Actes de CORIA-TALN-RJCRI-RECITAL 2025
Marseille, France, 30 juin au 4 juillet 2025

Accessibilité visuelle et éducation inclusive : Étude préliminaire sur la génération de textes alternatifs

Elise Lincker[1] Elisabeth Olamisan[2] Theodora Pazakou[2]
Michèle Gouiffès[2] Camille Guinaudeau[2] Frédéric Dufaux[3]
(1) CNAM, Cedric, Paris, France (2) Université Paris-Saclay, CNRS, LISN, Orsay, France
(3) CentraleSupélec, Laboratoire des signaux et systèmes, Orsay, France
prenom.nom@lisn.fr

RÉSUMÉ

Tout contenu numérique devrait garantir l'accessibilité visuelle en incluant des textes alternatifs aux images. En l'absence de système et de métrique d'évaluation adaptés, nous présentons nos recherches préliminaires sur la génération et l'évaluation de textes alternatifs, d'abord dans un contexte générique. Dans une démarche d'inclusion scolaire, nous mettons en lumière les limites des systèmes existants et les contraintes à prendre en compte pour envisager un système applicable aux manuels scolaires.

ABSTRACT

Visual Accessibility and Inclusive Education : A Preliminary Study on Alt-Text Generation

All digital content should ensure visual accessibility by including alternative text for images. In the absence of suitable generation systems and evaluation metrics, we present our preliminary research on the generation and evaluation of alternative text, first in a generic context. In order to foster inclusive education, we highlight the limitations of existing systems and the challenges that must be addressed to develop a system applicable to school textbooks.

MOTS-CLÉS : Texte alternatif, Accessibilité visuelle, Éducation inclusive.

KEYWORDS: Alternative text, Alt text, Visual accessibility, Inclusive education.

ARTICLE : **Accepté à** IA-ÉDU@CORIA-TALN 2025.

1 Introduction

L'accessibilité visuelle des contenus numériques constitue un enjeu majeur pour garantir l'inclusion des personnes déficientes visuelles (DV), aveugles et malvoyantes, dans l'accès à l'information. Il est donc essentiel de rendre les contenus visuels accessibles, soit en produisant des images tactiles pour les supports adaptés, soit, le plus souvent, en proposant un texte alternatif. Ce dernier est une description textuelle destinée à remplacer l'image pour les DV. Il peut être transmis avec l'ensemble du contenu textuel d'un document numérique à un dispositif d'accessibilité visuelle, tel qu'un lecteur d'écran ou une plage Braille. Des directives, comme celles de W3C pour l'accessibilité du Web [1], préconisent des descriptions concises, objectives et contextualisées qui mettent en évidence le contenu prédominant afin d'en faciliter la compréhension. Cependant, il n'existe pas de consensus universel sur les normes de l'accessibilité visuelle, et celles-ci sont peu appliquées.

1. https://www.w3.org/WAI/tutorials/images/

Actes de l'atelier Intelligence Artificielle générative et ÉDUcation :
Enjeux, Défis et Perspectives de Recherche 2025 (IA-ÉDU), pages 1–9.

Les manuels scolaires numériques restent majoritairement inaccessibles aux élèves DV, en raison d'une conception non inclusive et d'un manque de compatibilité avec les outils d'assistance (Castillan *et al.*, 2018). Bien que des recommandations existent pour intégrer l'accessibilité dès la conception [2], elles sont rarement appliquées, rendant nécessaires des adaptations a posteriori. Des organismes de transcription de documents pour un public DV transforment les contenus numériques en gros caractères, braille ou format audio comme DAISY. L'association autrichienne BookAccess propose notamment des directives spécifiques pour l'adaptation de manuels scolaires [3]. Le travail d'adaptation reste cependant largement manuel, ce qui le rend long, coûteux et dépendant d'une expertise humaine.

Cette étude préliminaire propose un cadre pour la génération et l'évaluation automatique de textes alternatifs, face à plusieurs défis : le contexte de l'image, les besoins spécifiques des DV, et l'absence de métrique d'évaluation appropriée. Nous identifions également les contraintes supplémentaires à prendre en compte en vue d'une application aux manuels scolaires.

2 Travaux connexes

Les études sur les préférences des DV mettent en évidence l'importance du contexte pour le texte alternatif. Il inclut notamment le scénario dans lequel apparaît l'image (l'objectif informationnel de l'utilisateur et la source de l'image) (Stangl *et al.*, 2021) ainsi que son contexte immédiat. Un même visuel peut ainsi nécessiter des descriptions différentes selon le scénario. Les DV préfèrent les descriptions contextualisées, jugées plus pertinentes et utiles lorsqu'elles répondent à leurs attentes en fonction du scénario (Gubbi Mohanbabu & Pavel, 2024; Kreiss *et al.*, 2022a).

Cependant, les modèles et métriques d'évaluation existants ne gèrent pas le contexte. Alors qu'une légende complète une image, le texte alternatif la remplace pour garantir l'accessibilité visuelle. Cette différence fondamentale empêche l'adoption directe des modèles de génération de légendes dans un cadre d'accessibilité, en raison du manque de prise en compte du contexte et de leur pré-entraînement sur des corpus non adaptés. À l'inverse, les modèles vision-langage (VLMs) tendent à produire des descriptions trop longues, comportant des spéculations et des hallucinations (Lincker *et al.*, 2025).

La génération de textes alternatifs reste un domaine encore peu exploré ou limité aux images scientifiques (Chintalapati *et al.*, 2022; Williams *et al.*, 2022; McCall & Chagnon, 2022). Pour des images de Twitter, Srivatsan *et al.* (2024) combinent CLIP (Radford *et al.*, 2021) et un réseau de mapping qui concatène les plongements visuels aux vecteurs contextuels, formant un préfixe pour la génération autorégressive d'une description. Une approche similaire a été adoptée par AutoAD (Han *et al.*, 2023) pour générer des audiodescriptions. Zur *et al.* (2024) affinent CLIP pour favoriser les descriptions aux légendes, mais sans amélioration de la pertinence visuelle, le contexte n'étant pas utilisé. Enfin, l'outil AltAuthor (Song *et al.*, 2025) aide les développeurs web à intégrer du texte alternatif. Il comprend la classification de l'image selon son rôle pour déterminer si elle nécessite un texte alternatif, la génération conforme aux normes le cas échéant, et une interface d'édition.

Toutefois, la transposition de ces approches au contexte scolaire reste difficile, en particulier en l'absence de jeux de données adaptés, et peu de travaux abordent la gestion des images pour l'éducation inclusive. Yadav *et al.* (2025) proposent une classification des illustrations dans les manuels d'Étude de la Langue selon leur rôle pédagogique : essentielles, informatives ou décoratives. Cette typologie oriente les adaptations en fonction du type de handicap et pourrait constituer un point de départ pour la génération de texte alternatif, à l'instar de l'approche d'AltAuthor.

2. https://www.firah.org/fr/access-man.html 3. https://www.bookaccess.at/

3 Méthodologie proposée

3.1 Données

Il n'existe actuellement aucun ensemble de données de contenus visuels pédagogiques adaptés aux DV. Des organismes proposent des adaptations de manuels scolaires à destination des élèves DV, mais ces contenus ne sont pas diffusés en raison de restrictions liées aux droits d'auteur. Dans d'autres domaines, des jeux de données ont été créés à partir de paires (image, texte) extraites automatiquement de pages web (Sharma *et al.*, 2018; Schuhmann *et al.*, 2022; Srivatsan *et al.*, 2024). Cependant, les descriptions sont souvent rédigées par des contributeurs non spécialistes de l'accessibilité visuelle, ce qui entraîne une qualité variable, rarement suffisante pour répondre aux besoins des DV. Par ailleurs, aucun corpus francophone de textes alternatifs n'est actuellement disponible.

Nos premières expérimentations s'appuieront sur les jeux de données en anglais Concadia (Kreiss *et al.*, 2022b) et AD2AT (Lincker *et al.*, 2025), qui se distinguent par la fiabilité de leurs textes alternatifs. Concadia contient 96 918 images issues de pages Wikipedia, accompagnées de leurs textes alternatifs, légendes et paragraphes contextuels. Les descriptions ont été filtrées pour garantir leur qualité. AD2AT est construit à partir d'audiodescriptions de films produites par des professionnels, dont la modalité a été transformée : de la vidéo avec audiodescription à l'image avec texte alternatif. Les audiodescriptions précédant l'image cible sont utilisées comme contexte. La partie AD2AT-MD contient 37 266 images extraites du dataset d'audiodescriptions MPII-MD (Rohrbach *et al.*, 2015). La deuxième section AD2AT-VIW est construite sur Visuals Into Words (Matamala & Villegas, 2016), un même film audiodécrit par dix descripteurs professionnels. Elle comprend 28 images, chacune associée à 1 à 10 textes alternatifs. Elle constitue une base de test pour évaluer la variabilité et la qualité des descriptions.

3.2 Génération

Les premières approches, reposant soit sur des modèles de génération de légendes, soit sur la formulation d'instructions à un VLM, ont montré leurs limites en contexte d'accessibilité, produisant souvent des descriptions génériques, trop longues ou spéculatives (Lincker *et al.*, 2025). Nous proposons d'affiner LLaVa (Liu *et al.*, 2023) par un nouveau réglage des instructions, en appliquant Low Rank Adaptation (LoRA) (Hu *et al.*, 2022) et Direct Preference Optimization (DPO) (Rafailov *et al.*, 2023). Cette approche permet d'éviter l'affinage complet ou l'apprentissage par renforcement, coûteux en ressources. DPO vise à aligner les LLMs sur les préférences humaines et nécessite que chaque exemple d'entraînement soit associé à une paire de références contrastées : l'une positive et l'autre négative.

Notre architecture, illustrée en Figure 1, repose sur un modèle LLaVa figé, dans lequel seuls les poids des matrices de faible rang A et B (matrices LoRA) sont ajustés pendant l'apprentissage. L'image et l'instruction (*prompt*), qui inclut le contexte, sont passés deux fois au modèle : avec une référence positive et avec une négative. La préférence du modèle est mesurée par la log-vraisemblance de chaque réponse : Score $= \log P_{\theta+\Delta_\theta}(\text{response} \mid \text{input})$, avec θ le modèle figé et $\Delta\theta$ les poids de LoRA. Le problème est posé comme une classification binaire, où la fonction de perte s'appuie sur la différence entre les deux scores : Loss $= -\log \sigma(\text{score\_ref\_pos} - \text{score\_ref\_neg})$ où $\sigma(x)$ est la fonction sigmoïde.

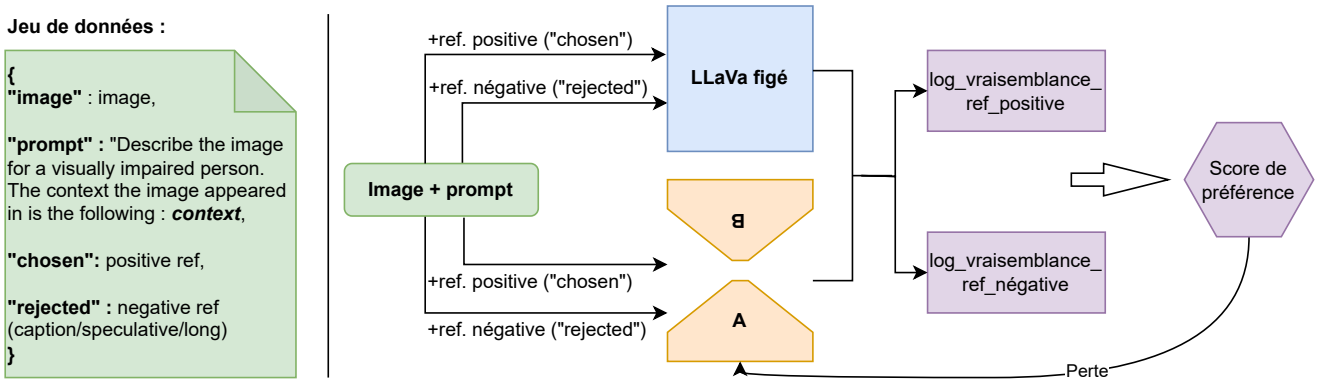


FIGURE 1 – Architecture proposée : Réglage des instructions de LLaVa avec LoRA et DPO

Pour l'entraînement, les références positives correspondent aux descriptions des jeux de données Concadia et AD2AT. Les légendes de Concadia servent de références négatives. En l'absence d'exemples négatifs pour AD2AT, nous en générons avec LLaVa. En ajustant uniquement les matrices LoRA, le modèle apprend à préférer des descriptions courtes, précises et adaptées au contexte, en rejetant les sorties trop longues ou spéculatives.

3.3 Evaluation

Les métriques classiques utilisées pour évaluer la génération de texte (BLEU (Papineni *et al.*, 2002), ROUGE (Lin & Hovy, 2003), METEOR (Banerjee & Lavie, 2005), CIDEr (Vedantam *et al.*, 2015), SPICE (Anderson *et al.*, 2016)) reposent sur la similarité avec des descriptions de référence produites par des humains. Cependant, elles ne tiennent compte ni du contexte d'apparition de l'image, ni des préférences des DV. De plus, une image dans un même contexte peut avoir plusieurs textes alternatifs acceptables, et ne doit pas nécessairement être évaluée en fonction d'une seule référence. Les résultats présentés dans (Kapur & Kreiss, 2024) soulignent la nécessité de développer une métrique sans référence qui prenne en compte les besoins spécifiques des DV. Les approches sans référence, telles que CLIPScore (Hessel *et al.*, 2021) et SPURTS (Feinglass & Yang, 2021), tentent de s'affranchir de la dépendance aux références. Toutefois, elles évaluent principalement la similarité brute entre l'image et le texte, sans prendre en compte la richesse de l'adéquation sémantique, la structure discursive, la cohérence logique ou les enjeux d'accessibilité (Hu *et al.*, 2023; Ahmadi & Agrawal, 2024). ContextRef (Kreiss *et al.*, 2024) constitue un banc d'essai pertinent, basé sur des paires image–texte contextualisées et des évaluations humaines, mais il se concentre sur une analyse de corrélation plutôt que sur une approche d'évaluation intégrée.

Pour pallier ces limites, nous proposons un cadre d'évaluation léger et interprétable. Le score final, **AltScore**, combine six dimensions : la *qualité d'ancrage visuel (QAV)*, la *cohérence discursive (CD)*, la *fluidité et cohérence linguistique (FCL)*, l'*imageabilité (IMG)*, la *pertinence contextuelle (PC)* et la *plausibilité des concepts de sens communs (PCC)*.

$$\text{AltScore} = \lambda_1 \cdot \text{QAV} + \lambda_2 \cdot \text{CD} + \lambda_3 \cdot \text{FCL} + \lambda_4 \cdot \text{IMG} + \lambda_5 \cdot \text{PC} + \lambda_6 \cdot \text{PCC}$$

Qualité d'ancrage visuel. Nous extrayons des masques de segmentation au niveau des objets à l'aide d'un modèle de segmentation tel qu'EfficientSAM (Xiong *et al.*, 2024). Chaque région segmentée ainsi que l'image entière sont encodées indépendamment à l'aide d'un modèle vision-langage compact (par exemple, SmolVLM (Marafioti *et al.*, 2025)). Les rôles sémantiques sont extraits du texte

alternatif à l'aide de techniques standards d'étiquetage des rôles sémantiques (Chen *et al.*, 2025). Chaque représentation vectorielle des rôles sémantiques est comparée aux représentations enrichies des régions visuelles, en calculant la similarité maximale (MaxSim) pour chaque rôle. Le *contexte local* fait référence aux régions segmentées spécifiques, tandis que le *contexte global* correspond à l'image dans son ensemble. La fusion des deux permet une évaluation fine de l'alignement entre le texte et le contenu visuel.

Cohérence discursive. Lorsque le texte alternatif comporte plusieurs phrases, nous en évaluons la cohérence interne. Nous vérifions que les entités (acteurs, objets) sont référencées de manière cohérente, qu'aucune contradiction n'est introduite, et que les rôles sémantiques à travers les phrases forment une description cohérente de la scène. La résolution des coréférences et l'analyse des rôles sémantiques sont utilisées pour soutenir cette étape de validation.

Évaluation de la fluidité et de la cohérence linguistique. La qualité linguistique de surface est évaluée selon une dimension inspirée de GRUEN (Zhu & Bhat, 2020), combinant trois sous-scores : la grammaticalité, la non-redondance et la cohérence thématique.

Imageabilité. Nous estimons la capacité évocatrice du texte alternatif en suivant une version adaptée de la méthode Tell as You Imagine (Umemura *et al.*, 2021). Des scores d'imageabilité issus de lexiques psycholinguistiques sont attribués aux mots clés et agrégés le long de l'arbre syntaxique. La méthode est adaptée pour pénaliser l'abstraction et encourager une visualisation concrète.

Pertinence contextuelle. Nous encodons le contexte textuel immédiat de l'image (via MiniLM (Wang *et al.*, 2020)) afin de pondérer la pertinence des régions visuelles : un texte proche et pertinent les renforce, tandis qu'un contexte éloigné ou non pertinent a peu d'effet.

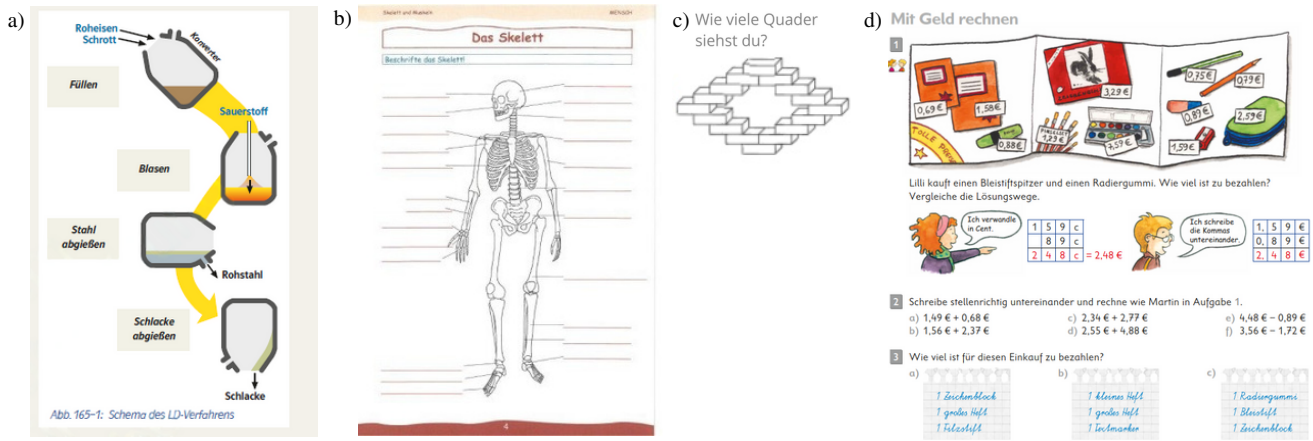
Validation des concepts de sens commun. Nous ajoutons une validation optionnelle de la plausibilité des rôles sémantiques. Les rôles sémantiques (acteurs, actions, objets) extraits du texte alternatif sont encodés sous forme de plongements de phrases, puis comparés à des représentations pré-encodées de scénarios plausibles issus du sens commun. Les combinaisons sémantiques présentant une faible similarité avec des événements typiques sont signalées comme potentiellement invraisemblables (Kapur & Kreiss, 2024).

Nous prévoyons d'évaluer dans quelle mesure **AltScore** est aligné avec les préférences des utilisateurs aveugles ou malvoyants, en corrélant ses résultats avec des jugements humains.

3.4 Application aux manuels scolaires

Les illustrations dans les manuels scolaires jouent un rôle crucial dans l'apprentissage. Lorsqu'une activité contient une image, l'adaptation diffère selon le rôle de celle-ci. L'élève doit pouvoir résoudre l'activité, grâce à une description de l'image ou un ajustement de la tâche. La figure 2 présente des extraits de manuels avec une adaptation des images pour les élèves DV. L'image a) est un schéma informatif essentiel à la compréhension d'une leçon ; elle doit être décrite en détail. L'image b) est un schéma à compléter par l'élève : sa description seule n'ayant aucun sens pour un élève DV, la tâche a été reformulée. L'image c) ne peut être décrite sans dévoiler la solution ; on indique la présence d'un graphique, et, si possible, un support tactile est proposé. Enfin, l'activité d) contient plusieurs images. La première fournit des informations essentielles à la réalisation des exercices et est remplacée par un texte alternatif clair et concis. Les énoncés des exercices à la suite sont également complétés pour inclure les informations visuelles, afin de ne pas alourdir la charge cognitive de l'élève.

Dans un objectif d'automatisation, il est nécessaire d'identifier les activités associées aux images, puis de filtrer celles-ci en fonction de leur nature et de leur rôle, en s'appuyant par exemple sur la classification proposée par Yadav *et al.* (2025). Les images décoratives, redondantes avec le texte, ou



Adaptations traduites de l'allemand vers l'anglais :

a) [[Graphic: Fig. 165-1: Diagram of the basic oxygen process:
Converter with 2 adjacent openings in different rotated positions
1. filling - of pig iron and scrap through the large opening (pointing diagonally upwards)
2. blowing - of oxygen through the large opening (pointing upwards)
3. tapping steel - pouring out raw steel from the small opening (pointing diagonally downwards)
4. tapping slag - from the large opening (now pointing downwards)]]

b) [[Which bones belong:
-) to the skull: []
-) to the trunk: []
-) to the limbs: []]]

c) [[Graphic: Geometric figure]]
[[Task solvable with tactile material]]

d) [[Graphic: Prices
small notebook: €0.69
large notebook: €1.58
...
pencil sharpener: €1.59]]
...
How much must be paid for this purchase?
a) 1 drawing pad (€3.29), 1 large notebook (€1.58), 1 felt-tip pen (€0.75)
[]

FIGURE 2 – Exemples d'images et activités de manuels scolaires et propositions d'adaptations par BookAccess. Source : https://www.bookaccess.at/

purement illustratives sans lien avec l'activité seront omises. En fonction de l'objectif pédagogique, il s'agit ensuite d'inclure une description de l'image et/ou de modifier la tâche pour la rendre accessible à un élève DV, si cela est possible. Nous envisageons un autre système de génération entièrement adapté au contexte scolaire. Les VLMs génériques bruts s'avèrent insuffisants dans ce cadre, d'autant plus qu'il n'existe pas de données d'entraînement dans un cadre scolaire. Par exemple, pour l'image du squelette (Figure 2 b)), même en contextualisant l'image dans l'instruction, LLaVa produit une description non adaptée comme « A skeleton is labeled with various body parts such as the skull, ribcage, pelvis, and legs. The skeleton stands on one leg. The image is in black and white. » Un tel système devrait prendre en compte les besoins des élèves DV et les objectifs pédagogiques. Il serait paramétrable à la fois en niveau de détail (selon la nature et le rôle de l'image) et en vocabulaire (adapté au niveau scolaire de l'enfant).

4 Conclusion

Dans une démarche inclusive, ce travail exploratoire propose un cadre pour la génération de textes alternatifs aux images tenant compte du contexte et des besoins des DV, en exploitant l'IA générative et en contournant ses limites. En l'absence de métrique d'évaluation adaptée, nous introduisons un score sans référence, avec pour perspective de l'aligner aux préférences des DV. Enfin, nous mettons en évidence la complexité de l'adaptation des contenus visuels dans les documents pédagogiques, en fonction de leur rôle et en raison de l'absence de jeux de données adaptés. Nos travaux futurs visent à appliquer et optimiser les méthodes proposées dans un contexte générique, puis à développer un système plus complexe spécifique aux activités scolaires.

Remerciements

Ce travail a été financé par le projet ANR-21-CE38-0014, l'institut DATAIA et le LISN.

Références

AHMADI S. & AGRAWAL A. (2024). An examination of the robustness of reference-free image captioning evaluation metrics. In *Findings of the Association for Computational Linguistics : EACL 2024*, p. 196–208.

ANDERSON P., FERNANDO B., JOHNSON M. & GOULD S. (2016). SPICE : Semantic propositional image caption evaluation. In *Computer Vision–ECCV 2016 : 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, p. 382–398.

BANERJEE S. & LAVIE A. (2005). METEOR : An automatic metric for mt evaluation with improved correlation with human judgments. In *ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*.

CASTILLAN L., LEMARIÉ J. & MOJAHID M. (2018). Numérique, handicap visuel et accessibilité des apprentissages. contenus pédagogiques numériques : quelle accessibilité pour les élèves présentant une déficience visuelle ? *Éducation & Formation*, p. 90–102.

CHEN H., ZHANG M., LI J., ZHANG M., ØVRELID L., HAJIČ J. & FEI H. (2025). Semantic role labeling : A systematical survey. *arXiv preprint arXiv :2502.08660*.

CHINTALAPATI S. S., BRAGG J. & WANG L. L. (2022). A dataset of alt texts from HCI publications : Analyses and uses towards producing more descriptive alt texts of data visualizations in scientific papers. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility*, p. 1–12.

FEINGLASS J. & YANG Y. (2021). SMURF : SeMantic and linguistic UndeRstanding Fusion for caption evaluation via typicality analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, p. 2250–2260.

GUBBI MOHANBABU A. & PAVEL A. (2024). Context-aware image descriptions for web accessibility. In *Proceedings of the 26th International ACM SIGACCESS Conference on Computers and Accessibility*, p. 1–17.

HAN T., BAIN M., NAGRANI A., VAROL G., XIE W. & ZISSERMAN A. (2023). AutoAD : Movie Description in Context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, p. 18930–18940.

HESSEL J., HOLTZMAN A., FORBES M., LE BRAS R. & CHOI Y. (2021). CLIPScore : A Reference-free Evaluation Metric for Image Captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, p. 7514–7528.

HU A., CHEN S., ZHANG L. & JIN Q. (2023). InfoMetIC : An informative metric for reference-free image caption evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 3171–3185.

HU E. J., SHEN Y., WALLIS P., ALLEN-ZHU Z., LI Y., WANG S., WANG L. & CHEN W. (2022). LoRA : Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

KAPUR R. & KREISS E. (2024). Reference-based metrics are biased against blind and low-vision users' image description preferences. In *Proceedings of the Third Workshop on NLP for Positive Impact*, p. 308–314.

KREISS E., BENNETT C., HOOSHMAND S., ZELIKMAN E., MORRIS M. R. & POTTS C. (2022a). Context matters for image descriptions for accessibility : Challenges for referenceless evaluation metrics. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, p. 4685–4697.

KREISS E., FANG F., GOODMAN N. & POTTS C. (2022b). Concadia : Towards image-based text generation with a purpose. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, p. 4667–4684.

KREISS E., ZELIKMAN E., POTTS C. & HABER N. (2024). ContextRef : Evaluating Referenceless Metrics For Image Description Generation. In *The Twelfth International Conference on Learning Representations*.

LIN C.-Y. & HOVY E. (2003). Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 human language technology conference of the North American chapter of the association for computational linguistics*, p. 150–157.

LINCKER E., GUINAUDEAU C. & SATOH S. (2025). AD2AT : Audio description to alternative text, a dataset of alternative text from movies. In *Proceedings of the 31st International Conference on Multimedia Modeling*, p. 58–71.

LIU H., LI C., WU Q. & LEE Y. J. (2023). Visual Instruction Tuning. *Advances in Neural Information Processing Systems*, **36**, 34892–34916.

MARAFIOTI A., ZOHAR O., FARRÉ M., NOYAN M., BAKOUCH E., CUENCA P., ZAKKA C., ALLAL L. B., LOZHKOV A., TAZI N. *et al.* (2025). Smolvlm : Redefining small and efficient multimodal models. *arXiv preprint arXiv :2504.05299*.

MATAMALA A. & VILLEGAS M. (2016). Building an audio description multilingual multimodal corpus : the VIW project. *Multimodal Corpora : Computer vision and language processing*, (11).

MCCALL K. & CHAGNON B. (2022). Rethinking alt text to improve its effectiveness. In *International Conference on Computers Helping People with Special Needs*, p. 26–33.

PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2002). BLEU : a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, p. 311–318.

RADFORD A., KIM J. W., HALLACY C., RAMESH A., GOH G., AGARWAL S., SASTRY G., ASKELL A., MISHKIN P., CLARK J., KRUEGER G. & SUTSKEVER I. (2021). Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, p. 8748–8763.

RAFAILOV R., SHARMA A., MITCHELL E., MANNING C. D., ERMON S. & FINN C. (2023). Direct Preference Optimization : Your Language Model is Secretly a Reward Model. *Advances in Neural Information Processing Systems*, **36**, 53728–53741.

ROHRBACH A., ROHRBACH M., TANDON N. & SCHIELE B. (2015). A dataset for movie description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, p. 3202–3212.

SCHUHMANN C., BEAUMONT R., VENCU R., GORDON C., WIGHTMAN R., CHERTI M., COOMBES T., KATTA A., MULLIS C., WORTSMAN M., SCHRAMOWSKI P., KUNDURTHY S., CROWSON K., SCHMIDT L., KACZMARCZYK R. & JITSEV J. (2022). Laion-5b : An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, **35**, 25278–25294.

SHARMA P., DING N., GOODMAN S. & SORICUT R. (2018). Conceptual captions : A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 2556–2565.

SONG H., SHIN M., KIM Y., JANG K., CHOI J., JUNG H. & SUH B. (2025). Altauthor : A context-aware alt text authoring tool with image classification and lmm-powered accessibility compliance. In *Companion Proceedings of the 30th International Conference on Intelligent User Interfaces*, p. 124–128.

SRIVATSAN N., SAMANIEGO S., FLOREZ O. & BERG-KIRKPATRICK T. (2024). Alt-text with context : Improving accessibility for images on twitter. In *The Twelfth International Conference on Learning Representations*.

STANGL A., VERMA N., FLEISCHMANN K. R., MORRIS M. R. & GURARI D. (2021). Going beyond one-size-fits-all image descriptions to satisfy the information wants of people who are blind or have low vision. In *Proceedings of the 23rd international ACM SIGACCESS conference on computers and accessibility*, p. 1–15.

UMEMURA K., KASTNER M. A., IDE I., KAWANISHI Y., HIRAYAMA T., DOMAN K., DEGUCHI D. & MURASE H. (2021). Tell as you imagine : Sentence imageability-aware image captioning. In *Proceedings of the 27th International Conference on MultiMedia Modeling, Part II*.

VEDANTAM R., LAWRENCE ZITNICK C. & PARIKH D. (2015). CIDEr : Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, p. 4566–4575.

WANG W., WEI F., DONG L., BAO H., YANG N. & ZHOU M. (2020). Minilm : Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in neural information processing systems*, **33**, 5776–5788.

WILLIAMS C., DE GREEF L., HARRIS III E., FINDLATER L., PAVEL A. & BENNETT C. (2022). Toward supporting quality alt text in computing publications. In *Proceedings of the 19th International Web for All Conference*, p. 1–12.

XIONG Y., VARADARAJAN B., WU L., XIANG X., XIAO F., ZHU C., DAI X., WANG D., SUN F., IANDOLA F., KRISHNAMOORTHI R. & CHANDRA V. (2024). EfficientSAM : Leveraged Masked Image Pretraining for Efficient Segment Anything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, p. 16111–16121.

YADAV S., LINCKER E., HURON C., MARTIN S., GUINAUDEAU C., SATOH S. & SHUKLA J. (2025). Towards inclusive education : Multimodal classification of textbook images for accessibility. In *Proceedings of the 31st International Conference on Multimedia Modeling*, p. 212–225.

ZHU W. & BHAT S. (2020). GRUEN for evaluating linguistic quality of generated text. In *Findings of the Association for Computational Linguistics : EMNLP 2020*, p. 94–108.

ZUR A., KREISS E., D'OOSTERLINCK K., POTTS C. & GEIGER A. (2024). Updating clip to prefer descriptions over captions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, p. 20178–20187.