Contents

1. Motivation 2

2. Methodology 2

2.1 Convolutional Neural Networks (CNN) – LSTM 2

2.2 Transformers 3

2.3 Evaluation 6

2.3.1 Stress-test scenarios 6

2.3.2 Performance metrics 6

2.4 Limitations 6

3. Findings 6

4. Conclusion 11

5. References 12

6. Appendices 15

1. Motivation

Artificial Intelligence (AI) exponential advancements have emerged as a catalyst for substantial improvements in life quality (Alowais et al., 2023). This report delves into various image captioning techniques underpinning a wide range of applications for the visually impaired. These image-to-text innovations, integrated with text-to-speech, will enhance the cognitive and perceptual experiences of visually impaired users in both social (e.g. route navigation) and business settings (e.g., product description on an ecommerce platform), and other applications set out in Appendix 1.

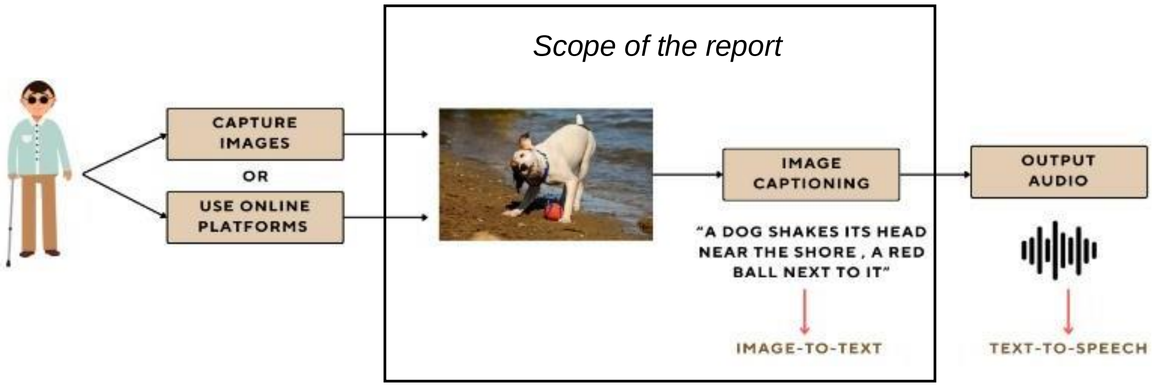


Figure 1. Image to Speech in Aiding Blind People

2. Methodology

2.1 Convolutional Neural Networks (CNN) – Long Short-Term Memory Networks (LSTM)

Image-to-text sits at the intersection of image-recognition and natural-language-processing, both of which have been revolutionized by the Transformer architecture. Prior to them, most image-to-text models mixed CNN and LSTMs. Examples include Google Neural Image Caption (NIC) (Vinyals et al., 2015) and its improved version Show and Tell or Show, Attend and Tell (Xu et al., 2016).

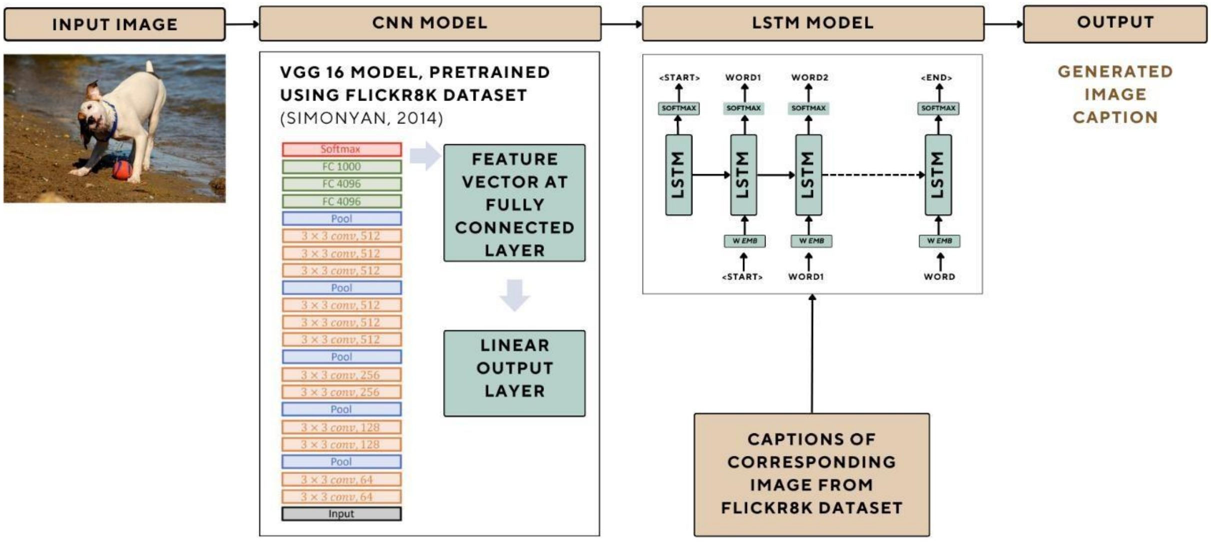


Figure 2. Flow Diagram of CNN and LSTM Model

This report replicates the CNN-LSTM structure by deploying a pre-trained VGG16 model (CNN with 16 layers deep) for feature extraction, which are tokenized and passed through to a pretrained LSTM decoder for text generation (Figure 2). Additionally, a transfer learning approach for a small sample is also included in the attached code for learning and demonstration purposes.

2.2 Transformers

Transformer models, with self-attention and cross-attention mechanisms and pre-training on large and diverse image-text pairs, effectively weight the importance of different parts of image input, enhance contextual understanding and facilitate more complex and nuanced language generation (Figure 3).

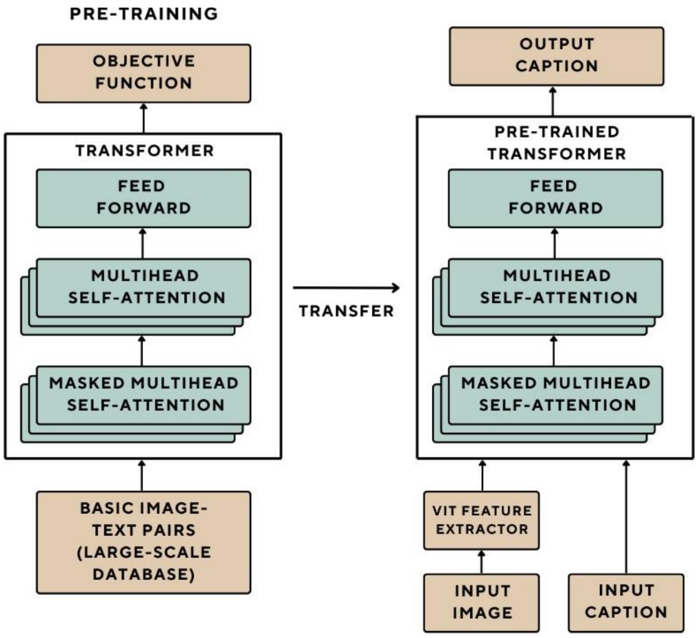


Figure 3. Visual-language Pre-training Transformer-based Captioning

Since limited computational capacities place constraints to fine-tuning, this study evaluates performance of the following recent open-sourced models: Salesforce BLIP (Radford et al. 2021) and BLIP-2 (Li et al., 2023), Microsoft GIT-based COCO (Wang et al., 2022), and Moondream2 (Vikhyatk, 2024). Architecture and differentiators of each are summarised in Table 1 with further details in Appendices 2-5.

Table 1. Transformer-based Model Comparison

Model	**BLIP**	**BLIP-2**	**GIT based COCO**	**Moondream2**
Release/Owner	2022/Salesforce	2023/Salesforce	2022/Microsoft	2024/Moondream.ai
Differentiator	Uses CapFit (bootstrap the captions, a **captioner** generates synthetic captions and a **filter** removes noisy image-text pairs). Allows training on larger bootstrapped dataset.	Uses **frozen pre-trained image encoders** and **frozen large language** models, reducing the number of trainable parameters.	Uses **one image encoder** and **one text decoder** under a single language modelling task while scaling up the pre-training data.	Trained on 1.8 billion parameters and still a **lightweight** model.
Model Architecture	Multimodal mixture of **ViT** image encoder, **BERT** text encoder and BERT text decoder (Figure A.3)	Multimodal mixture of a **CLIP**-like image encoder, a **LLM**, and a **Querying Transformer** (Qformer) acting as an information bottleneck between encoder and LLM (Figure A.5)	Multimodal mixture of **Swin-like vision** transformer image encoder and **BERT** text decoder (Figure A.6)	Uses **SigLIP** processing one pair of image/text at a time, simplifying the encoding process (Zhai et al., 2023) and **Phi-1.5** which was trained using quality data (textbooks) (Li et al., 2023).
Beyond Image captioning	Enhances image comprehension and responsiveness to image-related queries	Zero-shot image-to-text generation including visual conversation, visual knowledge reasoning, common-sense reasoning, and storytelling.	Can extend its capability on video tasks, offering a cost sensitive alternative for video captioning	As a multimodal model it can work with text-to-text applications.
Advantages	Flexible, fast, easy to implement	Enables visual knowledge reasoning and conversation; achieves state-of-the-art performance on tasks	Simple, low run time	Small, lightweight, high precision
Disadvantages	Limited capability	High complexity, has not been tested in the real-world applications	Unclear on how to control the generated caption and how to perform in-context learning without parameter update	Higher run time

2.3 Evaluation

2.3.1 Stress-test scenarios

Inspired by Park et al. (2023), stress-tests are undertaken to evaluate model performance on the following scenarios, in the context of assistive technologies for the vision impaired:

1. Image in the dark
2. Image in the rain
3. Multi-object image
4. Motion snapshot
5. Facial emotion image
6. Fooling image (trick the viewer into perceiving non-existent elements)
7. Rotating object
8. Adversarial image (intentionally created to mislead AI system)

2.3.2 Performance metrics

The model performance is assessed by (i) qualitative human judgement and (ii) scalable quantitative metrics including Bilingual Evaluation Understudy (BLEU) and Recall Oriented Understudy for Gisting Evaluation (ROUGE). BLEU measures translation quality disregarding grammar whereas ROUGE evaluates summarization, including coherence and grammar (Cui et al., 2018). ROUGE is favoured for its recall focus (i.e., how much the words in the labelled caption appeared in the model generated captions) and lack of conciseness penalty, aligning well with caption evaluation.

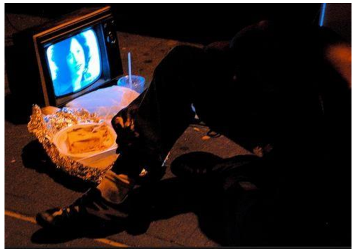
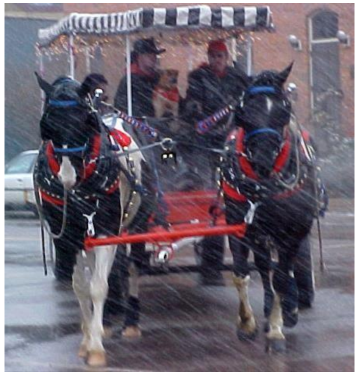
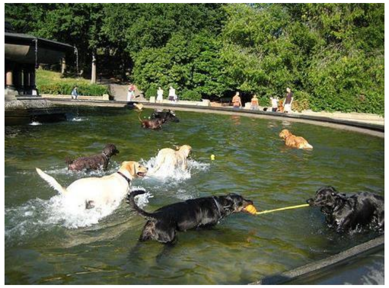
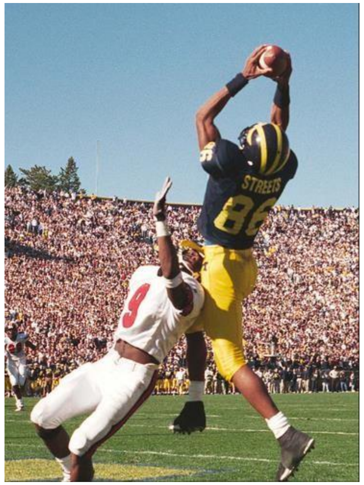
2.4 Limitations

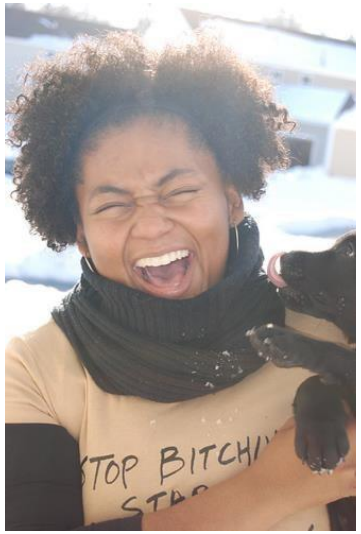
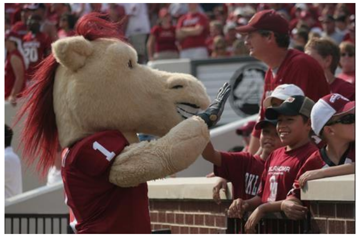
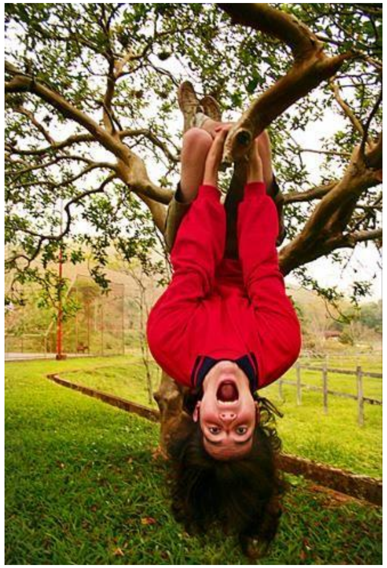
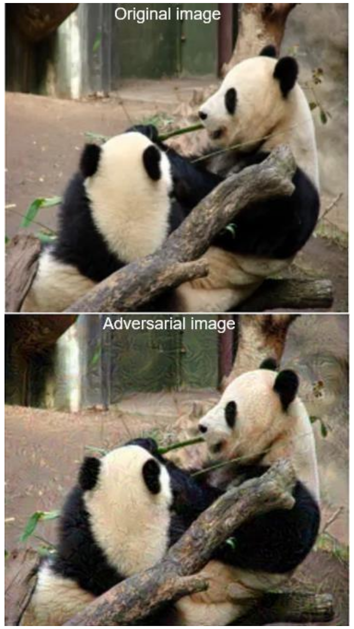
Due to computational constraints, analysis was conducted on a small sample for learning. Findings may not be representative of model performance, which should be evaluated on appropriate sample size and domain-specific contents.

3. Findings

Moondream2 outperformed other models in most stress-testing scenarios, indicating superior caption quality (Table 2 with further details in Appendix 7-8). Additional considerations include the ingeniousness of text, crucial for assisting visually impaired individuals.

Table 2. Image Captioning Result

	1. Darkness	2. Rainy	3. Multi objects	4. Motion
Picture				
Actual Caption	A person eats takeout while watching a small television	A cart containing two men be pull by horse in the rain	A few dogs swim in a lake	A football player in a full stadium jumping and receiving a football
Best generated caption	A person is sitting on the ground with a television in front of them	A couple of horses pulling a cart in the rain	A group of dogs playing in a pool with a yellow stick in their mouths	A football player in a jersey with the number 36 is jumping to catch a football
Model	Moondream2	Git Coco	Moondream2	Moondream2
Justification	Highest Rouge	Highest Rouge & BLEU, most detailed	Highest Rouge, most detailed	Highest Rouge & BLEU, most detailed

	5. Facial Emotion	6. Fooling image	7. Rotation	8. Adversarial
Picture				
Actual Caption	A laugh woman in a scarf	A horse mascot give high five to some football fan	A boy in a red top be hang upside down from a tree	Two panda on a tree
Best generated caption	Woman holding a black dog in her arms and laughing	Mascot greets fans at a game	A person hanging upside down from a tree in a park	Two panda bears are playing with a tree branch
Model	BLIP	Git Coco	Moondream2	Moondream2
Justification	Despite a lower Rouge than Moondream2 (20 vs. 30) BLIP is the only model captioning the laughing emotion	Despite a lower Rouge than Moondream2 (20 vs. 27), Git Coco recognizes the mascot	Highest Rouge & BLEU	Highest Rouge & resulting in the same caption for both images

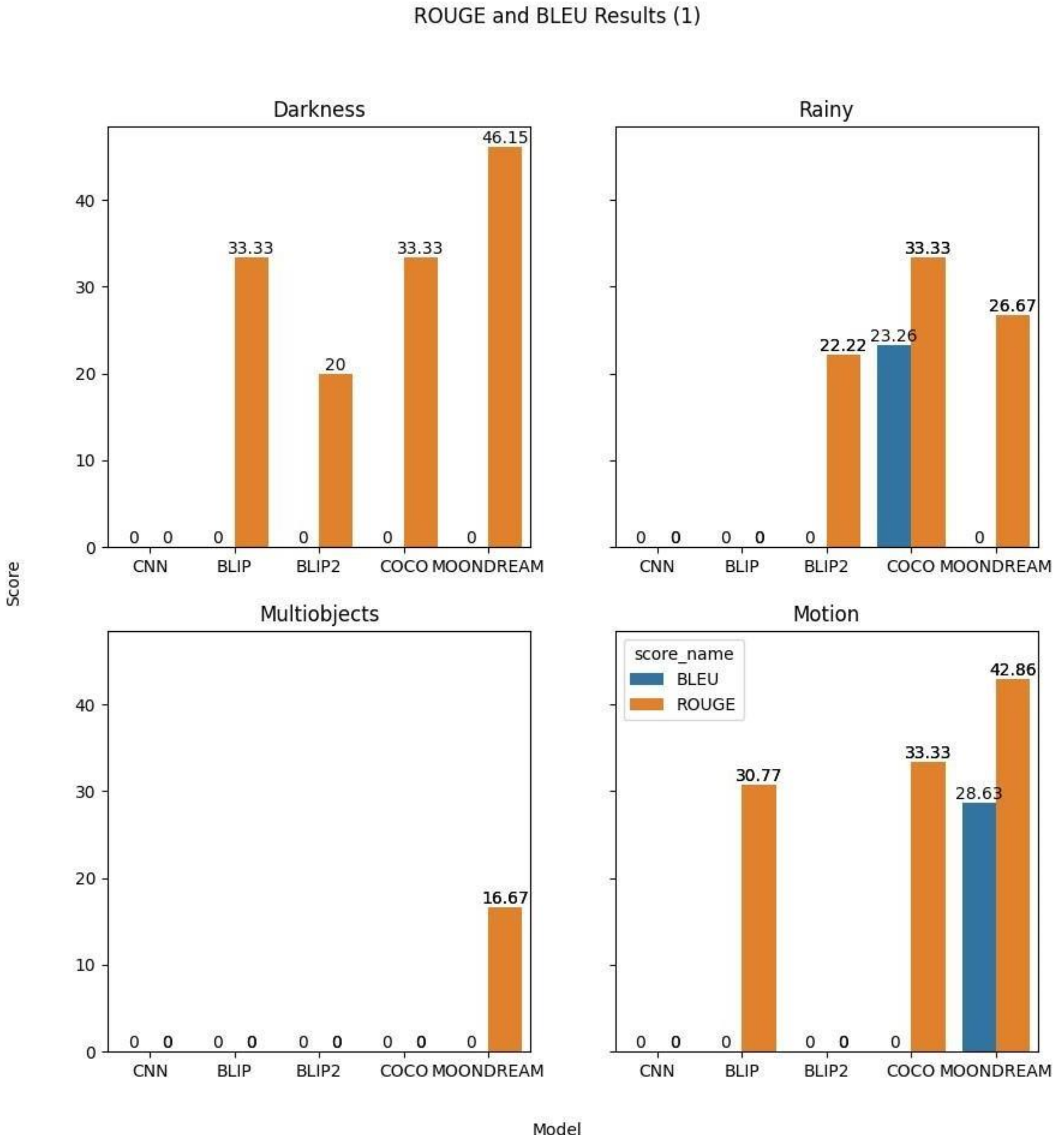


Figure 4. ROUGE and BLEU Result (1)

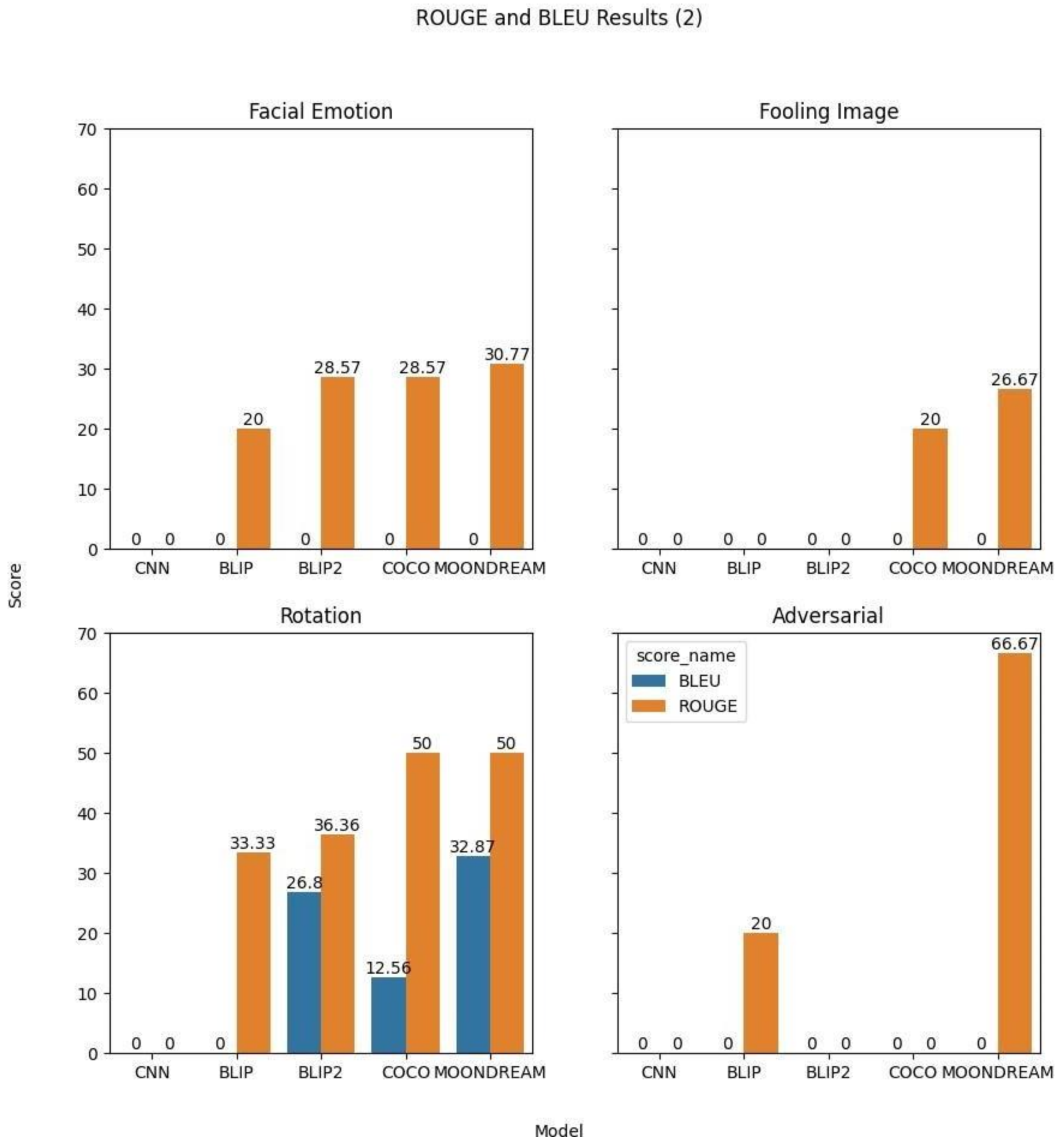


Figure 5. ROUGE and BLEU Result (2)

4. Conclusion

This report highlights the potential of image captioning technologies, especially for aiding the visually impaired. Among the evaluated models, Moondream2 excels in stress tests, demonstrating robust capabilities in realistic scenarios. Whilst the report provides an educational overview, advantages and disadvantages of each image technique and evaluation considerations, limitations in our approach due to computational constraints suggest that future research is required to refine these technologies, expand on evaluation techniques and datasets, and enhancing usability in practical applications.

5. References

Alowais, S. A., Alghamdi, S. S., Alsuhebany, N., Alqahtani, T., Alshaya, A. I., Almohareb, S. N., Aldairem, A., Alrashed, M., Bin Saleh, K. and Badreldin, H. A. (2023). 'Revolutionizing healthcare: the role of artificial intelligence in clinical practice'. BMC Medical Education, 23(1), 689. doi: 10.1186/s12909-023-04698-z

Cui, Y., Yang, G., Veit, A., Huang, X. and Belongie, S. (2018). 'Learning to evaluate image captioning'. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5804-5812. doi: 10.48550/arXiv.1806.06422

Goyal, S., Chattopadhyay, C. and Bhatnagar, G. (2021). 'Knowledge driven Description Synthesis for Floor Plan Interpretation'. International Journal on Document Analysis and Recognition (IJDAR), 24, pp.19-32. doi: 10.1007/s10032-021-00367-3

Gunasekar, S., Zhang, Y., Aneja, J., Mendes, C.C.T., Del Giorno, A., Gopi, S., Javaheripi, M., Kauffmann, P., de Rosa, G., Saarikivi, O. and Salim, A. (2023). 'Textbooks are all you need'. doi: 10.48550/arXiv.2306.11644

Hugging Face. (2023). Choosing a metric for your task. Available at: https://huggingface.co/docs/evaluate/en/choosing_a_metric (Accessed: 7 May 2024)

Kartheek, B. (2022). Image Caption Generator Using CNN and LSTM. Available at: https://www.kaggle.com/code/balajikartheek/image-caption-generator-using-cnn-and-lstm (Accessed 10 May 2024).

Koch, D., Despotovic, M., Leiber, S., Sakeena, M., Döller, M. and Zeppelzauer, M. (2019). 'Real Estate Image Analysis: A Literature Review'. Journal of Real Estate Literature, 27(2), pp.269–300. doi:10.22300/0927-7544.27.2.269

Magalhães, G. V., Santos, R. L. S., Vogado, L. H. S., Paiva, A. C. de., and Neto, P. de A. dos S. (2024). 'XRaySwinGen: Automatic medical reporting for X-ray exams with multimodal model'. Heliyon, e27516. doi: 10.1016/j.heliyon. 2024.e27516

Michelecafagna26. (2022). CIDEr score.

Available at: https://github.com/michelecafagna26/cider (Accessed: 11 May 2024)

Li, Y., Bubeck, S., Eldan, R., Del Giorno, A., Gunasekar, S. and Lee, Y.T. (2023). 'Textbooks Are All You Need II: phi-1.5 technical report' doi: 10.48550/arXiv.2309.05463

Li, J., Li, D., Xiong, C. and Hoi, S. (2022). 'BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation'. International Conference on Machine Learning (ICML). doi: 10.48550/arXiv:2201.12086.

Li, J., Li, D., Savarese, S. and Hoi, S. (2023). ‘BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models’. International Conference on Machine Learning (ICML),814, pp.19730-19742.doi: 10.48550/arXiv.2301.12597.

OpenAI. (2023). CLIP (Contrastive Language-Image Pretraining), Predict the most relevant text snippet given an image. Available at: https://github.com/OpenAI/CLIP (Accessed: 11 May 2024)

Park, S., Um, D., Yoon, H., Chun, S., Yun, S., & Choi, J. Y. (2023). RoCOCO: Robustness Benchmark of MS-COCO to Stress-test Image-Text Matching Models. ArXiv Preprint ArXiv:2304.10727.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G. and Sutskever, I. (2021). ‘Learning Transferable Visual Models From Natural Language Supervision’, Cornell University Library, arXiv.org, Ithaca.

Rogge, N. (2023). Transformers-Tutorials.

Available at: https://github.com/NielsRogge/Transformers-Tutorials/tree/master/BLIP-2 (Accessed: 9 May 2024)

Salesforce. (2023). BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. Available at: https://huggingface.co/Salesforce/blip-image-captioning-large (Accessed: 9 May 2024)

Salesforce. (2024). BLIP-2, OPT-2.7b, pre-trained only. Available at: https://huggingface.co/Salesforce/blip2-opt-2.7b (Accessed: 9 May 2024)

Simonyan, K. and Zisserman, A. (2014). ‘Very Deep Convolutional Networks for Large-Scale Image Recognition’. International Conference on Learning Representations (ICLR). doi: https://doi.org/10.48550/arXiv.1409.1556

Suno. (2023). Metric Evaluation. Available at: https://huggingface.co/suno/bark-small (Accessed: 11 May 2024)

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R. & Bengio, Y. (2016). ‘Show, Attend and Tell: Neural Image Caption Generation with Visual Attention’, Cornell University Library, arXiv.org, Ithaca.

Vikhyatk. (2024). Moondream2. Available at: https://huggingface.co/vikhyatk/moondream2 (Accessed: 9 May 2024)

Vikhyatk. (2024). Moondream. Available at: https://github.com/vikhyat/moondream (Accessed: 9 May 2024)

Wang, J., Yang, Z., Hu, X., Li, L., Lin, K., Gan, Z., Liu, Z., Liu, C. and Wang, L. (2022). ‘GIT: A Generative Image-to-text Transformer for Vision and Language’. arXiv:2205.14100 [cs]. [online] Available at: https://arxiv.org/abs/2205.14100. (Accessed: 10 May 2024)

Wang, Y., Xu, J. and Sun, Y. (2022). ‘End-to-End Transformer Based Model for Image Captioning’. AAAI Conference on Artificial Intelligence, 36(3), pp. 2585–2594. doi: 10.48550/arXiv.2203.15350

Web Content Accessibility Guidelines 2.1. (2023). W3C World Wide Web Consortium Recommendation. Available at: https://www.w3.org/TR/WCAG21/ (Accessed: 10 May 2024)

Zhai, X., Wang, X., Mustafa, B., Steiner, A., Keysers, D., Kolesnikov, A. and Beyer, L. (2022). ‘Lit: Zero-Shot Transfer with Locked-Image Text Tuning’. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 18123-18133. doi: 10.1109/CVPR52688.2022.01759

Zhai, X., Mustafa, B., Kolesnikov, A. and Beyer, L. (2023). ‘Sigmoid Loss for Language Image Pre-Training'. IEEE/CVF International Conference on Computer Vision (ICCV), pp. 11941-11952. doi: 10.1109/ICCV51070.2023.01100

6. Appendices

Appendix 1: Other Applications' Frameworks and Architecture

Image-to-text generation technology has found myriad applications in multitude of sectors. Such as, in healthcare, XRaySwinGen, a multimodal model, generates written medical reports from X-ray pictures (Magalhães et al., 2024) (see Figure A.1 for model architecture overview). X-ray textual descriptions can reveal abnormalities, fractures, and other medical issues that are often missed, helping healthcare providers quickly interpret findings and make informed patient care decisions.

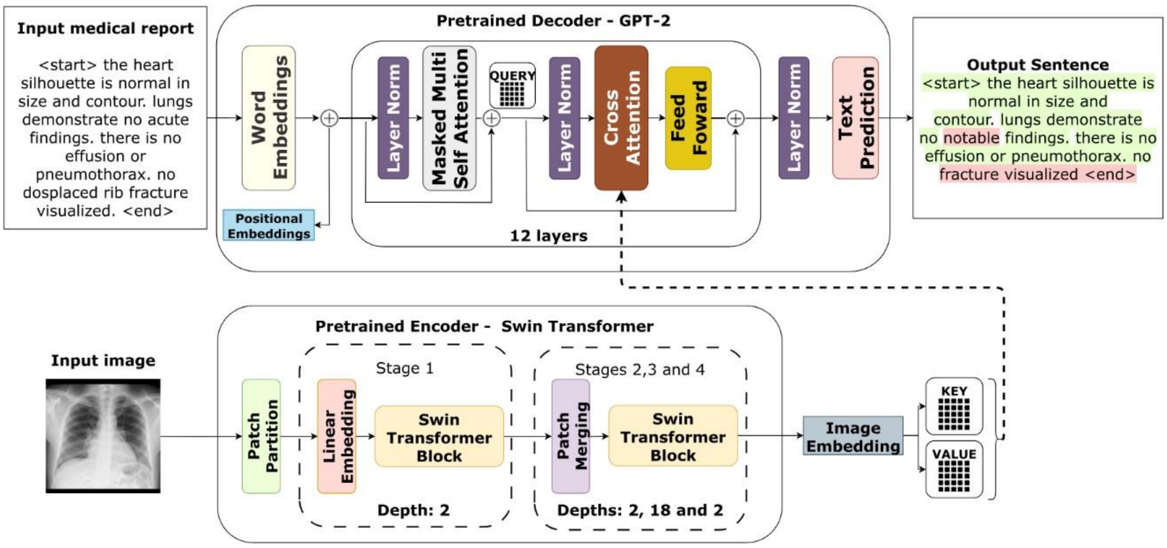
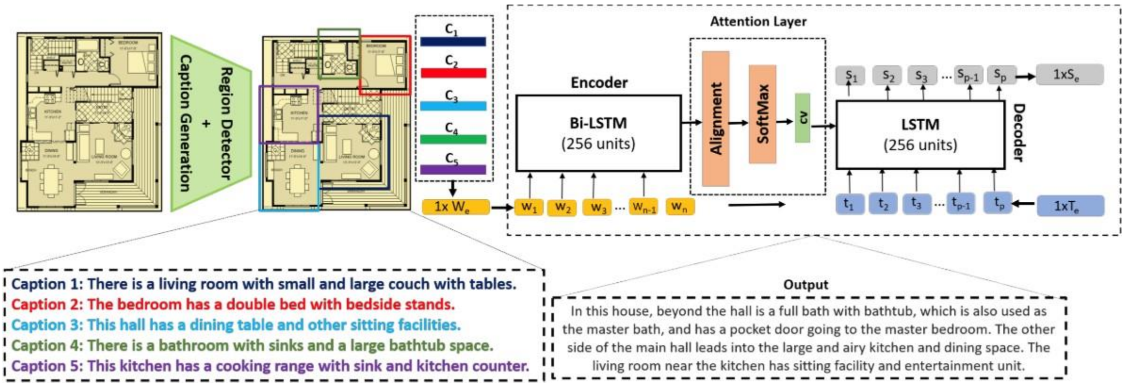

Input medical report
<start> the heart silhouette is normal in size and contour. lungs demonstrate no acute findings. there is no effusion or pneumothorax. no dosplaced rib fracture visualized. <end>
Pretrained Decoder - GPT-2
Word Embeddings
Positional Embeddings
Layer Norm
Masked Multi Self Attention
QUERY
Layer Norm
Cross Attention
Feed Foward
Layer Norm
Text Prediction
12 layers
Output Sentence
<start> the heart silhouette is normal in size and contour. lungs demonstrate no notable findings. there is no effusion or pneumothorax. no fracture visualized <end>
Input image
Pretrained Encoder - Swin Transformer
Patch Partition
Stage 1
Linear Embedding
Swin Transformer Block
Depth: 2
Stages 2,3 and 4
Patch Merging
Swin Transformer Block
Depths: 2, 18 and 2
Image Embedding
KEY
VALUE

Figure A.1. Overview of XRaySwinGen Multimodal Model Architecture

Real estate listings use image-to-text generation to extract useful information from photographs (Koch et al., 2019). Key characteristics like room measurements and features are automatically identified in text. This gives buyers and tenants more information about home features and layouts.

Description Synthesis from Image Cue (DSIC) and Transformer Based Description Generation (TBDG) use contemporary deep neural networks for visual feature extraction to produce textual descriptions from floor plan images (Goyal et al., 2021). (See Figure A.2 for TBDG framework overview).



Caption 1: There is a living room with small and large couch with tables.
Caption 2: The bedroom has a double bed with bedside stands.
Caption 3: This hall has a dining table and other sitting facilities.
Caption 4: There is a bathroom with sinks and a large bathtub space.
Caption 5: This kitchen has a cooking range with sink and kitchen counter.

In this house, beyond the hall is a full bath with bathtub, which is also used as the master bath, and has a pocket door going to the master bedroom. The other side of the main hall leads into the large and airy kitchen and dining space. The living room near the kitchen has sitting facility and entertainment unit.

Figure A.2. Overview of TBDG of Generating Paragraph Description from Input Floor Plan Image.

Appendix 2: Transformer-based Model Overview

Table A.1. Transformer-based Model Comparison (Complete Version of Table 1)

Model	BLIP	BLIP-2	Microsoft GIT based COCO	Moondream2
Release/Owner	2022 / Salesforce	2023 / Salesforce	2022 / Microsoft	2024 / Moondream.ai
Differentiator	BLIP improves the quality of pretrained web data by bootstrapping the captions, where a captioner generates synthetic captions and a filter removes noisy image-text pairs (CapFilt). The model then can be trained on a much larger bootstrapped dataset (vs. limited human labelled dataset in other models)	BLIP2 bootstraps vision-language pre-training from off-the-shelf frozen pre-trained image encoders and frozen large language models, significantly reducing the number of trainable parameters and associated computation costs.	GIT simplifies the architecture as one image encoder and one text decoder under a single language modeling task while scaling up the pre-training data and the model size to boost the model performance	Moondream2 implements SigLIP for the image processing evaluating one pair of image/text during training, simplifying the encoding process (Zhai et al., 2023). It also uses Phi-1.5 for decoding, which was trained using textbooks, meaning higher quality texts (Li et al. 2023).
Model Architecture	Multimodal mixture of ViT image encoder, BERT text encoder and BERT text decoder (Figure A.3)	Multimodal mixture of a CLIP-like image encoder, a large language model, and a Querying Transformer (QFormer) acting as an information bottleneck between the frozen image encoder and the frozen LLM, where it feeds the most useful visual feature for the LLM to output the desired text (Figure A.5)	Multimodal mixture of a Swin-like vision transformer image encoder and BERT text decoder (Figure A.6)	Multimodal mixture of SigLIP for encoder, and Phi-1.5 as decoder.
Beyond Image captioning	BLIP can be used to enhance models' ability to understand and respond to questions about images. Moreover, BLIP can aid in translation tasks where the input or output involves both textual and visual information.	BLIP2's instructed zero-shot image-to-text generation capability may have a wide range of application including visual conversation, visual knowledge reasoning, visual common-sense reasoning, and storytelling, adding more nuances to assistive solutions or visually impaired people.	Despite not having video-dedicated encoders, GIT can extend its capability on video tasks, offering a cost sensitive alternative for video captioning	Being a multimodal model, it can work with text-to-text applications as well. However, its primary intended usage is for image descriptions.

Model	BLIP	BLIP-2	Microsoft GIT based COCO	Moondream2
Pre-training data	The bootstrapped dataset from web data and labelled COCO dataset	BLIP with 129M images (including COCO), LAION400M with 115M images (Li et.al., 2023)	GIT was trained on 0.8 billion image-text pairs GIT based was trained on 10 million image-text pairs	SigLIP was trained on LiT dataset, which uses data from YFCC100m and CC12M (Zhai et al. 2022). Phi-1.5 uses training data from Phi-1, consisting in a mix of StackOverflow code, and textbooks to emulate “textbook quality” data (Gunasekar et al. 2023).
Fine-tuning data for image captioning	N/A	COCO	COCO	N/A
Parameters	N/A	188 million	GIT 0.7 billion	1.86 billion
Running time	2 minutes	3 minutes	4 minutes	10 minutes

Appendix 3: BLIP Architecture and Overview

Bootstrapping Language-Image Pre-training (BLIP) is a BERT-based pre-training method, also known as Multimodal mixture of Encoder-Decoder (MED). It facilitates unified vision-language understanding and generation tasks by jointly training on large-scale multimodal datasets, as illustrated in Figure A.3. This model card describes a pre-trained model for image captioning, trained on the COCO dataset. The model architecture is based on a ViT large backbone.

BLIP, also known as Captioning and Filtering (shown in Figure A.4), is flexible, fast (only takes 2 minutes to run the model in the context of this study), and easy to implement, which makes it adaptable to many industries and applications. Its capabilities are restricted compared to newer models.

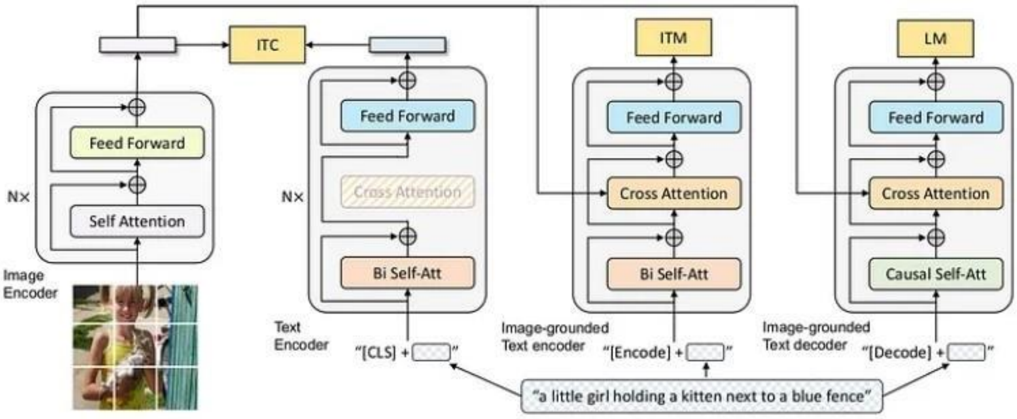


Figure A.3. Pre-training Model Architecture and Objectives of BLIP (Li et.al., 2022)

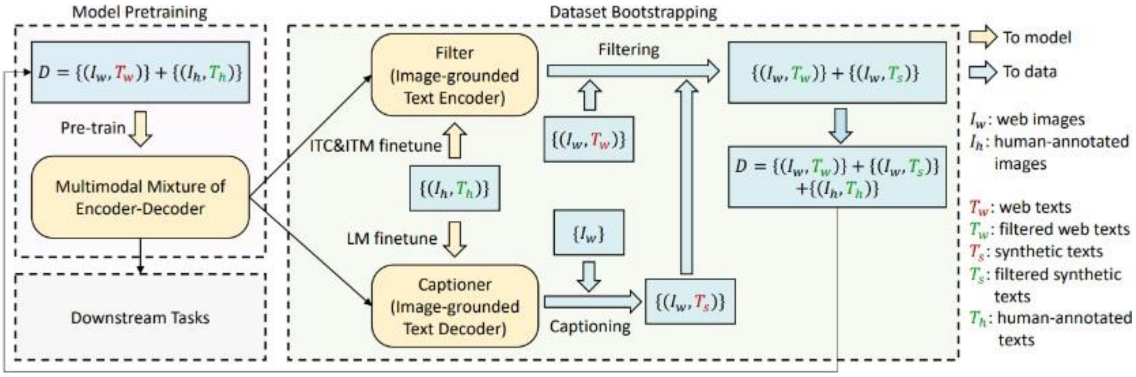

Model Pretraining
$D = \{(I_w, T_w)\} + \{(I_h, T_h)\}$
Pre-train
Multimodal Mixture of Encoder-Decoder
Downstream Tasks
Dataset Bootstrapping
Filter (Image-grounded Text Encoder)
ITC&ITM finetune
$\{(I_h, T_h)\}$
LM finetune
Captioner (Image-grounded Text Decoder)
Filtering
$\{(I_w, T_w)\}$
$\{I_w\}$
Captioning
$\{(I_w, T_s)\}$
$\{(I_w, T_w)\} + \{(I_w, T_s)\}$
$D = \{(I_w, T_w)\} + \{(I_w, T_s)\} + \{(I_h, T_h)\}$
To model
To data
I_w: web images
I_h: human-annotated images
T_w: web texts
T_w: filtered web texts
T_s: synthetic texts
T_s: filtered synthetic texts
T_h: human-annotated texts

Figure A.4. Learning Framework of BLIP (Li et.al., 2022)

BLIP-2, an improved version of BLIP, helps LLMs process and understand images for zero-shot image-to-text generation. Figure A.5 shows BLIP-2 pre-training using the Querying Transformer in two stages: vision-language representation learning with a frozen image encoder and generative learning with a frozen LLM. This model performs well in visual knowledge reasoning and communication tasks, especially visual question answering. However, BLIP-2 has not been tested in the real-world applications.

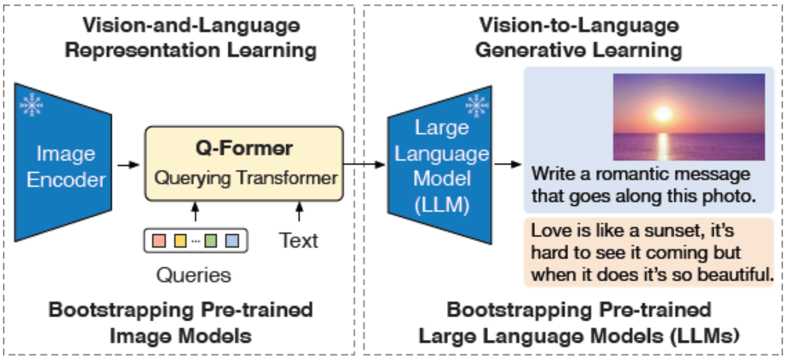

Vision-and-Language
Representation Learning
Image Encoder
Q-Former
Querying Transformer
Queries
Text
Bootstrapping Pre-trained
Image Models
Vision-to-Language
Generative Learning
Large Language Model (LLM)
Write a romantic message that goes along this photo.
Love is like a sunset, it's hard to see it coming but when it does it's so beautiful.
Bootstrapping Pre-trained
Large Language Models (LLMs)

Figure A.5. Overview of BLIP-2's Framework (Li et.al., 2023)

Appendix 4: GIT Overview

It adopts a smaller variant of GIT (Generative Image to Text Transformation) model that was pre-trained on 10 million image-text pairs then finetuned on the COCO dataset. Advantages of GIT is its simple architecture with one image encoder and one text decoder under a single language modelling task, reducing the running time, and proven to be beneficial to handle additional category data. However, as the model focuses on the pretraining-and-finetuning strategy, it is unclear on how to control the generated caption and how to perform in-context learning without parameter update.

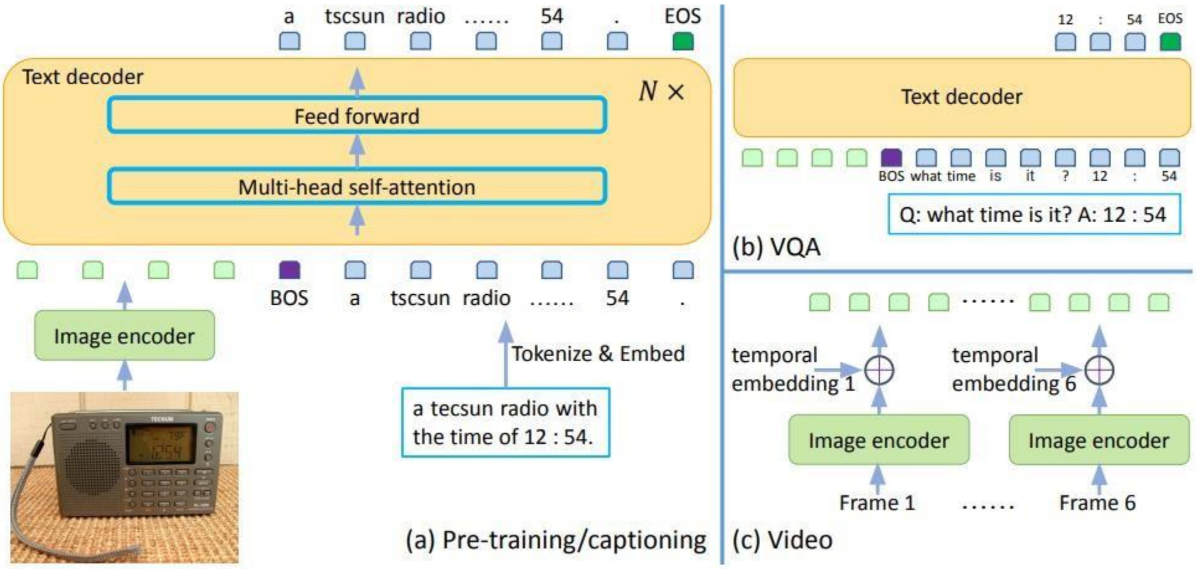

a tscsun radio ...... 54 . EOS
Text decoder
N ×
Feed forward
Multi-head self-attention
BOS a tscsun radio ...... 54 .
Image encoder
Tokenize & Embed
a tecsun radio with the time of 12 : 54.
(a) Pre-training/captioning
12 : 54 EOS
Text decoder
BOS what time is it ? 12 : 54
Q: what time is it? A: 12 : 54
(b) VQA
temporal embedding 1
temporal embedding 6
Image encoder
Image encoder
Frame 1
......
Frame 6
(c) Video

Figure A.6. Overview of GIT's Framework (Wang et.al., 2022)

Appendix 5: Moondream2 Overview

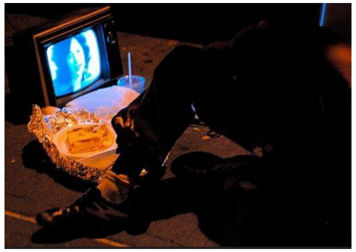
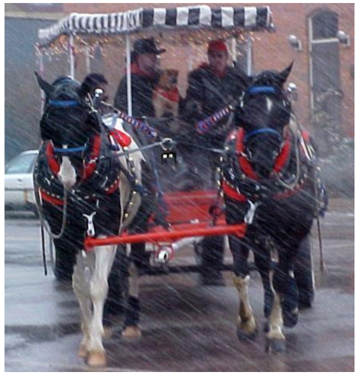
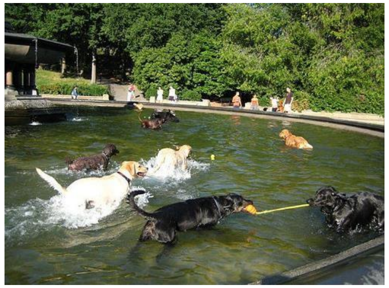
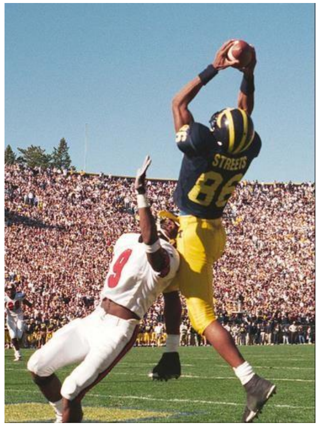
Moondream2 is a small lightweight multimodal model that has extensive visual descriptions, including poetic styles. It was released during early 2024, its open-source nature allows for customisation. However, it took 10 minutes to run the same 10 photos, compared to 2-4 minutes for other models. Its major benefit is that it produces better captions than other models despite taking longer to process. Thus, recommended for precision at lower runtimes. Since it was recently introduced, upgrades are frequent, therefore production and implementation require careful attention.

As for its architecture, it was trained on 1.86 billion parameters and uses transformers for encoding and decoding. The image encoder uses SigLIP (Sigmoid loss for Language-Image Pre-training), processing each “image-text” pair independently instead of considering all images at the same time, turning it into a binary classification problem (Zhai et al., 2023).

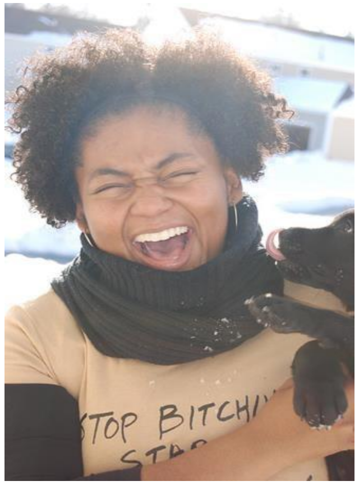
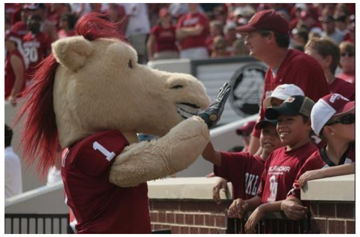
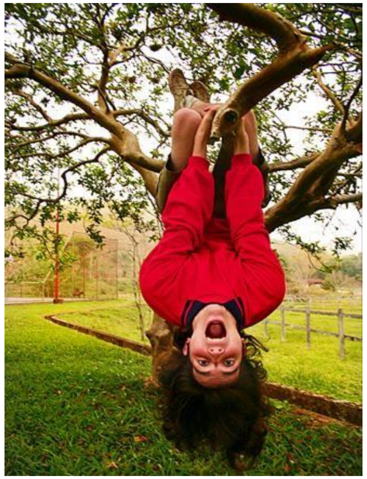
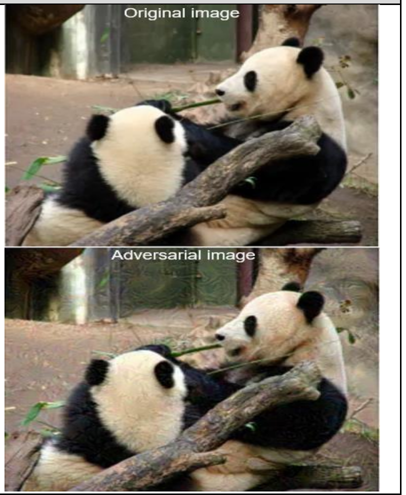
Decoding uses Microsoft's Phi-1.5, trained on textbooks and “textbook-like” data mixed with internet selected prompts for diversity, resulting in a smaller language model of 1.3 billion parameters, performing similarly to models five times larger (Li et al., 2023).

Appendix 6: Model Captions & Qualitative Judgment

Table A.2. Image Captioning Result (Complete Version of Table 2)

	1. Darkness	2. Rainy	3. Multi objects	4. Motion
Picture				
Actual Caption	A person eats takeout while watching a small television	A cart containing two men be pull by horse in the rain	A few dogs swim in a lake	A football player in a full stadium jumping and receiving a football
CNN-LSTM based	startseq of throw the boy costume in in in…	startseq climb white air as hats in in in…	startseq purple boy running woman while the on red in in in in…	startseq dresses runs on tikes' dresses runs blue bench in in in…
BLIP	someone is sitting on the ground watching a television with a cat on it	horses pulling a carriage with people in it on a rainy day	dogs playing in a pool of water with a ball in it	arafed football player catching a ball while another player tries to catch it
BLIP-2	person sitting on the ground	a man and his dog	a group of dogs in a pool	a man in a yellow and white uniform
GIT COCO	a person sitting on the ground watching tv	a couple of horses pulling a cart in the rain.	a group of dogs swimming in a pond.	american football player catches a touchdown pass over american football player.
Moondream2	A person is sitting on the ground with a television in front of them.	A horse pulling a carriage with two horses and a man in a red hat.	A group of dogs playing in a pool with a yellow stick in their mouths.	A football player in a jersey with the number 36 is jumping to catch a football.

Group 1

	5. Facial Emotion	6. Fooling image	7. Rotation	8. Adversarial
Picture				
Actual Caption	A laugh woman in a scarf	A horse mascot give high five to some football fan	A boy in a red top be hang upside down from a tree	Two panda on a tree
CNN-LSTM based	startseq tries tries tries of on boy girl man boy with has the on child in in in…	startseq tattoos wrestling in in in in car out in in in…	startseq grass to the camera down and while man children jump in the four in in in…	startseq boy with team standing in dog boy with team standing in in in…/startseq boy with team is dogs below in large dogs dog walking with in in in
BLIP	woman holding a black dog in her arms and laughing	araffe dressed in red and white holding a tennis racket	araffed girl upside down on a tree branch in a park	panda bear eating bamboo in zoo enclosure with tree branch in foreground / pandas eating bamboo in a zoo enclosure with a tree branch
BLIP-2	a woman holding a dog	a man in a red shirt	a young woman hanging upside down from a tree	two pandas eating bamboo / two pandas eating bamboo
GIT COCO	a woman with a dog in her mouth.	mascot greets fans at a game.	a boy in a tree	pandas playing with a branch / two pandas are playing with a branch.
Moondream2	A woman with a dog in her arms and a shirt that says Stop Bitching Stars.	A man in a red shirt with the number 1 on it is shaking hands with a mascot.	A person hanging upside down from a tree in a park.	Two panda bears are playing with a tree branch / Two panda bears are playing with a tree branch.

Appendix 7: Model Quantitative Evaluation Scores (BLEU & ROUGE)

	1. Darkness	2. Rainy	3. Multi objects	4. Motion	5. Facial Emotion	6. Fooling image	7. Rotation	8A. Original	8B. Adversarial
BLEU score									
CNN-LSTM based	0	0	0	0	0	0	0	0	0
BLIP	0	0	0	0	0	0	0	0	0
BLIP-2	0	0	0	0	0	0	26.8	0	0
GIT COCO	0	23.26	0	0	0	0	12.56	0	0
Moondream2	0	0	0	28.63	0	0	32.87	0	0
ROUGE score									
CNN-LSTM based	0	0	0	0	0	0	0	0	0
BLIP	33.33	0	0	30.77	20	0	33.33	33.33	20
BLIP-2	20	22.22	0	0	28.57	0	36.36	0	0
GIT COCO	33.33	33.33	0	33.33	28.57	20	50	0	0
Moondream2	46.15	26.67	16.67	42.86	30.77	26.67	50	66.67	66.67