https://doi.org/10.2352/EI.2026.38.3.MOBMU-322

Capabilities of Image-to-Text Transformation Models for Enabling Visually Impaired to Perceive Complex Imaging Visuals at Conferences and Scientific Journals

Ruthra Bellan[1], Frank Wittig[1], Reiner Creutzburg[1, 2, 3]

[1]SRH University of Applied Sciences -Campus Berlin, Sonnenallee 221 A-F, D-12059 Berlin, Germany
[2]German University of Digital Science, Marlene-Dietrich-Allee 14, D-14772 Potsdam, Germany
[3]Technische Hochschule Brandenburg, Department of Informatics and Media, Magdeburger Str. 50, D-14770 Brandenburg, Germany

ruthra.bellan@srh.de, frank.wittig@srh.de, reiner.creutzburg@srh.de, creutzburg@th-brandenburg.de, reiner.creutzburg@german-uds.de

Abstract

Scientific figures (charts, composite panels, and data visualizations) are routinely inaccessible to visually impaired readers because screen readers cannot interpret visual content and published captions are often too brief or domain-specific to convey what the figure shows. Vision-language models (VLMs) offer a potential route to automated, accessible image description at scale. In this study, we evaluate five open-source, instruction-tuned VLMs (BLIP-2, LLaVA-1.5-7B, Moondream2, Qwen2-VL-2B, and Idefics3-8B) on a dataset of 245 scientific figures drawn from 32 papers presented at Electronic Imaging 2025. Generated captions are scored against author-provided ground-truth captions using four complementary metrics: BLEU, ROUGE-L, Sentence-BERT cosine similarity (SBERT), and RefCLIPScore. Moondream2 achieves the highest performance across all semantic metrics (RefCLIPScore = 1.025, SBERT = 0.392) despite being one of the smallest models evaluated (~1.86B parameters), offering the best balance of quality and speed (8.7 s per image). The four metrics tell a consistent story: Moondream2 scores low on lexical match but high on semantic similarity and image alignment, which is the expected pattern when detailed visual descriptions are compared against brief author captions. These findings are broadly paralleled in an evaluation of VLM-generated captions performed by a small sample of actual publication authors. Besides highlighting the suitability of the aforementioned VLMs in aiding visually impaired individuals, the explored approaches may also serve as orientation for familiarizing authors and publishers of scientific articles with the needs of assistive tech and the increasing expectations in accessibility regulations.

Keywords: vision-language models, accessibility, scientific figures, automated captioning, RefCLIPScore, visual impairment, blind

Introduction

The academic publishing ecosystem presents a persistent accessibility gap: millions of scientific figures published each year (graphs, flowcharts, photographic composites, and multi-panel visualizations) are accompanied only by brief author-written captions that are insufficient for visually impaired readers, as they are written for sighted domain experts rather than as standalone visual descriptions. Assistive technologies such as screen readers can parse the surrounding text of a paper but often are blind to the visual content itself. For visually impaired researchers, students, and conference attendees, this represents a fundamental barrier to scientific participation.

Manual alt-text authoring is rarely practiced in academic publishing. Accessibility guidelines such as WCAG 2.1 [1] recommend descriptive alternative text for all meaningful images, yet a survey of 250 leading journals found that no journal provided meaningful alt-text meeting accessibility standards, and PDFs from 85% of surveyed journals contained no alt-text whatsoever [2, 3]. The problem is structural: authors may face time pressure, editorial workflows might lack enforcement mechanisms, and writing an accessible description of a multi-panel experimental result requires both domain knowledge and an understanding of what visual detail a blind reader needs.

Automated captioning via vision-language models (VLMs) offers a plausible intervention. Recent instruction-tuned VLMs can follow natural-language prompts to describe images in detail, and several have demonstrated strong performance on general image captioning benchmarks [4]. However, scientific figures differ substantially from the natural images that populate most benchmarks; they may often feature domain-specific notation, dense multi-panel layouts, abstract data visualizations, and captions that often reference external context not visible in the figure itself. It is therefore not clear which VLMs perform best on scientific content, or whether existing evaluation metrics correctly rank models for this task.

This paper reports on a systematic empirical evaluation addressing these questions. We make the following contributions:

1. A curated evaluation dataset of 245 scientific figures from 32 EI 2025 conference papers, spanning 15 figure types, with author-provided ground-truth captions.
2. A systematic evaluation of five open-source VLMs under standardized conditions, revealing that Moondream2 achieves the best semantic alignment despite being the smallest model evaluated.
3. A demonstration that image-grounded metrics (RefCLIPScore, SBERT) provide more meaningful rankings than lexical metrics (BLEU, ROUGE-L) for scientific figure captioning, where reference captions are brief author labels rather than accessibility descriptions.
4. A brief summary of sample anecdotal VLM output assessments by a sample of selected conference contribution authors.

Related Work

Vision-Language Models for Image Captioning

Large-scale pretraining on image-text pairs has produced a generation of models capable of flexible image description. BLIP-2 [5] extends a frozen ViT-g vision encoder and a large language model via a lightweight Q-Former adapter, enabling efficient zero-shot visual question answering and captioning. LLaVA-1.5 [6] connects CLIP visual features to a large language model via a 2-layer MLP projection, showing that simple alignment suffices for strong instruction-following behavior. Moondream2 [7] is a compact (~1.86B) but capable model designed for efficient deployment. Qwen2-VL [8] introduces Naive Dynamic Resolution and Multimodal Rotary Position Embedding (M-RoPE) to handle varying image resolutions. Idefics3-8B [9] is an 8B-parameter open model building on the Idefics family, combining SigLIP-SO400M with Llama 3.1, with significant improvements in document

understanding achieved through training exclusively on open datasets.

Accessibility and Alternative Text

Research on automated alternative text generation has primarily addressed natural images [10] and web content [11]. Scientific figures present qualitatively different challenges: they are abstract, contain domain-specific notation, often derive meaning from spatial arrangement or color encoding, and are frequently composite, combining multiple sub-figures that each require independent interpretation. Prior work on chart captioning [12, 13] and figure summarization in scientific papers [14] has explored subsets of this problem, but comprehensive evaluation across figure types and models remains limited. To our knowledge, this is the first study to evaluate instruction-tuned VLMs on captioning of scientific figures drawn from a single conference proceedings dataset, with direct attention to accessibility use cases.

Evaluation Metrics for Generated Captions

After caption generation by the aforementioned VLMs, these outputs were assessed with regards to textual similarity and visual alignment. Caption evaluation relied on BLEU [15] and ROUGE-L [16] methods, which measure n-gram overlap with reference text. These metrics are known to correlate poorly with human judgement when reference captions are short or stylistically different from generated text [17]. Sentence-BERT (SBERT) [18] computes dense embedding cosine similarity, offering a more semantic comparison. CLIPScore [19] measures the alignment between an image and a generated caption using a pretrained CLIP model [20], bypassing the reference caption entirely. RefCLIPScore extends this by taking the harmonic mean of CLIPScore(image, generated) and CLIPScore(generated, reference), rewarding both visual grounding and reference alignment. For scientific figure captioning, where reference captions are often minimal and not written for accessibility, we argue that RefCLIPScore and SBERT are more meaningful than BLEU/ROUGE-L, and present evidence supporting this claim.

Dataset

Source Material

Our dataset is drawn from 32 technical papers presented at IS&T Electronic Imaging 2025 (EI 2025), spanning eight conference tracks: Autonomous Vehicles and Machines (AVM), Human Vision and Electronic Imaging (HVEI), Image Quality and System Performance (IQSP), Imaging Sensors and Systems (ISS), Mobile and Ubiquitous Multimedia (MOBMU), Media Watermarking, Security, and Forensics (MWSF), Stereoscopic Displays and Applications (SDA), and Image Processing and Analysis (IPAS / IMAGE). The multi-track sourcing provides naturally diverse content: automotive sensor imagery, perceptual psychophysics results, colorimetric measurements, face-recognition pipelines, and more.

Extraction and Preprocessing

Figures were extracted from PDF source files using PyMuPDF, rasterized at 300 DPI. A spatial clustering algorithm grouped image regions based on vertical proximity and size similarity, preserving composite and multi-panel figures as single units while separating distinct figures on the same page. A two-pass extraction pipeline was applied: the first pass captured figure regions with broad padding to ensure full coverage; the second pass re-cropped each region with tight 2-pixel padding to eliminate adjacent text bleed. For papers where automated extraction produced quality issues, figures were manually cropped from the source PDFs.

Captions were extracted automatically using regex-based text block analysis, matching figure numbers to their corresponding captions. Manual correction was applied in cases where automated caption extraction captured surrounding body text. Ground-truth captions were taken directly from the figure captions in the published papers and were not rewritten for accessibility. This is a deliberate design choice which facilitated assessment as to how well models capture visual content relative to what authors themselves considered the key description of each figure.

The final dataset comprises 245 figures drawn from all 32 papers, with a mean of 7.7 figures per paper (ranging between 2 and 15).

Figure Type Distribution

All 245 figures were visually classified into 15 categories by manual inspection (Table 1). Classification was performed by visual examination of each image by two authors, with caption-based keyword matching as a fallback. Notably, 53% of figures had their label corrected from caption-based to visual-based classification, confirming that automated caption classification alone is insufficient.

Table 1: Figure type distribution across the 245 validated figures.

Figure Type	Count	%
Composite / multi-panel	64	26.1%
Diagram (flowchart / architecture)	41	16.7%
Result visualization	26	10.6%
Bar chart	26	10.6%
Line chart / spectral plot	17	6.9%
Sample image	16	6.5%
Photograph	16	6.5%
Scatter plot	10	4.1%
Screenshot	7	2.9%
3D plot / surface / point cloud	6	2.4%
Heatmap	5	2.0%
Illustration / 3D render	4	1.6%
Table image	3	1.2%
Box plot / violin plot	2	0.8%
Map / floor plan	2	0.8%

Composite figures were the most encountered type (26.1%), reflecting the ubiquity of multi-panel result presentations in computer vision and imaging research. The distribution is naturally imbalanced, with rare categories such as maps, box plots, and table images each comprising fewer than five figures.

Models

We evaluated five open-source, instruction-tuned VLMs that were publicly available at the time of the study (Table 2). Models were selected based on successful inference under our evaluation setup, producing coherent outputs in response to a natural language instruction prompt.

Table 2: Models included in the evaluation.

Model	Parameters	Architecture	Reference
BLIP-2	~3.9B	Frozen ViT-g + Q-Former + OPT-2.7B	[5]
LLaVA-1.5-7B	7B	CLIP ViT-L/14 + Vicuna-7B (MLP proj.)	[6]
Moondream2	~1.86B	SigLIP + Phi-1.5	[7]
Qwen2-VL-2B	2B	ViT + Qwen2-2B (M-RoPE)	[8]
Idefics3-8B	8B	SigLIP-SO400M + Llama-3.1-8B	[9]

All models were run on Google Colab with an A100 GPU (40 GB VRAM). LLaVA-1.5-7B, Qwen2-VL-2B, and Idefics3-8B were loaded in 4-bit quantization to reduce memory footprint. BLIP-2 and Moondream2 were loaded in float16 without quantization. A standardized instruction prompt was used across all models:

"Describe this scientific figure in detail. Include what type of visualization it is, what it shows, and any visible text, labels, or data."

The max_new_tokens was set to 512 for BLIP-2, LLaVA-1.5-7B, and Moondream2, and to 1024 for Qwen2-VL-2B and Idefics3-8B, where preliminary testing showed outputs were being truncated at 512. Using a shared prompt and comparable token budget across models is essential for a fair comparison; prior pilot experiments with inconsistent prompting confirmed that prompt differences substantially alter output length and metric scores.

Evaluation Methodology

Caption Cleaning

Raw model outputs were preprocessed before scoring to remove artefacts introduced by the evaluation setup rather than the model's visual understanding. Four specific cleaning steps were applied:

1. Prompt prefix removal (BLIP-2). BLIP-2 was invoked in VQA mode, causing it to prepend the question string to its output. This prefix was stripped before scoring.
2. Markdown formatting removal (LLaVA, Qwen2-VL, Idefics3). These models frequently produce bold text, section headers, and bulleted lists. Markdown symbols were stripped; underlying content was preserved.
3. Repetition loop truncation (BLIP-2). Without a repetition penalty, BLIP-2 occasionally enters decoding loops that repeat sentences. Outputs were truncated after the second occurrence of any sentence.
4. Whitespace normalization (all models). Multiple spaces, tabs, and newlines were collapsed to single spaces.

Caption length differences between models were not corrected. A model generating a ~440-word description versus one generating ~15 words reflects genuine differences in output behavior that affect both accessibility quality and metric scores and should in the authors' opinion be reported and interpreted, not silenced.

Metrics

In order to facilitate evaluation of textual similarity and visual alignment of the VLM-generated captions, four metrics were chosen for each image-caption pair, as none of the approaches below achieves comprehensive capture of the aforementioned objectives by themselves:

- BLEU [1] measures precision-based n-gram overlap (1-4 grams) between the generated and reference caption, with a brevity penalty. It is highly sensitive to exact wording and length, and low scores are expected across all models given the mismatch between verbose generated descriptions and short reference captions.
- ROUGE-L [2] instead uses an F-score based on the longest common subsequence, making it slightly more tolerant of length differences than BLEU.
- Beyond lexical overlap, SBERT cosine similarity [3][24], was computed between sentence embeddings produced by all-MiniLM-L6-v2. This captures semantic similarity independent of surface phrasing and is expected to track actual description quality more closely than n-gram metrics.
- The primary metric chosen was RefCLIPScore using ViT-B/32 CLIP features. Agrawal et al. defines it as the harmonic mean of CLIPScore(image, generated caption) and CLIPScore(generated caption, reference caption) [4]:

$$\text{CLIPScore}(a, b) = w \cdot \max\left(\cos(f_a, f_b), 0\right),$$

where

- a, b: The image and text being compared
- f_a, f_b: The normalized feature embeddings extracted using the CLIP image or text encoders
- $\cos(f_a, f_b)$: The cosine similarity between the two embeddings
- w = 2.5: A scaling factor as established by Hessel et al. [19] used to bring the score into a more interpretable range.

A RefCLIPScore above 1.0 indicates that the generated caption achieves strong alignment to both the image content and the reference caption simultaneously, which can legitimately occur when reference captions are short or highly abstract. This metric rewards both visual grounding and textual consistency with the reference, making it the most appropriate ranking criterion for this task. A GT-CLIPScore baseline of 0.7137 was computed as the mean CLIPScore(image, reference caption) across the dataset. Because the ground-truth captions are fixed, this value is identical for all models and can therefore serve as a reference point: any model's CLIPScore above 0.7137 indicates that the generated caption aligns more closely with the image than the original paper caption does.

Together, these four metrics provide complementary perspectives on caption quality. BLEU and ROUGE-L measure how closely the generated text matches the exact wording of the reference caption, while penalizing verbosity and rewarding brevity. SBERT operates at the semantic level, capturing whether the generated description conveys the same meaning as the reference regardless of phrasing. RefCLIPScore adds the image itself into the evaluation, while rewarding captions that are visually accurate and not merely textually similar to the reference. A model that scores well across all four metrics would produce concise, semantically faithful, and visually grounded descriptions, but in practice, as we show in the subsequent metric correlation subsection, these metrics

often disagree, which by itself is informative about what each model is doing.

Results

Aggregate Performance

Table 3 reports mean scores across all 245 figures. RefCLIPScore is used as the primary ranking criterion, with SBERT as a secondary metric.

Table 3: Mean scores across 245 figures. Best score per column in bold. RefCLIPScore standard deviation reported for all models. GT baseline is the mean CLIPScore of ground-truth captions vs. images (fixed, same for all models).

Rank	Model	BLEU	ROUGE-L	SBERT	CLIPScore	RefCLIPScore	RefCLIPScore StDEV	Inference Time
1	**Moondream2**	0.0113	0.1030	**0.3922**	*0.8061*	**1.0245**	0.132	8.7 s
2	**Idefics3-8B**	0.0031	0.0400	0.3724	*0.7669*	1.0182	0.108	56.2 s
3	**LLaVA-1.5-7B**	0.0059	0.0892	0.2920	*0.7639*	0.9908	0.121	8.0 s
4	**BLIP-2**	**0.0116**	**0.1197**	0.2433	*0.6672*	0.9451	0.121	**2.9 s**
5	**Qwen2-VL-2B**	0.0030	0.0497	0.2090	*0.5751*	0.8109	0.188	25.7 s
-	*GT Baseline*	-	-	-	***0.7137***	-	-	-

The ranking by RefCLIPScore diverges substantially from the n-gram metric rankings. BLIP-2 places first on both BLEU (0.0116) and ROUGE-L (0.1197) but fourth on RefCLIPScore (0.9451). SBERT rankings perfectly agree with RefCLIPScore rankings across all five models, providing convergent evidence for the semantic ordering. Moondream2 and Idefics3-8B are the only two models with mean RefCLIPScore above 1.0, indicating that their captions achieve strong alignment to both the image content and the reference caption simultaneously. Pairwise Wilcoxon signed-rank tests (paired, non-parametric) were conducted on RefCLIPScore across the 245 figures. All pairwise ranking differences are statistically significant ($p < 0.001$), with one exception: Moondream2 vs. Idefics3-8B ($p = 0.441$). These two models performed at a statistically equivalent level on RefCLIPScore. All other adjacent-rank separations, including LLaVA vs. Idefics3 and BLIP-2 vs. LLaVA, are highly significant, confirming that the overall ranking is robust.

Performance by Figure Type

Table 4 presents mean RefCLIPScore per figure type across all five models. We focus on RefCLIPScore as the primary metric here, as it jointly accounts for image fidelity and reference similarity, making it more informative than BLEU or ROUGE-L for analyzing performance differences across visual categories.

Table 4. Mean RefCLIPScore by figure type (n = number of figures; bold = best per row).

Figure Type	n	BLIP-2	LLaVA	Moondream2	Qwen2-VL	Idefics3	Best
Composite / multi-panel	64	0.929	0.962	0.995	0.780	0.977	Moondream2
Diagram (flowchart/arch.)	41	1.002	1.063	1.097	0.807	1.055	Moondream2
Result visualization	26	0.949	0.981	1.020	0.789	1.049	Idefics 3-8B
Bar chart	26	0.957	1.040	1.100	0.994	1.121	Idefics 3-8B
Line chart / spectral plot	17	0.966	0.958	1.023	0.910	1.013	Moondream2
Sample image	16	0.906	0.958	0.894	0.724	0.963	Idefics 3-8B
Photograph	16	0.929	0.976	0.961	0.643	0.974	LLaVA-1.5
Scatter plot	10	0.987	1.020	1.051	0.908	1.011	Moondream2
Screenshot	7	0.823	0.905	0.969	0.751	0.989	Idefics 3-8B
3D plot / surface	6	0.975	0.913	0.963	0.712	0.977	Idefics 3-8B
Heatmap	5	0.875	0.998	1.067	0.827	1.000	Moondream2
Illustration / 3D render	4	0.854	0.880	1.002	0.824	0.979	Moondream2
Table image	3	0.901	1.046	1.153	0.811	1.068	Moondream2
Box plot / violin plot	2	0.812	1.103	1.126	1.012	1.062	Moondream2
Map / floor plan	2	1.056	0.986	1.044	0.885	0.999	BLIP-2

Several patterns are notable:

Composite figures (26.1% of the dataset, n = 64) proved the most challenging category overall: no model exceeded a RefCLIPScore of 1.0, with Moondream2 performing best at 0.995. These multi-panel figures typically combine heterogeneous content such as photographs, charts, and diagrams in a single image requiring simultaneous description of several visual elements. The consistently sub-1.0 scores across all models indicate that generated

captions describe these figures less completely than the ground truth captions aligning with the respective image.

Moondream2 led on the largest number of individual figure types (8 of 15), including diagrams (1.097), line charts (1.023), scatter plots (1.051), heatmaps (1.067), illustrations (1.002), table images (1.153), and box plots (1.126). Idefics3-8B led on five types: bar charts (1.121), result visualizations (1.049), sample images (0.963), screenshots (0.989), and 3D plots (0.977). LLaVA-1.5-7B led on photographs (0.976) and BLIP-2 led on maps (1.056). Qwen2-VL-2B did not lead on any figure type.

Among the three most common figure types – composites (26.1%), diagrams (16.7%), and result visualizations and bar charts (10.6% each) – diagrams saw the strongest performance, with Moondream2 (1.097), LLaVA (1.063), and Idefics3 (1.055) all comfortably exceeding the GT baseline of 0.714. This suggests that architectural and flowchart-style diagrams, which have relatively structured visual layouts, are better matched to current VLM capabilities than composite or result-visualization figures.

Qwen2-VL-2B scored below 1.0 in 14 of 15 figure type categories, and lowest overall on photographs (0.643), 3D plots (0.712), and sample images (0.724). This consistent underperformance across visual domains contrasts with its moderate ROUGE-L score (0.050) at the aggregate level, illustrating a case where lexical overlap with the ground truth caption does not translate into image-grounded accuracy. This divergence reinforces the argument that RefCLIPScore is a more appropriate primary metric for evaluating accessibility descriptions than n-gram-based metrics.

Caption Length

Table 5: Mean generated caption word count by model.

Source	Mean Words	Ratio to GT
Ground truth	21.9	1.0×
BLIP-2	14.5	0.7×
LLaVA-1.5-7B	80.8	3.7×
Moondream2	91.9	4.2×
Qwen2-VL-2B	216.5	9.9×
Idefics3-8B	439.5	20.0×

BLIP-2 is the only model that approaches ground-truth caption length, producing an average of 14.5 words compared to the ground-truth mean of 21.9 words (0.7x). While slightly shorter than the ground truth, it is far closer than any other model. This explains its superior BLEU and ROUGE-L scores: n-gram precision metrics favors length-matched outputs, regardless of whether the content accurately describes the figure. Moondream2 achieves the best semantic performance (RefCLIPScore, SBERT) with a moderate approximately 92-word output, 4.2 times the ground-truth length. Notably, this is achieved at less than a quarter of the verbosity of Idefics3-8B (approximately 440 words, 20.0 times), demonstrating that the most verbose model does not produce the best-quality descriptions.

Metric Correlation

To assess how well each metric correlates with the primary image-grounded metric (RefCLIPScore), we computed per-model Pearson correlations across the 245-figure dataset and observed ranges across the five models as tabulated below:

Table 6: Metric Pearson Correlation to RefCLIPScore

Metric	r with RefCLIPScore
BLEU	0.13 - 0.30
ROUGE-L	0.03 - 0.37
SBERT	0.33 - 0.47
CLIP Score	0.51 - 0.96

The low correlations between BLEU/ROUGE-L and RefCLIPScore confirm that n-gram metrics are not reliable proxies for visual grounding quality in this domain. The wide per-model range for ROUGE-L (0.03 - 0.37) is notable: Qwen2-VL-2B shows near-zero correlation ($r = 0.034$), reflecting that its generated captions have some lexical overlap with ground truth but do not correspond well to image content. SBERT shows a moderate positive correlation across all models, providing useful semantic signal but remaining an incomplete substitute for image-grounded evaluation.

CLIPScore shows a wide range (0.51 - 0.96), with the variation driven by model behavior: Qwen2-VL-2B and BLIP-2 show high correlation ($r = 0.96$ and $r = 0.88$ respectively), while Moondream2 shows the weakest ($r = 0.51$). This is expected for a consistently high-performing model: when image-text alignment scores are uniformly strong across figures, per-image variation in RefCLIPScore is driven more by the reference similarity component, reflecting differences in ground truth caption quality across figures, than by CLIPScore alone, so the two measures naturally diverge. For Qwen2-VL-2B and BLIP-2, image-text alignment is the dominant source of per-image variance, causing CLIPScore and RefCLIPScore to track closely.

These results empirically validate the use of RefCLIPScore as the primary metric for scientific figure captioning evaluation.

Inference Time and Practical Deployment

From an accessibility tooling perspective, inference speed is practically important: a tool that takes over 55 seconds per figure would be unusable in interactive applications, especially in mobile contexts. BLIP-2 is by far the fastest (mean 2.9 s/figure), driven by its compact output. LLaVA-1.5-7B and Moondream2 are comparable (approximately 8.0 - 8.7 s/figure). Qwen2-VL-2B is slower (approximately 26 s/figure) despite being a 2B-parameter model, likely due to its highly verbose output generation. Idefics3-8B is the slowest at approximately 56 s/figure, consistent with its 8B parameter count and longest outputs.

Given that Moondream2 achieves the best RefCLIPScore (1.025) at a competitive speed (8.7 s/figure) and with a model practical to run on consumer hardware (approximately 1.86B parameters), it represents the most attractive option for real-world accessibility deployment among those evaluated.

Human Author Evaluation

Motivation

Automated metrics measure surface similarity (BLEU, ROUGE-L) or embedding space proximity (SBERT, RefCLIPScore) against reference captions. However, the ultimate goal is accessibility: descriptions that enable a visually impaired person to understand a scientific figure without seeing it. This quality dimension (informativeness, faithfulness, and usability) is not directly captured by any automated metric. To complement the automated evaluation, we conducted a pilot human evaluation to investigate the perceptual accuracy of the model outputs via an author survey, with a larger-scale study planned as future work.

Pilot Study Design

A pilot survey was conducted with a subset of EI 2025 paper authors all of whose figures appeared in our dataset. For this pilot study, selection criteria were limited conference contributor in the authors' personal network in Germany. Being authors to their publications, participants can be considered as domain experts possessing the background knowledge needed to accurately evaluate the VLM captions for their scientific imagery. Surveying platform was LimeSurvey Version 5.4.15, hosted on a university server that was administered by one of the authors. Each author received an individual access token particular to themselves and their paper(s) in question via e-mail in order to prevent participation of respondents beyond the selection criteria.

The survey was structured to begin with an initial question group to determine respondent gender & age, academic background & domain, and country of primary residence. Subsequently, the platform presented each participant with the first image from their own paper(s) and their own ground-truth caption from their paper, followed by the VLM generated captions from all employed models. Authors were not shown the identity of the employed models.

Participants then rated each caption on eight dimensions using a 5-point Likert scale: Self-perceived difficulty for visually impaired readers to experience the image, Factual Accuracy & Precision, Factual Completeness, Expressive Sophistication & Tone, Orthography (i.e. correct grammar, syntax, vocabulary), Contextomy (i.e. inclusion of proper prior and subsequent information), Avoidance of Bias and Misemphasis, as well as Abstraction Capability. In addition, participants had to state whether the respective image was to be experienced along a specific vision path, or whether it could be perceived in random order or direction. Survey participants were also given the opportunity to provide free-response commentary for each image. were queried. If the publication featured additional imagery, the authors were asked whether they sought to continue evaluating the next image. The subsequent respective evaluation question group would then only be shown, if they responded in the affirmative.

In the survey's landing page, respondents were notified that they could abort the survey at any time, all questions were optional, and that deletion requests could be made to the survey administrator at any time without repercussions.

Preliminary Results

Due to the low response rate (n = 8), detailed correlation analyses were omitted, and results are therefore to be considered as anecdotal. Respondents were all in the 18-45 age range, had primarily background in engineering, with additional mentions in formal sciences, natural and social sciences, and interdisciplinary fields, holding master or doctorate degrees.

Consolidated results are tabulated as follows:

Table 7: Mean human evaluation ratings per model across seven quality dimensions (scale: -2 (very poor) to +2 (very good)).

Dimension	BLIP-2	Idefics3-8B	LLaVA-1.5-7B	Moondream2	Qwen2-VL-2B
Factual Accuracy & Precision	-0.50	+1.20	+0.40	+1.20	-0.80
Factual Completeness	-1.50	+1.60	0.00	+0.80	-1.00
Expressive Sophistication & Tone	+0.67	+1.00	+0.80	+1.20	+0.75
Orthography	+1.33	+1.20	+1.60	+1.80	+1.20
Contextomy	0.00	+1.40	+1.25	+1.60	-0.25
Avoidance of Bias & Misemphasis	+1.00	+1.20	+0.50	+0.75	0.00
Abstraction Capability	+1.00	+0.80	+1.00	+1.00	-0.33
Overall Score	-0.25	+1.20	+0.74	+1.21	-0.18

In general, free-text responses were more favorable for the Moondream2 outputs; some criticisms voiced by the authors included omission of image elements (in particular solely "visual components of the diagram without addressing its actual content", misinterpretation of content types, as well as repetitions of statements in the generated captions.

The pilot findings will be used to refine the survey instrument and figure selection criteria for a larger-scale study targeting all conference participants providing contact information is available. Full human evaluation results, including correlation analysis between human ratings and automated metrics, will be reported in a follow-up study. Ultimately, long-term evaluation by visually impaired individuals would provide a holistic effectiveness evaluation involving all stakeholders.

Discussion

Moondream2 as the Best Practical Model

Moondream2 achieves the highest mean RefCLIPScore (1.025) and SBERT (0.392) among all models evaluated and is statistically indistinguishable from Idefics3-8B on RefCLIPScore ($p = 0.441$, Wilcoxon signed-rank). Despite this equivalence at the metric level, Moondream2 also holds advantages for practical deployment: it is six times faster (8.7 s vs. 56.2 s per figure) and requires four times fewer parameters (approximately 2B vs. 8B) than alternative models. This suggests that architecture and training data choices matter more than raw parameter count for this task. Moondream2's SigLIP backbone provides strong visual feature extraction, and its Phi-1.5 language component is a compact base model pre-trained on high-quality textbook-style data [22], providing strong reasoning capabilities. The combination produces outputs that are detailed enough to ground well to CLIP features without excessive verbosity.

Practically, Moondream2 can be run on a single consumer GPU (or even CPU for small batches), making it the most feasible option for deployment in publication workflows, document processors, or conference management systems, suggesting suitability also for implementation on mobile devices.

Why BLEU and ROUGE-L Mislead on This Task

The discrepancy between BLIP-2's n-gram ranking (1st) and its RefCLIPScore ranking (4th) arises from a structural mismatch between how these metrics work and what the task requires. BLIP-2's outputs average approximately 14 words, shorter than the ground-truth mean of approximately 22 words, but far closer than any other model. N-gram precision scores improve with length matching. But short BLIP-2 descriptions (e.g., *"noise limit, saturation, and contrast"*) convey far less to a visually impaired reader than Moondream2's medium-length structured description,

which identifies the figure type, interprets the axes, and describes the data trend. Proclivity of the former models to misinterpret imagery was also noted by the surveyed authors.

This finding has methodological implications for future work: researchers evaluating captioning systems for scientific figures should not rely solely on BLEU or ROUGE-L as primary metrics, particularly when reference captions are brief. SBERT and RefCLIPScore are more appropriate, and human evaluation remains essential for validating accessibility quality.

Exceeding the GT-CLIPScore Baseline

Two models, Moondream2 (1.025) and Idefics3-8B (1.018), have mean RefCLIPScore above 1.0, indicating that their captions achieve strong alignment to both the image content and the reference caption simultaneously. This is not surprising given the nature of scientific captions: they are written by authors as brief, reference-laden labels (*"Illustration of object contrast within the dynamic range"*, *"GPS-Map Route starts and ends directly at the hotel"*), which presuppose visual context and domain knowledge. A VLM generating a detailed description of visible elements can score higher against CLIP than a minimal author caption. This finding reinforces the argument that purpose-written accessibility descriptions would substantially differ from figure captions as they appear in papers.

Composite Figures as a Persistent Challenge

Composite / multi-panel figures are both the most common type in our dataset (26.1%) and the type on which all models perform worst, with no model exceeding a mean RefCLIPScore of 1.0. The challenge is structural: a composite figure may place results from different experiments side by side, display both input and output images side by side, or aggregate sub-figures with independent axes and legends. Current VLMs tend to describe these figures as a whole, providing a general summary, rather than systematically enumerating sub-panel content. Addressing this limitation likely requires multi-step reasoning (for example, detect panels, describe each, then synthesize) or specialized training on multi-panel figure data.

Verbosity vs. Quality

The observation that the most verbose model (Idefics3, approximately 440 words) does not achieve the highest quality (it ranks 2nd on RefCLIPScore) while the most concise non-BLIP-2 model (Moondream2, approximately 92 words) achieves the best, suggests that raw output length is not a useful proxy for description quality. Longer outputs increase the chance of hallucination, generating plausible-sounding text that does not correspond to image content, and may also reduce CLIP alignment if the hallucinated content pulls the text embedding away from the image. For accessibility applications, overly verbose descriptions may also reduce usability for screen reader users who must listen to entire captions sequentially.

Qwen2-VL-2B Underperformance

Qwen2-VL-2B underperforms substantially across all metrics (RefCLIPScore = 0.811, compared to the next-lowest 0.945 for BLIP-2), despite being a competitive general-purpose VLM on standard benchmarks [9]. We attribute this to three factors: (1) extremely verbose outputs (approximately 217 words) with heavy markdown structure, which our cleaning pipeline partially addresses but cannot fully compensate for; and (2) a possible domain gap: Qwen2-VL appears tuned for document-centric tasks with rich visual text, but its outputs for abstract scientific figures tend toward structured enumeration of visible elements rather than synthesized descriptions; and (3) a tendency toward hallucination: human domain experts who participated in our pilot evaluation flagged Qwen2-VL-2B outputs as containing content not present in the figures, suggesting that its verbose outputs introduce fabricated detail alongside accurate description.

Limitations

Ground-truth captions are not pure accessibility descriptions. All reference captions were taken directly from the papers. They are typically short (meaning 22 words), assume domain knowledge, and often reference context external to the figure (e.g. other figures or methods described in the text). This limits the meaningfulness of BLEU/ROUGE-L scores: a model that writes an accurate, detailed description of figure content may score low simply because it uses different vocabulary than the paper caption. SBERT and RefCLIPScore partially address this, but the ideal evaluation would use purpose-written accessibility descriptions as references.

Dataset imbalance: While 15 figure types are represented, many appear fewer than 10 times (box plots, maps, table images, illustrations). Per-type conclusions for rare categories should be treated as exploratory rather than statistically robust.

Human evaluation is preliminary: The pilot survey ($n = 8$) provides early directional signal but is insufficient for statistically robust conclusions. A larger-scale human evaluation study is planned and will be reported separately.

Models were evaluated at fixed settings: We used a single prompt and fixed token budget across all models. It is possible that prompt tuning or few-shot examples would improve performance, particularly for models like Qwen2-VL-2B that appear sensitive to prompting style. Our results characterize zero-shot performance under a standardized setting, which is the relevant scenario for automated deployment.

Single GPU configuration: BLIP-2 and Moondream2 were run in float16; LLaVA-1.5-7B, Qwen2-VL-2B, and Idefics3-8B were loaded in 4-bit quantization on an A100 GPU. Results may differ under different quantization levels, inference frameworks, or hardware.

Conclusion

Our results show that Moondream2 and Idefics3-8B achieve statistically equivalent RefCLIPScore performance (1.025 vs. 1.018, $p = 0.441$), with all other pairwise ranking differences being highly significant ($p < 0.001$). Moondream2 is the recommended deployment choice on practical grounds: it matches Idefics3-8B in image-grounded quality while being 6 times faster (8.7 s vs. 56.2 s per figure) and requiring 4 times fewer parameters (approximately 2B vs. 8B).We demonstrate that standard n-gram metrics (BLEU, ROUGE-L) provide misleading rankings for this task, a finding with direct implications for how caption generation systems should be evaluated in scientific accessibility research. RefCLIPScore and SBERT are more appropriate primary metrics when reference captions are brief and not written for accessibility.

Composite and multi-panel figures remain the most challenging category for all models, despite being the most common figure type in conference papers. Addressing this gap will likely require specialized architecture or training approaches for multi-panel scientific content.

A pilot human evaluation study has been conducted ($n = 8$ domain-expert participants), with a larger-scale follow-up study planned to provide statistically robust validation of automated metric rankings against human accessibility judgements.

Our findings might hold significance for publisher guidelines with regards to both length and level of detail for alt-text descriptions increasingly demanded of authors in the evolving field of accessibility law and regulations which aim to serve the integration of visually impaired individuals in scientific discourse.

Acknowledgements

The European Union partially supported this work through ERASMUS MUNDUS, Project CyberMACS (Project No. 101082683, (https://cybermacs.eu).

References

[1] Kirkpatrick, A., O'Connor, J., Campbell, A., & Cooper, M. (Eds.). (2018). Web Content Accessibility Guidelines (WCAG) 2.1. W3C Recommendation. World Wide Web Consortium.

[2] Crane, M. A., Nguyen, M., Lam, A., Berger, Z. D., Paulus, Y. M., Romley, J. A., & Faden, R. (2023). Figure accessibility in journals: analysis of alt-text in 2021–23. *The Lancet*, 402(10419), 2287–2289. https://doi.org/10.1016/S0140-6736(23)02348-6

[3] Kumar, A., & Wang, L. L. (2024). Uncovering the New Accessibility Crisis in Scholarly PDFs. In *Proceedings of ASSETS '24*. ACM.

[4] Agrawal, H., Desai, K., Wang, Y., Chen, X., Jain, R., Johnson, M., Batra, D., Parikh, D., Lee, S., & Anderson, P. (2019). nocaps: novel object captioning at scale. In *Proceedings of ICCV* (pp. 8948–8957).

[5] Li, J, Li, D., Savarese, S., & Hoi, S. (2023). BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *Proceedings of ICML 2023*.

[6] Liu, H., Li, C., Li, Y., & Lee, Y. J. (2024). Improved baselines with visual instruction tuning. In *Proceedings of CVPR 2024* (pp. 26296–26306).

[7] vikhyatk. (2024). Moondream2 (revision 2024-04-02) [software]. HuggingFace Model Repository. Retrieved from https://huggingface.co/vikhyatk/moondream2 on 25-11-09

[8] Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., … & Zhou, J. (2024). Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution.

[9] Laurençon, H., Marafioti, A., Sanh, V., & Tronchon, L. (2024). Building and better understanding vision-language models: insights and future directions.

[10] Gurari, D., Zhao, Y., Zhang, M., & Bhattacharya, N. (2020). Captioning Images Taken by People Who Are Blind. In *Proceedings of ECCV* (pp. 417–434).

[11] Wu, S., Wieland, J., Farivar, O., & Schiller, J. (2017). Automatic Alt-text: Computer-generated Image Descriptions for Blind Users on a Social Network Service. In *Proceedings of CSCW '17* (pp. 1180–1192). ACM.

[12] Kantharaj, S., Leong, R. T., Lin, X., Masry, A., Thakkar, M., Hoque, E., & Joty, S. (2022). Chart-to-Text: A Large-Scale Benchmark for Chart Summarization. In *Proceedings of ACL 2022* (pp. 4005–4023).

[13] Obeid, J., & Hoque, E. (2020). Chart-to-Text: Generating Natural Language Descriptions for Charts by Adapting the Transformer Model. In *Proceedings of INLG 2020* (pp. 138–147).

[14] Hsu, T.-Y., Giles, C. L., & Huang, T.-H. (2021). SciCap: Generating Captions for Scientific Figures. In *Findings of EMNLP 2021* (pp. 3258–3264).

[15] Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting of the ACL*, 311–318.

[16] Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. *Proceedings of the ACL Workshop: Text Summarization Branches Out*, 74–81.

[17] Fabbri, A. R., Kryscinski, W., McCann, B., Xiong, C., Socher, R., & Radev, D. (2021). SummEval: Re-evaluating Summarization Evaluation. *Transactions of the Association for Computational Linguistics*, 9, 391–409.

[18] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-Networks. *Proceedings of EMNLP-IJCNLP 2019*, 3982–3992.

[19] Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., & Choi, Y. (2021). CLIPScore: A reference-free evaluation metric for image captioning. *Proceedings of EMNLP 2021*, 7514–7528.

[20] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., … & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *Proceedings of ICML 2021*, 8748–8763.

[21] 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.

[22] Li, Y., Bubeck, S., Eldan, R., Del Giorno, A., Gunasekar, S., & Lee, Y. T. (2023). Textbooks Are All You Need II: phi-1.5 technical report. arXiv:2309.05463.

Author Biographies

Ruthra Bellan is a special purpose instructor teaching artificial intelligence and mathematics in the School of Technology and Architecture at SRH University of Applied Sciences –Campus Berlin. She holds a M.Sc. in Big Data and AI and a B.Sc. in Computer Science. Prior to working in academia, she held positions as NLP developer in a Berlin AI company and has more than a decade of experience in data science and quality assurance at a multinational IT consulting & service company. Her research interests include vision language models, brain-machine interfaces, and autonomous systems.

Frank Wittig is a research associate in the School of Business and Law at SRH University of Applied Sciences –Campus Berlin where he teaches graduate-level courses in operations & project management, intellectual property protection, and managerial responsibilities. He is also vice-chairman of the works council of the campus branches in Berlin, Dresden, and Hamburg. He holds an M.A. in International Strategic Management and a B.Sc. in Biochemistry & Molecular Biology. His research interests include labor relations, occupational safety, hospital operations management, and negotiation simulations.

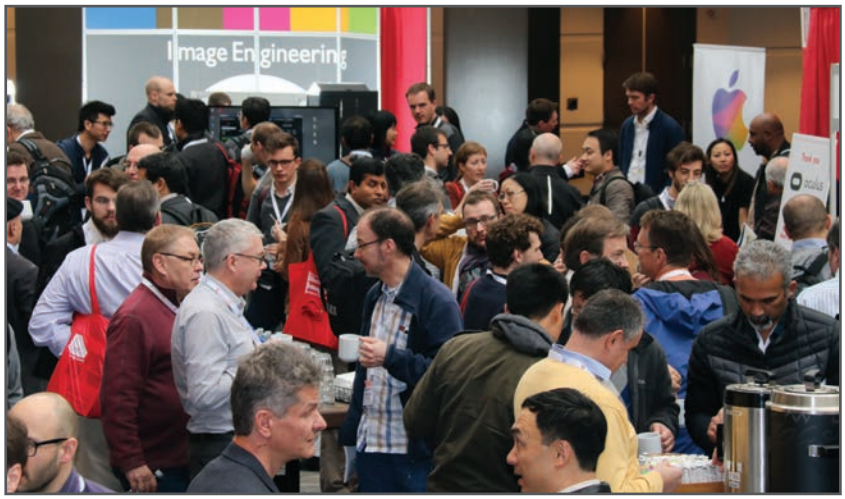
Reiner Creutzburg is a Retired Professor of Applied Computer Science at the Technische Hochschule Brandenburg in Brandenburg, Germany. Since 2019, he has been a Professor of IT Security in the School of Technology & Architecture at SRH University of Applied Sciences -Campus Berlin. In 2025, he was appointed as a Senior Professor of Cybersecurity at the newly founded German University of Digital Science in Potsdam, Germany. He is a member of IEEE and SPIE and has served as chairman of

the Multimedia on Mobile Devices (MOBMU) Conference at the Electronic Imaging Conferences since 2005. In 2019, he was elected as a member of the Leibniz Society of Sciences to Berlin e.V. His research interests include Cybersecurity, Digital Forensics, Open-Source Intelligence (OSINT), Multimedia Signal Processing, e-learning, Parallel Memory architecture, and Modern Digital Media and Imaging Applications.

JOIN US AT THE NEXT EI!

electronic IMAGING

Imaging across applications . . . Where industry and academia meet!



• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

ADVANCING IMAGING SCIENCE 75 1947 IS&T 2022 imaging.org