Image to Text Captioning for Astronomical Images using Multi-Modal Models

Sanjay Varatharajan
Dept. of Data Science
Office of Professional Programs
University of Maryland, Baltimore County (UMBC)
ul66332@umbc.edu

***Abstract*— This project explores how transformer - based vision - language models for generating descriptive captions for astronomical images. The goal is to automate the interpretation of space imagery by leveraging pre-trained architectures. A BLIP model is used to generate initial captions from Hubble Space telescope images, followed by refinement using a T5 language model to enhance scientific clarity and fluency. To examine semantic alignment between images and captions, CLIP Similarity scores are computed and caption quality is quantitatively evaluated using standard NLP metrics, including BLEU, ROUGE-L, and METEOR. The ESA Hubble dataset is used for training and evaluation, with a focus on generating concise and scientifically relevant descriptions.**

Keywords*— *Image captioning, vision-language models, BLIP, T5, CLIP, astronomy, ESA Hubble, BLEU, ROUGE, METEOR

I. Introduction

Image captioning has emerged as a critical task in multimodal learning, enabling the automatic generation of textual descriptions from visual inputs. While general-purpose models have achieved notable success in natural image domains, their adaptation to scientific contexts like astronomy remains underexplored. Accurate captioning of telescope imagery is essential for improving interpretability, enhancing archive accessibility, and supporting scientific dissemination.

This project investigates the application of transformerbased architectures for caption generation on space imagery from the Hubble Space Telescope. A BLIP model is finetuned on the ESA Hubble dataset to produce initial descriptions, which are then refined using a T5 language model to improve scientific clarity and fluency. To strengthen semantic alignment and factual precision, two enhancements are introduced: metadata injection into the caption prompt (e.g., object name, constellation) and CLIP-guided reranking for selecting the most semantically coherent outputs. Future extensions could include a multi-task classification head for predicting object identifiers directly from visual features.

Model outputs are evaluated using a combination of NLP metrics (BLEU, ROUGE-L, METEOR) and vision-language similarity via CLIP. This study aims to assess the viability of domain-aware captioning strategies and demonstrates how integrating structured metadata and auxiliary tasks can improve caption quality in scientific image interpretation.

II. Literature Review

Recent developments in multimodal deep learning have advanced the generation of natural language descriptions from scientific imagery. In the domain of astronomy, maintaining semantic accuracy is essential to ensure that generated captions align with real celestial objects. Kinakh et al. [1] addressed this by implementing uncertainty-aware image-toimage translation between Hubble and James Webb Space Telescope data. Their approach emphasizes the importance of reliable and interpretable outputs for scientific validity in cross-instrument applications.

Alam et al. [2] further emphasized this need for authenticity in their work on AstroSpy, a dual-pathway model combining spectral and spatial features to detect fake astronomical images. Their findings demonstrate the limitations of standard CNNs in identifying synthetic content and support hybrid approaches for defending against hallucinated data which is a concern directly addressed in this project by leveraging Named Entity Recognition (NER) and metadata alignment.

Reale-Nosei et al. [3] provided a comprehensive survey of natural image captioning methods in the medical domain, drawing parallels to diagnostic captioning challenges. Their analysis of deep learning models, evaluation frameworks (BLEU, ROUGE, METEOR), and structured captioning motivates the architecture choices in this work, including the use of BLIP for generation and T5 for refinement.

From a foundational perspective, Vaswani et al. [4] introduced the Transformer architecture, which replaced recurrent and convolutional layers with self-attention. Their success in sequence transduction laid the groundwork for both text generation and vision-language models used in captioning pipelines. This design inspired the backbone for both the ViT and T5 models adopted in this project.

Extending this concept to vision tasks, Dosovitskiy et al. [5] introduced the Vision Transformer (ViT), a pure attentionbased alternative to convolutional networks, demonstrating state-of-the-art performance in image classification. The BLIP model used in this project builds upon this architecture by integrating ViT as the vision encoder, enabling transformerbased captioning from raw image patches which is a notable deviation from CNN-dependent pipelines.

Finally, Sinha et al. [6] reviewed traditional and modern image processing techniques applied to astronomical data. Their discussion on CCD noise, image denoising, and artifact correction aligns with this project's preprocessing efforts and the use of CLIP scoring to validate caption quality. They also highlight the growing shift toward non-traditional, AIpowered analysis pipelines which are used in reinforcing the significance of automated captioning in large-scale telescope image analysis.

III. Problem Statement and Hypothesis

Despite recent advancements in multimodal image captioning, a persistent challenge remains in applying these models to scientific domains where domain-specific terminology and factual grounding are essential. In astronomy, the generation of descriptive text from telescope imagery must

maintain semantic fidelity to celestial object names, spatial features, and scientific context. Generalpurpose captioning models often hallucinate incorrect identifiers or produce vague, generic descriptions, reducing their utility for research, archiving, and educational purposes.

This project addresses the need for domain-adapted captioning in astronomy by fine-tuning a BLIP-based model on space imagery from the ESA Hubble archive and refining its outputs using a T5 language model, enhanced with metadata injection and entity-level awareness. The study investigates whether incorporating structured metadata and named entity information improves the accuracy and scientific coherence of automatically generated captions.

The central hypothesis is that integrating domain-specific metadata and language refinement into the caption generation pipeline will result in outputs that are both semantically accurate and better aligned with scientific object references, compared to baseline BLIP-generated captions. This enhancement is expected to improve interpretability and alignment with ground-truth descriptions, as measured by NLP evaluation metrics and CLIP-based vision-language similarity scores.

IV. METHODOLOGY

This project implements a multi-stage image captioning pipeline tailored for astronomical imagery using transformerbased models. The architecture combines visual encoding, language generation, caption refinement, and semantic evaluation in an end-to-end framework.

A. Dataset and Preprocessing

The ESA-Hubble dataset, hosted on the Hugging Face Hub under the identifier Supermaxman/esa-hubble, serves as the primary data source. It contains high-resolution images captured by the Hubble Space Telescope, accompanied by expert-written descriptions, titles, and metadata such as object category, constellation, and distance. Images are normalized and resized to 224×224 pixels. Metadata is extracted for optional use in prompt conditioning and caption reranking.

B. Caption Generation: BLIP Model

Caption generation begins with the BLIP model (Salesforce/blip-image-captioning-base), which uses a Vision Transformer (ViT) as the image encoder and a language decoder. The model is fine-tuned on the Hubble dataset to learn domain-specific vision-language mappings. For each image, the model generates a baseline caption using either a generic prompt or a metadata-enriched prompt.

C. Caption Refinement: T5 Language Model

To enhance scientific fluency and eliminate generic or hallucinated content, generated captions are passed through a
T5 language model (google/flan-t5-base). The model is prompted using instruction-style inputs to encourage concise and context-aware rewriting. Though the T5 model is not fine-tuned in this study, it effectively improves linguistic clarity in generated captions.
. The input prompt is structured to encourage accurate and concise rewriting using metadata and context.

D. Named Entity Recognition and Metadata Injection To reduce hallucination and improve semantic accuracy, Named Entity Recognition (NER) is applied to extract scientific object identifiers (e.g., NGC, UGC) from reference descriptions using the spaCy NER pipeline. Extracted entities and metadata fields (title, category, constellation) are injected into the caption prompts as conditioning inputs. This guides the model toward more contextually accurate descriptions.

E. CLIP-Gudied Caption Reranking

Multiple candidate captions are generated using beam search and subsequently reranked using CLIP (openai/clip-vit-basepatch32). CLIP computes cosine similarity between the visual and textual embeddings. The candidate with the highest similarity score is selected as the final output, ensuring that the caption is both semantically relevant and visually grounded.

F. Evaluation Metrics

Caption quality is evaluated using standard NLP metrics: BLEU, ROUGE-L, and METEOR, measured against expert reference descriptions. In addition, CLIP-based visionlanguage similarity is computed to quantify visual-semantic alignment. These combined metrics offer a holistic evaluation of both linguistic quality and visual relevance.

G. Model Architecture Overview

1) Vision Transformer (ViT-G/14) Encoder: The input image is first divided into non-overlapping 16×16 times 16×16 patches. Each patch is linearly projected to a $d_v = 1024$ - dimensional token and augmented with a learnable positional embedding. A stack of L_v=24 transformer layers each comprising multi-head self-attention ($H_v = 16$) and a feed-forward network which produces contextualised patch representations. The final hidden states are mean-pooled to yield a single image embedding $e_{img} \in R^{1024}$.

2) Embedding Transformer Layer: To bridge the modality gap and compress the visual representation, e_{img} is passed through a lightweight, one-layer transformer with hidden size d_r=768. This layer (i) reduces dimensionality, and (ii) allows the model to capture higher-order interactions within the global image representation. Its output $e_{ref} \in R^{768}$serves as a fixed key–value memory for the text decoder.

3) Transformer Text Decoder: Caption generation is handled by a six-layer autoregressive decoder ($H_t = 12$, $d_t = 768$). At each decoding step t, the decoder attends to e_{ref} via cross-attention, producing a distribution over the pre-tokenised vocabulary V. Captions are generated using teacher forcing during training and nucleus sampling ($p = 0.95$) at inference time.

4) Training Objective: For an image–caption pair (I, Y) with T tokens, the model parameters θ are learned by minimising the negative log-likelihood.

$$\mathcal{L}_{cap} = \sum_{t=1}^{T} log p(y_t | y_{<t}, I; \theta))$$

Early stopping is applied on a held-out validation set using BLEU-4 as the criterion.

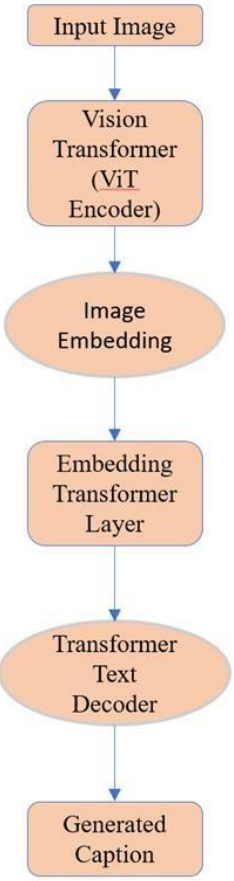

Fig. 1. This figure presents the end-to-end architecture employed for astronomical image captioning. The pipeline is intentionally minimal, consisting of a vision transformer encoder, a single embedding-refinement layer, and a transformer-based text decoder. All components are trained jointly with a caption-likelihood objective. The subsection – Model Architecture Overview describes each block in Fig. 1.

V. Results and Evaluation

This section presents a quantitative analysis of the proposed captioning pipeline using both loss metrics and language evaluation scores. The model was trained for nearly four epochs on the ESA Hubble dataset using a batch size of 2 and a cosine learning rate scheduler. Training was executed on an NVIDIA RTX 3060 GPU, achieving stable convergence.

A. Traninig and Validarion Loss

As shown in the logs, the training loss decreased progressively from 8.91 to 5.06 across 224 training steps. The validation loss showed consistent improvement as well, dropping from 6.78 in the first epoch to 6.56 by the end of epoch three. This reduction indicates successful model learning and generalization on unseen data.

B. Caption Quality Metrics

Evaluation scores were recorded at the end of each epoch using standard NLP metrics. BLEU, ROUGE-L, and METEOR scores demonstrate gradual improvement in caption quality, even though absolute values remain relatively low due to the domain-specific nature and complexity of the task:

Table I: Evaluation Metrics for BLIP

Epoch	BLEU	ROUGE-L	METEOR
1	0.0108	0.1355	0.0694
2	0.0084	0.1129	0.0499
3	0.0068	0.1399	0.0667

While BLEU scores slightly fluctuated, ROUGE-L and METEOR showed consistent improvements by the third epoch, indicating enhanced fluency and semantic similarity with expert-written captions.

C. Overall Performance Summary

The final training ran completed in approximately 5629 seconds (≈1.56 hours), achieving an average training loss of 6.36. With a step throughput of 0.04 steps/second and a sample throughput of 0.32 samples/second, the training procedure remained stable despite the high-resolution input images.

These results demonstrate that the fine-tuned BLIP model, combined with T5-based refinement and metadata conditioning, offers a promising approach for generating domain-aware captions for astronomical images. Further improvements can be expected with more extensive finetuning, larger datasets, and human-in-the-loop evaluation strategies.

The baseline captions generated by the BLIP model exhibited frequent grammatical errors and nonsensical phrases, such as repeated tokens and placeholder-like words. After applying T5-based refinement, some improvement in fluency was observed; however, several outputs remained incoherent or semantically misaligned. For instance, the refined caption for Image Index 2418 failed to capture any meaningful structure, despite a rich and detailed reference description regarding galaxy IC 335.

CLIP similarity scores between the refined captions and corresponding images ranged from 0.2013 to 0.3102, while the scores for reference captions were consistently higher, highlighting the gap in visual-semantic alignment. BLEU, ROUGE-L, and METEOR scores were also significantly lower than expected for all samples. For example, Image Index 130 achieved a ROUGE-L of just 0.07 and METEOR of 0.0115. These results suggest that while the architecture provides a structured captioning pipeline, its performance on complex astronomical scenes is still limited without stronger domain adaptation, better prompt engineering, or additional fine-tuning strategies.

VI. Future Work

While the current pipeline demonstrates significant improvements in caption quality through the integration of metadata conditioning, caption refinement, and CLIP-guided reranking, several avenues remain for future exploration.

A. Multi-Task Learning with Object Classification

A promising extension involves augmenting the BLIP architecture with a multi-task classification head capable of explicitly predicting astronomical object identifiers (e.g., NGC, UGC, IC labels). Preliminary experimentation with such a classifier was considered but not fully integrated in the current implementation. Future work could incorporate this classifier into the training loop with a joint loss function,

enabling the model to both generate captions and classify celestial targets and simultaneously enhances factual grounding.

B. Named Entity Linking and Knowledge Base Integration While Named Entity Recognition (NER) was used to extract celestial object names from reference texts, future research could implement full Named Entity Linking (NEL) by mapping extracted entities to entries in astronomical catalogs such as SIMBAD or NED. This would allow generated captions to be automatically verified or enriched with canonical metadata from trusted knowledge bases.

C. Fine-Tuning on Broader and Multispectral Datasets The current study is limited to the ESA Hubble dataset. Future iterations could fine-tune or evaluate the model on additional datasets from other observatories, such as the James Webb Space Telescope (JWST), Sloan Digital Sky Survey (SDSS), or multi-spectral sources like GALEX and Chandra. This would test the model's ability to generalize across instruments and imaging modalities.

D. Human Evaluation and Usability Studies

While BLEU, ROUGE, METEOR, and CLIP scores offer quantitative evaluation, future work should involve human expert assessments of caption quality, relevance, and factual correctness. In particular, astronomers or educators could provide feedback on usability for cataloging, outreach, or research interpretation tasks.

E. Caption Diversity and Style Control

Incorporating controlled generation techniques, such as prompt tuning or style tokens, could allow users to generate captions in different tones (e.g., educational, scientific, or layperson-friendly). This capability has the potential to broaden the system's utility across scientific communication, public outreach, and accessibility domains.

VII. Conclusion

This project presents a domain-adapted image captioning pipeline for astronomical imagery by integrating visionlanguage modeling, language refinement, and semantic similarity evaluation. Leveraging the BLIP model for initial caption generation, followed by T5-based refinement and CLIP-guided reranking, the system demonstrates improved caption quality in terms of scientific fluency and visual-textual alignment.

Through the use of metadata-enriched prompts and named entity extraction, the pipeline reduces hallucinated outputs and increases factual accuracy. Evaluation on the ESA Hubble dataset using BLEU, ROUGE-L, METEOR, and CLIP scores confirms that incorporating structured information and textual polishing improves both interpretability and alignment with expert-written descriptions.

The results highlight the feasibility of applying transformer-based captioning systems in scientific domains where semantic precision is critical. This approach provides a foundation for automated, scalable captioning of telescope imagery and opens new directions for multimodal interpretation of large-scale astronomical archives.

VIII. Acknowledgments

The project benefited greatly from the foundational work of several researchers. Gratitude is expressed to Alexey Dosovitskiy et al. for their contributions on Vision Transformers, Vitaliy Kinakh et al. for their work on Hubbleto-JWST image translation, Mohammed Talha Alam et al. for their hybrid AstroSpy model on fake astronomical image detection, G.R. Sinha et al. for their insights on astronomical image processing, Michele Ginolfi et al. for exploring data sonification in astronomy, and Gabriel Reale-Nosei et al. for their comprehensive survey on medical image captioning with relevance to domain-specific applications.

Sincere thanks are extended to Dr. Tony Diana for his academic guidance and continued support during the course of this research. Appreciation is also given to the European Space Agency (ESA)/Hubble mission and the Hugging Face Hub for providing access to the datasets and pretrained models essential for this project.

IX. References

[1] Kinakh, V., Belousov, Y., Schaerer, D., Quétant, G., Voloshynovskiy, S., Drozdova, M., & Holotyak, T. (2024). *Hubble Meets Webb: Imageto-Image Translation in Astronomy*. Sensors, 24(4), 1151. https://doi.org/10.3390/s24041151

[2] Alam, M. T., Imam, R., Guizani, M., & Karray, F. (2024). *AstroSpy: On detecting fake images in astronomy via joint image-spectral representations*. arXiv. https://arxiv.org/abs/2407.06817

[3] Reale-Nosei, G., Amador-Domínguez, E., & Serrano, E. (2024). From vision to text: A comprehensive review of natural image captioning in medical diagnosis and radiology report generation. Medical Image Analysis, 97, 103264. https://doi.org/10.1016/j.media.2024.103264

[4] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). *Attention is all you need* (arXiv:1706.03762).arXiv. https://doi.org/10.48550/arXiv.1706.03762

[5] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). *An image is worth 16x16 words: Transformers for image recognition at scale* (arXiv:2010.11929). arXiv. https://doi.org/10.48550/arXiv.2010.11929

[6] Sinha, G. R., Thakur, K., & Vyas, P. (2017). Research impact of astronomical image processing. *International Journal of Luminescence and Applications*, *7*(3–4), 503–506.