JURNAL ILMIAH TEKNOLOGI DAN REKAYASA
Vol. 31, No. 2, August 2026, 110 – 127
p-ISSN: 1410-9093, e-ISSN: 2089-8088
https://ejournal.gunadarma.ac.id/index.php/tekno | tekrek@gunadarma.ac.id

Image captioning bilingual berbasis BLIP-2 dan MarianMT dengan integrasi *text-to-speech* untuk literasi visual siswa

[1]Rizka Oktaviani, [2*]Fauziah

[1,2]Informatika, Fakultas Teknologi Komunikasi dan Informatika, Universitas Nasional
[1,2]Jl. Sawo Manila No.61, Jakarta Selatan 12520, Daerah Khusus Ibukota Jakarta, Indonesia
[1]rizka.oktaviani007@gmail.com, [2]fauziah@civitas.unas.ac.id

Abstract

Visual literacy is the ability to understand and interpret information from visual representations, which involves linguistic processing. In the context of elementary education, this ability is crucial for helping students associate visual objects with bilingual vocabulary. However, most image captioning research remains monolingual, few studies accommodate Indonesian, and none have integrated audio representations into a single multimodal learning system. This study develops a bilingual image captioning system based on Bootstrapping Language-Image Pre-training 2 (BLIP-2) and Marian Neural Machine Translation (MarianMT), integrated with Text-to-Speech (TTS) within a FastAPI-based web application. English captions are generated using a pretrained BLIP-2 model, then translated into Indonesian using MarianMT, and converted to audio using Google Text-to-Speech (gTTS). Additionally, a Quantized Low-Rank Adaptation (QloRA) fine-tuning experiment was conducted to compare model performance. The dataset consists of 6400 animal images relevant to the elementary school learning context, with caption quality evaluated using the Metric for Evaluation of Translation with Explicit Ordering (METEOR). The results show that the pretrained BLIP-2 model delivers relatively stable performance with a METEOR score of 0.3765 for English and 0.3295 for Indonesian. These scores indicate that the generated captions are semantically relevant, although the overall performance is still relatively moderate compared to advanced image captioning systems. Functional testing of the prototype involving five elementary school students showed that the system is capable of generating bilingual captions and audio in real time and is easy to use. This multimodal integration supports the visual–verbal association process and has the potential to enrich students' bilingual vocabulary, although no controlled experimental testing has yet been conducted to quantitatively measure improvements in literacy.

Keywords: bilingual, BLIP-2, image captioning, MarianMT, visual literacy

Abstrak

Literasi visual merupakan kemampuan memahami dan menafsirkan informasi dari representasi gambar yang melibatkan pemrosesan linguistik. Dalam konteks pendidikan sekolah dasar, kemampuan ini penting untuk membantu siswa mengasosiasikan objek visual dengan kosakata bilingual. Namun, sebagian besar penelitian *image captioning* masih bersifat monolingual, belum banyak yang mengakomodasi bahasa Indonesia, serta belum mengintegrasikan representasi audio dalam satu sistem pembelajaran multimodal. Penelitian ini mengembangkan sistem *image captioning* bilingual berbasis *Bootstrapping Language-Image Pre-training* 2 (BLIP-2) dan *Marian Neural Machine Translation* (MarianMT) yang terintegrasi dengan *Text-to-Speech* (TTS) dalam aplikasi web berbasis FastAPI. *Caption* bahasa Inggris dihasilkan menggunakan model BLIP-2 *pretrained*, kemudian diterjemahkan ke bahasa Indonesia menggunakan MarianMT, dan dikonversi menjadi audio menggunakan Google *Text-to-Speech* (gTTS). Selain itu, dilakukan eksperimen *fine-tuning Quantized Low-Rank Adaptation* (QloRA) untuk membandingkan performa model. *Dataset* yang digunakan terdiri dari 6.400 citra hewan yang relevan dengan konteks pembelajaran sekolah dasar, dengan evaluasi kualitas *caption* menggunakan *Metric for*

* Corresponding Author

Evaluation of Translation with Explicit Ordering (METEOR). Hasil menunjukkan bahwa model BLIP-2 *pretrained* memberikan performa yang relatif stabil dengan skor METEOR sebesar 0,3765 untuk bahasa Inggris dan 0,3295 untuk bahasa Indonesia. Skor tersebut menunjukkan bahwa *caption* yang dihasilkan telah mampu merepresentasikan informasi visual secara semantik, meskipun performanya masih tergolong moderat dibandingkan sistem image captioning tingkat lanjut. Uji fungsional terhadap prototipe yang melibatkan lima siswa sekolah dasar menunjukkan bahwa sistem mampu menghasilkan *caption* bilingual dan audio secara responsif serta mudah digunakan. Integrasi multimodal ini mendukung proses asosiasi visual–verbal dan berpotensi memperkaya kosakata bilingual siswa, meskipun belum dilakukan pengujian eksperimental terkontrol untuk mengukur peningkatan literasi secara kuantitatif.

Kata kunci: Bilingual, BLIP-2, *image captioning*, literasi visual, MarianMT

1. Pendahuluan

Literasi visual merupakan kemampuan memahami dan menafsirkan informasi dari representasi gambar yang melibatkan pemrosesan linguistik [1]. Dalam konteks pendidikan dasar, kemampuan ini penting karena siswa cenderung belajar melalui kombinasi stimulus visual dan teks. Perkembangan *Artificial Intelligence* (AI) mendorong pemanfaatan teknologi multimodal dalam pendidikan, salah satunya *image captioning* yang mampu menghasilkan deskripsi teks secara otomatis dari citra [2], [3]. Pendekatan ini sejalan dengan teori *dual coding* yang menyatakan bahwa kombinasi visual dan verbal dapat meningkatkan pemahaman, retensi memori, serta membantu siswa mengasosiasikan objek visual dengan kosakata [4]–[6].

Meskipun pendekatan multimodal memiliki potensi dalam pembelajaran, tantangan literasi masih menjadi permasalahan global. Laporan Bank Dunia, *United Nations Children's Fund* (UNICEF), dan *United Nations Educational, Scientific and Cultural Organization* (UNESCO) menunjukkan bahwa lebih dari 70% anak usia 10 tahun di negara berkembang mengalami *learning poverty*. Selain itu, capaian literasi Indonesia berdasarkan *Programme for International Student Assessment* (PISA) 2022 masih berada di bawah rata-rata *Organisation for Economic Co-operation and Development* (OECD) [7]. Kondisi ini mendorong kebutuhan teknologi pembelajaran multimodal yang mendukung literasi visual dan bahasa secara simultan.

Perkembangan *image captioning* telah bergeser dari pendekatan berbasis *Convolutional Neural Network–Long Short-Term Memory* (CNN–LSTM) menuju arsitektur Transformer dan *vision–language model* (VLM). Salah satu model yang banyak digunakan adalah BLIP-2, yang mengintegrasikan *frozen image encoder* dan *large language model* (LLM) melalui *Querying Transformer* (Q-*Former*) untuk menghasilkan *caption* yang lebih kontekstual dan koheren [8]-[15]. Adaptasi domain selanjutnya dapat ditingkatkan melalui teknik *parameter-efficient fine-tuning* seperti *Low-Rank Adaptation* (LoRA) dan QLoRA [16]–[18].

Meskipun demikian, sebagian besar penelitian *image captioning* masih bersifat monolingual. Penelitian *image captioning* pada bahasa Indonesia sebagai *low-resource language* masih terbatas dan umumnya menggunakan pendekatan konvensional [19]–[21]. Pendekatan *generate-then-translate* berbasis *Neural Machine Translation* (NMT) seperti MarianMT menjadi solusi efisien untuk membangun sistem bilingual tanpa pelatihan model terpisah [22], [23].

Dalam konteks pembelajaran multimodal, representasi informasi tidak hanya disajikan dalam bentuk visual dan teks, tetapi juga audio. Oleh karena itu, integrasi TTS berpotensi meningkatkan aksesibilitas pembelajaran melalui representasi audio yang mendukung asosiasi visual–verbal siswa [13], [24], [25], termasuk pada siswa dengan kemampuan membaca rendah [26], [27]. Namun, integrasi TTS dalam *pipeline* VLM modern masih relatif terbatas [28].

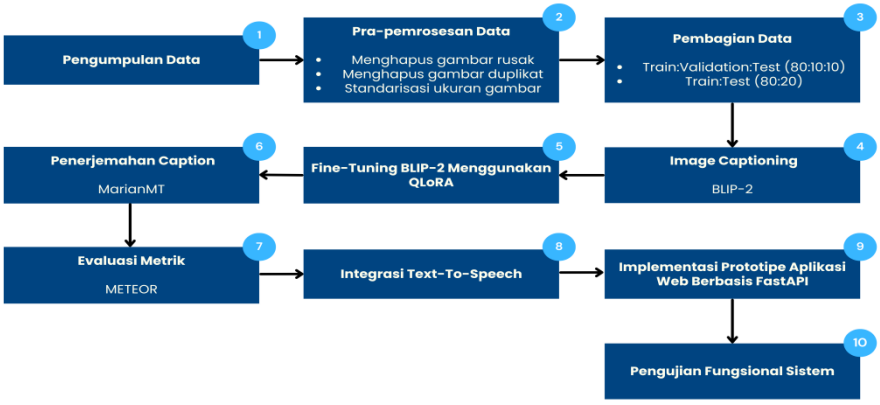
Dari sisi evaluasi, metrik berbasis *n-gram* seperti *Bilingual Evaluation Understudy* (BLEU) dan *Recall-Oriented Understudy for Gisting Evaluation* (ROUGE) masih umum digunakan [29]–[31], meskipun kurang sensitif terhadap kesetaraan makna. Oleh karena itu, penelitian ini menggunakan METEOR karena mempertimbangkan sinonim, *stemming*, dan struktur kalimat sehingga lebih representatif dalam mengevaluasi kualitas semantik *caption* [12], [32], [33].

Penggunaan *dataset* citra hewan yang kontekstual dengan materi pembelajaran sekolah dasar juga menjadi penting, karena objek yang familiar dapat memperkuat proses asosiasi visual–verbal siswa usia 7 hingga 12 tahun. Berdasarkan uraian tersebut, masih terdapat beberapa *research gap* dalam pengembangan *image captioning* bilingual. Penelitian *image captioning* bilingual Inggris–Indonesia berbasis VLM modern masih terbatas. Selain itu, integrasi BLIP-2, MarianMT, dan TTS dalam satu sistem multimodal masih jarang dikaji, sementara penggunaan METEOR untuk evaluasi kesetaraan semantik *caption* juga masih belum banyak dieksplorasi.

Penelitian ini memberikan kontribusi dalam pengembangan sistem *image captioning* bilingual berbasis BLIP-2 dan MarianMT yang terintegrasi dengan TTS dalam satu *pipeline* multimodal berbasis web. Melalui pendekatan tersebut, sistem diharapkan dapat mendukung proses asosiasi antara representasi visual, teks, dan audio, sehingga berpotensi meningkatkan pemahaman serta memperkuat kosakata bilingual siswa dalam pembelajaran berbasis multimodal.

2. Metode Penelitian

Penelitian ini mengusulkan *pipeline* multimodal *image captioning* bilingual dengan integrasi penerjemahan otomatis dan TTS untuk mendukung literasi visual siswa. Gambar 1 menunjukkan alur metodologi penelitian yang mencakup tahapan pengumpulan data, pra-pemrosesan, pelatihan model, hingga evaluasi sistem.



Gambar 1. Alur metodologi penelitian

2.1 Pengumpulan Data

Penelitian ini menggunakan *dataset* citra yang berfokus pada objek hewan yang relevan dengan materi pembelajaran sekolah dasar dan mendukung pengembangan literasi visual siswa. *Dataset* utama diperoleh dari Kaggle dengan judul *Animal Image Dataset (90 Different Animals)* yang dipublikasikan oleh Sourav Banerjee [34]. Penelitian ini juga menggunakan *dataset* tambahan dari buku ajar IPAS sekolah dasar melalui Sistem Informasi Perbukuan Indonesia (SIBI) Kemendikbud dan beberapa sumber terbuka lainnya untuk memastikan kesesuaian konteks pembelajaran.

Secara keseluruhan, *dataset* yang digunakan berjumlah sekitar 6.400 citra yang dipilih berdasarkan kejelasan objek, kualitas visual, dan relevansi materi. *Dataset* ini digunakan dalam proses pelatihan, evaluasi, dan pengujian sistem *image captioning* bilingual yang diusulkan.

2.2 Pra-pemrosesan Data

Tahap pra-pemrosesan data dilakukan untuk menjamin kualitas dan konsistensi citra sebelum digunakan sebagai masukan pada model BLIP-2. Proses ini bertujuan untuk meminimalkan gangguan data yang berpotensi memengaruhi kinerja model serta memastikan kesesuaian format citra dengan spesifikasi model yang digunakan.

Pada tahap ini, citra rusak dan duplikat dihapus untuk menghindari redundansi data. Selanjutnya, seluruh citra diseragamkan ukurannya menjadi 384 × 384 piksel sesuai spesifikasi *input* BLIP-2. Setelah pra-pemrosesan, jumlah *dataset* berkurang dari 6.400 menjadi 6.399 citra yang selanjutnya digunakan pada tahap pelatihan, evaluasi, dan pengujian sistem *image captioning* bilingual.

2.3 Pembagian Data

Dataset citra yang telah dipraproses dibagi menjadi dua skema pembagian untuk pelatihan dan evaluasi guna menguji stabilitas serta konsistensi performa model BLIP-2, baik pada kondisi *pretrained* maupun setelah *fine-tuning* menggunakan QLoRA, sebagaimana ditunjukkan pada Tabel 1.

Tabel 1. Distribusi data

Subset	**Skema 80:10:10**		**Skema 80:20**	
	Persentase	**Jumlah Citra**	**Persentase**	**Jumlah Citra**
Pelatihan	80%	5.119	80%	5.119
Validasi	10%	639		
Uji	10%	641	20%	1.280

Pada *subset* pelatihan dan validasi, *caption* bahasa Inggris dihasilkan menggunakan model BLIP-2 *pretrained* sebagai pseudo-label, kemudian diterjemahkan ke dalam bahasa Indonesia menggunakan MarianMT untuk membentuk *dataset* bilingual. Sementara itu, *subset* data uji tidak digunakan dalam proses pelatihan maupun *fine-tuning*. *Caption* pada data uji dihasilkan setelah proses pembagian data selesai dan digunakan sebagai *pseudo-ground truth* dalam evaluasi menggunakan metrik METEOR.

Skema 80:10:10 digunakan untuk *fine-tuning* QLoRA dengan memanfaatkan data validasi guna memantau proses pelatihan dan mencegah *overfitting*. Sementara itu, pada skema 80:20, data tidak dibagi ke dalam *subset* validasi sehingga seluruh sisa data setelah pelatihan digunakan sebagai data uji. Kedua skema ini digunakan untuk membandingkan performa model BLIP-2 *pretrained* dan hasil *fine-tuning* QLoRA, serta menganalisis konsistensi generalisasi model pada kondisi dengan dan tanpa validasi selama proses pelatihan.

2.4 Proses *Image Captioning* Menggunakan BLIP-2

Pada tahap ini, sistem menggunakan model BLIP-2 untuk menghasilkan *caption* bahasa Inggris dari citra masukan. BLIP-2 merupakan VLM yang menghubungkan *image encoder* dan *large language model* melalui Q-*Former*, sehingga mampu menangkap keterkaitan visual dan bahasa secara efektif.

Citra yang telah melalui tahap pra-pemrosesan diekstraksi oleh *image encoder* dan diproyeksikan ke ruang bahasa melalui Q-*Former* sebelum diproses oleh *language model* untuk menghasilkan deskripsi kontekstual. *Caption* yang dihasilkan lalu digunakan sebagai

pseudo-label pada proses *fine-tuning* QLoRA dan sebagai *pseudo-ground truth* pada tahap evaluasi.

2.5 *Fine-Tuning* Menggunakan QLoRA

Proses *fine-tuning* dilakukan pada model BLIP-2 menggunakan QLoRA untuk menyesuaikan model dengan karakteristik *dataset*. QLoRA merupakan metode *parameter-efficient fine-tuning* yang memperbarui sebagian kecil parameter adaptasi tanpa melatih ulang seluruh model, sehingga lebih hemat komputasi dan memori.

Parameter utama BLIP-2 dibekukan, sedangkan parameter adaptasi QLoRA diperbarui agar model lebih sesuai dengan representasi visual–tekstual pada domain penelitian. Evaluasi dilakukan menggunakan dua skema pembagian data, yaitu 80:10:10 dan 80:20, untuk membandingkan kualitas *caption* setelah penerapan QLoRA.

2.6 Penerjemahan *Caption* Menggunakan MarianMT

Pada tahap ini, *caption* bahasa Inggris yang dihasilkan oleh BLIP-2, baik *pretrained* maupun hasil *fine-tuning* QLoRA, diterjemahkan ke bahasa Indonesia menggunakan MarianMT berbasis Transformer. Proses ini menggunakan pendekatan *generate-then-translate* untuk membangun sistem bilingual tanpa melatih model terpisah untuk setiap bahasa.

Caption hasil terjemahan digunakan sebagai keluaran bilingual sistem dan bagian dari evaluasi kualitas *caption*. Integrasi MarianMT memungkinkan sistem menghasilkan deskripsi gambar dalam bahasa Inggris dan bahasa Indonesia secara koheren untuk mendukung pembelajaran bilingual.

2.7 Evaluasi *Caption* Menggunakan METEOR

Evaluasi kualitas *caption* dilakukan menggunakan metrik METEOR karena mampu menilai kesetaraan makna secara lebih komprehensif dibandingkan metrik berbasis *n-gram*, dengan mempertimbangkan kecocokan kata, variasi morfologis, sinonim, dan struktur frasa. Evaluasi diterapkan pada *caption* bahasa Inggris hasil BLIP-2 serta *caption* bahasa Indonesia hasil terjemahan MarianMT menggunakan data pengujian yang tidak terlibat dalam proses pelatihan maupun *fine-tuning*. *Image captioning* merupakan *generative task*, sehingga evaluasi tidak menggunakan *confusion matrix*, melainkan metrik kesamaan semantik seperti METEOR.

Secara matematis, metrik METEOR dihitung menggunakan kombinasi *precision* dan *recall* sebagaimana dirumuskan pada Persamaan (1). Nilai penalti dihitung menggunakan Persamaan (2) untuk mempertimbangkan fragmentasi urutan kata. Nilai akhir METEOR diperoleh dengan mengombinasikan kedua komponen tersebut sebagaimana ditunjukkan pada Persamaan (3).

$$F_{mean} = \frac{10PR}{R + 9P} \tag{1}$$

$$Penalty = 0.5\left(\frac{chunks}{matches}\right)^3 \tag{2}$$

$$METEOR = (1 - Penalty) \times F_{mean} \tag{3}$$

Pada Persamaan (1)–(3), *Precision* (*P*) menunjukkan proporsi kata pada *caption* prediksi yang sesuai dengan *caption* referensi, sedangkan *Recall* (*R*) menunjukkan proporsi kata pada *caption* referensi yang berhasil muncul pada *caption* prediksi. Variabel *matches* menyatakan jumlah kata yang cocok antara *caption* prediksi dan referensi setelah mempertimbangkan proses *stemming* dan sinonim. Variabel *chunks* menunjukkan jumlah segmen kecocokan kata yang tidak berurutan, di mana semakin banyak jumlah *chunks* maka

nilai penalti akan semakin besar sehingga skor METEOR menurun. Sementara itu, *Penalty* digunakan untuk merepresentasikan penalti akibat ketidaksesuaian urutan atau struktur kata antara *caption* prediksi dan *caption* referensi.

Nilai METEOR berada pada rentang 0 hingga 1, di mana nilai yang mendekati 1 menunjukkan kesesuaian yang semakin tinggi antara *caption* prediksi dan referensi, baik dari sisi pemilihan kata maupun struktur kalimat. Secara umum, METEOR menghitung kesesuaian antara *caption* yang dihasilkan dan *caption* referensi berdasarkan kombinasi *precision* dan *recall* dengan penalti untuk perbedaan urutan kata, sebagaimana dirumuskan pada Persamaan (1)–(3). Nilai METEOR yang diperoleh digunakan untuk membandingkan performa antar konfigurasi model dan menganalisis kualitas *caption* bilingual yang dihasilkan.

Sebagai ilustrasi, digunakan *caption* referensi “a pair of golden retriever dogs looking at a fence” dan *caption* prediksi “two dogs standing next to a fence”. Kedua *caption* memiliki tiga kata yang cocok, yaitu “dogs”, “a”, dan “fence”, sehingga diperoleh nilai *matches* = 3. Berdasarkan jumlah kata yang cocok tersebut, nilai *precision* dan *recall* dihitung menggunakan Persamaan (1), kemudian diperoleh nilai F-*mean* sebesar 0,309.

Selain kesamaan kata, METEOR juga mempertimbangkan urutan kemunculan kata melalui *fragmentation penalty*. Pada contoh ini, kecocokan kata membentuk dua *chunks*, yaitu “dogs” dan “a fence”, sehingga menghasilkan nilai *penalty* sebesar 0,148 berdasarkan Persamaan (2). Selanjutnya, skor akhir METEOR yang dihitung menggunakan Persamaan (3) adalah sebesar 0,263.

Hasil tersebut menunjukkan bahwa meskipun terdapat kesamaan objek utama antara caption prediksi dan referensi, perbedaan struktur kalimat dan urutan kata tetap memengaruhi nilai evaluasi. Dengan demikian, METEOR tidak hanya menilai kesamaan kata, tetapi juga mempertimbangkan keselarasan struktur kalimat secara semantik.

2.8 Integrasi TTS

Untuk mendukung pembelajaran multimodal dan meningkatkan aksesibilitas, *caption* dalam bahasa Inggris dan bahasa Indonesia diintegrasikan dengan teknologi TTS. Pada penelitian ini, konversi teks menjadi audio dilakukan menggunakan gTTS yang menghasilkan suara sintetis dengan pelafalan natural.

Caption hasil *image captioning* dan penerjemahan digunakan sebagai *input* TTS untuk menghasilkan audio dalam kedua bahasa. Integrasi ini memungkinkan informasi visual disampaikan tidak hanya melalui teks, tetapi juga melalui audio, sehingga membantu siswa dengan kemampuan membaca terbatas atau preferensi belajar auditori.

Audio yang dihasilkan menjadi bagian dari prototipe pembelajaran multimodal yang menggabungkan gambar, teks, dan suara. Pendekatan ini bertujuan memperkaya pengalaman belajar sekaligus mendukung peningkatan literasi visual dan bahasa siswa.

2.9 Implementasi Prototipe Aplikasi Web Berbasis FastAPI

Tahap akhir penelitian dilakukan dengan membangun aplikasi web interaktif berbasis arsitektur *client–server*, menggunakan FastAPI sebagai *backend* dan HTML, CSS, serta JavaScript sebagai *frontend*. Sistem menyediakan mode gambar acak berbasis *dataset* yang telah diproses sebelumnya (*precomputed*) dan mode unggah gambar secara *real-time* untuk menghasilkan *caption* bilingual dan audio melalui integrasi BLIP-2, MarianMT, dan gTTS dalam satu *pipeline* multimodal.

Backend menangani proses inferensi, penerjemahan, dan konversi audio, sedangkan *frontend* menampilkan gambar, *caption* bilingual, dan pemutar audio dalam antarmuka berbasis web. Sistem mendukung format JPG dan PNG dengan ukuran maksimum sekitar 5

MB serta menggunakan *temporary storage* tanpa *database*, sehingga data tidak disimpan dalam jangka panjang.

2.10 Pengujian Fungsional Sistem

Pengujian dilakukan secara terbatas terhadap lima siswa sekolah dasar berusia 7 hingga 12 tahun di lingkungan tempat tinggal peneliti. Pemilihan peserta dilakukan secara non-formal dan bersifat sukarela. Pengujian menggunakan pendekatan *User Acceptance Test* (UAT) untuk mengevaluasi aspek kemudahan penggunaan, kejelasan *output*, dan respons sistem dalam mendukung pemahaman visual.

Pengujian difokuskan pada fitur utama, yaitu unggah gambar, proses *image captioning* bilingual, penerjemahan bilingual menggunakan MarianMT, serta pemutaran audio TTS. Setiap fitur diuji berdasarkan kesesuaian antara *input* dan *output* (*black-box testing*), tanpa mempertimbangkan struktur internal sistem. Selain itu, dilakukan observasi terhadap interaksi pengguna untuk menilai kemudahan penggunaan dan respons sistem secara umum. Namun, pengujian ini masih bersifat terbatas dan belum menggunakan instrumen evaluasi kuantitatif terstandar.

2.11 Lingkungan Implementasi Sistem

Implementasi sistem dalam penelitian ini dilakukan menggunakan bahasa pemrograman Python 3.10 dengan dukungan *library* PyTorch, Hugging Face Transformers, FastAPI dan gTTS. Model BLIP-2 dan MarianMT dijalankan menggunakan *framework* Hugging Face berbasis PyTorch, sedangkan FastAPI digunakan untuk mendukung layanan berbasis web secara *real-time*.

Eksperimen dijalankan pada perangkat dengan dukungan GPU NVIDIA GeForce RTX 3060 yang terintegrasi dengan CUDA untuk mempercepat inferensi dan *fine-tuning* model. Pada skenario *fine-tuning*, digunakan pendekatan QLoRA dengan membekukan parameter utama model dan hanya memperbarui parameter adaptasi. *Hyperparameter* yang digunakan meliputi *learning rate* 2×10^{-5}, *epoch* 100, dan *batch size* 8 untuk menjaga keseimbangan antara efisiensi komputasi dan performa model.

3. Hasil dan Pembahasan

Bagian ini membahas hasil evaluasi sistem *image captioning* bilingual berbasis BLIP-2 dan MarianMT melalui analisis kuantitatif dan kualitatif. Sebelum evaluasi, dilakukan *cleaning caption* berupa normalisasi teks, penghapusan *caption* tidak relevan, serta eliminasi *caption* yang terlalu pendek atau tidak informatif untuk memastikan kualitas data pelatihan dan pengujian.

Berdasarkan hasil implementasi dan pengujian, sebagian besar citra berhasil diproses tanpa kegagalan inferensi. Meskipun terdapat beberapa kasus *output* yang kurang optimal akibat kualitas citra atau kompleksitas objek visual, *pipeline* sistem secara umum berjalan baik mulai dari *image captioning*, penerjemahan, hingga konversi audio. Oleh karena itu, evaluasi difokuskan pada kualitas semantik *caption* menggunakan metrik METEOR.

3.1 Hasil *Image Captioning* Menggunakan BLIP-2 *Pretrained*

Pada subbagian ini disajikan hasil *image captioning* yang dihasilkan oleh model BLIP-2 *pretrained* yang digunakan sebagai *baseline* sebelum dilakukan proses *fine-tuning*. Evaluasi pada tahap ini bersifat kualitatif, dengan menampilkan beberapa contoh citra uji beserta *caption* bahasa Inggris yang dihasilkan secara langsung oleh model.

a brown bear in the water

a pair of golden retriever dogs looking at a fence

hummingbird flying near flowers

Gambar 2. Hasil *image captioning* menggunakan model BLIP-2 *pretrained*

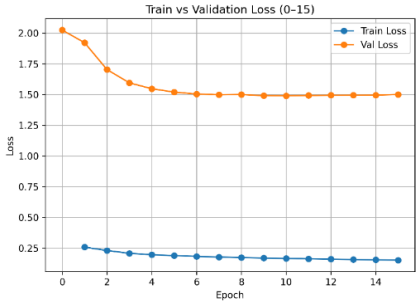
Berdasarkan visualisasi pada Gambar 2, model BLIP-2 *pretrained* mampu mengidentifikasi objek utama dan menghasilkan *caption* yang sesuai dengan konteks visual. Namun, beberapa *caption* yang dihasilkan masih bersifat umum dan belum menangkap detail visual yang lebih spesifik. Hal ini menunjukkan bahwa meskipun model *pretrained* memiliki kemampuan dasar yang baik, masih terdapat potensi peningkatan kualitas deskripsi melalui proses *fine-tuning* menggunakan QLoRA.

3.2 Hasil *Fine-Tuning* Model BLIP-2 Menggunakan QLoRA

Bagian ini menjelaskan hasil *fine-tuning* BLIP-2 menggunakan QLoRA, yang memungkinkan penyesuaian model secara efisien dengan kebutuhan komputasi dan memori lebih rendah, sehingga cocok untuk VLM berskala besar.

a. Hasil *Fine-Tuning* pada Skema 80:10:10

Hasil *fine-tuning* BLIP-2 menggunakan QLoRA dengan pembagian data 80:10:10 disajikan dalam bentuk grafik *training loss* dan *validation loss* pada Gambar 3.



Gambar 3. Grafik *train* dan *validation loss* QLoRA pada skema 80:10:10

Berdasarkan Gambar 3, *training loss* menurun seiring bertambahnya *epoch*, menunjukkan bahwa model mampu mempelajari pola data latih, sementara *validation loss* hanya menurun hingga titik tertentu dan kemudian stagnan, yang mengindikasikan potensi *overfitting* apabila pelatihan dilanjutkan tanpa pembatasan *epoch*.

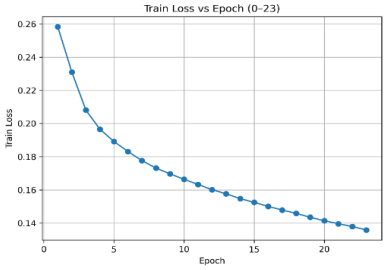
Tabel 2. Rekapitulasi nilai *train loss* dan *validation loss* pada skema 80:10:10

Epoch	*Train Loss*	*Validation Loss*
1	0,258	1,922
5	0,189	1,520
10	**0,166**	**1,490**
15	0,152	1,500

Berdasarkan Tabel 2, nilai *validation loss* terendah diperoleh pada *epoch* ke-10 dengan nilai 1,490. Setelah *epoch* tersebut, tidak terjadi penurunan *validation loss* yang signifikan. Oleh karena itu, *epoch* ke-10 ditetapkan sebagai *epoch* optimal untuk skema pembagian data 80:10:10.

b. Hasil *Fine-Tuning* pada Skema 80:20

Hasil *fine-tuning* BLIP-2 menggunakan QLoRA dengan pembagian data 80:20 ditunjukkan pada Gambar 4.



Gambar 4. Grafik *train loss* QLoRA pada skema 80:20

Berdasarkan Gambar 4, nilai *training loss* menunjukkan penurunan yang konsisten dari *epoch* awal hingga akhir pelatihan. Tidak terlihat fluktuasi yang signifikan pada kurva *loss*, yang menandakan bahwa model masih berada pada fase pembelajaran dan belum menunjukkan indikasi *overfitting* yang kuat.

Tabel 3. Rekapitulasi nilai *train loss* pada skema 80:20

Epoch	***Train Loss***
1	0,258
5	0,189
10	0,166
15	0,152
20	0,141
23	**0,135**

Berdasarkan Tabel 3, nilai *training loss* terendah dicapai pada *epoch* ke-23 dengan nilai 0,135. Hasil ini menunjukkan bahwa penambahan jumlah *epoch* pada skema 80:20 masih memberikan peningkatan kemampuan model dalam mempelajari data latih.

c. Perbandingan *Fine-Tuning* pada Skema 80:10:10 dan 80:20

Secara umum, *fine-tuning* dengan QLoRA meningkatkan adaptasi BLIP-2 terhadap *dataset*. Skema 80:10:10 menunjukkan pembelajaran lebih terkontrol melalui *validation loss* yang optimal pada *epoch* ke-10 dan mencerminkan generalisasi stabil, sedangkan skema 80:20 hanya memperlihatkan penurunan *train loss* tanpa indikator validasi sehingga peningkatan generalisasi tidak dapat dipastikan.

3.3 Hasil Penerjemahan *Caption* Menggunakan MarianMT

Caption bilingual dibentuk melalui penerjemahan *caption* bahasa Inggris ke bahasa Indonesia menggunakan MarianMT. *Caption* bahasa Inggris terlebih dahulu dihasilkan oleh BLIP-2, baik pada kondisi *pretrained* maupun setelah *fine-tuning* QLoRA, kemudian diterjemahkan sehingga menghasilkan pasangan *caption* Inggris–Indonesia.

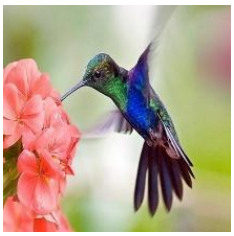
Berdasarkan visualisasi pada Gambar 5, *caption* bahasa Indonesia hasil terjemahan MarianMT umumnya mampu mempertahankan makna utama dari *caption* bahasa Inggris. Meskipun terdapat perbedaan struktur kalimat dan pilihan kata yang cenderung literal, makna utama tetap terjaga dan relevan untuk pembelajaran visual. *Caption* bilingual ini kemudian digunakan sebagai referensi (*pseudo-ground truth*) dalam visualisasi dan evaluasi kuantitatif menggunakan METEOR.

EN:
a brown bear in the water

ID:
beruang coklat di dalam air

EN:
a pair of golden retriever dogs looking at a fence

ID:
sepasang anjing golden retriever melihat pagar



EN:
hummingbird flying near flowers

ID:
burung kolibri terbang di dekat bunga

Gambar 5. Hasil *captio*n bilingual menggunakan MarianMT

3.4 Visualisasi Kualitatif Hasil *Image Captioning*

Visualisasi hasil *captioning* pada Gambar 6 menunjukkan contoh *caption* referensi dalam bahasa Inggris dan bahasa Indonesia yang digunakan sebagai *pseudo-ground truth.* Visualisasi ini digunakan untuk menilai kemampuan kualitatif tiga konfigurasi model, yaitu BLIP-2 *pretrained*, BLIP-2 hasil *fine-tuning* QLoRA skema 80:10:10, dan QLoRA skema 80:20, menggunakan sampel representatif dari data uji yang disajikan bersama gambar dan *caption* bilingual sebagai referensi evaluasi.

GT EN:
a pair of golden retriever dogs looking at a fence
GT ID:
sepasang anjing golden retriever melihat pagar

Gambar 6. Referensi Inggris dan Indonesia

Pred EN Pretrained:
two dogs standing next to a fence

Pred ID Pretrained:
dua anjing berdiri di samping pagar

Pred EN QLora 80:10:10:
a golden retriever and black labrador dog standing next to a fence in a field with trees and grasses in the background

Pred ID QLora 80:10:10:
seekor anjing emas dan seekor labrador hitam berdiri di sebelah pagar di lapangan dengan pohon dan rumput di latar belakang

Pred EN QLora 80:20:
two dogs standing next to a fence in a field with trees and buildings in the background

Pred ID QLora 80:20:
dua anjing berdiri di samping pagar di lapangan dengan pohon dan bangunan di latar belakang

Gambar 7. Prediksi Inggris dan Indonesia

Visualisasi pada Gambar 7 menunjukkan perbedaan karakteristik *output* antar model. Model BLIP-2 *pretrained* cenderung menghasilkan deskripsi yang lebih sederhana dan fokus pada objek utama, namun dengan keterbatasan dalam menangkap detail tambahan. Model QLoRA dengan skema 80:10:10 menghasilkan deskripsi yang lebih panjang dengan penambahan atribut visual, yang pada beberapa kasus mampu memberikan informasi lebih rinci dibandingkan model *pretrained*. Sementara itu, model QLoRA skema 80:20 juga menghasilkan deskripsi yang lebih panjang, namun tidak semua atribut tambahan yang dihasilkan sepenuhnya didukung oleh konteks visual pada citra. Hal ini menunjukkan adanya variasi kualitas *output* antar model serta peningkatan kompleksitas deskripsi tidak selalu berbanding lurus dengan kesesuaian terhadap konten visual.

3.5 Evaluasi Kuantitatif *Caption* Bilingual Menggunakan METEOR

Evaluasi kuantitatif dilakukan untuk membandingkan performa BLIP-2 *pretrained* dan model hasil *fine-tuning* QLoRA skema 80:10:10 serta 80:20 menggunakan metrik METEOR pada *caption* bahasa Inggris dan bahasa Indonesia, sebagaimana dirangkum pada Tabel 4.

Tabel 4. Perbandingan skor METEOR

Model	Bahasa Inggris	Bahasa Indonesia
Pretrained	0,3765	0,3295
QLoRA 80:10:10	0,3818	0,3065
QLoRA 80:20	0,3696	0,2946

Model QLoRA skema 80:10:10 menghasilkan skor METEOR tertinggi pada bahasa Inggris sebesar 0,3818, yang menunjukkan peningkatan kesesuaian linguistik dibandingkan model *pretrained*. Namun, pada bahasa Indonesia, skor METEOR mengalami penurunan menjadi 0,3065, yang mengindikasikan adanya penurunan konsistensi hasil terjemahan setelah proses *fine-tuning*.

Sebaliknya, model BLIP-2 *pretrained* menunjukkan performa yang lebih seimbang antara bahasa Inggris dan bahasa Indonesia, dengan skor masing-masing 0,3765 dan 0,3295. Hal tersebut mengindikasikan bahwa model *pretrained* memiliki stabilitas yang lebih baik dalam menghasilkan *caption* bilingual.

Sementara itu, QLoRA skema 80:20 mencatat performa terendah pada kedua bahasa, yaitu 0,3696 untuk bahasa Inggris dan 0,2946 untuk bahasa Indonesia. Meskipun *training loss* menunjukkan penurunan selama proses pelatihan, hasil evaluasi METEOR menunjukkan bahwa peningkatan tersebut tidak secara langsung berbanding lurus dengan kualitas *caption* yang dihasilkan pada data uji.

Berdasarkan hasil tersebut, pemilihan model untuk implementasi sistem tidak hanya mempertimbangkan skor METEOR tertinggi pada satu bahasa, tetapi juga keseimbangan performa antar bahasa serta konsistensi *output* model. Meskipun model QLoRA dengan skema 80:10:10 menghasilkan skor METEOR yang lebih tinggi pada bahasa Inggris, model BLIP-2 *pretrained* dipilih sebagai model utama karena menunjukkan performa yang lebih stabil pada kedua bahasa. Dalam konteks pembelajaran siswa sekolah dasar, keseimbangan performa antar bahasa menjadi faktor penting untuk memastikan kesesuaian *output* dalam sistem bilingual.

Skor METEOR pada kisaran 0,2946–0,3818 menunjukkan bahwa model telah mampu menghasilkan *caption* yang relevan secara semantik terhadap objek utama pada citra, meskipun kualitas deskripsi masih berada pada tingkat moderat dibandingkan sistem *image captioning* dengan anotasi manual dan *dataset* berskala besar. Keterbatasan ini dapat dipengaruhi oleh penggunaan *pseudo-ground truth*, keterbatasan variasi *dataset*, serta pergeseran struktur semantik akibat proses penerjemahan otomatis menggunakan MarianMT.

Pada beberapa kasus, *fine-tuning* juga menghasilkan atribut tambahan yang tidak sepenuhnya sesuai dengan konteks visual citra (*hallucination*).

Hasil tersebut menunjukkan bahwa pendekatan VLM berbasis BLIP-2 memiliki potensi dalam menghasilkan *caption* bilingual untuk mendukung pembelajaran multimodal, meskipun peningkatan masih diperlukan dalam menangkap detail visual dan menjaga konsistensi struktur kalimat.

3.6 Integrasi TTS untuk Mendukung Pembelajaran Multimodal

Integrasi TTS mengonversi *caption* bahasa Inggris dan bahasa Indonesia menjadi audio untuk mendukung pembelajaran multimodal melalui kombinasi gambar, teks, dan suara. Proses ini dilakukan setelah tahap *captioning* dan penerjemahan, sehingga tidak memengaruhi pelatihan maupun evaluasi model.

Pada *backend* FastAPI, sistem memproses gambar, menghasilkan *caption* menggunakan BLIP-2, menerjemahkannya dengan MarianMT, lalu mengubahnya menjadi audio melalui gTTS. Audio hasil gTTS ditampilkan bersama *caption* bilingual melalui antarmuka web seperti ditunjukkan pada Gambar 8.

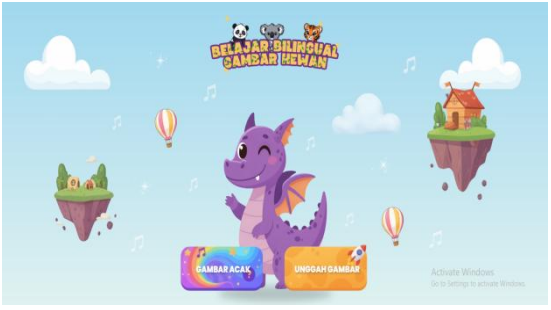
Gambar 8. Tampilan web dengan *caption* bilingual dan audio TTS

Frontend menampilkan gambar, *caption* bilingual, dan pemutar audio dalam satu antarmuka terpadu. Integrasi ini memungkinkan pengguna membaca sekaligus mendengarkan deskripsi gambar, sehingga meningkatkan pemahaman dan keterlibatan belajar.

Meskipun audio hasil gTTS dapat diputar dengan baik dan memiliki pelafalan yang cukup jelas selama pengujian fungsional, penelitian ini belum melakukan evaluasi kualitas audio secara formal, seperti pengukuran *Mean Opinion Score* (MOS) atau evaluasi persepsi pengguna terhadap *naturalness* dan *intelligibility* suara yang dihasilkan. Oleh karena itu, kualitas audio pada penelitian ini masih dievaluasi secara fungsional dan belum dianalisis secara kuantitatif.

3.7 Hasil Implementasi Aplikasi Web Berbasis FastAPI

Subbagian ini menampilkan hasil implementasi sistem *image captioning* bilingual berbasis web. Tampilan awal aplikasi ditunjukkan pada Gambar 9.



Gambar 9. Tampilan awal aplikasi

Aplikasi menyediakan dua mode utama, yaitu mode gambar acak berbasis *dataset* (*precomputed*) dan mode unggah gambar dengan inferensi *real-time*. Kedua mode ini dirancang untuk menunjukkan pemanfaatan sistem baik pada data yang telah diproses sebelumnya maupun pada gambar baru dari pengguna. Tampilan mode gambar acak yang menampilkan *caption* bilingual beserta audio hasil TTS ditunjukkan pada Gambar 10.

Gambar 10. Tampilan mode gambar acak dengan *caption* bilingual dan audio

Pada mode unggah gambar, pengguna dapat memilih gambar dari galeri atau kamera untuk diproses secara langsung. Proses pemilihan sumber gambar ditunjukkan pada Gambar 11, sedangkan pratinjau gambar sebelum inferensi ditampilkan pada Gambar 12.

Gambar 11. Tampilan pemilihan sumber gambar dari galeri atau kamera

Gambar 12. Tampilan pratinjau gambar sebelum proses inferensi

Setelah gambar dikirim ke *backend*, sistem menjalankan *pipeline* utama yang meliputi *captioning* menggunakan BLIP-2 *pretrained*, translasi dengan MarianMT, serta konversi audio menggunakan gTTS. Hasil *caption* bilingual dan audio ditampilkan secara dinamis melalui browser, seperti ditunjukkan pada Gambar 13.

Gambar 13. Tampilan mode unggah gambar dengan *caption* dan audio

Hasil implementasi menunjukkan bahwa sistem mampu menghasilkan *caption* bilingual dan audio yang berfungsi dengan baik pada proses inferensi berbasis web. Fitur tambahan berupa suara hewan saat gambar diklik turut memperkaya pengalaman pembelajaran visual dan auditori.

3.8 Hasil Pengujian Prototipe Aplikasi

Untuk memverifikasi fungsionalitas sistem, dilakukan pengujian menggunakan pendekatan *black-box testing* pada seluruh fitur utama. Hasil pengujian tersebut ditunjukkan pada Tabel 5.

Tabel 5. Hasil *black-box testing*

Fitur yang diuji	Hasil yang diharapkan	Hasil
Mode gambar acak	*Caption* dan audio berhasil ditampilkan	Berhasil
Unggah gambar	Gambar berhasil diterima sistem	Berhasil
Image captioning bahasa Inggris	*Caption* berhasil dihasilkan	Berhasil
Penerjemahan bahasa Indonesia	*Caption* terjemahan berhasil ditampilkan	Berhasil
Audio TTS bahasa Inggris	Audio dapat diputar	Berhasil
Audio TTS bahasa Indonesia	Audio dapat diputar	Berhasil

Selain itu, dilakukan UAT secara terbatas terhadap lima siswa sekolah dasar untuk mengevaluasi kemudahan penggunaan serta kejelasan *output* sistem. Hasil UAT ditunjukkan pada Tabel 6.

Tabel 6. Hasil UAT

Pernyataan	Jumlah responden setuju
Sistem mudah digunakan	5
Caption mudah dipahami	5
Audio dapat diputar dengan baik	5
Tampilan aplikasi menarik	5
Sistem membantu memahami gambar	5

Berdasarkan hasil *black-box testing* dan UAT, seluruh fitur utama sistem berfungsi sesuai spesifikasi. Sistem mampu menghasilkan *caption* bilingual dan audio secara konsisten serta memproses gambar pengguna secara *real-time* dengan respons yang stabil. Antarmuka yang sederhana memudahkan pengguna dalam mengoperasikan sistem dan memahami *output* yang dihasilkan. Integrasi gambar, teks, dan audio menunjukkan potensi sistem sebagai prototipe media pembelajaran multimodal untuk mendukung literasi visual dan bahasa siswa sekolah dasar.

4. Kesimpulan

Penelitian ini berhasil mengembangkan sistem *image captioning* bilingual berbasis BLIP-2 dan MarianMT yang terintegrasi dengan gTTS dalam satu *pipeline* multimodal berbasis web. Sistem mampu menghasilkan deskripsi gambar dalam bahasa Inggris dan bahasa Indonesia dalam bentuk teks dan audio secara responsif. Hasil evaluasi menggunakan METEOR menunjukkan bahwa model BLIP-2 *pretrained* memberikan performa yang lebih seimbang antara bahasa Inggris 0,3765 dan bahasa Indonesia 0,3295, dibandingkan konfigurasi *fine-tuning* QLoRA, yang menunjukkan penurunan skor pada bahasa Indonesia. Meskipun *fine-tuning* meningkatkan variasi linguistik pada bahasa Inggris, model *pretrained* menunjukkan konsistensi performa yang lebih baik antar bahasa dalam menghasilkan *caption* bilingual, sehingga lebih sesuai untuk implementasi sistem akhir. Secara fungsional, sistem mampu menghasilkan *output* pada sebagian besar data uji dan berjalan dengan responsif, sehingga layak sebagai prototipe media pembelajaran digital berbasis multimodal untuk mendukung literasi visual dan bahasa siswa sekolah dasar. Meskipun demikian, skor METEOR yang diperoleh masih menunjukkan bahwa kualitas *caption* berada pada tingkat moderat dan belum sepenuhnya mampu menghasilkan deskripsi visual yang detail dan konsisten pada seluruh citra uji.

Penelitian selanjutnya dapat diarahkan pada penggunaan anotasi manual sebagai *ground truth*, optimalisasi MarianMT melalui *fine-tuning* domain-spesifik, evaluasi audio menggunakan MOS, serta penambahan metrik evaluasi seperti CIDEr, SPICE, atau BERTScore untuk memperkaya analisis semantik. Selain itu, pemanfaatan VLM multilingual yang lebih besar dan eksperimen pendidikan terkontrol juga diperlukan untuk meningkatkan kualitas *caption* bilingual serta mengukur dampak sistem terhadap peningkatan literasi siswa secara kuantitatif.

Ucapan Terima Kasih

Penulis mengucapkan terima kasih kepada Fakultas Teknologi Komunikasi dan Informatika Universitas Nasional atas dukungan akademik yang diberikan selama pelaksanaan penelitian.

Pernyataan

Kontribusi Penulis. R.O.: konseptualisasi, implementasi sistem, analisis data, penulisan naskah. F.: supervisi penelitian, validasi metodologi, peninjauan dan penyuntingan naskah.
Pendanaan. Penelitian ini tidak menerima pendanaan eksternal.
Konflik Kepentingan. Penulis menyatakan tidak terdapat konflik kepentingan terkait publikasi artikel ini.
Ketersediaan Data. *Dataset* utama pada penelitian ini berasal dari Kaggle dengan judul *Animal Image Dataset (90 Different Animals)*. Selain itu, digunakan *dataset* tambahan dari buku ajar IPAS melalui SIBI Kemendikbud serta sumber terbuka lainnya untuk kebutuhan penelitian dan tidak didistribusikan kembali. *Dataset* gabungan yang telah diproses dapat diperoleh dari penulis korespondensi melalui permintaan yang wajar.
Persetujuan Etik. Penelitian melibatkan pengujian terbatas terhadap lima siswa sekolah dasar secara sukarela tanpa pengumpulan data sensitif maupun intervensi eksperimental, sehingga tidak memerlukan persetujuan etik formal.
Penggunaan Kecerdasan Buatan (AI). Penulis menyatakan bahwa penggunaan alat berbasis kecerdasan buatan (AI) hanya terbatas pada aspek penyuntingan bahasa dan tidak mempengaruhi substansi ilmiah penelitian.

Daftar Referensi

[1] T. Xian, Z. Zhou, W. Zhou, and Z. Zhang, "Refining visual token sequence for efficient image captioning," *Neural Netw.*, Art. no. 107759, vol. 191, Nov. 2025, doi: 10.1016/j.neunet.2025.107759.

[2] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *Proc. 39th Int. Conf. Mach. Learn. (ICML)*, vol. 162, Baltimore, MD, USA, Jul. 17–23, 2022, pp. 12888–12900.

[3] M. Cornia, L. Baraldi, A. Tal, and R. Cucchiara, "Fully-attentive iterative networks for region-based controllable image and video captioning," *Comput. Vis. Image Underst.*, vol. 237, Art. no. 103857, Dec. 2023, doi: 10.1016/j.cviu.2023.103857.

[4] H. Fadhilah and N. P. 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, May 2025, doi: 10.14710/jmasif.16.1.73127.

[5] Z. Gao *et al.*, "AnDR-BLIP2: Enhanced semantic understanding framework for industrial image anomaly detection and report generation," *J. Franklin Inst.*, vol. 362, no. 12, Art. no. 107816, Aug. 2025, doi: 10.1016/j.jfranklin.2025.107816.

[6] R. Castro, I. Pineda, W. Lim, and M. E. Morocho-Cayamcela, "Deep learning approaches based on transformer architectures for image captioning tasks," *IEEE Access*, vol. 10, pp. 33679–33694, 2022, doi: 10.1109/ACCESS.2022.3161428.

[7] R. Yilmaz and F. G. K. Yilmaz, "The effect of generative artificial intelligence (AI)-based tool use on students' computational thinking skills, programming self-efficacy and motivation," *Comput. Educ. Artif. Intell.*, vol. 4, Art. no. 100147, Jun. 2023, doi: 10.1016/j.caeai.2023.100147.

[8] A. K. Poddar and R. Rani, "Hybrid architecture using CNN and LSTM for image captioning in Hindi language," in *Procedia Comput. Sci.*, vol. 218, pp. 686–696, 2023, doi: 10.1016/j.procs.2023.01.049.

[9] J. A. Alzubi, R. Jain, P. Nagrath, S. Satapathy, S. Taneja, and P. Gupta, "Deep image captioning using an ensemble of CNN and LSTM based deep neural networks," *J. Intell. Fuzzy Syst.*, vol. 40, no. 4, pp. 5761–5769, 2021, doi: 10.3233/JIFS-189415.

[10] I. A. Albadarneh, B. H. Hammo, and O. S. Al-Kadi, "Attention-based transformer models for image captioning across languages: An in-depth survey and evaluation," *Comput. Sci. Rev.*, vol. 58, Art. no. 100766, Nov. 2025, doi: 10.1016/j.cosrev.2025.100766.

[11] M. A. Al-Malla, O. Hamdoun, and N. Ghneim, "A comprehensive survey on deep learning approaches for image captioning: A systematic review," *J. Big Data*, vol. 13, Art. no. 48, Feb. 2026, doi: 10.1186/s40537-026-01377-w.

[12] L. Xu, Q. Tang, J. Lv, B. Zheng, X. Zeng, and W. Li, "Deep image captioning: A review of methods, trends and future challenges," *Neurocomputing*, vol. 546, Art. no. 126287, Aug. 2023, doi: 10.1016/j.neucom.2023.126287.

[13] M. Cho, S. Kim, D. Choi, and Y. Sung, "Enhanced BLIP-2 optimization using LoRA for generating dashcam captions," *Appl. Sci.*, vol. 15, no. 7, Art. no. 3712, Apr. 2025, doi: 10.3390/app15073712.

[14] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proc. 40th Int. Conf. Mach. Learn. (ICML)*, vol. 202, Honolulu, HI, USA, Jul. 2023, Art. no. 814, pp. 19730–19742.

[15] S. Tyagi *et al.*, "Novel advance image caption generation utilizing vision transformer and generative adversarial networks," *Computers*, vol. 13, no. 12, Art. no. 305, Nov. 2024, doi: 10.3390/computers13120305.

[16] B. Brzęk and G. Dziczkowski, "Efficient medical question answering through QLoRA fine-tuning and knowledge distillation," *Procedia Comput. Sci.*, vol. 270, pp. 1468–1477, 2025, doi: 10.1016/j.procs.2025.09.268.
[17] F. L. Chen *et al.*, "VLP: A survey on vision-language pre-training," *Mach. Intell. Res.*, vol. 20, pp. 38–56, Jan. 2023, doi: 10.1007/s11633-022-1369-5.
[18] R. Patil, P. Khot, and V. Gudivada, "Analyzing LLAMA3 performance on classification task using LoRA and QLoRA techniques," *Appl. Sci.*, vol. 15, no. 6, Art. no. 3087, Mar. 2025, doi: 10.3390/app15063087.
[19] R. Mulyawan, A. Sunyoto, and A. H. Muhammad, "Automatic Indonesian image captioning using CNN and transformer-Based model approach," in *Proc. 5th Int. Conf. Inf. Commun. Technol. (ICOIACT)*, Yogyakarta, Indonesia, 2022, pp. 355–360, doi: 10.1109/ICOIACT55506.2022.9971855.
[20] M. Lupaşcu, A.-C. Rogoz, M. S. Stupariu, and R. T. Ionescu, "Large multimodal models for low-resource languages: A survey," *Inf. Fusion*, vol. 131, Art. no. 104189, Jul. 2026, doi: 10.1016/j.inffus.2026.104189.
[21] P. Choudhury, S. Nair, P. Guha, and S. Nandi, "Image captioning in low resource Assamese language with semantic information prior and spatially encoded transformer model," *Expert Syst. Appl.*, vol. 297, pt. C, Art. no. 129479, Feb. 2026, doi: 10.1016/j.eswa.2025.129479.
[22] Z. Tan *et al.*, "Neural machine translation: A review of methods, resources, and tools," *AI Open*, vol. 1, pp. 5–21, 2020, doi: 10.1016/j.aiopen.2020.11.001.
[23] F. I. Maulana, Y. Heryadi, G. P. Kusuma, and W. Budiharto, "Data augmentation English-Indonesia-Madurese parallel corpus dataset using neural machine translation," *Data Brief*, vol. 62, Art. no. 112046, 2025, doi: 10.1016/j.dib.2025.112046.
[24] S. A. Roslyn, A. Negha, and R. V. Sekhar, "Enhancing accessibility for visually impaired users: A BLIP2-powered image description system in Tamil," in *Proc. Int. Conf. Adv. Data Eng. Intell. Comput. Syst. (ADICS)*, Chennai, India, 2024, pp. 1–6, doi: 10.1109/ADICS58448.2024.10533565.
[25] J.-H. Kim, S.-W. Park, J.-H. Huh, S.-H. Jung, and C.-B. Sim, "Human scene understanding mechanism-based image captioning for blind assistance," *IEEE Access*, vol. 13, pp. 81933–81947, 2025, doi: 10.1109/ACCESS.2025.3564991.
[26] A. Alsayed, M. Arif, T. M. Qadah, and S. Alotaibi, "A systematic literature review on using the encoder-decoder models for image captioning in English and Arabic languages," *Appl. Sci.*, vol. 13, no. 19, Art. no. 10894, Sep. 2023, doi: 10.3390/app131910894.
[27] Y. Tao *et al.*, "CMDF-TTS: Text-to-speech method with limited target speaker corpus," *Neural Netw.*, vol. 188, Art. no. 107432, Aug. 2025, doi: 10.1016/j.neunet.2025.107432.
[28] O. Ondeng, H. Ouma, and P. Akuon, "A review of transformer-based approaches for image captioning," *Appl. Sci.*, vol. 13, no. 19, Art. no. 11103, Oct. 2023, doi: 10.3390/app131911103.
[29] K. R. Narejo *et al.*, "Optimizing sentiment integration in image captioning using transformer-based fusion strategies," *Comput. Mater. Continua (CMC)*, vol. 84, no. 2, pp. 3407–3429, Jul. 2025, doi: 10.32604/cmc.2025.065872.
[30] A. Thobhani *et al.*, "A survey on enhancing image captioning with advanced strategies and techniques," *CMES Comput. Model. Eng. Sci.*, vol. 142, no. 3, pp. 2247–2280, Mar. 2025, doi: 10.32604/cmes.2025.059192.
[31] J. Peng and T. Tang, "A unified visual and linguistic semantics method for enhanced image captioning," *Appl. Sci.*, vol. 14, no. 6, Art. no. 2657, Mar. 2024, doi: 10.3390/app14062657.

[32] A. S. Al-Shamayleh, O. Adwan, M. A. Alsharaiah, A. H. Hussein, Q. M. Kharma, and C. I. Eke, "A comprehensive literature review on image captioning methods and metrics based on deep learning technique," *Multimed. Tools Appl.*, vol. 83, pp. 34219–34268, Apr. 2024, doi: 10.1007/s11042-024-18307-8.

[33] M. Kaur and H. Kaur, "An efficient CNN-LSTM based framework for improved image captioning," in *Procedia Comput. Sci.*, vol. 258, pp. 3601–3607, 2025, doi: 10.1016/j.procs.2025.04.615.

[34] S. Banerjee, "Animal Image Dataset (90 Different Animals)," Kaggle, 2021. [Online]. Available: https://www.kaggle.com/datasets/iamsouravbanerjee/animal-image-dataset-90-different-animals. (Accessed: Oct. 24, 2025).