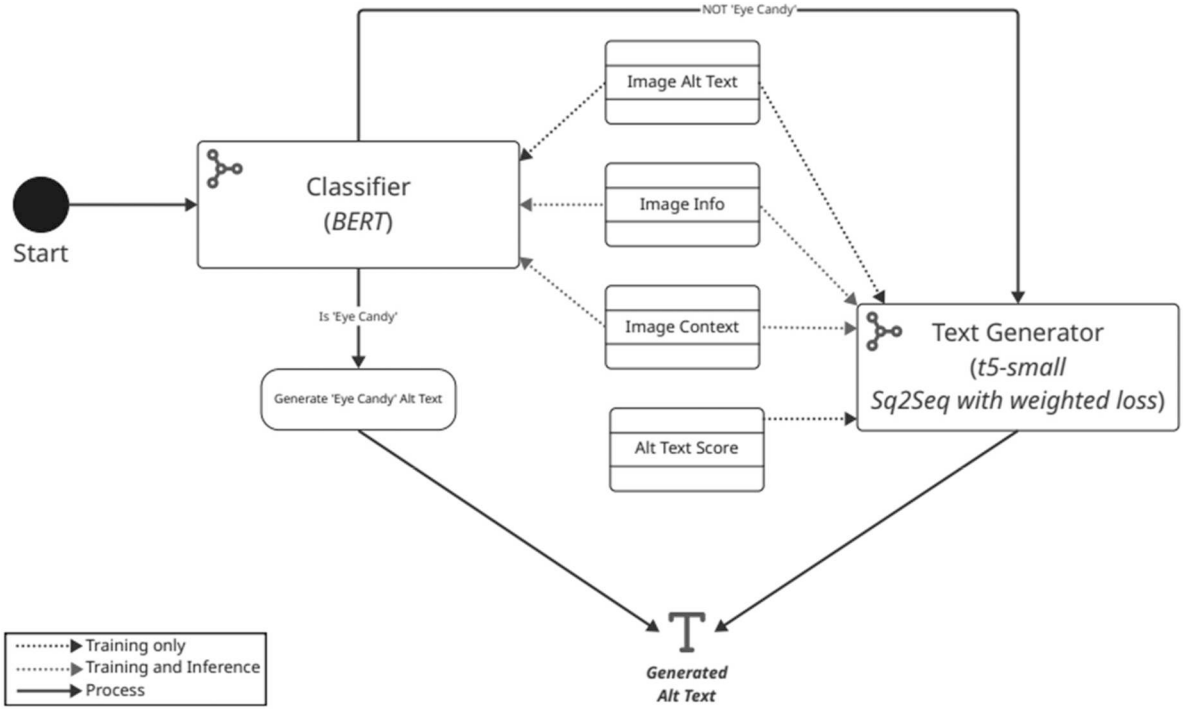

Figure 2. HumanALT-O-matic model architecture.

instead reported as an indicative assessment. Nevertheless, to support the analysis of context-dependent routing (RQ2), we additionally report indicative classification performance of the routing component in Section 4.1.5.

For non-decorative images, the T5-Small model uses a sequence-to-sequence (Seq2Seq) architecture with a weighted loss function to generate alt text. This translates the input of image information, context prompts, and GWAP-generated data (player-authored alt text and rating scores) into automatically generated alt text. The weighted loss function further treats alt text descriptions with higher player rating scores with higher importance. This feature was aimed at closely capturing the aggregated rating scores from the GWAP, translating them into a blueprint for the model to generate alt text descriptions close to average human-level quality. Importantly, 30% of the dataset was used to train the T5-Small Seq2Seq model, while 10% of this (approx. 3-4% of the whole dataset) was used for validation. Then, in inference, the model was applied to the entire dataset to generate an alt text description for all image-context pairs in the cleaned dataset.

3.5.2. ContextALT-O-matic: investigating the impact of context on quality

Whereas HumanALT-O-matic leveraged human ratings to approximate average human-level quality, this second model was designed to test the role of context itself by training it to consider context during alt text generation. Two otherwise identical variants were trained that differed only in whether the training data contained contextual information. For this, an image-as-tokens V2L model, which can natively parse images as well as textual data in its token stream, was fine-tuned on the GWAP-derived dataset via contrastive preference learning (Direct Preference Optimisation, DPO). The training objective was to prefer alt texts with higher human-given rating scores over those with poorer ratings.

The model selected for training, Qwen2.5-VL-3B-Instruct (Qwen Team Feb. 2025), already produced alt text-like outputs before fine-tuning. This makes the GWAP preference data well-matched for post-training: it shapes generation towards human-preferred alt texts while leveraging the base model's ability to read rich natural-language context. In other words, the model primarily needs to learn how to apply context correctly rather than to parse it. The GWAP-generated dataset is suitable for reward-modelling techniques such as Reinforcement Learning for Human Feedback (RLHF) and for contrastive preference objectives such as DPO (Rafailov et al. 2023), both of which steer models towards human-preferred outputs. DPO was adopted due to its simplicity, its stability on small- to medium-sized datasets, and its ability to train without a separate reward model, making it well-suited for the narrow task domain of alt text generation. Figure 3 below shows the architecture of ContextALT-O-matic.

The DPO dataset consists of pairs of image-context-alt text tuples: one with a higher human-given rating score and one with a lower rating score. Image and context are held constant within each pair, letting the model learn which differences in alt text lead to higher ratings. To examine whether training on context prompts aids context-aware alt text generation, two otherwise identical models were trained on the same DPO recipe, differing only in the exclusion of context prompts from the latter, which resulted in a **context-aware** and a **no-context** model. Both were trained with Low-Rank Adaptation (LoRA) for three epochs, at a learning rate of 1e-4. The goal of this setup was to evaluate whether context-aware training helps the model better utilise contextual information, rather than solely relying on its inherent understanding. While the base model can make use of contextual information without task-specific fine-tuning, subpar performance of the base model could improve through contrastive learning. Conversely, if the base model can already fully utilise context prompts, significant improvements are unlikely. To better understand the role of context during training, the altC context-presence (see Section 2.2) was measured through binary classification on the outputs of the context-aware and no-context models. To accommodate dataset sparsity, 90% of the GWAP-generated data were used for training and 10% held out for validation. The context-presence evaluation dataset was generated from a random subsample of image-context pairs that received no human-authored alt text in the GWAP.

4. Results: model performance evaluation

In order to validate the effectiveness of the models, we conducted a twofold evaluation: first, via a user study to assess whether we can approximate average human-level alt text quality while automating the process (RQ1, RQ2); and, second, via an investigation of the binary presence of our context definition elements (Section 2.2) in automatically generated alt text descriptions (RQ3). Thus, the research questions in Section 1 were addressed by our evaluation, with RQ1 and RQ2 relating to the HumanALT-O-matic model (Section 3.5.1) and RQ3 to the Context-ALT-O-matic model (Section 3.5.2).

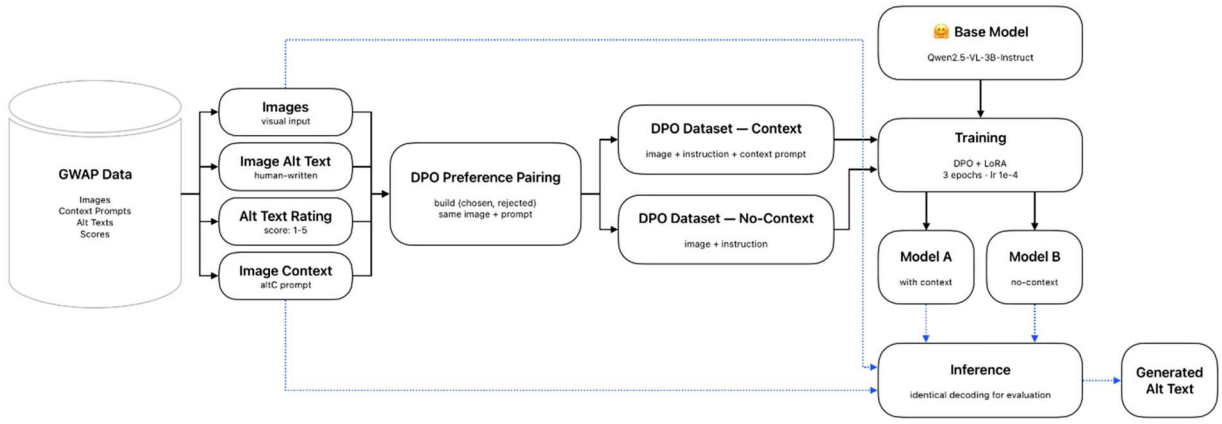

Figure 3. ContextALT-O-matic model architecture.

4.1. User study evaluation (RQ1, RQ2)

To address RQ1 and RQ2, we conducted a comparative user study evaluating the alt text outputs generated by the proposed pipeline. The system combines a routing component and a T5-Small-based generation model, fine-tuned and trained on the GWAP-generated data (human-curated alt text descriptions and rating scores), with comparisons made between a control configuration and a context-aware trained configuration. Both models generate captions from the same input which is formed of image processing data and image context. The same generation prompt was used for both models. The primary aim of the study was to assess whether our HumanALT-O-matic model can approximate average human-level alt text quality (RQ1). The study further examined the role of context-dependent routing in influencing alt text suitability (RQ2), making it the first model to attempt this **crucial routing pathway** towards improving web navigation via screen readers. The study design, data collection and analysis procedures are detailed in the following subsections.

4.1.1. Study design, sampling strategy and procedure

Ethics approval was granted by the researchers' institutional Research Ethics Committee (Ref: 41665-A-Feb/2025- 53701-1). The study adopted a quantitative, exploratory design to evaluate the impact of the context-driven GWAP output (alt text descriptions and peer ratings) on participants' quality perceptions of alt text. In this vein, participants were asked to complete an online survey where they rated (using a 1–5 Likert scale) a sample of alt text descriptions generated by the trained and the control models for the same image-context pairs. The sample comprised alt text descriptions for 20 image-context pairs generated by both the control and the trained models, resulting in 40 image-context-alt text tuples. The sampling criteria were as follows:

- Prioritise alt text descriptions with the highest rating scores in TagALTlong
- No duplicate images or contexts
- A good mix of context prompt elements
- A subset (30%) of image-context pairs was marked as 'eye candy' (decorative images) in the GWAP to enable the investigation of context-dependent routing and its effect on alt text generation
- A close population distribution

For trustworthiness, the alt text descriptions that had received the most rating scores by players of the GWAP were selected. To avoid redundancy, no image or context prompt was used more than once, and the selected context prompts presented a variety of context prompt elements (*e.g.* links, logos, non-functional images, or images found on a health/social media webpage). In line with RQ2, 30% of the image-context pairs selected for the sample had been marked as eye candy by players of the GWAP to illustrate the role of context-dependent routing between decorative ('eye candy') and non-decorative images within the pipeline. The objective was also to use a sample that represented the

Table 2. Distribution difference between sample and population.

	Overall mean rating	Overall std. dev.	Non-decorative mean	Non-decorative std. dev.	Decorative mean	Decorative std. dev.
Sample	3.2	0.63	3.16	0.66	3.29	0.58
Population	3.23	0.72	3.24	0.81	3.19	0.58

Table 3. Gender representation and age average difference between sample and population.

	Age average	Male representation	Female representation	Prefer not to say representation	Other representation
Sample	29.76	10 (58.82%)	6 (35.29%)	1 (5.88%)	0 (0)
Population	29.79	68 (60.71%)	39 (34.82%)	4 (3.57%)	1 (0.89%)

population well; therefore, the distribution in the sample aimed for a difference within two decimal points with the distribution in the population, while the average mean rating score was deemed more important than the standard deviation for this difference (see Table 2 below). As can be appreciated in this table, the sample is a close representative of the population, as a difference within one decimal point was achieved.

The survey adopted a within-subjects design, where every participant rated all 40 image-context-alt text tuples (20 unique image-context pairs; each pair had one alt text description generated by the control model and one by the trained model). The survey was administered via Microsoft Forms; specifically, participants were provided with a private link to access the survey on this platform. In line with the objective of the study, participants were able to access and complete the survey in a naturalistic, unsupervised setting in their own time and using their own devices. No manipulation or experimental intervention was introduced; instead, evaluation relied exclusively on data captured via survey completion sessions. All data were generated exclusively via participants' survey completions. Upon visiting the survey, they were first prompted to read the Participant Information Sheet (PIS) and Consent form, and following that, they were asked to provide electronic consent by agreeing to take part in the study. Only participants who provided consent were able to continue and interact with the survey. At the top of all pages of the survey, participants were provided with clear instructions on how to rate each alt text description for suitability based on each image and its specific context, which were the same instructions found in TagALTlong. On average, each survey session lasted approximately less than 20 min, depending on the participant.

4.1.2. Participants

The recruited participants were a subset of participants from our original pool of players who evaluated TagALTlong. Although these individuals are not domain experts, they had previously authored and evaluated alt text descriptions as part of our initial data collection phase. Their prior involvement ensured familiarity with the goals of the task and with typical qualities of suitable alt text. Further, using a small, purposive sample enabled us to obtain focused, comparative judgements between human-authored and model-generated outputs, leveraging participants' existing knowledge of the task at hand. This is in line with established practice in human-centred AI research, where depth of insight and task-specific familiarity are often prioritised over large sample sizes when the goal is formative evaluation (Shneiderman 2022) and it aligns with past research supporting the use of small, representative samples when the objective is to assess whether outputs meet users' expectations (Nielsen 2000). Participation was voluntary and no financial incentives were offered. The first researcher contacted former TagALTlong players, who had provided consent to be contacted for a follow-up study upon registering on TagALTlong. These players were invited via email to participate in this online survey, while aiming for a sample that closely represented the population in terms of gender and age distribution (see Table 3 below). Seventeen (N = 17) unique former TagALTlong players (13.6%) accepted the request to participate in the survey in June 2025.

The above table shows a fair per-gender representation of the population in the selected sample of participants and a difference within one decimal point in terms of average participant age. For clarity, gender has been used in this research in accordance with the definition of the World Health Organization (World Health Organization 2024).

4.1.3. Data normality and distribution

First, the distribution of the data was evaluated based on the assumption of normality to determine the appropriate statistical tests (parametric/non-parametric) for ensuring the validity of the findings. The study participants (N = 17) rated a total of N = 680 alt text descriptions split across two independent groups (Control: N = 340 rating scores; Trained: N = 340 rating scores) for 20 unique image-context pairs. To evaluate normality, the Shapiro-Wilk (Control – W: .877, $p < .001$; Trained – W: .904, $p < .001$) and Kolmogorov-Smirnov (Control – W: .177, $p < .001$; Trained – W: .174, $p < .001$) tests were used; both tests confirmed significant deviations from normality in both groups. The data were also not normally distributed at the image-context pair level, with 16 and 14 out of the 20 image-context pairs showing significant deviations from normality ($p < .05$) in the control and trained groups, respectively.

Therefore, non-parametric statistical tests were used for both the overall data and the paired testing at the image-context pair level, where the alt text generated by the trained model for each unique image-context pair was compared to its control model counterpart.

4.1.4. Training effectiveness on generated alt text quality (RQ1)

The performance of the model that was fine-tuned and trained on the GWAP-generated dataset compared to pure image processing in terms of approximating human-level quality alt text was investigated next. Specifically, the null hypothesis tested was as follows: 'The distribution of rating scores is the same across categories of Control and Trained'. To achieve this, the Mann-Whitney *U* test was used to compare the aggregated rating scores across the 20 image-context pairs in both groups (control and trained) independently (see Figure 4 below).

The above figure reveals significantly higher rating scores in the trained group (Mean rank = 383.92) compared to the control group (Mean rank = 297.08), indicating the strong, positive effect of training on the perceived quality of generated alt text descriptions. The null hypothesis was, therefore, rejected by the independent-samples Mann-Whitney *U* test ($U = 72{,}564$, $p < .001$), and a further null hypothesis was tested; i.e. 'The median of differences between Control and Trained equals 0', using a related-samples Wilcoxon Signed-Rank test (see Figure 5 below).

The distribution of differences shown in the above figure rejects the null hypothesis; i.e. the median of such differences equals 0; instead, revealing a strong, positive skewness, with 168 positive differences (higher rating scores in Trained) compared to 92 negative differences (higher rating scores in Control). Eighty ties, i.e. identical rating scores for the same image-context pair for both the alt text generated by the trained and the control (baseline) model, were observed, highlighting identical low rating scores across the two groups. These results indicate a significant net improvement rate of 64.6% (ignoring ties) between alt text descriptions generated by the trained model compared to the control model, highlighting the improved performance of the former, which was fine-tuned and trained on the GWAP-generated dataset, subsequently demonstrating the value of the dataset. Therefore, the related-samples Wilcoxon Signed-Rank test supports the results of the independent-samples Mann-Whitney *U* test, indicating significantly higher rating scores in the trained group ($Z = 5.803$, $p < .001$). However, it must be noted that although these tests were non-parametric, aligning well with the data not being normally distributed (see previous section), a limitation with non-parametric tests is that they make no assumptions about the distribution of the population. This was addressed in this work by the selection of a sample that is representative of the population based on the strategy outlined in Section 4.1.2.

Accordingly, the effectiveness of training was also investigated at the image-context pair level, i.e. the rating scores for the alt text descriptions generated by the control model for each image-context pair (C01-C20) were compared to their counterparts from the trained model (T01-T20). To achieve this, Cohen's d effect sizes were calculated for each CT pair to measure the magnitude and the direction of differences between quality perceptions of the two models' outputs. These training effect sizes were mapped into a forest plot alongside their 95% confidence intervals (CI) for each pair, which are included for replicability (see Figure 6).

The above plot is revealing in several ways. First, CT pairs are sorted from strongest to weakest absolute Cohen's d effect size values (absolute magnitude); thus, the bigger the difference between the control and trained model, the higher the CT pair is shown in the plot. Second, the line of no effect (i.e. the vertical barred line at zero in the plot) allows for a clear interpretation of when the perceived quality of the output of each model outperformed the other. Specifically, the trained model outperforms the control model when Cohen's d is negative (below zero) (11 cases), and vice versa (above zero) (9 cases). Further, the Wilcoxon Signed-Rank test, which is well-suited for paired data, was used indicating statistically significant ($p < .05$) and non-significant ($p \geq 0.5$) differences in rating scores for alt text descriptions generated by the trained and control models. Twelve out of twenty cases were statistically significant (blue circle points in Figure 6 above), with nine of those being cases where the trained model outperformed the control model. Therefore, the trained model substantially improved alt text descriptions generated by the control model in 75% of statistically significant cases (as per participants' quality perceptions), highlighting the effectiveness of using a human-curated output to train such models for generating better alt text. To further investigate the validity of the indicated magnitude and direction of the difference between quality perceptions across CT pairs, the three strongest Cohen's d effect sizes where the trained model was preferred over the control model and vice versa, as well as three illustrative non-statistically significant cases, were examined (see Table 4 below).

The qualitative data (alt text descriptions) in the above table highlight the most significant improvements in cases where the control model generated descriptions

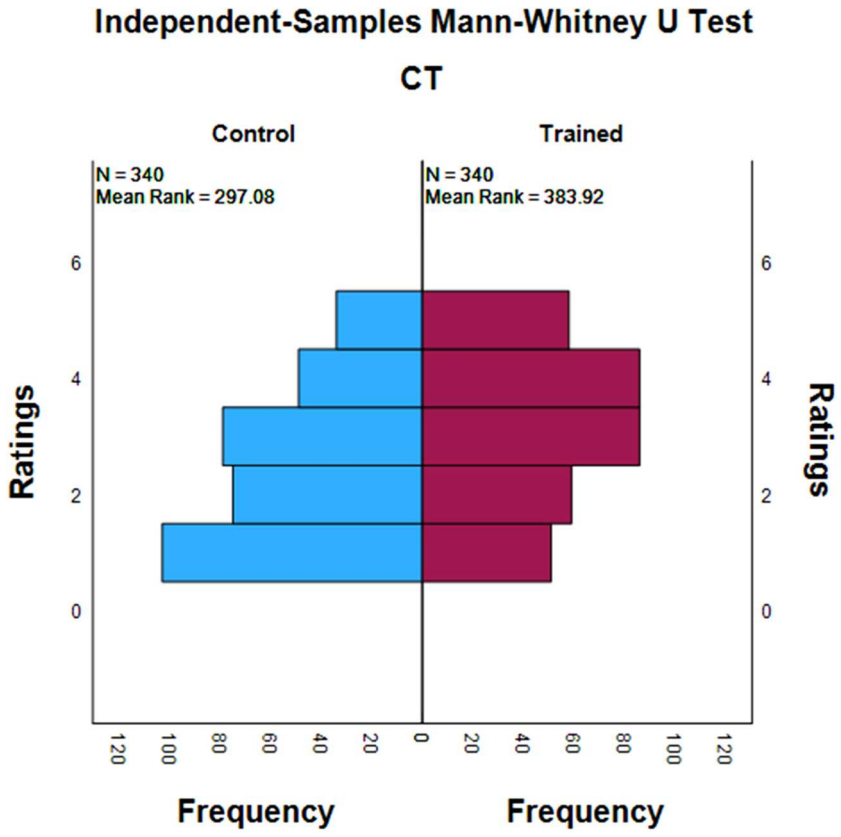

Figure 4. Distribution of rating scores across the control (left) and the trained (right) groups.

more closely resembling keyword tags than alt text (CT07, CT14) or non-specific descriptions (CT10), wherein the trained model integrated rich contextual information. Conversely, alt text generated by the control model was preferred in cases where the trained model added confusing or irrelevant information that did not match what the image depicted (CT15, CT01) or when the model misinterpreted an important image, such as a graph, for a decorative ('eye candy') image (CT02). Whereas it appears that the trained model was preferred over the control model (and vice versa) based on the former's ability to correctly interpret and integrate rich contextual information in the alt text descriptions, it is important to note that the three cases where the control model was preferred were also the only statistically significant cases, as opposed to 9 out of 11 total cases for the trained model being statistically significant. In non-statistically significant cases, it appears that contextual information integrated by the trained model was trivial (CT16) or violated instructions given to participants, such as including 'This is a picture' (CT11). There was also an image-context pair (CT05) where the Wilcoxon test confirmed no statistical difference ($Z = .000$; $p = 1.000$), despite the pair appearing as a case where the control model was preferred in the forest plot (see Figure 6) based on its Cohen's d effect size (.045). It is also important to note that in cases where the trained model was preferred, the alt text descriptions generated by the control model were lowly rated (average Likert score 1-2), which is when the magnitude of the difference between the two models is most evident. Conversely, in cases where the control model was preferred, the alt text already had acceptable ratings (average Likert score 3-4) before training.

For practical applications, the trained model showed promise for meaningfully improving accessibility, as it improved poor-quality alt text (average Likert scores 1.59-1.82) to acceptable alt text (average Likert scores 3.41–3.94). The trained model's potential to improve poor-quality descriptions generated by the control model has, thus, been highlighted, and the former was also preferred in 9 out of 12 (75%) statistically significant cases. It is therefore necessary to gather human-curated data (alt text descriptions and rating scores) at a larger scale to train such models to further investigate their potential to improve the quality of generated descriptions.

4.1.5. Context-dependent routing of decorative (eye candy) images (RQ2)

Finally, the role of context-dependent routing for decorative ('eye candy') images was investigated by

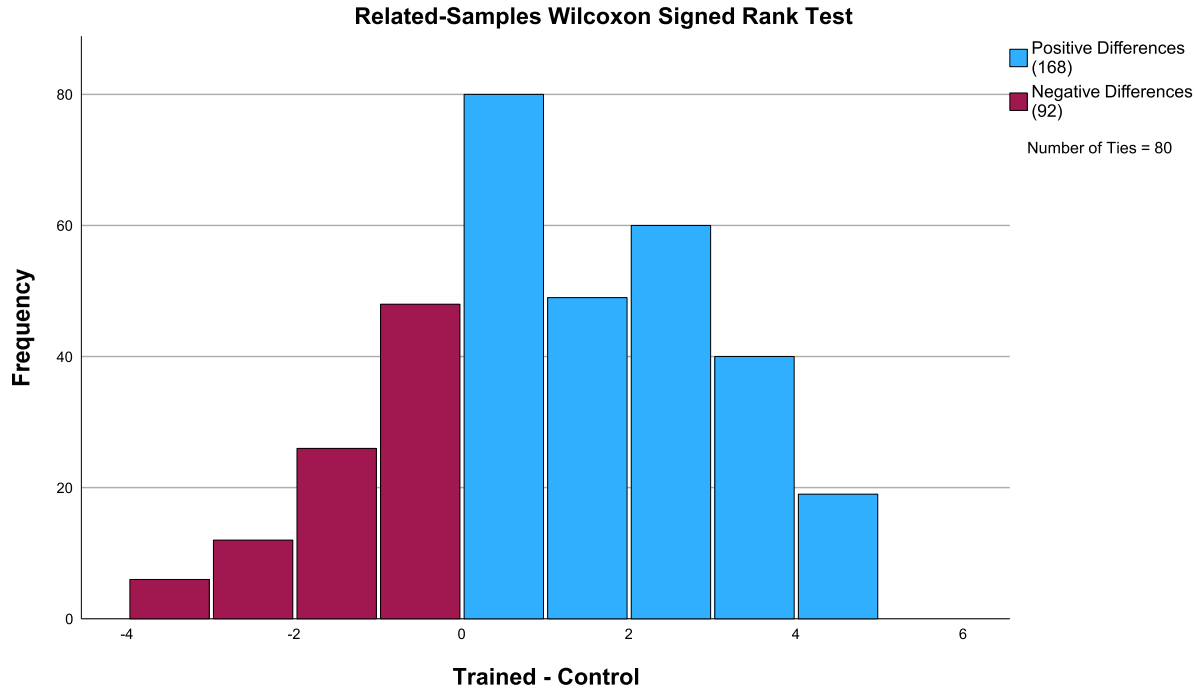

Figure 5. Distribution of rating score differences between the control and the trained groups.

analysing its downstream alt text quality. In this pipeline, routing is operationalised through the identification of image-context pairs marked as decorative in TagALTlong, which are handled differently during the generation process (see Table 5).

The table highlights how context-dependent routing influences the alt text outputs generated by the trained model for image-context pairs marked as decorative ('eye candy') in TagALTlong. In cases T17 and T18, the trained model produces outputs aligned with the intended handling of decorative images (i.e. indicating their decorative nature), resulting in mean ratings comparable to those observed in TagALTlong. In contrast, in case T20, the model generates a more descriptive

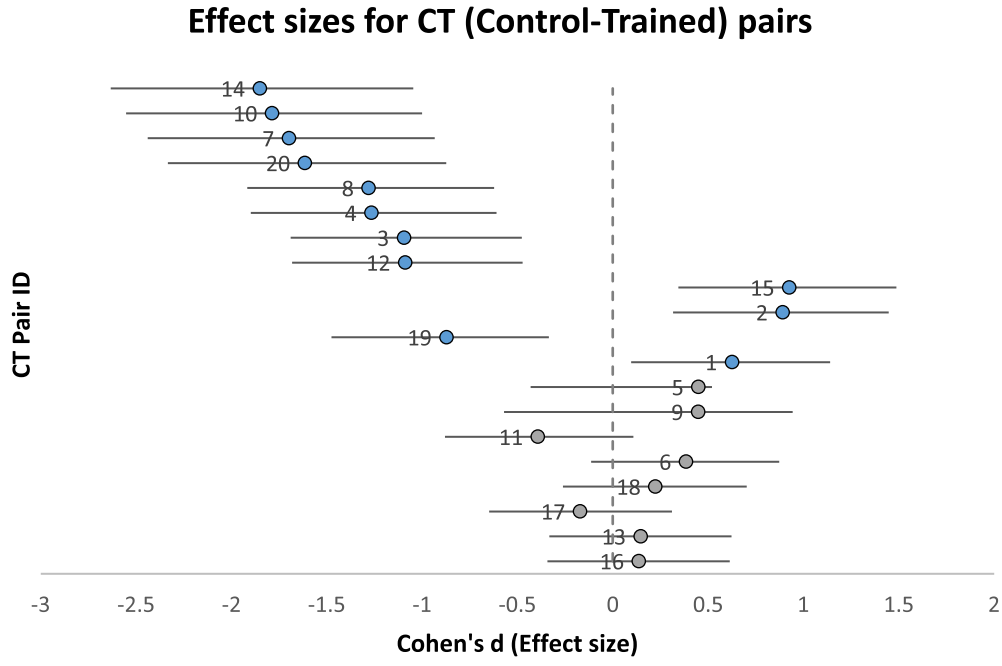

Figure 6. Forest plot of Cohen's d effect sizes for CT (Control-Trained) pairs with error bars representing 95% CIs (sorted by absolute magnitude). Colour-coded for statistical significance (blue: $p < .05$; grey: $p \geq 0.5$).

Table 4. Comparison of CT pairs with the most notable model performance differences.

Category	Pair ID	Control alt text	Trained alt text	Control avg. Likert score	Trained avg. Likert score	Wilcoxon Z	Cohen's d	Significance (*p*)
Trained > Control	CT14	Person, tree, sky	A picture of a young woman holding a watermelon slice in front of her face.	1.59	3.94	−3.443	−1.850	<.001
	CT10	Two bottles of vitamin c3	Vitamin C and B12 tablets	1.82	3.71	−3.555	−1.786	<.001
	CT07	Person, bicycle, umbrella, road, sky	A man riding a bike in a valley	1.59	3.41	−3.360	−1.697	<.001
Control > Trained	CT15	Woman standing on a ledge with her head tilted	A young woman with her head tilted on a ledge in front of a social media page	3.12	2.06	−2.859	.926	.004
	CT02	Graph showing the number of students in each class	This photo is marked as 'Eye Candy'.	3.00	1.59	−2.699	.892	.007
	CT01	Coffee beans and a coffee pot	A photo of coffee beans showing how to make your own coffee.	3.82	2.94	−2.080	.626	.038
No difference	CT05	Zeimm & Rosen hand cream	This is a hand cream	2.71	2.65	.000	.045	1.000
Non-significant	CT11	Dog, sea, sky, rock	This is a picture of a dog running on the beach with his mouth open.	2.65	3.23	−1.667	-.392	.95
	CT16	Black and white photo of a tree	A black and white photo of a tree with a bright blue sky.	3.00	2.82	-.5667	.137	.571

output that was rated higher than its TagALTlong counterpart, suggesting that the routing mechanism may, in some instances, lead to alternative interpretations of image function. Across the remaining cases (T15, T16, T19), the generated outputs received lower mean ratings, indicating variability in how contextual information is incorporated into the generation process. Training without contextual inputs may lead to inconsistencies when context is introduced at inference time, which may contribute to the lower performance observed in the no-context condition. These inconsistencies suggest that the model's handling of contextual cues remains sensitive to the quality and scale of the training data. The results are deemed reliable, as the sample of image-context pairs that were marked as decorative in TagALTlong were selected for having been rated by the highest numbers of unique raters (see section 4.1.1), and they were subsequently rated by all 17 survey participants. However, the results should be interpreted as indicative of the potential of context-driven modelling to influence alt text suitability, rather than as evidence of consistent performance across all cases. These findings directly address RQ2 by demonstrating how context-dependent routing influences the suitability of generated alt text across different image-context conditions.

Finally, it should be noted that this analysis focuses on the suitability of the resulting alt text produced by the integrated pipeline, rather than on the predictive accuracy of the classifier in isolation. Nevertheless, to provide indicative performance of the routing component, we further report classification metrics computed using the same filtering and data preparation criteria applied during model training and validation. In this process, image-context pairs with strong agreement among GWAP participants were treated as indicative of decorative ('eye candy') images, while lower-agreement instances were excluded to reduce ambiguity. Under these conditions, the classifier achieved **a precision of 0.84, recall of 0.78, and F1-score of 0.81**. As this evaluation reflects the same assumptions used during training, it therefore provides an indicative rather than independent benchmark of classification performance. Overall, these findings provide initial support for the proposed approach of leveraging human-curated datasets to inform context-aware alt text generation. However, further evaluation using larger and more diverse datasets is required to assess the robustness and generalisability of the approach.

4.2. Context presence evaluation (RQ3)

The absence of context in training datasets has been highlighted as a key deficiency relating to the poor quality of alt text generated by SoTA V2L models (see also Section 2). To investigate whether structured context prompts can address this gap, two variants of our second model, ContextALT-O-matic (Qwen2.5-VL-3B-Instruct), were trained on the GWAP-generated dataset: one with contextual information included in the training data (context-aware) and one without (no-context). The contextual factor presence metric was used as a diagnostic measure to examine how contextual inputs influenced the content of model outputs, rather than as an independent indicator of quality. Both

Table 5. Illustrative examples of alt text outputs for image-context pairs marked as decorative (eye candy) in TagALTlong, demonstrating the effect of context-dependent routing on generated descriptions.

Pair ID	Generated alt text (trained model)	Context prompt	TagALTlong mean	Trained mean
T15	A young woman with a head tilted on a ledge infront of a social media page	This photograph is found on a social media webpage with the goal of selling. The intention of the image is to guide (e.g. instructions or steps).	3.46	2.06
T16	A black and white photo of a tree with a bright blue sky.	This painting is a button on a social media webpage with the goal of advertising. The intention of the image is to illustrate (e.g. showing a product).	3.5	2.82
T17	This painting is marked as 'Eye Candy' (decorative image)	This painting is found on an educational webpage with the goal of entertaining. The intention of the image is to complement (e.g. enhancing the text context).	3.56	3.59
T18	This photo is marked as 'Eye Candy' (decorative image)	This photograph is found on a travel webpage with the goal of advertising. The intention of the image is to illustrate (e.g. showing a product).	3.86	3.3
T19	The image is of a signpost on a street directing to a library. The signpost is on the left and there's a tree on the left.	This photograph is found on an education webpage with the goal of entertaining. The intention of the image is to illustrate (e.g. showing a product).	3.14	2.94
T20	A woman with blonde hair standing in front of an auditorium. Button Operation	This photograph is a button on a health webpage with the goal of informing. The intention of the image is to elicit emotion (e.g. compassion).	2.2	3.4

Note: This table is illustrative and not intended as a quantitative evaluation of classification performance.

variants were otherwise identical in architecture, training procedure, and hyperparameters (see Section 3.5.2), ensuring that any observed differences in output can be attributed to the presence or absence of contextual information during training. To measure the extent to which context is reflected in generated alt text, both model variants were provided with the context prompt when generating alt text for this evaluation. This ensures that the comparison reflects differences in how each model utilises contextual information as a result of training, rather than differences in access to context during generation. This design supports the analysis of RQ3 by isolating the effect of context-aware training on the incorporation of contextual cues in generated alt text.

The first researcher carried out an explanatory analysis of how the binary presence of each element of the altC definition in the descriptions produced by each variant is reflected in the outputs. For each generated alt text description, each of the five altC elements (image type, webpage topic, webpage purpose, image function, and image intent) is assigned a value of one if present or zero if absent, yielding a context presence score ranging from zero (no elements present) to five (all elements present). We further validated whether context presence scores correlated with human-perceived alt text quality (see Figure 7 below).

Deferring to the above figure, a positive monotonic relationship was observed between higher context presence scores and higher quality perceptions of alt text descriptions by players. This was further substantiated with the calculation of Spearman's rho (0.467, $p < .001$), which is preferred for ordinal data, indicating a positive, moderate and statistically significant correlation between contextual factor presence and human quality ratings, indicating that outputs incorporating contextual elements were more likely to be rated as higher quality by human evaluators. With the validation of the context presence metric, we performed context presence evaluation on a held-out set of 399 image-context pairs that received no human-authored alt text in the GWAP, ensuring that the model was evaluated on unseen inputs. The average context presence score of the no-context variant was 1.64, while the context-aware variant achieved an average score of 3.22, indicating that the model learns to incorporate contextual cues when they are present during training, rather than relying solely on its pretrained visual and linguistic knowledge. Table 6 below demonstrates this shift. Descriptions with low context presence scores (0-2) decreased from 339 to 82, while those with high scores (3-5) increased from 60 to 317, corresponding to a 64-percentage point reduction in alt text with low context presence.

Additionally, Table 7 provides illustrative examples of this effect across representative image-context pairs. These results indicate that structured context prompts, when included in the training data, meaningfully improve the ability of the model to generate alt text that reflects contextual attributes from the altC framework.

5. Concluding discussion and future work

This section discusses the overall findings and presents the identified implications and contributions of this work. It also highlights our study limitations, as well as avenues for future work.

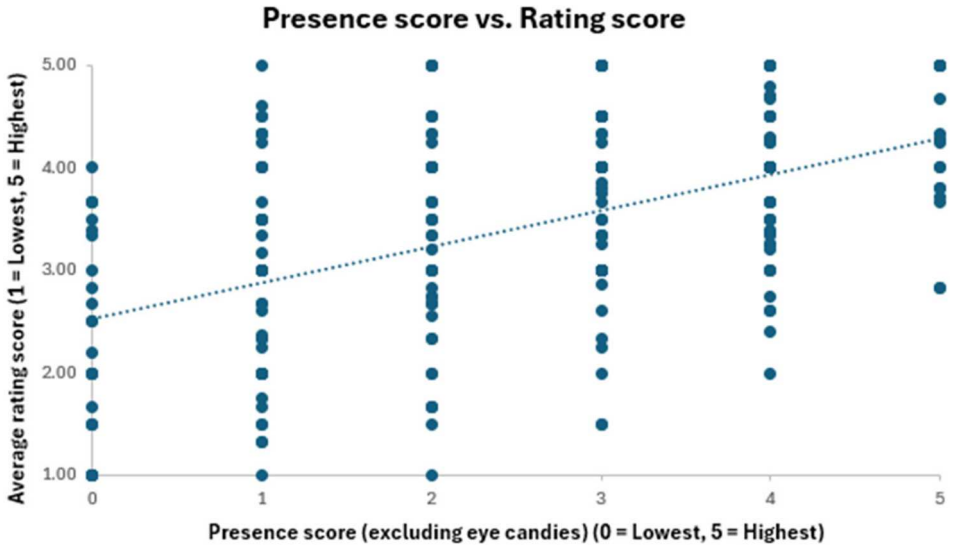

Figure 7. Correlation between context presence score and average rating scores for player-authored alt text from the GWAP dataset.

5.1. Overall findings and contributions

This study set out to examine whether a context-driven approach can meaningfully improve the automated generation of suitable alt text. By integrating a structured definition of alt text context (Section 2.2) into a reproducible transformer-based modelling pipeline, we demonstrated clear benefits for approximating human-level alt text quality, while automating the generation of alt text through ML, even when using modest model architectures within a reproducible pipeline (including a routing component and a T5-Small generator). Therefore, the aim of this work was not to introduce new SoTA models, but to demonstrate that incorporating an explicit context definition as model input under controlled conditions leads to measurably improved human-perceived alt text quality. Specifically, across our evaluation, the model (HumanALT-O-matic) trained with contextual data was perceived to be of significantly higher quality by human raters than its control counterpart (Mann-Whitney $U = 72.564$, $p < .001$; net improvement rate of 64.6%). It must be noted that a comparison with SoTA V2L captioning models, such as PaLI-17B and InternVL-Chat, was not feasible as these models were only evaluated on Computer Vision (CV) benchmark metrics (e.g. CIDEr, VQAv2 ZS acc), and as such, they can only measure how closely alt text can resemble captions in benchmark datasets, or how many questions can be answered about the content the image depicts. This was outside the scope of this work, which was to evaluate the suitability and contextual richness of alt text. Furthermore, while the IDEFICS model was trained on a context-rich dataset (CIDEr score 91.8), this was prone to limitations relating to being based on a gargantuan web-scraped dataset, and the generated alt text was not measured for suitability.

Table 6. Context presence scores for alt text generated by the context-aware and no-context variants of ContextALT-O-matic (N = 399 per variant).

Context presence score	No-Context (n)	Context-Aware (n)	Improvement (Δ n)	Improvement (Δ % points)
0	34	4	−30	−7.5
1	151	18	−133	−33.3
2	154	60	−94	−23.6
3	48	158	+110	+27.6
4	10	121	+111	+27.8
5	2	38	+36	+9.0
Total	399	399	N/A	N/A
Average score	1.64	3.22	N/A	N/A

Further, in this work, we defined and used a set of evaluation measures, including human-perceived alt text quality, training effectiveness, and context presence in alt text, owing to captions and alt text being fundamentally different (Ramos et al. 2023), and the need for alt text to draw on contextual information beyond the visual depiction in an image (Chen et al. 2023). These were applied for the subsequent performance evaluation of the models. Specifically, we distinguished between outcome-based validation and explanatory analysis, whereby human-perceived quality ratings served as the primary outcome measure, directly assessing the perceived suitability and usefulness of generated alt text. Differences in ratings between context-aware and non-context-aware outputs, therefore provided empirical evidence regarding whether structured contextual inputs improve human-perceived quality. The contextual factor presence analysis was used as an

Table 7. Paired examples of alt text generated by the context-aware and no-context variants of ContextALT-O-matic where the context-aware variant scored higher, lower, or equal.

Alt text of No-Context variant	Alt text of Context-Aware variant	No-Context presence score	Context-Aware presence score
A simple line drawing of an airplane in flight, with blue dashed lines indicating speed or motion.	An airplane icon with blue dashed lines indicating speed on a health website button designed to inform and evoke compassion.	0	5
A golden retriever wearing glasses sits on a desk, looking at an open magazine featuring a photo of another dog.	A golden retriever wearing glasses sits attentively at a desk, looking up at the camera while an open magazine lies in front of it. This image is used as a button on a business website to advertise a product or service related to pet care or lifestyle.	1	5
A street sign indicating directions to a library and shoppers' parking areas, with arrows pointing left and right respectively.	Sign pointing to the left for library parking and up for shoppers parking, with recycling symbol.	4	3
A fencer in a white uniform and mask, lunging forward with their epee, ready to strike. The background shows a blurred audience and fencing equipment.	A fencer in full gear, ready to strike, with a playful twist added to the image.	2	1
A person's feet stand on a sidewalk with the words 'DREAM UNTIL IT'S YOUR REALITY' written in blue chalk, emphasising the importance of pursuing one's dreams.	A person's feet stand on a sidewalk, with the words 'DREAM UNTIL IT'S YOUR REALITY' written in blue chalk above them. The image evokes a sense of motivation and aspiration.	5	5
Neon sign spelling 'COFFEE' on a concrete pillar in an urban setting.	Neon sign that reads 'COFFEE' mounted on a concrete pillar, likely indicating the presence of a coffee shop or cafe.	0	0

explanatory measure rather than as an independent indicator of quality. This analysis quantifies the extent to which predefined contextual elements appear in generated descriptions and is intended to illuminate how contextual conditioning influences output. Importantly, we observed a moderate positive association between contextual factor presence and human quality ratings (Spearman's rho = 0.467), suggesting that descriptions incorporating structured contextual elements were more likely to be rated as higher quality. This supports the interpretation that contextual modelling contributes meaningfully to perceived alt text suitability, while avoiding the assumption that factor presence alone determines quality.

These findings support the central premise of this work: that alt text is not merely a literal description of visual content, but a communicative artefact shaped by purpose, user needs and situational context. Embedding these contextual attributes into the modelling process yields measurable improvements, reinforcing the importance of integrating human-centred considerations into automated accessibility solutions. Beyond these empirical findings, this work also contributes conceptually and methodologically to accessibility-focused research. First, it presents a complete, reproducible, context-integrated pipeline for alt text generation that operationalises a formal definition of alt text context. The definition itself was originally proposed in our work elsewhere, but its independent application and evaluation here provide the first evidence that such a conceptual framework has practical value when embedded into model design. Second, the work demonstrates how lightweight transformer architectures such as BERT and T5-Small can be effectively adapted for specialised accessibility tasks. This highlights that substantive improvements in alt text quality do not require excessively large or proprietary models; rather, they require architectures that incorporate contextual signals aligned with accessibility principles. It is thus noted that this work does not aim to benchmark against large proprietary models (e.g. GPT-4o, Claude 3.5, or Gemini), and its focus is instead on demonstrating the value of structured contextual modelling within transparent, reproducible pipelines using architectures that are publicly available (e.g. BERT, T5-Small). Finally, the study proposes an evaluation approach that bridges technical performance and human-centred quality considerations, an area still underdeveloped within automated description research, thereby offering a framework that future studies may adopt or extend.

5.2. Limitations and future work

Our work presents limitations that should also be acknowledged. The alt text context definition that underpins this work, while fully restated in this paper to ensure transparency and completeness, is itself undergoing peer review, and its final form may evolve. This dependency, however, does not affect the interpretation of the present findings, although future refinements may warrant revisiting some modelling assumptions. Additionally, the dataset used for fine-tuning and evaluation was derived from a single annotation source (i.e. the GWAP), which may limit the diversity of linguistic styles, cultural interpretations, and contextual expectations captured in the data.

Broader sampling could reveal additional contextual nuances. Moreover, although we employed both quantitative and qualitative analyses, the evaluation did not include screen-reader users or people with visual impairments. As a result, the findings speak to technical and semantic improvements rather than lived accessibility experience.

Further, the role of the eye candy classifier was examined through its effect on downstream alt text quality supported by an initial classifier inference-based evaluation (precision: 0.84, recall: 0.78, and F1-score: 0.81). However, this evaluation was conducted using the same data preparation criteria applied during training and validation and therefore represents an indicative assessment rather than an independent benchmark of classification performance. A formal evaluation of the eye candy classifier against a dedicated benchmark dataset and benchmarking conditions therefore represents an important direction for future work. Additionally, the reliance on T5-Small also imposes limitations on generative richness, and the absence of multimodal visual input means that the model cannot capture image-level nuances that lie beyond textual context alone. We would also like to note that the image analysis pipeline uses established open-source components rather than the most recent proprietary systems. While this supports reproducibility and deployability, future work may explore integrating stronger segmentation and OCR models to improve absolute performance. Finally, and regarding ContextALT-O-matic, the evaluation focuses on comparing models trained with and without contextual inputs and does not include a formal baseline evaluation of the untrained model. As a result, it is not possible to fully assess potential regression effects relative to the base model though future work formally evaluating the untrained model as an independent baseline would be beneficial. This represents a limitation of the current study, although the comparative improvements observed between the evaluated configurations remain sufficient to address the primary research objective.

These limitations point naturally towards avenues for future work. First, a priority for subsequent research is to conduct user studies with visually impaired participants to assess whether the contextual improvements identified here translate into perceptible gains in clarity, usefulness and cognitive load. Second, opportunities can also arise to refine and extend our alt text context definition, potentially adding task specificity, user preferences or impairment-related parameters. Third, future modelling work could incorporate multimodal inputs, thereby enriching the contextual grounding of generated alt text descriptions. Fourth, there is also substantial potential for personalised alt text solutions that adapt to user characteristics and reading preferences, as well as for exploring the performance of larger generative models or multimodal large language models. Finally, deploying context-aware alt text generation in real-world platforms could surface practical considerations related to latency, explainability, and user trust, while our proposed context definition (Section 2.2) could be used in proprietary systems (GPT-4o, Gemini, etc.) to guide or augment the evaluation of contextual prompting strategies.

In conclusion, this work demonstrates that contextual grounding is essential for producing suitable, accessible alt text descriptions through automated methods. By integrating a structured definition of context into well-defined model pipelines, we show that even modest transformer models can generate higher-quality descriptions when informed by user-centred and situational factors. The results underscore the broader argument that accessibility technologies should be guided by human-centred principles rather than purely algorithmic optimisation. Through its conceptual framing, empirical validation and methodological approach, this work advances the foundation for developing human-aligned, context-aware alt text generation systems that better serve the needs of visually impaired users.

Note

1. If decorative (i.e. not functional), then 'is found' is automatically used in the prompt.

Acknowledgements

We want to sincerely thank all the participants for contributing their time and insight on alt text descriptions to our online survey. Their insight was invaluable to gauge the value of the human-generated dataset in improving automatically generated alt text.

Author contributions

CRediT: **Nikolaos Droutsas:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Validation, Visualization, Writing – original draft, Writing – review & editing; **Fotios Spyridonis:** Project administration, Supervision, Validation, Writing – review & editing; **Damon Daylamani-Zad:** Methodology, Resources, Software, Supervision, Validation, Writing – review & editing; **Philipp E. Glass:** Methodology, Resources, Software, Validation; **Gheorghita Ghinea:** Validation, Writing – review & editing.

Disclosure statement

No potential conflict of interest was reported by the authors.

Data availability statement

The data collection protocol using a GWAP is currently under review for separate publication; data and detailed collection protocols will be made publicly available upon acceptance and can be provided upon reasonable request. For the purposes of open access, the authors have applied a Creative Commons Attribution (CC BY) Licence to any Accepted Author Manuscript version arising from this submission.

ORCID

Nikolaos Droutsas http://orcid.org/0009-0002-1418-7542
Fotios Spyridonis http://orcid.org/0000-0003-4253-365X

References

Aliady, W., and M. Poesio. Aug. 2024. "Master the Linguistic Landscape: Puzzle Integration in a 3D NLP Game," *2024 IEEE Conference on Games (CoG)*, 1–8. https://doi.org/10.1109/CoG60054.2024.10645588.

Anonymous. Under review. *Painting Dragons and Dotting the Eyes: A Game-Based Study of How Context Shapes Human-Generated Alt Text*.

Apostolopoulos, I., E. Folmer, and G. Bebis. Jan. 2013. "Improving Accessibility of Virtual Worlds by Automatic Object Labeling." *Lecture Notes in Computer Science* 8034: 254–265. https://doi.org/10.1007/978-3-642-41939-3_25.

Bi, T., X. Xia, D. Lo, J. Grundy, T. Zimmermann, and D. Ford. May 2022. "Accessibility in Software Practice: A Practitioner's Perspective." *ACM Transactions on Software Engineering and Methodology* 31 (4): 1–26. https://doi.org/10.1145/3503508.

Birhane, A., V. U. Prabhu, and E. Kahembwe. Oct. 2021. "Multimodal Datasets: Misogyny, Pornography, and Malignant Stereotypes," *arXiv* (Cornell University). https://doi.org/10.48550/arxiv.2110.01963

Chamberlain, J., K. Fort, U. Kruschwitz, M. Lafourcade, and M. Poesio. 2013. "Using Games to Create Language Resources: Successes and Limitations of the Approach." In *Theory and Applications of Natural Language Processing*, edited by I. Gurevych and J. Kim, 3–44. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-642-35085-6_1.

Changpinyo, S., P. Sharma, N. Ding, and R. Soricut. Jun. 2021. "Conceptual 12M: Pushing Web-Scale Image-Text Pre-training To Recognize Long-Tail Visual Concepts," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3557–3567. https://doi.org/10.1109/cvpr46437.2021.00356

Chen, W., H. Hu, Y. Li, N. Ruiz, X. Jia, M. W. Chang, and W. W. Cohen. Dec. 2023. "Subject-driven Text-to-Image Generation via Apprenticeship Learning." *Advances in Neural Information Processing Systems* 36:30286–30305. https://dl.acm.org/doi/proceedings/10.55553666122

Chen, X., X. Wang, S. Changpinyo, A. J. Piergiovanni, P. Padlewski, D. Salz, S. Goodman, et al. May 2023. "PaLI: A Jointly-Scaled Multilingual Language-Image Model." *International Conference on Learning Representations (ICLR)*. https://openreview.net/forum?id=mWVoBz4W0u.

Chen, Z., J. Wu, W. Wang, W. Su, G. Chen, S. Xing, M. Zhong, et al. Jun. 2024. "InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 24185–24198. https://doi.org/10.1109/CVPR52733.2024.02290

Curtis, V. Dec. 2015. "Motivation to Participate in an Online Citizen Science Game: A Study of Foldit." *Science Communication* 37 (6): 723–746. https://doi.org/10.1177/1075547015609347.

Desai, K., G. Kaul, Z. Aysola, and J. Johnson. Dec. 2021. "RedCaps: Web-curated Image-text Data Created by the People, for the People," *Advances in Neural Information Processing Systems*, vol. 34, 13218–13232. https://proceedings.neurips.cc/paper/2021/hash/c990269324e93f619b883049b49f99e8-Abstract.html.

Devlin, J., M. W. Chang, K. Lee, and K. Toutanova. Jun. 2019. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, vol. 1, 4171–4186. doi:10.18653/v1/N19-1423.

Droutsas, N., F. Spyridonis, D. Daylamani-Zad, and G. Ghinea. Sep. 2024. "Web Accessibility Barriers and Their Cross-Disability Impact in ESystems: A Scoping Review." *Computer Standards & Interfaces* 92:103923. https://doi.org/10.1016/j.csi.2024.103923.

Hanley, M., S. Barocas, K. Levy, S. Azenkot, and H. Nissenbaum. Jul. 2021. "Computer Vision and Conflicting Values: Describing People with Automated Alt Text," *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 543–554. https://doi.org/10.1145/3461702.3462620

Harris, C. May, 2020. "ClueMeIn: Obtaining More Specific Image Labels Through a Game," ACL Anthology. https://aclanthology.org/2020.gamnlp-1.2/.

Kicikoglu, O. D., R. Bartle, J. Chamberlain, S. Paun, and M. Poesio. May, 2020. "Aggregation Driven Progression System for GWAPs." *ACL Anthology*. https://aclanthology.org/2020.gamnlp-1.11/.

Kreiss, E., C. Bennett, S. Hooshmand, E. Zelikman, M. R. Morris, and C. Potts. Dec. 2022. "Context Matters for Image Descriptions for Accessibility: Challenges for Referenceless Evaluation Metrics," *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 4685–4697. https://doi.org/10.18653/v1/2022.emnlp-main.309.

Laurençon, H., L. Saulnier, L. Tronchon, S. Bekman, A. Singh, A. Lozhkov, T. Wang, et al. Dec. 2023. "Obelics: An Open Web-Scale Filtered Dataset of Interleaved Image-Text Documents," Advances in Neural Information Processing Systems, vol. 36, 71683–71702. https://proceedings.neurips.cc/paper_files/paper/2023/hash/e28820063e2617f699131976241b777a-abstract-Conference.html.

Lee, H.-N., and V. Ashok. Apr. 2022. "Impact of Out-of-Vocabulary Words on the Twitter Experience of Blind

Users," CHI Conference on Human Factors in Computing Systems, 1–20. https://doi.org/10.1145/3491102.3501958.

Lee, K., M. Joshi, I. R. Turc, H. Hu, F. Liu, J. M. Eisenschlos, U. Khandelwal, P. Shaw, M. W. Chang, and K. Toutanova. Jul. 2023. "Pix2Struct: Screenshot Parsing as Pretraining for Visual Language Understanding." *Proceedings of the 40th International Conference on Machine Learning (ICML)*, Vol. 202, 18893–18912. https://proceedings.mlr.press/v202/lee23g.html.

Lengua, C., V. Rubano, and F. Vitali. Jan. 2022. "Aligning Accessibility Design to Non-disabled People's Perceptions," *022 IEEE 19th Annual Consumer Communications & Networking Conference (CCNC)*, 1–6. https://doi.org/10.1109/ccnc49033.2022.9700592

Leotta, M., F. Mori, and M. Ribaudo. Aug. 2022. "Evaluating the Effectiveness of Automatic Image Captioning for web Accessibility." *Universal Access in the Information Society* 22 (4): 1293–1313. https://doi.org/10.1007/s10209-022-00906-7

Lin, T.-Y., M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. 2014. "Microsoft COCO: Common Objects in Context." In *Lecture Notes in Computer Science*, 740–755. https://doi.org/10.1007/978-3-319-10602-1_48

Luccioni, A. S., and J. D. Viviano. Aug. 2021. "What's in the Box? An Analysis of Undesirable Content in the Common Crawl Corpus," *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 182–189. https://doi.org/10.18653/v1/2021.acl-short.24

Mack, K., E. Cutrell, B. Lee, and M. R. Morris. Oct. 2021. "Designing Tools for High-Quality Alt Text Authoring," *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility*, 1–14. https://doi.org/10.1145/3441852.3471207.

Madge, C. J. 2020. "Gamifying Language Resource Acquisition." Ph.D. dissertation, Dept. Comput. Sci., Queen Mary Univ. of London, London, UK. [Online]. https://qmro.qmul.ac.uk/xmlui/handle/123456789/68617.

Mangiatordi, A., and M. Lazzari. Jan. 2018. "Combined Use of Artificial Intelligence and Crowdsourcing to Provide Alternative Content for Images on Websites," 2018 15th IEEE Annual Consumer Communications & Networking Conference (CCNC), 1–6. https://doi.org/10.1109/CCNC.2018.8319312.

Miranda, D., and J. Araujo. Apr. 2022. "Studying Industry Practices of Accessibility Requirements in Agile Development," *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*, 1309–1317. https://doi.org/10.1145/3477314.3507041

Muehlbradt, A., and S. K. Kane. Mar. 2022. "What's in an ALT Tag? Exploring Caption Content Priorities through Collaborative Captioning." *ACM Transactions on Accessible Computing* 15 (1): 1–32. https://doi.org/10.1145/3507659.

Nguyen, N. C., H. V. Pham, Z. Wei, R. Thawonmas, P. Paliyawan, and T. Harada. Jun. 2019. "Using GWAP to Retrieve Informative Description for Ukiyo-e Images on Live Streaming Platform," *2019 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia)*, 156–157. https://doi.org/10.1109/ICCE-Asia46551.2019.8941600.

Nguyen, N. C., R. Thawonmas, P. Paliyawan, and H. V. Pham. Aug. 2020. "JUSTIN: An Audience Participation Game with a Purpose For Collecting Descriptions for Artwork Images," 2020 *IEEE Conference on Games (CoG)*, 344–350. https://doi.org/10.1109/cog47356.2020.9231771.

Nielsen, J. Mar. 2000. "Why You Only Need to Test with 5 Users." Nielsen Norman Group. https://www.nngroup.com/articles/why-You-only-need-to-test-with-5-users/.

Noreskal, L., G. Feuilloley, and S.-P. Charbel. 2024. "An AI Solution for Web Accessibility and Images Classification." *2024 IEEE Thirteenth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, 1–6. https://doi.org/10.1109/IPTA62886.2024.10755614.

Petrie, H., K. Höckner, and W. Rosenberger. 2022. "Digital Accessibility: Readability and Understandability." *in Lecture notes in computer science* 13342: 3–5. https://doi.org/10.1007/978-3-031-08645-8_1.

Poesio, M., J. Chamberlain, U. Kruschwitz, L. Robaldo, and L. Ducceschi. Apr. 2013. "Phrase Detectives." *ACM Transactions on Interactive Intelligent Systems* 3 (1): 1–44. https://doi.org/10.1145/2448116.2448119

Qwen Team. Feb. 2025. "Qwen2.5-VL Technical Report." *arXiv* preprint arXiv:2502.13923. doi:10.48550/arXiv.2502.13923.

Rafailov, R., A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, and C. Finn. Dec. 2023. "Direct Preference Optimization: Your Language Model Is Secretly a Reward Model." *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, 53728–53741. https://papers.nips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.

Raffel, C., N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. 2020. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer." *Journal of Machine Learning Research* 21 (140): 1–67. https://jmlr.org/papers/v21/20-074.html

Ramos, R., B. Martins, D. Elliott, and Y. Kementchedjhieva. Jun. 2023. "Smallcap: Lightweight Image Captioning Prompted with Retrieval Augmentation," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2840–2849. https://doi.org/10.1109/CVPR54622.2023.00278

Risi, M. 2025. "Comparative Study on Gen-AI Models to Write Alt Text." In *Technology for Inclusion and Participation for All: Recent Achievements and Future Directions*, edited by K. Mavrou and P. Encarnação, 217–223. Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-032-01628-7_27

Schuhmann, C., R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, et al. Dec. 2022. "LAION-5B: An Open Large-Scale Dataset for Training Next Generation Image-Text Models." *Advances in Neural Information Processing Systems* 35:25278–25294. https://proceedings.neurips.cc/paper_files/paper/2022/hash/a1859debfb3b59d094f3504d5ebb6c25-Abstract-Datasets_and_Benchmarks.html

Shen, Y., H. Zhang, Y. Shen, L. Wang, C. Shi, S. Du, and Y. Tao. Feb. 2025. "AltGen: AI-Driven Alt Text Generation for Enhancing EPUB Accessibility," *Proceedings of the 2025 International Conference on Artificial Intelligence and Computational Intelligence*, 78–83. https://doi.org/10.1145/3730436.3730449.

Shneiderman, B. 2022. *Human-Centered AI*. Oxford, UK: Oxford University Press. https://doi.org/10.1093/oso/9780192845290.001.0001

Steinmayr, B., C. Wieser, F. Kneißl, and F. Bry. Jul. 2011. "Karido: A GWAP for Telling Artworks Apart," *2011 16th International Conference on Computer Games (CGAMES)*, 193–200. https://doi.org/10.1109/CGAMES.2011.6000338.

Tuite, K. 2014. "GWAPs: Games with a Problem." In *FDG*.

von Ahn, L., S. Ginosar, M. Kedia, and M. Blum. Apr. 2007. "Improving Image Search with PHETCH." *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP*, IV-1209–IV-1212. https://doi.org/10.1109/ICASSP.2007.367293..

von Ahn, L., S. Ginosar, M. Kedia, R. Liu, and M. Blum. Apr. 2006. "Improving Accessibility of the Web with a Computer Game," *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, 79–82.. https://doi.org/10.1145/1124772.1124785.

W3C. 2025. *Understanding Success Criterion 1.1.1: Non-text Content | WAI | W3C*. Accessed February 24, 2025. https://www.w3.org/WAI/WCAG21/Understanding/non-text-content.html.

WebAIM. 2023. *WebAIM: The WebAIM Million - 2020 - An Annual Accessibility Analysis of the top 1,000,000 Home Pages*. Accessed September 29, 2023 from https://webaim.org/projects/million/2020.

WebAIM. 2024. *WebAIM: The WebAIM Million - The 2024 Report on the Accessibility of the Top 1,000,000 Home Pages*. Accessed April 30, 2024. https://webaim.org/projects/million/.

WebAIM. 2025. *WebAIM: The WebAIM Million - The 2025 Report on the Accessibility of the Top 1,000,000 Home Pages*. Accessed July 29, 2025. https://webaim.org/projects/million/.

World Health Organization. 2024. "Gender and Health." World Health Organization. https://www.Who.int/health-topics/gender.

Zong, J., C. Lee, A. Lundgard, J. Jang, D. Hajas, and A. Satyanarayan. Jun. 2022. "Rich Screen Reader Experiences for Accessible Data Visualization." *Computer Graphics Forum* 41 (3): 15–27. https://doi.org/10.1111/cgf.14519