AD2AT: Audio Description to Alternative Text, a Dataset of Alternative Text from Movies

Elise Lincker, Camille Guinaudeau, Shin'Ichi Satoh

AD2AT: Audio Description to Alternative Text, a Dataset of Alternative Text from Movies

Elise Lincker[1,2][0009−0005−1104−1785], Camille Guinaudeau[2,3][0000−0001−7249−8715], and Shin'ichi Satoh[2][0000−0001−6995−6447]

[1] Cedric, CNAM, Paris, France
[2] National Institute of Informatics, Tokyo, Japan
[3] University Paris-Saclay / JFLI, CNRS, Tokyo, Japan
elise.lincker@lecnam.net guinaudeau@lisn.fr satoh@nii.ac.jp

Abstract. Alternative text (alt text) is often mistaken for image captions. However, alt text is intended to replace an image, whereas a caption supports an image. Effective alt text is essential for enhancing visual accessibility for blind and low vision (BLV) individuals. While there has been substantial research in image captioning, this work often falls short in assessing visual accessibility needs. In this paper, we introduce **AD2AT**, a dataset of alt text derived from professionally tailored audio descriptions in movies. Our dataset, comprising over 3,800 text-image pairs, represents a first step toward advancing the alt text generation task and serves as a valuable resource for a range of vision-language applications. Through a qualitative analysis, we demonstrate the limitations of state-of-the-art image captioning and text generation models in producing effective alt text. We provide insights into improving alt text generation and call for future work on developing robust, context-aware models and evaluation metrics that align with accessibility guidelines, to better serve BLV users across different domains.

Keywords: Alt text · Alternative text · Audio description · Image-to-text generation · Visual accessibility.

1 Introduction

The motivation behind alternative text (alt text) generation lies in the need for effective annotations that enhance visual accessibility for blind and low vision (BLV) individuals. Although there is no universal consensus on visual accessibility standards, guidelines from the W3C's Web Accessibility Initiative (WAI)[4] and other image and video description standards advocate for concise, objective, context-aware descriptions that highlight the predominant content (e.g., objects, people, text, scenery) to aid understanding. Prior work has also emphasized that BLV people's preferences for image descriptions vary with the image's context, source, and user goals [33].

[4] https://www.w3.org/WAI/tutorials/images/
https://www.w3.org/WAI/media/av/description/

Alt text is often confused with image captions, which already has established research and available datasets. However, alt text is intended to *replace* an image, whereas a caption *supports* an image. The primary purpose of alt text is to ensure visual accessibility for BLV users, within a given context. Research has demonstrated that standard image captioning is insufficient for visual accessibility needs [12]. Also, alt text on the internet is often unreliable and text designed *for* visual accessibility is difficult to access. Therefore, obtaining effective alt text data and advancing alt text generation is crucial to address these gaps and improve the overall accessibility of visual content for all users.

Alt text generation is a challenging task that has not been explored deeply enough and deserves attention in the multimodal community. In this paper, we introduce AD2AT (Audio Description to Alternative Text), a dataset of over 3,800 images paired with alt text, derived from audio-described movies. We take advantage of existing audio description (AD) datasets and provide a finer annotation to match one description to one image. Providing new image-text pairs, our dataset may also be used to enhance other visual-language tasks, including image and video captioning, AD generation, character identification, visual storytelling and visual entailment. Additionally, we provide a qualitative analysis comparing our gold-standard data with automatically generated captions and descriptions, highlighting the limitations of state-of-the-art text generation models and evaluation metrics in addressing accessibility needs for BLV users.

2 Related Work

2.1 Alt-text Datasets

"Alt" tags on websites could serve as a valuable starting point for creating alt text datasets. Conceptual Captions [30], its extension Conceptual 12M [3], and LAION-5B [29] involve crawling images and their associated alt texts from web pages. Similarly, the Wikipedia-based dataset Concadia [13] and the Twitter-based dataset used in [32] follow this methodology, while also going a step further by providing the context in which the image is situated in. However, "alt" tags are often either empty or similar to the image title or caption. Indeed, after an automatic filtering step, only 3% of image-alt text candidate pairs are retained in Conceptual Caption, and 10% in LAION-5B. Alongside these challenges, other research focus on alt text for data visualizations in scientific publications [4, 22].

Above all, alt text attributes are frequently written by untrained individuals who may lack an understanding of the needs of BLV users, resulting in descriptions that are often inadequate to effectively convey the necessary information. To the best of our knowledge, there is no dataset of alt text written by experts available.

2.2 AD Datasets

Movies and TV programs may have professionally generated AD in a supplementary audio track. Descriptive Video Services (DVS) provide AD to make visual media accessible to BLV users.

Online platforms such as AudioVault[5] and Blind Mice Movie Vault[6] provide free audio-described files, produced by DVS, for hundreds of films and TV series. Nevertheless, they only provide audio files without transcription or video, mixed with the original audio soundtrack, which makes it difficult to separate the AD. These files are meant to be used by visually impaired individuals alongside the original movie video. Most of them are in English, but some are available in other languages.

Several datasets built upon movies' DVS have been released. AutoAD [8] propose a text-only dataset derived from AudioVault that covers ADs and subtitles from over 7,000 movies.

The Movie Audio Description (MAD) [31], Montreal Video Annotation (M-VAD) [34] and Movie Description (MPII-MD) [26] datasets are built directly from DVDs' DVS data. M-VAD and MPII-MD were merged to form a comprehensive dataset for the Large Scale Movie Description Challenge (LSMDC) [27]. These studies implement an automatic approach for DVS segmentation to isolate AD from the original soundtrack and align them with video content. However, this alignment is frequently imprecise, as AD is inserted between dialogues and original narrative audio, which may cause a misalignment between the spoken AD and corresponding visual content. Among these datasets, only MPII-MD employs a manual sentence-video alignment process to ensure that each AD accurately matches the corresponding video clip.

Finally, the Visuals Into Words (VIW) project [20, 21] aims to research AD from a multilingual and multimodal corpus perspective. They built the VIW corpus upon a short film, featuring multiple AD tracks in English, Spanish and Catalan. For English, 10 different AD generated by professionals are available.

While there is no other open-access accessible video content made by professionals, YouDescribe is a free crowdsourced platform that allows users to add ADs to YouTube videos. Approximately 21% of the audio-described videos on YouDescribe are rated by viewers, which provides an indication of their quality in meeting the needs of BLV users. The team released You Described, We Archived (YuWA) [24], a dataset composed of AD data collected from YouDescribe.

3 Image Descriptions from Movies

AD2AT is a dataset of images paired with alt text, built around two AD datasets: MPII-MD and VIW. It includes fine-grained annotation for frames extracted from 12 Hollywood movies (MD) and the short film *"What Happens While"* (VIW). All annotations and the complete AD2AT-VIW portion are publicly accessible.[7] To download images from the MD part, separate access to MPII-MD dataset must be requested.[8]

[5] https://audiovault.net
[6] https://www.blindmicemegamall.com/bmm/shop/Movie_Vault
[7] https://github.com/eliselinc/AD2AT/
[8] https://www.mpi-inf.mpg.de/departments/computer-vision-and-machine-learning/research/vision-and-language/mpii-movie-description-dataset

3.1 Movie Description (MD)

We chose the MPII-MD dataset as a starting point for its unique advantage: each AD is manually aligned with its corresponding video clip, addressing the issue of mismatch between the time of speech and the corresponding visual content. Such annotation is indispensable to ensure accurate text-to-video correspondence. MPII-MD is divided into three groups: Group 0 consists of 39 movies aligned with scripts, Group 1 consists of 55 movies aligned with DVS, and Group 2 consists of 11 movies from Group 1 that are also aligned with scripts. The dataset provides CSV annotation files with the original AD, and anonymized AD, where characters' names have been replaced by the word "Someone" or "People" when plural. For each annotated video clip, 10 to 13 frames are also made available. For our research, we exclusively utilize the data from Group 1.

Our contribution involves refining the annotation to match one description to one image instead of one video clip. Since ADs inherently describe a video, they cannot always be directly associated with a single image. It is necessary to filter and modify the data accordingly.

First, we filter the images from the video clips. For each sentence-video clip pair, we compute the similarity between successive images using CLIP [25] ViT-B/32. Based on the obtained similarity scores, we determine how many different shots are part of the video clip, using a threshold of 0.8. We assume that video clips with multiple shots correspond to ADs that describe several actions and may not be easily associated with a single image; therefore, we retain only pairs with one or two different shots. Then, we select two candidate images: the frame in the middle of the video clip, and the frame with the highest similarity score to the AD.

Second, we filter the text. Often, only a fraction of the AD matches the image, requiring us to select the relevant parts. To facilitate this annotation task, we split the AD text into clauses using Stanford CoreNLP's constituency parsing [19]. We built an annotation tool that displays for each original annotation: the two selected candidate images, the current AD text, the four previous AD and the next AD. The annotator manually selects the image and parts of the text from any of those ADs that match the most. Sometimes, the annotator must apply minor changes to the text.

The guidelines for the annotation process are as follows: retain text that closely matches the image, making only minor modifications if necessary, even if the text doesn't cover every detail of the image. Discard the text-image pair if the image doesn't match the description despite some adjustments or if a single image fails to convey the meaning of the description (sometimes two or more frames are necessary, for example if 2 people from two different shots are involved or in case of a complicated action). Minor adjustments to the text include removing time, location, and manner adverbials as well as prepositions, and expressions that are not conveyed in the image (such as "then," "moments later," "begin/continue to," "back," "elsewhere," "slowly," etc.); modifying verbs or expressions related to movement if the image does not depict such movement; removing any textual elements not visible in the frame; changing verbs from past

tense to present tense as needed; and proceeding with minor sentence reformulations based on selected elements. Examples of annotated pairs, with various text modifications applied, are presented in section 3.4.

The annotation was meticulously carried out by two expert annotators, adhering closely to the defined objectives and guidelines, and staying faithful to the original text. Each movie was annotated by a single annotator, but we ensured a high inter-annotator agreement. The agreement was calculated on the first 30 minutes of the movie TITANIC, considering 2 labels: *kept* or *discarded*, regardless of the selected text. The annotators found agreement in 89% of cases (102 out of 114 pairs), with a Cohen's Kappa of 0.775. We do not calculate the exact agreement on the selected and modified text: even though all annotators should retain the same information and keep the text as close to the original AD as possible, alt text can be formulated in different ways and still be appropriate for the needs of BLV users. Out of the 55 pairs kept by both annotators, they selected the same frame in 91% of the cases.

By refining both the visual and textual data, our approach ensures a more accurate alignment between descriptions and images.

3.2 Visuals Into Words (VIW)

The VIW corpus provides AD in English, Spanish and Catalan, made by professionals and students, for the 14-minute movie *"What Happens While"*. They provide full audio-described videos along with ELAN annotation files, which include additional linguistic, morphosyntactic and visual information. However, unlike the MPII-MD data, AD is not aligned with video clips.

We select the 10 English versions, all produced by professionals, and leverage these multiple ADs to associate multiple text descriptions with the same image. As mentioned earlier, we cannot rely on timestamps to align text with video clips. Therefore, the VIW part of AD2AT is entirely manually annotated to match one text description with one frame. Keyframes of the entire movie are extracted using the keyframe extractor video-kf.[9] Then, we carefully select one frame for each AD, minimizing the number of images so that several texts correspond to the same keyframe while ensuring they accurately match the text. Minor adjustments to the text were made, following the MD annotation process.

Ultimately, we provide 28 annotated frames, each paired with 1 to 10 different alt texts from the 10 audio-described videos. One exception: the title frame, which appears twice in the movie, has 14 different descriptions. For the same image, the length of the descriptions and the elements included in the text vary greatly. Figure 1 shows two frames with two alt texts, both written by professionals but differing significantly in length. Since descriptions can vary depending on the describer, despite all following accessibility guidelines and describing the same image in the same context, a single description does not represent the full range of possibilities. As a result, AD2AT's VIW portion stands as gold-standard data, ideal for evaluation purposes.

[9] https://pypi.org/project/video-kf/

Shortest (8 tokens): She hikes up an overgrown grassy slope. Longest (27 tokens): Outside, dressed in a warm jacket and jeans, Jess climbs up an overgrown hillside with the sun glinting on her long, springy hair.

Shortest (8 tokens): James lies back with his shirt off. Longest (28 tokens): James lies back, bare-chested, on a towel in the sand. His brownish, stubbly beard and chest hair almost match the ground.

Fig. 1: Examples of AD2AT-VIW annotated frames with short and long alt texts.

3.3 Additional Information

For the VIW part, supplementary annotation information indicates whether the description matches the scene and/or the text on the frame.

ADs often contain information specific to the plot, including character names. To anonymize and adapt our data for other vision-language tasks, we provide a pre-processed variant. The "someone" variant adheres to MPII-MD's anonymization approach: by aligning the AD2AT's MD part with both versions of MPII-MD, we convert all character names to "Someone" or "People" in case of plural. For the VIW part, anonymization is performed manually.

Finally, since alt text is a description in context, for each image-alt text pair, we also provide the three preceding ADs from the movie. Note that this contextual information may not be relevant if there is a shift to a new scene, or it may be redundant if the selected alt text already covers parts of the preceding ADs.

3.4 Statistics and Examples

As summarized in Table 1, AD2AT-MD provides gold-standard annotation for 12 out of the 55 movies in MPII-MD, resulting in 3,607 text-image pairs. AD2AT-VIW provides multiple alt text for 28 images, resulting in 226 pairs.

Table 1: Annotation statistics.

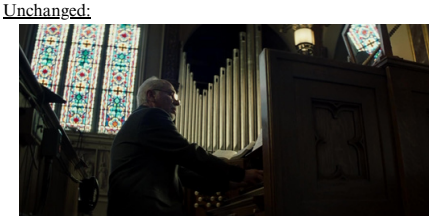
	MPII-MD	**AD2AT-MD (v1)**			**AD2AT-VIW**
		Total	**Unchanged**	**Modified**	**Discarded**	
Movies	55	12				1
Pairs	37,266	3,607	1,453	2,154	3,409	226

Figure 2 presents six AD2AT-MD annotated frames from various films, including examples where modifications were made to the original ADs. The dataset annotations are available in a CSV file, formatted as shown in Table 2.

Discarded:

Img: 1054_Harry_Potter_and_the_prisoner_of_azkaban_ 00.04.17.036-00.04.19.691/0008.jpg
Original AD: Lights start to flash and plates rattle.
Previous AD: Harry shakes with rage.

Unchanged:



Img: 1048_Gran_Torino_00.00.50.904-00.00.53.535/ 0007.jpg
Original AD: A balding man plays the organ.
Alt text: A balding man plays the organ.
Anonymized: A balding man plays the organ.

Modified:

Img: 1039_The_Queen_01.36.40.090-01.36.44.537/0007.jpg
Original AD: The Queen and Tony descend the stairs with a corgi scuttling ahead.
Alt text: The Queen and Tony descend the stairs.
Anonymized: People descend the stairs.

Img: 1015_27_Dresses_00.27.29.948-00.27.35.603/0007.jpg
Original AD: She rushes to the front door, peers through the peephole, then flings it open.
Alt text: She peers through the front door's peephole.
Anonymized: She peers through the front door's peephole.

Img: 1027_Les_Miserables_00.00.48.301-00.00.53.125/0002.jpg
Original AD: A French tricolor is under water.
Alt text: A French tricolor is under water. Text: 1815, twenty six years after the start of the French revolution, a king is once again on the throne of France.
Anonymized: A French tricolor is under water. Text: 1815, twenty six years after the start of the French revolution, a king is once again on the throne of France.

Img: 1005_Signs_01.12.33.590-01.12.36.962/0010.jpg
Original AD: Graham and his family look up and stare at.
Previous AD: At the other end of the table, the red lines on the baby monitor begin to flash.
Alt text: At the other end of the table, the red lines on the baby monitor flash. Graham and his family stare at it.
Anonymized: At the other end of the table, the red lines on the baby monitor flash. Someone and his family stare at it.

Fig. 2: Examples of AD2AT-MD annotated frames.

Table 2: Example of a dataset entry. The image file path corresponds to the MPII-MD format: video clip ID (film identifier, film title, and video clip timestamps), followed by the selected frame file.

image	text	textSomeone	names	context
1039_ The_Queen_ 00.03.24.200- 00.03.26.634/ 0006.jpg	The Queen looks at the paper.	Someone looks at the paper.	The Queen	The Queen lies in bed in a darkened room. A door opens behind the Queen. A woman enters and puts a newspaper onto a bedside table.

4 Limitations of Current Models and Evaluation Metrics

To underscore the relevance of our contribution, we carefully compare the ground truth descriptions from AD2AT-VIW with automatically generated captions and descriptions. The VIW subset offers multiple reference descriptions per image, allowing for a more comprehensive and robust evaluation compared to AD2AT-MD. Our analysis highlights both the advantages and limitations of existing state-of-the-art text generation models, assessing their performance in generating alt text in terms of content and detail.

We feed the 28 images of AD2AT-VIW to the following models for inference:
- Image captioning models: BLIPcaption [15], GIT [36], FuseCap [28];
- Vision-language instruction models: LLaVA [17, 18], InstructBLIP [5].

We test each model's recommended prompt as well as more detailed ones. Table 3 shows the combinations that yielded the best outcomes. Given the models' limitations in accurately identifying the exact text displayed on images, this analysis distinguishes between images with and without text.

Supplementary experiments on vision-language models included (1) few-shot prompting and (2) providing both the prompt and context retrieved from preceding AD within the same scene. However, none of the combinations tested on LLaVA and InstructBLIP effectively managed the contextual information, which is essential for accurate alt text generation.

Table 3: Model and prompt combinations

Model	**Prompt**
GIT	/
BLIPcaption	a photography of
FuseCap	a picture of
LLaVA	Write alternative text for this image.
InstructBlip	Write a description for the photo.
LLaVA/InstructBlip long prompt	Write alternative text for this image. Be concise while providing sufficient information for visually impaired people.

Overall, the captioning models GIT and BLIP generate accurate but too generic descriptions, lacking the nuanced information necessary for comprehensive image understanding. In contrast, FuseCap tends to provide excessive detail, sometimes including irrelevant elements, inappropriate for visual accessibility.

Although some parts of LLaVA's descriptions could be considered as correct, the generated texts are often excessively lengthy and, more critically, they contain assumptions. The descriptions frequently include qualifiers such as "appear to be," "possibly," "suggest," and "may", which indicate undesirable speculative interpretations, rather than factual and objective descriptions. With InstructBLIP, the generated text is comparable to an image caption. With the detailed

prompt, the prediction becomes much longer and contains assumptions, showing the same flaws as LLaVA. Both lack a balance between conciseness and precision.

Figure 3 shows a typical example from the first scene of the movie, highlighting the advantages and limitations of each model.

AD2AT-VIW references: A man parks a sleek white BMW on the road by the sea.
A white saloon car is parked on the seafront and the driver is on his mobile phone.
A bearded man in a suit parks his expensive car by the promenade while on the phone.
A slick, white car is parked on the grey asphalt. The male driver sits alone inside and speaks on his cell phone.
A car is parked by the beach, silver with sporty wheels. Inside the driver is on his mobile phone.
A man in a suit and tie and talking on a cellphone parks his white BMW by the seaside.
A white saloon is parked alongside the beach wall. The male driver is on his mobile. The man is in his mid 30's with a sandy beard.
BLIPcaption: a photography of a bmw z4
GIT: a white car on the road
FuseCap: a picture of a white car parked on a beach with a cloudy sky in the background, featuring a closed white door and a window
LLaVA: A white car is parked on the side of a road, with a man sitting inside. The car is facing the ocean, and the man appears to be enjoying the view. The car is parked on the side of the road, and the man is sitting in the driver's seat.
LLaVA-long: A white car is parked on the side of the road. The driver is sitting in the car, looking out the window. The car is parked near the ocean, providing a scenic view for the driver.
InstructBLIP: a white bmw convertible parked on the side of the road
InstructBLIP-long: A white BMW convertible car is parked near the ocean, with the driver's side door open. The car is situated on a road near the water, and the driver is visible in the open door. The car is parked close to the water, giving the impression of a scenic and relaxing drive.

Fig. 3: Example of AD2AT-VIW vs. automatically generated descriptions.

We evaluate the models' predictions using classic text generation metrics to get an indication of their performance. Table 4 reports the scores computed on images that do not include any text. If most scores seem low, one reason could be that the metrics are not well-suited to our task and the expected text, a point we discuss in Section 5.2. The correlation between scores remains consistent: InstructBLIP's predictions are the closest to an alt text, and the model achieves the highest BLUE scores. In contrast, GIT's descriptions contain only a minimum of information, which is far from sufficient for the needs of alt text, and this is reflected in the low scores. LLaVA (short prompt) and InstructBLIP (long prompt) achieve the highest CLIPScore, probably because the generated descriptions are the longest, and therefore contain more information likely to be similar to the image. However, they don't meet our needs due to the length of the generated text and the assumptions made in the predictions. None of these metrics is fully appropriate to the alt text generation task, and the development of a reference-free evaluation metric, based on text-image similarity but also on other features specific to alt text, is necessary.

Another significant limitation of current models concerns the extraction of text displayed on the image. Titles, captions, and other critical textual information must be incorporated in the alt text. Figure 4 shows two frames from *"What Happens While"*, both containing text, alongside references from AD2AT-VIW and generated descriptions. The top frame is the title frame, containing only three words and no background. The bottom one is taken from the middle of

Table 4: Comparison of evaluation scores for automatically generated texts.

Model	BLEU-4	BLEU-2	BLEU-1	ROUGE	METEOR	CIDEr	SPICE	CLIPScore
GIT	0.0701	0.2124	0.3009	0.2564	0.0952	0.1305	0.0613	0.6353
BLIPcaption	0.0990	0.3017	0.4677	0.2630	0.1098	0.2286	0.0563	0.6469
FuseCap	0.0615	0.1917	0.3205	0.2561	0.1444	0.0780	0.0850	0.6748
LLaVA	0.0520	0.1701	0.2857	0.2317	0.1626	0.0133	0.0961	0.7044
LLaVA-long	0.0715	0.2352	0.3871	0.2856	0.1696	0.1405	0.0996	0.6927
InstructBLIP	0.1088	0.3305	0.5097	0.3153	0.1120	0.2824	0.0616	0.6761
InstructBLIP-long	0.0441	0.1344	0.2246	0.1982	0.1532	0.0171	0.1034	0.7229
References								0.7013

the movie, displaying an image in the background with text superimposed on it. Models often imagine irrelevant objects or scenes when there is only a plain background. Despite some errors, LLaVA- and BLIP-based models detect the written text. InstructBLIP detects the text more accurately, but tends to continue the sentence. In addition, in cases where the text is superimposed on a photo, only BLIP-based models report the text.

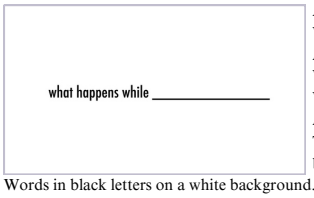
5 Discussion

5.1 Context-Aware Text Generation

Alt text for the same image may differ depending on the context in which the image is situated in. In our case, descriptions focus on elements relevant to the film's plot rather than describing the entire scene. In different contexts, descriptions might include other elements. On top of that, studies have shown that preferences among BLV individuals for image and video descriptions can vary significantly based on the context and source [11]. Exploring the creation of a dataset containing alt texts for identical images across diverse contexts represents a promising avenue for future research in visual accessibility, despite the substantial annotation efforts required.

In the meantime, models must generate accessible descriptions that account for the image context while avoiding redundancy. Recent context-aware approaches for image- and video-to-text generation [7, 8, 12, 32], in the realm of visual accessibility have been explored and warrant further investigation in conjunction with our approach.

AD2AT already provides character anonymization and context based on previous ADs. More contextual content can be inferred from a synopsis, preceding images, ADs, and dialogues. Future work will aim to improve context selection by retaining only relevant information, such as ADs from the same film scene. We will also work on location decontextualization and further character decontextualization by introducing visually impaired-friendly descriptions instead of names, pronouns or "Someone."



AD2AT-VIW references:
What happens while.
A title. What Happens While…
What happens while. The title ends with just a line.
Words: What Happens While… with a long, blank line at the end.
A title: What Happens While... with a long, blank line at the end.
The title, in white letters on a black background: what happens while, followed by a long black line.
A title reads: What happens while BLANK.
Caption: What happens while – blank.
On a white background: what happens while.
Words in black letters on a white background. what happens while, followed by a long black line.
Black text on a white background, reading: 'What happens while', followed by a long black line.
GIT: the line is the same as the line.
BLIPcaption: a photography of a white background with the words what happens wine
FuseCap: a picture of a white envelope sits on a surface, surrounded by various objects including a logo, text, and various other objects
LLaVA: The image is a black and white photo of a whiteboard with the words "most things happen while" written on it. The whiteboard is placed in the center of the image, and the text is written in a bold, capitalized font. The photo captures the essence of the message, emphasizing the idea that most events occur unexpectedly or coincidentally.
InstructBLIP: what happens while?
InstructBLIP-long: what happens while you're asleep?

AD2AT-VIW references: Wearing a singlet, the man has a greying beard and frizzy hair. White text reads: Rick, The retiree.
A man in late middle-age has strong features, a grizzled beard and wild, wiry, greying hair. He wears a vest top and a heavy chain around his neck. Rick. The Retiree.
A middle aged man with unkempt grey hair and heavy brows. He wears a green vest and a chain around his neck. In white letters: RICK, the retiree.
An unshaven older man with messy, longish, grey hair gazes around, looking bored. His eyes look moist and glassy. Words: Rick, the Retiree.
A bearded grey haired man with heavy green eyes. He has a chain around his neck, wears a vest. Caption: Rick, the retiree.
GIT: a man with a beard
BLIPcaption: a photography of a man with a beard and a beard, with the words rick rick rick rick rick rick
FuseCap: a picture of a man with a gray beard and hair, a large nose, and brown eyes wearing a gray shirt and silver necklace, with a gray strap visible in the foreground
LLaVA: The image features a man with a beard and mustache, wearing a green shirt. He appears to be looking at the camera, possibly posing for a picture. The man's facial hair and attire suggest that he might be a hipster or someone with a distinctive style.
LLaVA-long: A man with a beard and mustache is looking at the camera. He has a chain around his neck and is wearing a green shirt.
InstructBLIP: a man with a beard and a tattoo on his arm
InstructBLIP-long: a man with a beard and a tank top, with the words "rick the referee" displayed above him

Fig. 4: Examples of AD2AT-VIW vs. automatically generated descriptions.

5.2 Text Generation Evaluation Metrics

Metrics like BLEU [23], NIST [6] and ROUGE [16] focus on n-gram overlap between predictions and references, based on exact word matches, struggling to account for synonyms and paraphrases. The recall-oriented nature of ROUGE can also lead to issues with excessive content. METEOR [2] goes a step further by employing stemming to account for root and synonym correspondence. However, these *hard* metrics heavily rely on reference texts and often fall short in assessing the true semantic adequacy of generated text.

CIDEr [35] and SPICE [1], designed for image description tasks, improve on traditional metrics by focusing on semantic content rather than mere n-grams, but remain dependent on the diversity and representativeness of references. BERTScore [37], using contextual embeddings from BERT, captures semantic similarity more effectively, though it also relies on reference texts.

Reference-free metrics include CLIPScore [9], which uses CLIP embeddings to compute cosine similarity between images and texts. GRUEN [38], designed for text generation and adaptable to various tasks, evaluates generated text based

on 4 features: grammaticality, non-redundancy, focus, structure and coherence. Additionally, InfoMetIC [10] has been noted for returning incorrect words and unmentioned image regions in captions, improving model understanding and providing better correlation with human judgements.

Alt text must accurately convey essential visual content to visually impaired users, which goes beyond simple n-gram overlap or surface-level matching. While human evaluation by the target audience remains the ideal standard, there is a need to develop automatic evaluation methods that closely approximate this. Future work should aim to develop a reference-free metric that aligns with visual accessibility guidelines, while ensuring grammaticality, semantic accuracy and contextual relevance. Such a metric must integrate multiple features including visual features, context-awareness, and non-redundancy. A promising initiative in this direction is ContextRef [14], a benchmark for assessing referenceless metrics, which includes a context robustness check.

6 Conclusion

AD2AT is a new dataset of images paired with alternative text derived from movie audio descriptions. These audio descriptions, crafted by professional audio describers to enhance visual accessibility, are manually adapted to match images. The dataset includes over 3,800 text-image pairs manually annotated, along with anonymized versions and contextual information retrieved from preceding audio descriptions. Beyond alt text generation, we anticipate that AD2AT will also be valuable for a range of other vision-language tasks. Our comprehensive comparison of AD2AT-VIW gold-standard data with automatically generated captions and descriptions underscores the necessity to adapt existing state-of-the-art models and metrics to better address the specific needs of BLV individuals.

AD2AT represents a small yet significant step towards improving visual accessibility. Future research should focus on developing robust, context-aware alt text generation models and evaluation metrics that can better serve the diverse needs of visually impaired users across various domains.

Acknowledgements

This work was supported by the Japan Society for the Promotion of Science (JSPS) and the French National Research Agency (ANR) MALIN project (ANR-21-CE38-0014).

References

1. Anderson, P., Fernando, B., Johnson, M., Gould, S.: SPICE: Semantic propositional image caption evaluation. In: ECCV (2016)
2. Banerjee, S., Lavie, A.: 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 (2005)
3. Changpinyo, S., Sharma, P., Ding, N., Soricut, R.: Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In: IEEE/CVF CVPR (2021)
4. Chintalapati, S.S., Bragg, J., Wang, L.L.: A dataset of alt texts from hci publications: Analyses and uses towards producing more descriptive alt texts of data visualizations in scientific papers. In: ASSETS 2022 (2022)
5. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P.N., Hoi, S.: InstructBLIP: Towards general-purpose vision-language models with instruction tuning. Advances in Neural Information Processing Systems (2024)
6. Doddington, G.: Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In: HLT '02 (2002)
7. Han, T., Bain, M., Nagrani, A., Varol, G., Xie, W., Zisserman, A.: AutoAD II: The sequel-who, when, and what in movie audio description. In: IEEE/CVF CVPR (2023)
8. Han, T., Bain, M., Nagrani, A., Varol, G., Xie, W., Zisserman, A.: AutoAD: Movie description in context. In: IEEE/CVF CVPR (2023)
9. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: CLIPScore: a reference-free evaluation metric for image captioning. In: EMNLP 2021 (2021)
10. Hu, A., Chen, S., Zhang, L., Jin, Q.: InfoMetIC: An informative metric for reference-free image caption evaluation. In: ACL 2023 (2023)
11. Jiang, L., Jung, C., Phutane, M., Stangl, A., Azenkot, S.: "it's kind of context dependent": Understanding blind and low vision people's video accessibility preferences across viewing scenarios. In: CHI Conference on Human Factors in Computing Systems (2024)
12. Kreiss, E., Bennett, C., Hooshmand, S., Zelikman, E., Ringel Morris, M., Potts, C.: Context matters for image descriptions for accessibility: Challenges for referenceless evaluation metrics. In: EMNLP 2022 (2022)
13. Kreiss, E., Fang, F., Goodman, N., Potts, C.: Concadia: Towards image-based text generation with a purpose. In: EMNLP 2022 (2022)
14. Kreiss, E., Zelikman, E., Potts, C., Haber, N.: ContextRef: Evaluating referenceless metrics for image description generation. arXiv preprint arXiv:2309.11710 (2023)
15. Li, J., Li, D., Xiong, C., Hoi, S.: BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: ICML (2022)
16. Lin, C.Y., Hovy, E.: Automatic evaluation of summaries using n-gram co-occurrence statistics. In: HLT-NAACL 2003 (2003)
17. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: IEEE/CVF CVPR (2024)
18. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Advances in Neural Information Processing Systems (2023)
19. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J.R., Bethard, S., McClosky, D.: The Stanford CoreNLP natural language processing toolkit. In: ACL 2014 (2014)
20. Matamala, A.: The VIW project: Multimodal corpus linguistics for audio description analysis. Revista Española de Lingüística Aplicada (2019)

21. Matamala, A., Villegas, M.: Building an audio description multilingual multimodal corpus: the VIW project. Multimodal Corpora: Computer vision and language processing (2016)
22. Moured, O., Farooqui, S.A., Müller, K., Fadaeijouybari, S., Schwarz, T., Javed, M., Stiefelhagen, R.: Alt4blind: A user interface to simplify charts alt-text creation. In: International Conference on Computers Helping People with Special Needs (2024)
23. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: ACL '02 (2002)
24. Pitcher-Cooper, C., Seth, M., Kao, B., Coughlan, J.M., Yoon, I.: You Described, We Archived: A rich audio description dataset. Journal on Technology and Persons with Disabilities (2023)
25. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021)
26. Rohrbach, A., Rohrbach, M., Tandon, N., Schiele, B.: A dataset for movie description. In: IEEE CVPR (2015)
27. Rohrbach, A., Torabi, A., Rohrbach, M., Tandon, N., Pal, C., Larochelle, H., Courville, A., Schiele, B.: Movie description. International Journal of Computer Vision (2017)
28. Rotstein, N., Bensaïd, D., Brody, S., Ganz, R., Kimmel, R.: FuseCap: Leveraging large language models for enriched fused image captions. In: IEEE/CVF Winter Conference on Applications of Computer Vision (2024)
29. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in Neural Information Processing Systems (2022)
30. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: ACL 2018 (2018)
31. Soldan, M., Pardo, A., Alcázar, J.L., Caba, F., Zhao, C., Giancola, S., Ghanem, B.: MAD: A scalable dataset for language grounding in videos from movie audio descriptions. In: IEEE/CVF CVPR (2022)
32. Srivatsan, N., Samaniego, S., Florez, O., Berg-Kirkpatrick, T.: Alt-text with context: Improving accessibility for images on twitter. In: ICLR (2024)
33. Stangl, A., Verma, N., Fleischmann, K.R., Morris, M.R., Gurari, D.: Going beyond one-size-fits-all image descriptions to satisfy the information wants of people who are blind or have low vision. In: ASSETS 2021 (2021)
34. Torabi, A., Pal, C., Larochelle, H., Courville, A.: Using descriptive video services to create a large data source for video annotation research. arXiv preprint arXiv:1503.01070 (2015)
35. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: CIDEr: Consensus-based image description evaluation. In: IEEE CVPR (2015)
36. Wang, J., Yang, Z., Hu, X., Li, L., Lin, K., Gan, Z., Liu, Z., Liu, C., Wang, L.: GIT: A generative image-to-text transformer for vision and language. Transactions on Machine Learning Research (2022)
37. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: BertSCORE: Evaluating text generation with BERT. arXiv preprint arXiv:1904.09675 (2019)
38. Zhu, W., Bhat, S.: GRUEN for evaluating linguistic quality of generated text. In: Findings of the ACL: EMNLP 2020 (2020)