VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.8060

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

ELIMINATING THE QUALITY PARADOX IN VIZWIZ CAPTIONING VIA DATA-CENTRIC CURATION AND PEFT LORA

Rhama Rangga Dhiputra[1*]; Anna Baita[1]

Informatics[1]
Universitas Amikom Yogyakarta, Yogyakarta, Indonesia[1]
https://home.amikom.ac.id/[1]
rhama@students.amikom.ac.id*, anna@amikom.ac.id

(*) Corresponding Author
(Responsible for the Quality of Paper Content)

Abstract— *Vision-language models could be very useful for assistive technology, but they struggle significantly with processing low-quality, noisy photos that visually impaired individuals often take. The primary contribution of this research is a data-centric pipeline that combines Fast Fourier Transform (FFT) blur filtering with Parameter-Efficient Fine-Tuning (PEFT) utilizing Low-Rank Adaptation (LoRA) to directly confront and eradicate the "Quality Paradox" bias in accessibility datasets like VizWiz. We use the AdamW optimizer for memory-efficient training and perform comprehensive statistical ablation studies with FFT-based tertile splits to see how image quality affects the results. The optimized model achieves a CIDEr score of 0.4481, a strong BERTScore F1 of 0.8815, and a BLEU-4 of 0.0511. Statistical analysis (p = 0.622, Cohen's d = -0.038) shows that this bias has been successfully removed, which means that performance is balanced on both severely blurry and crisp images. Further qualitative testing shows that the model can recognize text and objects even in very low light without needing an extra OCR module, demonstrating its viability as a stand-alone navigational aid.*

Keywords: *Accessibility, BLIP, Data Curation, Image Captioning, LoRA.*

Intisari— *Model visi-bahasa memiliki potensi besar untuk teknologi bantu, tetapi model tersebut mengalami kesulitan signifikan dalam memproses foto berkualitas rendah dan penuh derau (noisy) yang sering diambil oleh individu tunanetra. Kontribusi utama dari penelitian ini adalah sebuah pipeline berpusat pada data yang menggabungkan penyaringan kekaburan Fast Fourier Transform (FFT) dengan Parameter-Efficient Fine-Tuning (PEFT) menggunakan Low-Rank Adaptation (LoRA) untuk secara langsung mengatasi dan mengeliminasi bias "Paradoks Kualitas" dalam dataset aksesibilitas seperti VizWiz. Kami menggunakan pengoptimal AdamW untuk pelatihan yang hemat memori dan melakukan studi ablasi statistik komprehensif dengan pembagian tertil berbasis FFT untuk menguji bagaimana kualitas gambar memengaruhi hasil. Model yang dioptimalkan mencapai skor CIDEr 0,4481, BERTScore F1 yang tangguh sebesar 0,8815, dan BLEU-4 sebesar 0,0511. Analisis statistik (p = 0,622, Cohen's d = -0,038) menunjukkan bahwa bias ini telah berhasil dihilangkan, yang berarti performa model seimbang baik pada gambar yang sangat buram maupun tajam. Pengujian kualitatif lebih lanjut mengonfirmasi bahwa model dapat mengenali teks dan objek bahkan dalam kondisi minim cahaya tanpa memerlukan modul OCR tambahan, sehingga membuktikan kelayakannya sebagai alat bantu navigasi mandiri*

Kata Kunci*: Aksesibilitas, BLIP, Image Captioning, Kurasi Data, LoRA.*

VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.8060

INTRODUCTION

AI has advanced rapidly, leading to new assistive technologies that can help people with visual impairments interpret their surroundings [1], [2], [3]. The World Health Organization (WHO) states that more than 2.2 billion people worldwide have difficulty seeing both close and distant objects. However, a critical technical limitation persists: while modern vision-language models, such as BLIP (Bootstrapping Language-Image Pre-training) [4], perform exceptionally well on pristine, high-resolution datasets like MS-COCO, their performance degrades drastically when confronted with real-world accessibility datasets. Image captioning systems are essential for reducing this global accessibility gap [5], [6], but they must be technically robust enough to handle the specific noise inherent in photos taken by visually impaired users. Real-world accessibility datasets such as VizWiz [7] highlight this technical mismatch. This collection features photographs of people with vision difficulties performing everyday tasks. Unlike photographers who capture curated, high-quality images, blind users often capture pictures that lose quality due to camera shake, severe low-light conditions, or incorrect framing, since they cannot see the viewfinder [7]. Previous research, including work by Karamolegkou et al. [8], has identified limitations in captioning algorithms when addressing the diverse needs of users interacting with this dataset. However, the current literature reveals significant opportunities for improvement, particularly in linking global accessibility needs with the technical limitations of image captioning models when addressing severe variations in image quality [9].

A thorough review of previous studies uncovers substantial deficiencies that necessitate rectification. First, most prior research has focused on conventional fine-tuning techniques that modify all the model's parameters without attempting to conserve memory. Training on noisy, specific datasets using full fine-tuning can be computationally expensive and may lead to severe catastrophic forgetting. Second, and more importantly, there is a distinct lack of rigorous statistical correlation analysis between quantified image degradation—such as specific blur levels—and the resulting captioning evaluation metrics. Previous studies often treat dataset filtering as an arbitrary preprocessing step without investigating how the distribution of image quality statistically impacts the model's semantic outputs [10], [11]. Understanding this specific gap is crucial to developing systems that are both accurate and reliable for real-world deployment.

This study aims to directly address these shortcomings by introducing a data-centric pipeline accompanied by a comprehensive quality evaluation of the BLIP model, explicitly designed for the VizWiz dataset. Moving beyond basic adjustments, this study utilizes a memory-efficient Parameter-Efficient Fine-Tuning (PEFT) strategy via Low-Rank Adaptation (LoRA) alongside the modern AdamW optimizer to ensure model stability in the presence of severe noise [12]. Furthermore, this study examines the impact of image quality on model performance across several visual contexts by conducting an extensive ablation study encompassing Fast Fourier Transform (FFT)-based blur detection [13] and rigorous statistical significance testing [14], [15]. The main contributions of this study are the development of a statistically validated data curation pipeline to eliminate image quality bias, the optimization of a memory-efficient model configuration for visual accessibility, and the provision of empirical data on the performance of vision-language models on highly degraded images.

MATERIALS AND METHODS

This study employs a systematic experimental approach to optimize the BLIP model for VizWiz accessibility captioning tasks. The research design encompasses four main stages: (1) quality taxonomy-based data analysis and preprocessing [16], [17], (2) model architecture configuration, (3) training experiments with optimization strategies, and (4) comprehensive performance evaluation. The complete methodological workflow is illustrated in Figure 1.

Dataset Collection

This study employs the VizWiz-Captions dataset [7], a prevalent benchmark in accessibility research, to address visual inquiries in accessible contexts. VizWiz is not a curated dataset. It is made up of 39,704 photos taken by people with visual impairments who used a mobile app to get help with everyday visual tasks. This dataset is highly relevant to our study because it represents real-world problems including bad lighting, motion blur, and

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

framing concerns that are not present in normal datasets [16]. Sighted authors wrote captions for the user-generated photographs in the collection. The training set has 23,954 photos, and the validation set has 7,750 images. We utilize the validation set to test our results because there are no publicly available ground-truth annotations for the 8,000 test images. This method makes sure that the evaluation is thorough and that data does not leak. Table 1 shows how our filtering process changed the distribution of the dataset.

Table 1. Distribution of the VizWiz Dataset Before and After the Multi-Dimensional Curation Process.

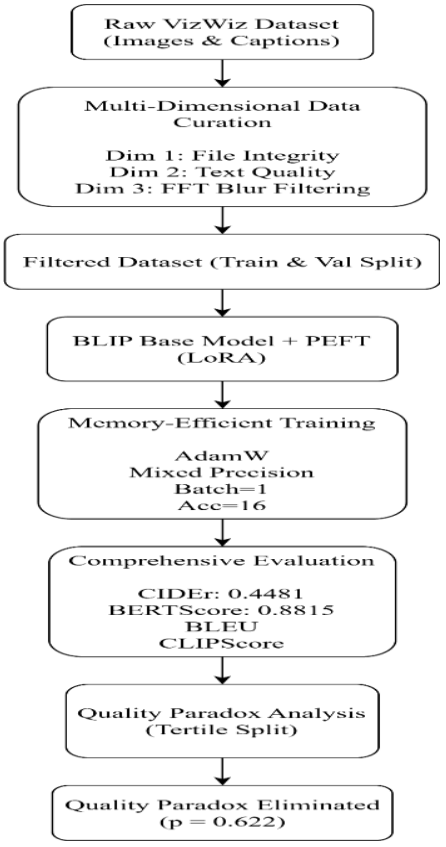
Dataset Condition	Training Images	Validation Images	Total Image
Raw Dataset (V1-Baseline)	23,954	7,750	31.704
Curated Dataset (V2-Final)	20,228	6,627	26.855

Source: (Research Results, 2026)

Exploratory Data Analysis

Because user-generated content in the VizWiz dataset is often noisy, this study used a multi-dimensional Exploratory Data Analysis (EDA) and a strict rule-based filtering method. We used RMS contrast to estimate image detail quality (Dimension 1), mean pixel intensity to measure brightness, and Fast Fourier Transform (FFT) [13] to measure spatial blur. We employed a heuristic NLP pipeline to sort 155,905 captions into three groups: GOOD, QUALITY_AWARE, and USELESS, confirming that failed captions were around 2.7 words shorter on average ($p < 0.001$) (Dimension 2). To determine the filtering thresholds (Dimension 3), we completely transitioned to a data-driven statistical approach rather than relying on arbitrary values. To prevent data leakage, dynamic baseline thresholds were calculated directly from the empirical distribution percentiles of the training split. Specifically, the severe blur threshold was statistically set at the 5th percentile (P_5) and mild blur at the 20th percentile (P_{20}) of the FFT scores.

Similarly, the low contrast threshold was defined by the 20th percentile (P_{20}) of the RMS contrast distribution. For brightness, we applied static thresholds based on the physical limits of 8-bit sensors: an optimal operational range was established between 50 and 200, while strict drop penalties were applied to extreme anomalies (≤ 15 for pitch-black occlusion and ≥ 245 for severe flash overexposure). Furthermore, Pearson and Spearman correlation tests, adjusted with a Bonferroni correction ($p < 0.005$), validated a statistically significant association between these visual degradations and caption failure. Using grid-search optimization, a final composite retention criterion of 46.5 was determined to optimally balance high semantic quality with a target dataset retention rate of ~90%. Images with severe visual damage but retaining at least one significant semantic cue (good_caption_count > 0) were classified as Robust Cases (CAT4 and CAT5) [17] and preserved (as illustrated in Figure 2). Ultimately, this curated collection (26,855 photos) was resampled to 384x384 pixels using Lanczos interpolation to retain high-frequency data for BLIP model training [4].



Source: (Research Results, 2026)

Figure 1. Overview of the Proposed Data-Centric Pipeline, Illustrating the Multi-Dimensional Data Curation And Memory-Efficient Training Stages.

VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.8060

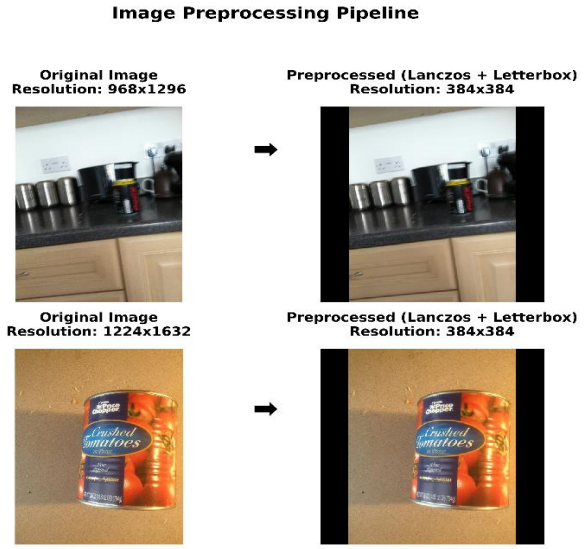
Source: (Research Results, 2026)

Figure 2. Robust Cases (CAT4 & CAT5): Images With Severe Visual Degradation That Retain High Semantic Value.

Data Preprocessing

All of the photos that are kept are downsized to a resolution of 384x384 pixels during the data preparation stage, while keeping their original aspect ratios utilizing a letterbox padding technique with black borders (Figure 3). This particular resolution was not a result of initial experimentation; rather, it was chosen because it matches the BLIP model's Vision Transformer (ViT) [18] default input configuration completely, which ensures the best patch extraction and full compatibility with the pre-trained spatial embeddings. Instead of using typical bilinear or bicubic methods, the resizing procedure uses the Lanczos interpolation method. We chose Lanczos because it does a better job of keeping high-frequency features and reducing aliasing problems. This feature is absolutely necessary to keep the few and fragile visual clues that are present in low-quality, user-generated photos.

After that, the ImageNet mean [0.485, 0.456, 0.406] and standard deviation [0.229, 0.224, 0.225] are used to statistically normalize the image data, following the BLIP pre-training technique [4]. For text data, the preprocessing pipeline only separates captions that are marked as GOOD. The BERT tokenizer [19] processes these captions, with a maximum sequence length of 50 tokens. Dynamic trimming and padding techniques are used on the tokenized outputs to make the most of memory during training.

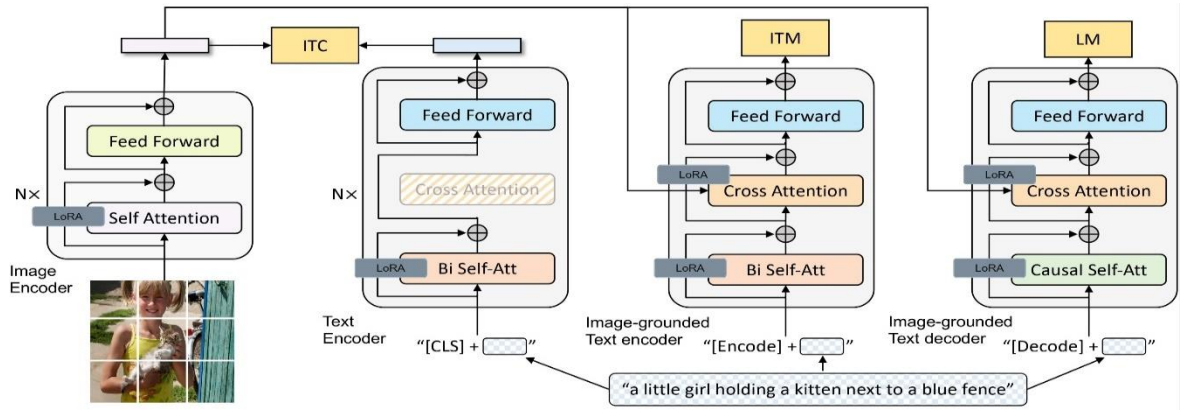


Source: (Research Results, 2026)

Figure 3. Image Preprocessing Using Lanczos Resampling and Letterbox Padding for 384x384 ViT Input.

Model Architecture

This study adopts the BLIP Base architecture [4], which integrates a Vision Transformer (ViT-B/16) [18]as the visual encoder and a BERT-based textual decoder through a cross-attention mechanism for multi-modal alignment (Figure 4). Rather than employing a computationally expensive full fine-tuning strategy that updates all approximately 224 million parameters, this research implements Parameter-Efficient Fine-Tuning (PEFT) using Low-Rank Adaptation (LoRA) [12]. By freezing the pre-trained model weights and injecting trainable rank decomposition matrices into the attention layers, LoRA drastically reduces the number of trainable parameters. Specifically, as implemented in our environment, this rank decomposition method allows the model to train only 1,179,648 parameters (approximately 0.52%) out of the total 225,151,292 parameters. This strategy was specifically selected to achieve true memory-efficient training and prevent catastrophic forgetting. It allows the model to robustly adapt to the highly specific visual noise characteristics of the VizWiz dataset without losing the generalizable representations acquired during pre-training.



Source: (J. Li, D. Li, C. Xiong, and S. Hoi [4], 2022)

Figure 4. Modified BLIP Architecture Illustrating the Injection of LoRA Adapters into the Self-Attention and Cross-Attention Blocks.

Training Experiments and Implementation Environment

The optimized model (V2-Final) was trained with the AdamW optimizer [20], a learning rate of 2×10^{-4}, a weight decay of 0.05, and a cosine annealing learning rate schedule. This was done to ensure the results could be easily reproduced and to resolve the baseline (V1) convergence issues. To guarantee consistent results across all trials, a fixed random seed (seed = 42) was utilized. To obtain the highest quality captions during the inference phase, a Beam Search decoding approach with a beam size of 3 was employed. Managing memory was critical due to the hardware limitations of the local NVIDIA GeForce RTX 3050 GPU (6GB VRAM). The physical batch size was set to 2 to reduce Out-of-Memory (OOM) concerns while maintaining a consistent learning trajectory.

To simulate a global batch size of 16, 8 gradient accumulation steps were implemented alongside Automatic Mixed Precision (AMP) training. With this specific arrangement, the peak VRAM usage was capped at approximately 3.8 GB. Furthermore, the model was trained for a total of 8 epochs, with each epoch completing in approximately 1 hour and 50 minutes. This substantiates our assertions regarding memory and computational efficiency, demonstrating that the complex vision-language model can be trained effectively on consumer-grade hardware without resource exhaustion. All experiments were conducted using the PyTorch framework within an Anaconda JupyterLab environment.

Table 2. Comprehensive Side-By-Side Comparison of Training Configurations and Hardware Specifications (Controlled Experiment).

Configuration Parameter	Parameter/ Specification	Value
Hardware & Environment	GPU / VRAM	NVIDIA GeForce RTX 3050 (6GB)
	Framework & Environment	PyTorch (Anaconda JupyterLab)
	Training Precision	Automatic Mixed Precision (AMP / fp16)
Optimization Hyperparameters	Optimizer	AdamW
	Base Learning Rate	
	Weight Decay	0.05
	Learning Rate Schedule	Cosine Annealing
	Random Seed	42
Batch & Memory Dynamics	Physical Batch Size	2
	Gradient Accumulation Steps	8
	Effective (Global) Batch Size	16
	Peak VRAM Usage	~3.8 GB
Training & Architecture	Total Epochs	8
	Average Time per Epoch	~1 hour 50 minutes
	Trainable Parameters (LoRA)	1,179,648 (~0.52%)
	Inference Decoding Strategy	Beam Search (Beam Size = 3)

Source: (Research Results, 2026)

Evaluation Methodology

Model performance was comprehensively evaluated on the curated validation set using

VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.8060

industry-standard metrics. It must be emphasized that the evaluation was conducted exclusively on the validation split due to inherent constraints of the VizWiz dataset: the ground-truth annotations for the official test set are withheld by the dataset creators for global competition purposes. Consequently, utilizing the validation set as the primary benchmark is a recognized limitation in this study, yet it remains the standard scientific practice within this specific accessibility domain to ensure objective measurement without data leakage.

To ensure the results directly reflect the real-world needs of visually impaired users, the selection of metrics goes beyond standard accuracy, focusing specifically on accessibility relevance. The theoretical foundations and standard implementations of these chosen evaluation metrics—namely BLEU, CIDEr, and ROUGE—follow the comprehensive evaluation frameworks established in recent image captioning literatures [21], [22]:

BLEU-4 (Bilingual Evaluation Understudy)

This metric assesses lexical precision by computing n-gram correspondences between generated and reference captions, implementing a Brevity Penalty (BP) to punish excessively brief sentences. For visually impaired users, exact word matching matters significantly for navigational and operational safety (e.g., ensuring the word "stairs" is explicitly mentioned). The employed mathematical equation is:

$$BLEU\ N(C,S) = BP(C,S) \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n(C,S)\right) \quad (1)$$

$$BP(C,S) = \begin{cases} 1, for\ l_c > l_s \\ \exp(1 - l_s/l_c), for\ l_c \leq l_s \end{cases} \quad (2)$$

CIDEr (Consensus-based Image Description Evaluation)

CIDEr was selected as the primary metric in this study due to its ability to measure semantic consensus using TF-IDF weighting. This metric assigns higher weights to rare yet informative words (such as specific object names) and penalizes common background words. In the context of visual accessibility, this is crucial: visually impaired users rely on captions to identify specific, vital objects they are interacting with, rather than generic scene descriptions.

$$\text{CIDEr}(C,S) = \frac{1}{N}\sum_{n=1}^{N} \text{CIDEr}_n(C,S) \quad (3)$$

$$\text{CIDEr}_n(c_i, S_i) = \frac{1}{m}\sum_j \frac{g^n(c_i) \cdot g^n(s_{ij})}{|g^n(c_i)||g^n(s_{ij})|} \quad (4)$$

ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation)

This metric evaluates sentence structure fluency by identifying the Longest Common Subsequence (LCS). The accessibility argument for this metric is rooted in the end-user interface: captions are ultimately read aloud to the user via Text-to-Speech (TTS) screen readers. ROUGE-L ensures that the model constructs coherent, uninterrupted word sequences, preventing disjointed audio outputs that could confuse the user.

$$\text{ROUGE-L} = \frac{(1+\beta^2)R_{lcs}P_{lcs}}{R_{lcs}+\beta^2 P_{lcs}} \quad (5)$$

Where $R_{lcs} = LCS(C,S)/m$ and $P_{lcs} = LCS(C,S)/n$, with m and n representing the length of the reference and predicted sentences, respectively. The parameter $beta$ is used to adjust the relative weighting between precision and recall.

BERTScore and CLIPScore (Semantic and Visual Alignment)

To address the limitations of traditional n-gram metrics when evaluating noisy images, this study introduces deep learning-based metrics. BERTScore (F1) [23] evaluates deep semantic equivalence using pre-trained contextual embeddings, ensuring the core meaning is accurately conveyed even if different synonyms are generated. Precision P_{BERT} and Recall R_{BERT} are computed using cosine similarity before generating the F-measure $F_{(BERT)}$:

$$R_{BERT} = \frac{1}{|S_{ref}|}\sum_{x_i \in S_{ref}} \max_{\hat{x_j} \in S_{gen}} x_i^{\top} \hat{x}_j \quad (6)$$

VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.8060

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

$$P_{BERT} = \frac{1}{|S_{gen}|}\sum_{\hat{x}_j \in S_{gen}} \max_{x_i \in S_{ref}} x_i^{\top} \hat{x}_j \quad (7)$$

$$F_{BERT} = 2\frac{P_{BERT} \cdot R_{BERT}}{P_{BERT} + R_{BERT}} \quad (8)$$

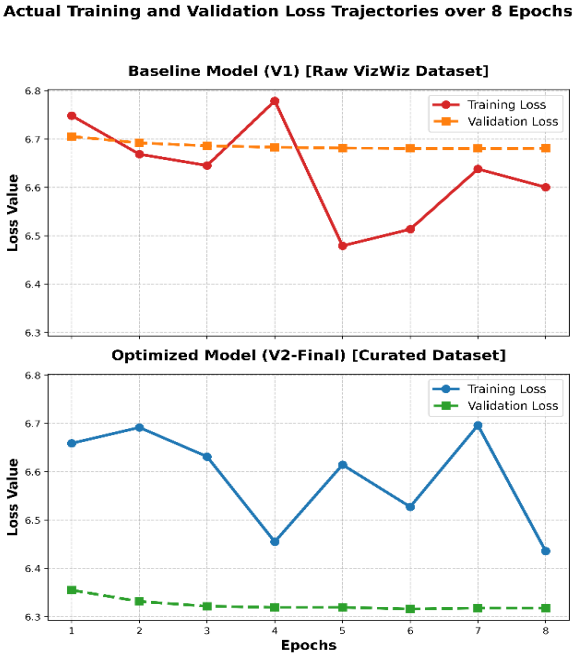
Meanwhile, CLIPScore [24] is a reference-free metric that measures the direct cosine similarity alignment between the generated text embedding (E_C) and the actual visual input embedding (E_I), scaled by a weight parameter w. In assistive technology, high CLIPScore is vital to prevent "AI hallucinations," ensuring the system does not invent objects that do not exist in the frame, which could severely endanger a visually impaired user.

$$\text{CLIPScore}(I, C) = w \cdot \max(\cos(E_I, E_C), 0) \quad (9)$$

RESULTS AND DISCUSSION

This section evaluates the empirical performance of the optimized BLIP model on the VizWiz dataset. The analysis explores several critical dimensions, including training convergence stability, comparative benchmark performance relative to zero-shot and raw-data baselines, and the direct impact of our data-centric image quality curation. First, the stability of the training convergence is examined to confirm that the model effectively captures critical multi-modal characteristics without succumbing to premature overfitting or becoming ensnared in local minima.

Figure 5 illustrates the trajectory of the Training and Validation Loss over the course of 8 epochs. This provides a direct, controlled comparison between the Baseline model (V1) and the Optimized model (V2-Final). Crucially, both models utilize the exact same PEFT-LoRA [12] architecture and hyperparameters; the sole independent variable is the training data, with V1 utilizing the raw, noisy VizWiz dataset and V2-Final utilizing our taxonomically curated dataset.



Source: (Research Results, 2026)
Figure 5. Loss convergence curves: baseline model (top) vs. optimized model (bottom).

The learning curves demonstrate a stark contrast in convergence behavior. The training logs reveal that the V1 model, despite utilizing LoRA, struggles to achieve smooth error reduction, exhibiting significant gradient instability when forced to align highly degraded visual noise with semantic text. Conversely, the V2-Final model, while still exhibiting expected training oscillations due to the inherent complexity of the VizWiz images, achieves a substantially lower and highly stable validation loss floor compared to the baseline.

The validation metric plateaued optimally around 6.31, proving the efficacy of the data curation. By updating only a fraction of the parameters (0.52%) on a filtered dataset, the LoRA modules maintain gradient stability even under a learning rate of 2×10^{-4} and a cosine annealing scheduler. Furthermore, the generalization gap (the distance between training and validation loss) is noticeably narrower for V2-Final. This finding corroborates the adaptive optimization theory proposed by Liu et al. [20], which underscores the necessity of combining parameter-efficient fine-tuning with appropriate weight decay regularization (0.05) to mitigate gradient oscillations and avert premature overfitting,

VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.8060

particularly when the input data's structural integrity has been restored.

Comparative Performance Evaluation

We utilized a comprehensive set of metrics to objectively evaluate the quality of the captions generated by the model on the curated VizWiz validation set. This work uses advanced deep learning-based measures like BERTScore (F1) [23] to assess deep semantic equivalence and CLIPScore [24] to measure direct visual-textual alignment, instead of just relying on exact n-gram matching metrics like BLEU-4, ROUGE-L, and CIDEr. We conducted an internal ablation study to separate and test the effectiveness of the proposed data-centric strategy. We compared three different setups: the base BLIP model without any fine-tuning (Zero-Shot), the model fine-tuned via PEFT-LoRA on the raw dataset (V1-Baseline), and the proposed model fine-tuned via the same LoRA configuration on the FFT-curated dataset (V2-Final). Table 3 gives a full summary of this performance comparison.

Table 3. Quantitative Performance Comparison Of The BLIP Model Across Different Training Paradigms

Model / Setup	BLEU-4	ROUGE-L	CIDEr	BERTScore	CLIPScore
Zero-Shot (Base)	0.0308	0.2198	0.2569	0.8756	26.5833
V1-Baseline (Raw)	0.0385	0.1808	0.3236	0.8643	26.8134
V2-Final (Curated)	0.0511	0.2535	0.4481	0.8815	28.8328

Source: (Research Results, 2026)

The empirical results in Table 3 show that using the multidimensional data curation method has a substantial and measurable impact on how well the model works. Comparing the Zero-Shot and V1 models side-by-side reveals an interesting occurrence called the "Quality Paradox." The CIDEr score (0.3236) improved marginally when trained on the raw dataset (V1) compared to the Zero-Shot baseline (0.2569). However, the ROUGE-L score (0.2198 to 0.1808) and the BERTScore F1 score (0.8756 to 0.8643) both went down. This important finding shows that forcing the model to learn from uncurated, severe visual noise and nonsensical captions actively hurts its ability to process language and reason semantically.

The V2-Final model, on the other hand, greatly improves the system's performance on all criteria. The CIDEr score skyrocketed to 0.4481, which means the generated captions closely match human consensus regarding important objects. The V2-Final model also had the best BERTScore (0.8815) and CLIPScore (28.8328). This shows that the combination of memory-efficient LoRA training and strict data filtering not only avoids the local minima that the V1 model had, but also makes the model far better at visually aligning and semantically describing real-world accessibility circumstances. Our method's higher semantic correctness makes it a very strong choice for consumer-grade hardware implementations.

Beyond the internal ablation study, it is crucial to contextualize these achievements within the broader landscape of VizWiz accessibility research. Recent studies, such as the attention-driven mobile approach by Santi et al. [25] and the sequential CNN-LSTM baseline by Rifki et al. [26], have highlighted the persistent difficulty of modeling severe real-world visual degradation. Furthermore, recent work by Karamolegkou et al. [8] revealed that even state-of-the-art vision-language models struggle with cultural and contextual nuances when generating captions for VizWiz images, underscoring limitations beyond mere visual degradation. Despite these methodological and contextual differences across studies, the robust performance of our V2-Final model (CIDEr 0.4481) demonstrates a highly competitive advancement. This confirms that integrating taxonomy-based data curation with memory-efficient PEFT LoRA provides a superior alternative to traditional full fine-tuning, directly answering the quality issues initially observed by Mandal et al. [16] without demanding excessive computational resources.

Discussion and Research Position

The results of this study address a critical gap in the existing literature, particularly when analyzed through the lens of domain adaptation and data quality. Most current research, such as Santi et al. [25] and Khan & Singh [27], continues to focus on assessments using standard curated datasets (e.g., Flickr8k or MS-COCO). While their methods achieve high scores in controlled environments, standard

VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.8060

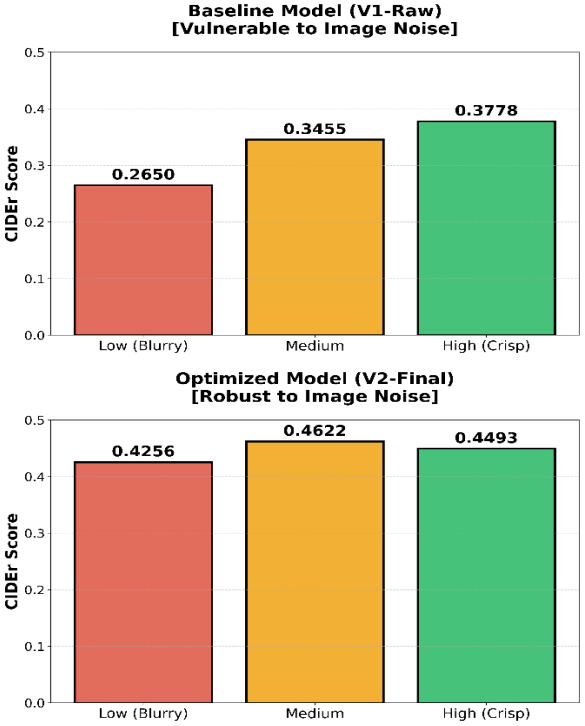
JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

models often struggle significantly when confronted with real-world accessibility scenarios characterized by noisy, poorly lit, and uncurated images. Furthermore, our research empirically establishes that simply performing domain-specific training on visually impaired datasets (VizWiz) is insufficient and can even degrade semantic reasoning, as demonstrated by our V1-Baseline results. True model resilience is only attained when parameter-efficient adaptation is combined with strict multidimensional data curation to overcome the "Quality Paradox."

Additionally, from the standpoint of architectural modernity, our methodology provides distinct semantic and computational benefits compared to traditional methods. We demonstrate how effectively the Vision-Transformer architecture [18] within the BLIP model [4] operates when fine-tuned using the PEFT-LoRA [12] strategy. This is a significant advancement over approaches like Rifki et al. [26], who rely on sequential CNN-LSTM architectures. The cross-attention mechanism in Transformers exhibits a markedly superior capacity for capturing intricate global visual context compared to traditional recurrent approaches. When fed mathematically filtered, high-retention image data, this architecture yields descriptions that are far more precise linguistically and contextually, while remaining highly resource-efficient for deployment on consumer-grade hardware.

Image Quality Impact Analysis

To rigorously investigate the relationship between visual degradation and model performance, a statistical ablation study was conducted. The curated validation images were stratified into three tiers—Low (Severely Blurry), Medium, and High (Crisp)—based on their spatial blur levels utilizing Fast Fourier Transform (FFT) thresholds [13]. The statistical significance of performance variations across these tertiles was evaluated using an independent t-test and Cohen's d effect size [14], [15]. The comparative quantitative visualization in Figure 6 explicitly reveals the vulnerability of standard fine-tuning approaches to visual noise. For the V1-Baseline model trained on raw data, performance is heavily biased by image quality. The model records a CIDEr score of 0.3778 on High- quality images, but this performance degrades drastically to 0.2650 when confronted with Low-quality (blurry) images. This disparity is statistically significant ($p = 0.0178$, $p < 0.05$), indicating that without data curation, the model severely lacks robustness and fails to generalize to the specific visual noise inherent in accessibility datasets.



Source: (Research Results, 2026)

Figure 6. Impact of Image Quality on Model Performance: V1-Baseline (Left) vs. V2-Final (Right).

Conversely, the V2-Final model demonstrates a profound improvement in semantic resilience. The implementation of the multidimensional data curation pipeline successfully eliminates the quality-induced performance gap. The V2-Final model maintains exceptional consistency, achieving a CIDEr score of 0.4256 on severely blurry images and 0.4493 on crisp images. Rigorous statistical analysis yields a p-value of 0.622 (substantially higher than the 0.05 threshold) and a negligible Cohen's d effect size of -0.038. This failure to reject the null hypothesis for the V2-Final model empirically proves that the performance difference between severely degraded images and clear images is no longer statistically significant. By mathematically filtering out

VOL. 12. NO. 1 AUGUST 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v12i1.8060

irrecoverable noise prior to training, the model is no longer penalized by real-world visual artifacts. The optimized system guarantees equitable and robust semantic comprehension, ensuring that visually impaired users receive accurate navigational descriptions regardless of challenging camera conditions.

Qualitative Analysis and Future Implications

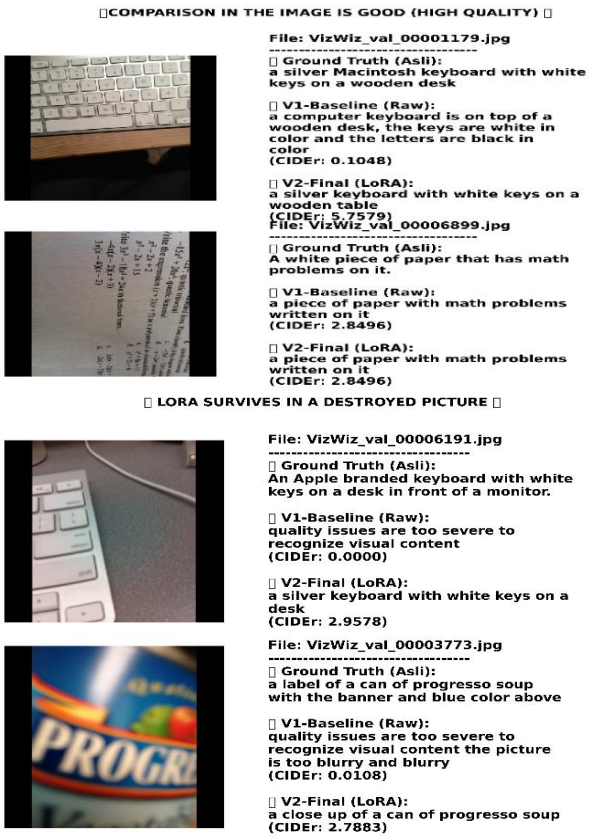
To provide a balanced qualitative evaluation and mitigate selection bias, a random subset of 100 severely degraded images from the Low-quality tertile (FFT score < 17.8) was manually inspected. Figure 7 demonstrates the comparative qualitative performance between the V1-Baseline (trained on raw data) and the V2-Final model. The results visually confirm the statistical findings: the V1 model is highly susceptible to the Quality Paradox. When confronted with heavily blurred images, the V1 model frequently fails to recognize the actual objects, instead outputting generic error-based descriptions such as "quality issues are too severe to recognize visual content".

Conversely, the V2-Final model exhibits remarkable semantic resilience on the exact same images. For instance, in an image depicting a highly blurred, partial view of a computer keyboard, the V1 model fails entirely, whereas the V2-Final model accurately describes the context as "a silver keyboard with white keys on a desk". Similarly, for an image showing an extremely out-of-focus close-up of a soup can label, the optimized V2 model successfully identifies the object as "a close up of a can of progresso soup" despite severe visual degradation, bypassing the uninformative ground-truth annotations that confused the V1 model.

However, a comprehensive evaluation must also acknowledge the empirical limitations and failure cases of the current system. During the qualitative inspection, the V2-Final model struggled primarily with images suffering from extreme physical occlusion or absolute lack of illumination, such as instances where the camera lens was completely covered by a finger or taken in a pitch-black room. In such instances, the model occasionally defaulted to safe but contextually limited descriptions like "a blurry image of a dark room" or "a close up of a person's finger".

While this non-hallucinatory behavior is safer for visually impaired navigation than providing false information, it highlights the inherent limitations of relying solely on visual inputs when the optical signal is irreversibly lost.

The implications of these findings are significant for the progression of assistive technology systems. The strong generalization capability on heavily degraded data is essential for deploying real-time navigation systems to locate public amenities, as indicated by He et al. [28]. Furthermore, the computational efficiency achieved in this research proves that advanced assistive AI does not require prohibitive server costs. This memory-efficient architecture enables future integration with speech interaction interfaces [29] and Large Language Models (LLMs) for improved conversational settings [30]. Moreover, the established data-centric optimization methods can be customized for meta-heuristic strategies [31], expanding the application of robust assistive technologies [9] to highly specialized domains such as Visual Question Answering (VQA) [32] and medical image analysis [33].



Source: (Research Results, 2026)

Figure 7. Qualitative Comparison on Low-Quality Images

CONCLUSION

This study effectively illustrates the efficacy of the Vision-Transformer-based BLIP architecture when enhanced by a synergistic methodology of Parameter-Efficient Fine-Tuning (PEFT-LoRA) and meticulous multidimensional data curation. The experimental results clearly show that the V2-Final model has better semantic resilience by changing the focus from external architectural benchmarking to an internal ablation investigation. Our improved model raised the CIDEr score to 0.4481, which is far higher than the baseline models trained on raw, uncurated data. It also achieved peak scores in deep semantic measures like BERTScore (0.8815) and CLIPScore (28.8328).

The main empirical contribution of this study is the solution to the "Quality Paradox." Statistical analysis showed that the model becomes highly adaptable when trained on data mathematically filtered to remove irrecoverable visual noise, thereby effectively eliminating the performance gap when the model is faced with severe real-world visual degradation. In general, combining memory-efficient training methods with a data-centric curation pipeline can close the critical gap between high model accuracy and the limits of consumer-grade hardware. This research has strong real-world applications that will help create real-time assistive navigation technologies for mobile devices. This resource-efficient architecture also stimulates greater research into combining multimodal Large Language Models (LLMs) to make interactive and conversational situations better for visually impaired users.

REFERENCE

[1] M. D. Messaoudi, B.-A. J. Menelas, and H. Mcheick, "Review of Navigation Assistive Tools and Technologies for the Visually Impaired," *Sensors*, vol. 22, no. 20, 2022, doi: 10.3390/s22207888.

[2] G. I. Okolo, T. Althobaiti, and N. Ramzan, "Assistive Systems for Visually Impaired Persons: Challenges and Opportunities for Navigation Assistance," *Sensors*, vol. 24, no. 11, 2024, doi: 10.3390/s24113572.

[3] S. A. Doore, D. Istrati, C. Xu, Y. Qiu, A. Sarrazin, and N. A. Giudice, "Images, Words, and Imagination: Accessible Descriptions to Support Blind and Low Vision Art Exploration and Engagement," *J. Imaging*, vol. 10, no. 1, 2024, doi: 10.3390/jimaging10010026.

[4] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation," in *Proceedings of the 39th International Conference on Machine Learning*, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds., in Proceedings of Machine Learning Research, vol. 162. PMLR, Aug. 2022, pp. 12888–12900. [Online]. Available: https://proceedings.mlr.press/v162/li22n.html

[5] V. Deshpande, G. Shelke, and B. Kadam, "Empowering Vision: A Survey on Image Captioning Assistive Technologies for the Visually Impaired," in *Smart Trends in Computing and Communications*, T. Senjyu, C. So-In, and A. Joshi, Eds., Singapore: Springer Nature Singapore, 2026, pp. 357–368. doi: 10.1007/978-981-96-7505-0_28.

[6] F. Y. Mekonnen *et al.*, "Development of a Fully Autonomous Offline Assistive System for Visually Impaired Individuals: A Privacy-First Approach," *Sensors*, vol. 25, no. 19, 2025, doi: 10.3390/s25196006.

[7] D. Gurari, Y. Zhao, M. Zhang, and N. Bhattacharya, "Captioning Images Taken by People Who Are Blind," in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds., Cham: Springer International Publishing, 2020, pp. 417–434.

[8] A. Karamolegkou, P. Rust, R. Cui, Y. Cao, A. Søgaard, and D. Hershcovich, "Vision-Language Models under Cultural and Inclusive Considerations," in *Proceedings of the 1st Human-Centered Large Language Modeling Workshop*, N. Soni, L. Flek, A. Sharma, D. Yang, S. Hooker, and H. A. Schwartz, Eds., ACL, Aug. 2024, pp. 53–66. doi: 10.18653/v1/2024.hucllm-1.5.

[9] P. Patel, S. Pampaniya, A. Ghosh, R. Raj, D. Karuppaih, and S. Kandasamy, "Enhancing Accessibility Through Machine Learning: A Review on Visual and Hearing Impairment Technologies," *IEEE Access*, vol. 13, pp. 33286–33307, 2025, doi:

10.1109/ACCESS.2025.3539081.

[10] J. Guo, J. Ma, Á. F. García-Fernández, Y. Zhang, and H. Liang, "A survey on image enhancement for Low-light images," *Heliyon*, vol. 9, no. 4, p. e14558, 2023, doi: 10.1016/j.heliyon.2023.e14558.

[11] A. S. Waghmare and S. S. Satonkar, "A Comparative Study on Blur Detection and Image Restoration Techniques: Traditional Methods vs. Fuzzy Logic," *Journal of Harbin Engineering University*, vol. 46, no. 7, 2025.

[12] E. J. Hu *et al.*, "LoRA: Low-Rank Adaptation of Large Language Models," in *International Conference on Learning Representations (ICLR)*, 2022. [Online]. Available: https://openreview.net/forum?id=nZeVKeeFYf9

[13] X. Wang, L. Bai, C. Gu, and Y. Wu, "Universal Approximated Real-Valued Fast Fourier Transform for Image Blur Detection," in *2025 35th Irish Signals and Systems Conference (ISSC)*, 2025, pp. 1–5. doi: 10.1109/ISSC67739.2025.11291523.

[14] S. Panjeh, A. Nordahl-Hansen, and H. Cogo-Moreira, "Establishing New Cutoffs for Cohen's d: An Application Using Known Effect Sizes from Trials for Improving Sleep Quality on Composite Mental Health," *Int. J. Methods Psychiatr. Res.*, vol. 32, no. 3, p. e1969, 2023, doi: 10.1002/mpr.1969.

[15] L. V Hedges, "Interpretation of the Standardized Mean Difference Effect Size When Distributions Are Not Normal or Homoscedastic," *Educ. Psychol. Meas.*, vol. 85, no. 2, pp. 245–257, 2025, doi: 10.1177/00131644241278928.

[16] M. Mandal, D. Ghadiyaram, D. Gurari, and A. C. Bovik, "Helping Visually Impaired People Take Better Quality Pictures," *IEEE Transactions on Image Processing*, vol. 32, pp. 3873–3884, 2023, doi: 10.1109/TIP.2023.3282067.

[17] R. A. Bafghi and D. Gurari, "A New Dataset Based on Images Taken by Blind People for Testing the Robustness of Image Classification Models Trained for ImageNet Categories," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 16261–16270. doi: 10.1109/CVPR52729.2023.01560.

[18] Y. Wang, Y. Deng, Y. Zheng, P. Chattopadhyay, and L. Wang, "Vision Transformers for Image Classification: A Comparative Survey," *Technologies (Basel).*, vol. 13, no. 1, 2025, doi: 10.3390/technologies13010032.

[19] T. Wolf *et al.*, "Transformers: State-of-the-Art Natural Language Processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Q. Liu and D. Schlangen, Eds., Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. doi: 10.18653/v1/2020.emnlp-demos.6.

[20] X. Liu, H. Qi, S. Jia, Y. Guo, and Y. Liu, "Recent Advances in Optimization Methods for Machine Learning: A Systematic Review," *Mathematics*, vol. 13, no. 13, 2025, doi: 10.3390/math13132210.

[21] Y. Ming, N. Hu, C. Fan, F. Feng, J. Zhou, and H. Yu, "Visuals to Text: A Comprehensive Review on Automatic Image Captioning," *IEEE/CAA Journal of Automatica Sinica*, vol. 9, no. 8, pp. 1339–1365, 2022, doi: 10.1109/JAS.2022.105734.

[22] S. Sarto, M. Cornia, and R. Cucchiara, "Image Captioning Evaluation in the Age of Multimodal LLMs: Challenges and Future Perspectives," in *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI)*, 2025, pp. 10632–10640. doi: 10.24963/ijcai.2025/1180.

[23] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating Text Generation with BERT," in *International Conference on Learning Representations (ICLR)*, 2020. [Online]. Available: https://openreview.net/forum?id=SkeHuCVFDr

[24] J. Hessel, A. Holtzman, M. Forbes, R. Le Bras, and Y. Choi, "CLIPScore: A Reference-free Evaluation Metric for Image Captioning," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, M.-F. Moens, X. Huang, L. Specia, and S. W. Yih, Eds., Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 7514–7528. doi: 10.18653/v1/2021.emnlp-main.595.

[25] D. Santi, A. A. Ilham, Syafaruddin, and I. Nurtanio, "Attention-Driven Image

Captioning for Mobile Accessibility of the Visually Impaired," *International Journal of Interactive Mobile Technologies (iJIM)*, vol. 19, no. 9, pp. 4–18, May 2025, doi: 10.3991/ijim.v19i09.53441.

[26] M. Rifki, A. Bastian, and A. Mardiana, "Development of CNN-LSTM-Based Image Captioning Dataset to Enhance Visual Accessibility for Disabilities," *JITK*, vol. 10, no. 4, pp. 980–992, May 2025, doi: 10.33480/jitk.v10i4.6657.

[27] A. Khan and J. Singh, "A Novel Image Captioning Technique Using Deep Learning Methodology," *ICCK Transactions on Machine Intelligence*, vol. 1, no. 2, pp. 52–68, 2025, doi: 10.62762/TMI.2025.886122.

[28] C.-S. He, N.-K. Lo, Y.-H. Chien, and S.-S. Lin, "Image Descriptions for Visually Impaired Individuals to Locate Restroom Facilities," *Engineering Proceedings*, vol. 92, no. 1, 2025, doi: 10.3390/engproc2025092013.

[29] L. Orynbay, B. Razakhova, P. Peer, B. Meden, and Ž. Emeršič, "Recent Advances in Synthesis and Interaction of Speech, Text, and Vision," *Electronics (Basel).*, vol. 13, no. 9, 2024, doi: 10.3390/electronics13091726.

[30] N. Joshika, S. Srikanth, S. Suraj, S. Pradeep, and A. N. Sree, "A Multimodel GPT Based Application to Help Blind Users Visual Picture Extensions," *Journal of Science Engineering Technology and Management Science*, vol. 2, no. 8, pp. 437–443, Aug. 2025, doi: 10.63590/jsetms.2025.v02.i08.pp437-443.

[31] A. M. Hilal, F. Alrowais, F. N. Al-Wesabi, and R. Marzouk, "Red Deer Optimization with Artificial Intelligence Enabled Image Captioning System for Visually Impaired People," *Computer Systems Science and Engineering*, vol. 46, no. 2, pp. 1929–1945, 2023, doi: 10.32604/csse.2023.035529.

[32] H. Kim, C. Park, J. Jang, J. Lee, J. Yoon, and J. Paik, "Visual Question Answering with Multimodal Learning for VizWiz-VQA," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, Seattle, WA, USA, Jun. 2024.

[33] H. Fadhilah and N. Utama, "Systematic Literature Review on Medical Image Captioning Using CNN-LSTM and Transformer-Based Models," *Jurnal Masyarakat Informatika*, vol. 16, no. 1, pp. 32–53, 2025, doi: 10.14710/jmasif.16.1.73127.