Technologies d'assistance pour les personnes malvoyantes basées sur la vision: avancées, limites et perspectives

Aela Le Sommer, Panagiotis Papadakis, Christophe Lohr

IMT Atlantique, Lab-STICC, UMR CNRS 6285, team RAMBO, F-29238 Brest, France

Résumé

Bien que les méthodes actuelles d'image captioning, de video summarization et de Visual Question-Answering soient des solutions prometteuses pour aider les personnes malvoyantes et aveugles dans leur vie quotidienne, elles se heurtent à de nombreux problèmes (biais, hallucinations, pertinence des descriptions et réponses générées, etc.). Cet article tente donc de dresser un aperçu de l'état de l'art des technologies basées sur la vision et le langage pour les personnes malvoyantes, en identifiant les principales problématiques et en proposant des pistes d'exploration.

Mots-clés

Malvoyants, Visual Question-Answering, Image Captioning, Technologies d'assistance, Vision par ordinateur

Abstract

Although current methods of image captioning, video summarization and Visual Question-Answering are promising solutions to help visually impaired and blind people in their daily lives, they face many problems (bias, hallucinations, relevance of the descriptions and answers generated, etc.). This article therefore attempts to provide an overview of the state-of-the-art of vision-language-based technologies for visually impaired people by identifying the main issues and proposing exploration avenues.

Keywords

Visually impaired, Visual Question-Answering, Image Captioning, Assistive technologies, Computer vision

1 Introduction

Le handicap peut prendre différentes formes. Il peut être moteur, sensoriel, cognitif ou bien psychique. Quelles que soient ses formes, il demande des aménagements pour garantir l'égalité des chances et l'inclusion de chacun. Dans leur vie quotidienne, les personnes aveugles et malvoyantes sont confrontées à de nombreux défis. En effet, elles rencontrent des difficultés considérables lorsqu'il s'agit de se déplacer dans des environnements intérieurs et extérieurs inconnus. Il leur est souvent difficile de comprendre l'environnement qui les entoure, surtout dans des environnements extérieurs qui sont en constante évolution. Même évoluant dans un environnement familier, ces personnes font face à des difficultés, notamment lorsqu'il s'agit de localiser un objet ou d'évaluer des risques et des dangers potentiels.

Les technologies d'aide au déplacement pour les personnes malvoyantes couvrent un large éventail de solutions, allant des cannes intelligentes aux dispositifs de guidage par GPS, en passant par les assistants vocaux et les capteurs ultrasons pour détecter les obstacles. L'article [26] dresse un large panorama de ces innovations. Les progrès récents en intelligence artificielle (IA) ont permis le développement de technologies basées sur la vision par ordinateur, capables d'analyser l'environnement et de fournir des retours en temps réel. Ces technologies offrent une perception plus riche et contextuelle de l'environnement, allant au-delà de la détection d'obstacles. Dans ce but, de nombreuses technologies basées sur l'image captioning (IC), la video summarization (VS) et le Visual-Question Answering (VQA) couplées avec des technologies de text-to-speech (TTS) ont vu le jour pour aider les malvoyants à se déplacer de manière plus confiante dans l'environnement qui les entoure (c.f. Sec. 2).

Les tâches d'IC, de VS et de VQA ont respectivement pour but de décrire le contenu d'une image avec du texte, de générer un court résumé du contenu d'une vidéo, et de répondre à des questions basées sur une image. La tâche de TTS, quant à elle, permet de convertir un texte en paroles, ce qui est particulièrement utile, par exemple, pour adapter les approches basées sur l'IC aux besoins des malvoyants.

Initialement, ces trois tâches étaient traitées avec des méthodes basées sur des règles et des approches statistiques. L'essor du deep learning a transformé ces approches en intégrant des architectures neuronales, notamment des réseaux de neurones convolutionels et récurrents. Avec l'arrivée des Transformers, de nouvelles approches ont vu le jour, permettant d'améliorer la compréhension sémantique, la précision et la richesse des descriptions en combinant efficacement vision et langage. Enfin, suite à l'avènement des grands modèles de langage (LLM), de nombreux modèles initialement orientés uniquement sur le texte sont progressivement devenus multimodaux en intégrant le traitement de nouvelles modalités comme les fichiers pdf, les images, les vidéos, etc. Des modèles de vision-language (VLM) sont alors apparus. Ces modèles composés généralement d'un encodeur de vision, d'un module de connexion de modalité et d'un LLM sont capables d'apprendre de manière simultanée à partir d'images et de textes.

L'évolution récente rapide des techniques de descriptions d'images ou de vidéos et des techniques qui permettent aux utilisateurs de poser des questions sur des images ou des vidéos offre des perspectives prometteuses pour aider les

App. Mobile	Reconnaissance d'objets	Lecture de textes	Description de scène	VQA	Reconnaissance faciale	Nb Téléchargements
Be My Eyes (Be My AI)			✓	✓		1M+*
Lookout	✓	✓	✓	✓		500k+*
TapTapSee	✓	✓				500k+*
Envision	✓	✓	✓	✓	✓	100k+*
Seeing AI		✓	✓	✓	✓	100k+*
Supersense	✓	✓				50k+*
Aipoly	✓	✓				-
Oorion	✓	✓				-
VizWiz				✓		-
VoiceVision			✓			-

* sur *Google Play* en février 2025

TABLE 1 – Exemples d'applications mobiles d'assistance basée sur la vision pour les personnes malvoyantes.

personnes malvoyantes à se déplacer. Dans ce contexte, ce travail tente d'offrir un aperçu des technologies basées sur la vision et le langage pour les personnes malvoyantes (Sec. 2), en identifiant les principales problématiques rencontrées dans la littérature (Sec. 3) et en proposant des pistes d'exploration pour y répondre (Sec. 4).

2 Technologies basées sur la vision et le langage pour les malvoyants

Les technologies basées sur la vision et le langage ont connu des avancées significatives ces dernières années leur permettant de renforcer l'autonomie des personnes malvoyantes. Ces solutions exploitent des techniques de vision par ordinateur et de traitement du langage naturel qui répondent à un ou plusieurs besoins : décrire l'environnement, répondre à des questions sur une image, identifier des objets ou encore lire du texte à voix haute.

Les applications mobiles jouent un rôle clé dans l'accessibilité à ces technologies grâce à leur faible coût et leur portabilité. Actuellement, sur le marché, plusieurs applications mobiles destinées aux personnes malvoyantes utilisent des modèles d'IA pour analyser l'environnement via l'appareil photo et fournir des descriptions audio. Le Tableau 1 donne un aperçu des fonctionnalités et de la popularité de ces applications. Be My Eyes est l'application la plus téléchargée. Elle se distingue des autres par sa fonctionnalité de mise en relation avec des volontaire humains, en plus de l'IA, offrant une assistance plus fiable et personnalisée.

Certains dispositifs physiques complètent ces applications. Parmi eux, *OrCam MyEye*, un dispositif sous forme d'une mini-caméra à fixer sur une monture de lunette, peut lire du texte, reconnaître des visages et des objets. D'autres solutions, comme les lunettes intelligentes dotées de caméras et d'IA telles qu'*Envision Glasses*, permettent de trouver des objets et des personnes, de décrire des scènes et permettent aux utilisateurs de poser des questions sur des images.

En recherche, avec l'explosion récente des performances des modèles d'IC, de VS et de VQA, de nombreux travaux basés sur ces modèles ont vu le jour pour aider les personnes malvoyantes dans leur vie quotidienne et notamment pour les aider à se déplacer de manière plus confiante dans des environnements inconnus. Les techniques d'IC et le VQA, associées à des modules de TTS, sont largement utilisées pour assister les personnes malvoyantes. Ces technologies permettent respectivement de générer des descriptions des images en temps réel et de répondre à des questions par rapport à une image, facilitant ainsi la compréhension de l'environnement. Dans le domaine de l'aide au déplacement, plusieurs approches ont été proposées, notamment des systèmes embarqués dans des lunettes intelligentes [24] ou des applications mobiles [6]. Ces solutions reposent sur des architectures neuronales avancées, comme les LSTM [9] et des Transformers [14]. Plus récemment, l'émergence des VLMs [12] et des modèles de langage multimodaux pré-entraînés [28] a permis d'accroître les performances des systèmes. Ces modèles permettent non seulement de générer des descriptions plus détaillées, mais aussi d'analyser des scènes plus complexes en tenant compte du contexte. Cependant, ces modèles étant sujet à des hallucinations (c.f. Sec. 3.2), certains auteurs intègrent des modules complémentaires (OCR, détection d'objets, correction d'angle,...) afin d'améliorer leur fiabilité pour qu'ils puissent être utilisés par des personnes malvoyantes [6, 12].

Comme mentionné précédemment, les systèmes actuels d'assistance pour les personnes malvoyantes reposent sur des modèles d'IC et de VQA, couplés à un module de TTS pour transformer les descriptions textuelles en audio. Bien que ces approches aient montré leur efficacité, ces approches restent limitées par leur architecture en plusieurs étapes. Cette séparation peut entraîner une perte d'information et un délai de traitement plus important. Pour pallier ces limites, une nouvelle tâche a émergé : l'image-to-speech end-to-end. Contrairement aux systèmes traditionnels, ces modèles convertissent directement l'image en parole [18, 7]. Cependant, bien qu'innovante, cette tâche présente plusieurs limites. Tout d'abord, l'absence d'une étape textuelle intermédiaire complique la vérification et l'évaluation des modèles. De plus, l'entraînement de ces modèles nécessite de grandes quantités de données annotées associant directement image et audio, ce qui peut être très difficile et couteux à obtenir.

Cependant, les technologies présentées précédemment restent toutefois limitées à l'analyse d'images statiques. L'exploitation de flux vidéo joue un rôle crucial pour aider les personnes malvoyantes à naviguer dans des environnements complexes. En effet, en prenant en compte la dimension temporelle, elle permet de capter les changements dans

un environnement dynamique. Cette capacité est particulièrement précieuse pour les malvoyants qui doivent continuellement s'adapter aux changements de leur environnement et anticiper d'éventuels dangers. Malheureusement, peu de travaux exploitent des flux vidéo pour les aider à se déplacer plus sereinement dans leur environnement. Certains chercheurs s'intéressent néanmoins à cette approche. Par exemple, afin de capturer plus largement l'environnement qui entoure les personnes malvoyantes, [25] introduit VIEW-QA, un nouvel ensemble de données de Video Question Answering capturé à l'aide d'une caméra portable égocentrique à 360 degrés. D'autres encore, comme [22], développent des systèmes intelligents portatifs basés sur le flux vidéo pour aider les personnes malvoyantes dans leur quotidien, leur permettant, entre autres, d'identifier les obstacles présents sur leur chemin.

3 Identification de problématiques majeures

Actuellement, les approches proposées pour aider les malvoyants basées sur l'IC et le VQA font face à de nombreuses problématiques qu'il semble nécessaire de surmonter, ou de moins d'atténuer pour améliorer leur fiabilité. Cette section aborde quatre problématiques majeures, classées par ordre de priorité : (i) les biais des modèles, (ii) les hallucinations, (iii) la qualité des photos prises par des personnes malvoyantes et (iv) la complexité d'évaluation des modèles d'IC et de VQA.

3.1 Biais

Malgré les avancées majeures dans les domaines de l'IC et du VQA, ces modèles restent très sensibles aux données sur lesquelles ils sont entraînés. Ils sont donc susceptibles d'hériter des biais provenant de ces ensembles de données. L'article [30] classe ces biais en trois catégories :

1. **Biais de labellisation** : Ils découlent des déséquilibres présents dans les annotations des jeux de données. Cette catégorie fait référence aux classes ou aux mots qui apparaissent plus fréquemment que d'autres ou aux mots qui apparaissent ensemble de manière récurrente dans le jeux de données. Par exemple, si les annotations mentionnent souvent "voiture" mais rarement "vélo", le modèle pourrait ne pas signaler un cycliste approchant d'un passage piéton. Les modèles VQA ont tendance à exploiter les régularités statistiques entre les occurrences de réponses et certains schémas dans la question. Bien qu'ils soient conçus pour fusionner les informations des modalités visuelles et textuelles, ils répondent souvent en n'utilisant que la question.

2. **Corrélations trompeuses faites par les modèles** : Certains mots et éléments visuels (objets, arrière-plan, etc.) qui apparaissent régulièrement ensemble peuvent orienter la réponse sans que le modèle ne regarde en détail la question et l'image [5]. Par exemple, si un jeu de données associe régulièrement les passages piétons aux feux tricolores, un modèle pourrait annoncer automatiquement la présence d'un feu dès qu'il détecte un passage piéton, ce qui pourrait entraîner des situations dangereuses pour les malvoyants.

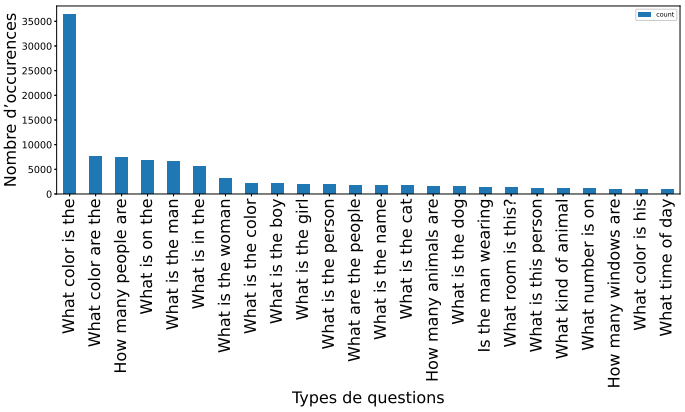

FIGURE 1 – Types de questions les plus posés dans VQAv2

3. **Biais sociaux** : Les paires image-texte peuvent contenir des préjugés humains qui entraînent des réponses discriminatoires ou stéréotypées.

D'autre part, à l'exception du jeu de données VizWiz [6], la majorité des autres jeux de données d'IC et de VQA n'ont pas été spécifiquement conçus pour répondre aux besoins des malvoyants. Par conséquent, ils contiennent des données qui ne sont pas toujours adaptées au développement de technologies destinées à cette population. En effet, les questions de nombreux jeux de données de VQA comme VQAv2 [1] impliquent une certaine connaissance préalable de la scène, ce qui, dans le cas d'applications pour les personnes malvoyantes, semble peu compatible. Par exemple, dans VQAv2, le type de question le plus fréquent est : "what color is the ..." (c.f. Figure 1), impliquant que l'utilisateur sache déjà qu'un certain objet est présent dans l'image.

3.2 Hallucinations

Les modèles d'IC et de VQA sont sujets à des hallucinations [19]. Dans ces domaines, les hallucinations correspondent à des cas où le contenu visuel d'une image et le texte généré pour la décrire ou répondre à une question sur cette image sont en inadéquation. De telles hallucinations affectent considérablement la fiabilité de ces modèles et, dans le cadre d'applications pour les personnes malvoyantes peuvent même s'avérer dangereuses.
Les hallucinations dans les modèles d'IC et de VQA peuvent provenir de plusieurs sources. Elles peuvent venir des biais dans les données d'entraînement, de l'incapacité des encodeurs de vision à ancrer avec précision les images, du désalignement entre les modalités visuelles et textuelles mais aussi pour les VLMs des LLMs, etc. Dans la littérature, plusieurs travaux ont tenté, en travaillant sur ces différentes sources d'hallucinations, de les atténuer. L'article [19] fait en particulier un état de l'art sur les hallucinations dans les VLMs dans lequel les auteurs présentent les différentes sources ainsi que diverses méthodes permettant d'atténuer ces hallucinations. Ces approches tournent principalement autour de l'optimisation des données d'apprentissage, du perfectionnement de divers modules et du post-traitement des sorties générées.
L'évaluation des hallucinations dans les modèles d'IC et

de VQA est une tâche complexe en raison de leur nature générative, qui peut entraîner l'ajout, la suppression ou la reformulation de détails par rapport à une sortie de référence. L'un des défis majeurs réside dans la difficulté d'établir, à l'aide des métriques existantes, une distinction claire entre les éléments fidèlement reformulés et ceux résultant d'hallucinations. En effet, ces modèles peuvent générer des descriptions plausibles mais incorrectes, rendant l'analyse encore plus complexe, notamment lorsque les différences sont subtiles ou contextuelles. C'est pourquoi, pour évaluer les hallucinations d'un modèle de VQA ou d'un VLM, certaines métriques comme CIEM [1], POPE[1] ou FGHE[1] formulent l'évaluation de l'hallucination comme une tâche de classification binaire qui invite les modèles à répondre par "Oui" ou par "Non". Cependant, les métriques comme CIEM, POPE, FGHE, CHAIR[1] ou encore NOPE[1] se focalisent sur la présence d'objets. À part certaines métriques comme FGHE et CIEM, elles ne prennent pas en compte les relations entre objets ni les attributs, ce qui peut conduire à des évaluations partielles. Les métriques reposant sur des modèles pré-entraînés, comme FAITHScore[1], utilisent des modèles de langage et de vision pour détecter les incohérences. Cependant, elles peuvent hériter des biais des modèles utilisés. Les métriques mentionnées ci-dessus comptent parmis les plus connues pour évaluer les hallucinations mais il en existent encore bien d'autres. Le dépôt GitHub[1] recense une grande partie d'entre-elles.

3.3 Qualité des images

Comme décrit dans la section précédente, de nombreuses technologies essaient de décrire et de répondre à des questions à propos d'une image prise directement par une personne malvoyante. Cependant, la plupart du temps, les images prises par ces personnes sont de mauvaise qualité et ne capturent souvent que partiellement la cible visée. En effet, de par leur déficience visuelle, il leur est très difficile, voire impossible, d'inspecter visuellement les images capturées et d'en garantir ainsi une certaine qualité et pertinence. Malheureusement, ces images de mauvaise qualité peuvent amener les modèles à générer des réponses peu fiables. La plupart des modèles actuels génèrent des réponses directes sans évaluer la suffisance des informations. En effet, peu de modèles d'IC et de VQA sont capables d'indiquer qu'ils ne peuvent pas répondre à la question en raison de la qualité de l'image ou de l'absence d'indices visuels nécessaires pour répondre à la question [28, 23].
Le tableau 2 met en évidence la diversité des jeux de données utilisés pour l'IC et le VQA, en termes de sources d'images et de types de scènes capturées. Il souligne notamment le caractère unique de VizWiz qui est l'un des seuls jeux de données contenant des images capturées directement par des personnes malvoyantes, contrastant avec les autres jeux de données, dont la plupart contiennent des images bien cadrées et de bonne qualité. Cependant, dans des applications réelles, l'hypothèse de bonne qualité des images ne tient plus. Un modèle entraîné uniquement sur des images de bonne qualité risque donc de mal généraliser sur ces cas réels et de fournir des réponses erronées. Paradoxalement, pour obtenir de bonnes performances en IC et en VQA, il est crucial de disposer de descriptions riches et précises, ce qui, dans le cas d'application pour les malvoyants, peut s'avérer compliqué. En effet, les images prises par ces dernières sont souvent de faible qualité, rendant l'annotation plus difficile et parfois moins détaillée.

1. https ://github.com/lhanchao777/LVLM-Hallucinations-Survey

3.4 Difficultés d'évaluation des modèles

L'évaluation des modèles d'IC et de VQA est particulièrement complexe en raison de la nature générative de ces modèles. En effet, une même image peut être décrite de multiples façons, et un modèle peut produire des réponses très différentes en fonction du contexte, de la formulation de la question ou même de biais présents dans les données d'entraînement. Les métriques classiques pour évaluer ces modèles comme ROUGE, BLEU, METEOR ou CIDEr [8] qui mesurent la correspondance entre la sortie du modèle et une ou plusieurs références humaines sans toujours capturer la pertinence sémantique ni l'utilité réelle d'une prédiction, semblent donc peu adaptées à cette nature générative. Pour pallier ces limites, d'autres métriques sont alors apparues. SPICE [8] par exemple évalue les descriptions en analysant leur contenu sémantique et leur relation avec la structure de l'image. BERTScore [8], basé sur des représentations de texte obtenues via des modèles de langage, permet de mieux mesurer la similarité sémantique entre une légende générée et une référence humaine. CLIPScore [8], quant à lui, exploite le modèle multimodal CLIP pour évaluer dans quelle mesure une légende est visuellement pertinente par rapport à l'image. Malgré ces avancées, ces métriques ne prennent pas en compte les biais du modèle, les hallucinations ou l'adéquation aux besoins des utilisateurs finaux. L'évaluation humaine reste donc indispensable. Une approche hybride combinant métriques traditionnelles et évaluation humaine permet d'obtenir un meilleur aperçu des performances réelles de ces modèles [13, 27].

4 Pistes d'exploration

Dans cette section, nous proposons quelques pistes d'exploration visant à atténuer les problématiques mentionnées dans la section précédente (Sec. 3). Compte tenu de la complexité de la tâche et des risques associés, nos futurs travaux viseront d'abord à comprendre la scène de manière la plus exhaustive possible et à informer simplement l'utilisateur, plutôt que de tenter de le guider directement.

4.1 Réduction de biais via des données synthétiques

Pour atténuer les biais évoqués dans la section 3.1, une approche consiste à générer des données synthétiques. En effet, générer de telles données permet de contourner les problèmes de confidentialité associés aux données du monde réel, d'équilibrer les jeux de données, de tester des scénarios rares et de réduire les coûts de collecte. Pour ce faire, on pourrait utiliser des plateformes de simulations 3D [4, 3] comme ThreeDWorld ou bien des modèles génératifs et de diffusions [21, 20]. Cependant, il semblerait que les images

Jeux de données	Tâches	Source des images	Nb images	Nb descript°/quest°	Apports potentiels pour des modèles destinés aux malvoyants
VizWiz [11]	IC, VQA	Photo prises par des malvoyants	+ 39k / +32k	+ 195k / +32k	Questions et images réelles issues d'utilisateurs malvoyants
MSCOCO [11]	IC	Flickr	+123k	+ 616k	Grandes diversités d'environnement (intérieur, extérieur, objets courant), Diversité des annotations
Flickr30k [11]	IC	Flickr	+ 31k	+ 158k	Images d'activités, d'événements et de scènes de la vie quotidienne
Conceptual Captions [11]	IC	Pages Web	+ 3.3M	+ 3.3M	Très grand jeux de données, Grande variété de types d'images (ex : images naturelles, de produits, dessins animés, dessins, etc.)
NoCaps [11]	IC	OpenImages (validation + test)	+ 15k	+ 166k	Conçu pour tester la généralisation sur de nouveaux objers, ce qui peut être utile pour décrire des scènes inconnues
Visual Genome [11, 1]	IC, VQA	YFCC100M ∩ MS-COCO	+ 108k	5.4M descriptions de régions / 1.7M	Très dense au niveau des annotations : descriptions de régions, objets, attributs, relations, graphes de régions, graphes de scènes et VQA
GQA [1]	VQA	Visual Genome	+ 113k	+ 22M	Exploite les représentations sémantiques des questions et les graphes de scènes pour répondre à la question, permettant un raisonnement structuré
VQAv2 [1]	VQA	COCO	+ 204k	+ 1.1M	Biais réduit : Chaque question est associée à une paire d'images similaires qui donnent lieu à deux réponses différente
OK-VQA [1]	VQA	COCO	+ 14k	+ 14k	Images du monde réel nécessitant des connaissances externes pour répondre aux questions
TDIUC [17]	VQA	MS-COCO + Visual Genome	+ 167k	+ 1.6M	Grande diversité de type de questions (12), comprenant des questions absurdes pour forcer le système à raisonner sur le contenu de l'image.

TABLE 2 – Principaux jeux de données d'IC et de VQA sur des scènes du monde réel

synthétiques générées par des modèles d'IA induisent des hallucinations en quantité plus élevée et avec une distribution plus uniforme que les images naturelles [10].

4.2 Distinction des niveaux de gravité des hallucinations

Dans les cas d'application pour les personnes malvoyantes, il serait intéressant de pouvoir distinguer différents niveaux de gravité des hallucinations générées. En effet, pour une personne malvoyante, que le modèle n'identifie pas correctement la couleur d'un lampadaire dans une description de scène a un impact limité ; par contre, que le modèle ne reconnaisse pas une voiture qui arrive lorsque la personne souhaite traverser la route est beaucoup plus grave.

4.3 Sélection d'images

Bien que certaines études [6, 29] aient montré un certain succès sur la tâche d'IC dans de mauvaises conditions, il semble que cette amélioration ait ses limites. En effet, il semble difficile pour un modèle de générer des descriptions ou des réponses utiles sur des images de mauvaises qualités. Dans les cas d'application pour les malvoyants il semblerait plus judicieux que le système notifie l'utilisateur de la mauvaise qualité de sa photo et l'aide à en prendre une nouvelle plutôt que de prendre le risque de générer une description inexacte qui peut amener l'utilisateur à prendre une mauvaise décision.

Pour pallier ce problème, certains articles évaluent la qualité de l'image avant de la décrire ou de répondre à une question [28, 23]. Si l'image est jugée de bonne qualité, le modèle génère une description ou répond à la question, sinon l'utilisateur est informé des défauts détectés et est invité à prendre une nouvelle photo. Cependant, un tel procédé dans la vie quotidienne peut s'avérer long et laborieux.

Une autre piste qui pourrait être explorée consiste à demander à l'utilisateur de capturer l'environnement qui l'entoure en prenant une série de photos, puis d'utiliser un modèle pour décrire la scène ou répondre à ses questions sur ces images. De cette façon, on pourrait limiter le nombre de photos de mauvaise qualité qui devront être reprises. En combinant les techniques de VQA appliquées à une série d'images d'une même scène [2] et de sélection de photos parmi une série d'images (techniques spécialisées dans l'évaluation de la qualité esthétique des images, qui recherchent la meilleure photo parmi une série de photos)[15, 16], on pourrait envisager un modèle en deux étapes. D'abord, il filtrerait les images en ne conservant que les images de bonne qualité puis il répondrait aux questions en se focalisant sur les images de la scène portant les informations nécessaires pour y répondre.

4.4 Prise en compte de l'utilité des réponses dans l'évaluation des modèles

Dans les cas d'application pour les malvoyants, l'évaluation des performances d'un modèle ne devrait pas se limiter à des scores automatiques, mais devrait également prendre en compte l'utilité des descriptions et des réponses générées [12]. En effet, une légende peut être objectivement correcte tout en étant peu utile ou peu informative pour un utilisateur malvoyant. Ici, l'objectif n'est pas seulement d'obtenir des sorties précises, mais également de s'assurer qu'elles apportent une réelle valeur ajoutée aux utilisateurs.

Références

[1] M. Agrawal, A. Jalal, and H. Sharma. A review on vqa : Methods, tools and datasets. In *International Conference on Computer Science and Emerging Technologies*, 2023.

[2] A. Bansal, Y. Zhang, and R. Chellappa. Visual Question Answering on Image Sets. In *Computer Vision – ECCV*, 2020.

[3] P. Cascante-Bonilla, K. Shehada, J. Smith, S. Doveh, D. Kim, R. Panda, G. Varol, A. Oliva, V. Ordonez,

R. Feris, and L. Karlinsky. Going Beyond Nouns With Vision & Language Models Using Synthetic Data. In *IEEE/CVF International Conference on Computer Vision*, 2023.

[4] P. Cascante-Bonilla, H. Wu, L. Wang, R. Feris, and V. Ordonez. Sim VQA : Exploring Simulated Environments for Visual Question Answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[5] C. Dancette. *Shortcut Learning in Visual Question Answering*. PhD thesis, Sorbonne Université, 2023.

[6] P. Dognin, I. Melnyk, Y. Mroueh, I. Padhi, M. Rigotti, J. Ross, Y. Schiff, R. Young, and B. Belgodere. Image Captioning as an Assistive Technology : Lessons Learned from VizWiz 2020 Challenge. *Journal of Artificial Intelligence Research* , 73, 2022.

[7] J. Effendi, S. Sakti, and S. Nakamura. End-to-End Image-to-Speech Generation for Untranscribed Unknown Languages. *IEEE Access*, 9, 2021.

[8] S. Elguendouze. *Explainable Artificial Intelligence approaches for Image Captioning*. PhD thesis, Université d'Orléans, 2024.

[9] R. Faurina, A. Jelita, A. Vatresia, and I. Agustian. Image captioning to aid blind and visually impaired outdoor navigation. *IAES International Journal of Artificial Intelligence*, 12(3), 2023.

[10] Y. Gao, J. Wang, Z. Lin, and J. Sang. AIGCs Confuse AI Too : Investigating and Explaining Synthetic Image-induced Hallucinations in Large Vision-Language Models. 2024.

[11] T. Ghandi, H. Pourreza, and H. Mahyar. Deep learning approaches on image captioning : A review. *ACM Comput. Surv.*, 56(3), 2023.

[12] Y. Hao, F. Yang, H. Huang, S. Yuan, S. Rangan, J. Rizzo, Y. Wang, and Y. Fang. A Multi-Modal Foundation Model to Assist People with Blindness and Low Vision in Environmental Interaction. *Journal of Imaging*, 10(5), 2024.

[13] Y. Hao, F. Yang, H. Huang, S. Yuan, S. Rangan, J. Rizzo, Y. Wang, and Y. Fang. A Multi-Modal Foundation Model to Assist People with Blindness and Low Vision in Environmental Interaction. *Journal of Imaging*, 10(5), 2024.

[14] R. Harshitha, B. Lakshmipriya, and V. Krishnamurthy. TransEffiVisNet – an image captioning architecture for auditory assistance for the visually impaired. *Multimedia Tools and Applications*, 2024.

[15] J. Huang, Y. Gong, Y. Shi, X. Zhang, J. Zhang, and Y. Yin. Focusing on Subtle Differences : A Feature Disentanglement Model for Series Photo Selection. *IEEE Transactions on Multimedia*, 26, 2024.

[16] J. Huang, Y. Gong, L. Zhang, J. Zhang, L. Nie, and Y. Yin. Modeling Multiple Aesthetic Views for Series Photo Selection. *IEEE Transactions on Multimedia*, 26, 2024.

[17] K. Kafle and C. Kanan. An analysis of visual question answering algorithms. In *IEEE International Conference on Computer Vision*, 2017.

[18] M. Kim, J. Choi, S. Maiti, J. Yeo, S. Watanabe, and Y. Ro. Towards Practical and Efficient Image-to-Speech Captioning with Vision-Language Pre-Training and Multi-Modal Tokens. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2024.

[19] H. Liu, W. Xue, Y. Chen, D. Chen, X. Zhao, K. Wang, L. Hou, R. Li, and W. Peng. A Survey on Hallucination in Large Vision-Language Models. 2024.

[20] Z. Liu, H. Liang, X. Huang, W. Xiong, Q. Yu, L. Sun, C. Chen, C. He, B. Cui, and W. Zhang. SynthVLM : High-Efficiency and High-Quality Synthetic Data for Vision Language Models. 2024.

[21] F. Ma, Y. Zhou, F. Rao, Y. Zhang, and X. Sun. Image Captioning with Multi-Context Synthetic Data. *AAAI Conference on Artificial Intelligence*, 38(5), 2024.

[22] R. Meghana and M. Kulkarni. Smart system with descriptive video service and spoken dialogue system for visually impaired individuals. In *International Symposium on VLSI Design and Test*, 2024.

[23] K. Ohata, S. Kitada, and H. Iyatomi. Feedback is Needed for Retakes : An Explainable Poor Image Notification Framework for the Visually Impaired. In *IEEE International Conference on Smart Communities : Improving Quality of Life Using ICT, IoT and AI*, 2022.

[24] C. Rane, A. Lashkare, A. Karande, and Y. Rao. Image Captioning based Smart Navigation System for Visually Impaired. In *International Conference on Communication information and Computing Technology*, 2021.

[25] I. Song, M. Joo, J. Kwon, and J. Lee. Video Question Answering for People with Visual Impairments Using an Egocentric 360-Degree Camera. 2024.

[26] Y. Thoo, M. Jeanneret Medina, J. Froehlich, N. Ruffieux, and D. Lalanne. A large-scale mixed-methods analysis of blind and low-vision research in acm and ieee. In *International ACM SIGACCESS Conference on Computers and Accessibility*, 2023.

[27] P. Xu, W. Shao, K. Zhang, P. Gao, S. Liu, M. Lei, F. Meng, S. Huang, Y. Qiao, and P. Luo. LVLM-eHub : A Comprehensive Evaluation Benchmark for Large Vision-Language Models. 2023.

[28] B. Yang, L. He, K. Liu, and Z. Yan. VIAssist : Adapting Multi-modal Large Language Models for Users with Visual Impairments. 2024.

[29] L. Yu, M. Nikandrou, J. Jin, and V. Rieser. Quality-agnostic image captioning to safely assist people with vision impairment. In *International Joint Conference on Artificial Intelligence*, 2023.

[30] B. Zhu and H. Zhang. Debiasing vision-language models for vision tasks : a survey. *Frontiers of Computer Science*, 19(1), 2025.