Emotional Enrichment of Arabic Captions for Art Images

Ahmad Sait
Deep Generative Modeling
KAUST
ahmed.sait@kaust.edu.sa

May 11, 2025

Abstract

Most image captioning systems produce literal, factual descriptions, which fail to capture the symbolic, emotional, and artistic qualities of visual art. This work presents an enhancement framework for Arabic captions, transforming plain modern-standard Arabic into poetic Fus'ha Arabic. Using the ArtELingo dataset, two pipelines were explored: direct stylistic rewriting via the ALLaM language model, and a translation-based pipeline using Hunayn, a fine-tuned poetic translator. Human and LLM-based evaluations strongly preferred ALLaM-generated captions for their grammatical fluency and expressive richness. These enhanced captions were used to train a BLIP-based vision-language model. Despite low lexical similarity scores under BLEU and METEOR, highlighting the limitations of such metrics for creative tasks, qualitative evaluations confirmed the superiority of the poetic captions. This framework offers a viable path toward culturally and emotionally rich image captioning in Arabic.

1 Introduction

Image captioning has historically focused on generating factual and objective descriptions of visual content: identifying people, objects, and actions with precision. While this approach suits natural photographs, it fails to capture the richness of artistic images, where emotion, symbolism, and cultural depth often carry more significance than literal elements.

Although recent multimodal models such as BLIP [1] are technically capable of generating stylistic or creative language, their outputs remain predominantly factual by default. This is largely due to the nature of the datasets they are

trained on, such as COCO [2] and Conceptual Captions [3], which prioritize literal descriptions, and the evaluation metrics used, which favor surface-level lexical overlap over creativity or interpretation.

This project addresses this limitation by building an enhanced Arabic image captioner that generates poetic and artistic descriptions of artworks using Fus'ha Arabic, the formal and literary register of the language. By leveraging the ArtELingo [4] dataset and fine-tuning on stylistic language, this work aims to close the gap between visual art and the language used to describe it. Code and processed datasets are available at:

Code: `github.com/AhmadSait/Emotional-Arabic-Captioning`

Processed data: `drive.google.com/...`

2 Related Work

ArtELingo is a dataset that has tackled the problem where image captions that past models were training on were factual and contained no artistry in its captions, especially for artistic images. Figure 1 shows the difference in emotional depth ArtELingo made sure it injected into its captions to stray away from being factual.

Figure 1: Factual vs ArtELingo caption of an artwork from WikiArt.

Not only did ArtELingo curate this dataset in English, they also did it for Arabic. However, the Arabic captions primarily reflect modern standard Arabic, rather than Fus'ha (pure and literary Arabic). Also the last sentence in

the above ArtELingo caption that translates to “Birds flying over a tree” is considered factual and does not have any artistic or emotional injection in it. To ensure comprehensive stylistic coverage, we work on building an enhanced image captioner that fully captions an image with a poetic and artistic Arabic description.

3 Methodology

To generate a stylistically richer, more poetic Arabic caption dataset, we followed a two-branch pipeline to enhance the original ArtELingo Arabic captions. The goal was to produce alternative versions of each caption using two different strategies, which were later evaluated in a user study to determine the preferred output.

3.1 Method 1: Direct Arabic Enhancement via ALLaM

In the first branch of the pipeline, the original Arabic captions from ArtELingo were directly enhanced using ALLaM [6], a large Arabic language model, which was prompted to improve each caption using the prompt:

> خذ الجملة المعطاة وأعد صياغتها باللغة العربية الفصحى بأسلوبٍ أكثر فصاحةً ورقيًّا، مع الحفاظ على نفس المعنى

This produced a new set of 50 Enhanced Arabic Captions written in a more expressive, literary style.

3.2 Method 2: Arabic to English to Enhanced Arabic via Hunayn

In the second branch, the original Arabic captions were first translated into English using ALLaM. These English captions were then stylistically enhanced using Hunayn [7], a poetic English-to-Arabic translator fine-tuned to produce Fus’ha Arabic enriched with classical vocabulary. Hunayn was fine-tuned end-to-end, updating all 75 million parameters across 10 epochs, with model checkpoints saved at each epoch. Evaluation showed that the second epoch’s checkpoint consistently yielded the most stylistically expressive and semantically coherent captions.

However, the model has displayed in multiple output generations why it could not be trusted; several output generations exhibited serious syntactic faults:

- فامرأة تضع المكياج على وجهها على وجه المكياج على نحو متفاوت، وتزيل المعاني، وتجعلها
- ولها باقة فيها ورد أبيض أحمر أخضر، وورد أخضر، وورد أبيض، وورد أحمر، وورد أخضر، وورد أخضر،
- فتجلس العجوز إلى جانب زوجها إلى جنب زوجها، وتمسك يديه، وتبصر بحزن شديد وكد عظيم،

After observing such errors, additional experiments were conducted to make sure that Hunayn was generating syntactically correct sentences, but at the cost of sacrificing innovation. Therefore, a larger variant of the Hunayn model was evaluated, a fine-tuned "Helsinki-NLP/opus-mt-tc-big-en-ar."

3.2.1 Training Big Hunayn

Linear probing was applied to a larger variant of the Helsinki model fine-tuned on the Hunayn dataset. The classification head of this model, containing 16 million parameters, was unfrozen and trained. After 7 epochs, where each epoch took 53 minutes, the model slightly caught on to the dataset. Here is a comparison between vanilla Helsinki's output without any fine-tuning (top) and the fine-tuned one (bottom):

- الوقت مورد ثمين لا يمكن استعادته. استخدم وقتك بحكمة
- الوقت مورد ثمين لا يمكن استعادته. اغتنم وقتك بحكمة

Instead of using استخدم, Big Hunayn used اغتنم which is a better word to use in this sentence.

A second set of 50 Enhanced Arabic Captions were generated via a translation-enhancement route.

3.3 User Study

To evaluate the effectiveness of the enhanced Arabic sentences generated by AL-LaM and Hunayn, multiple pre-processing steps were needed to form the surveys for 5 native Arabic speakers. After these pre-processing steps, a structured user study was conducted as shown in Figure 2.

The evaluation criteria that human evaluators used to evaluate each generated caption on a 1–5 scale consisted of:

- Grammatical Correctness: Sentence is well-formed in Fus'ha, with proper grammar, structure, and syntax.

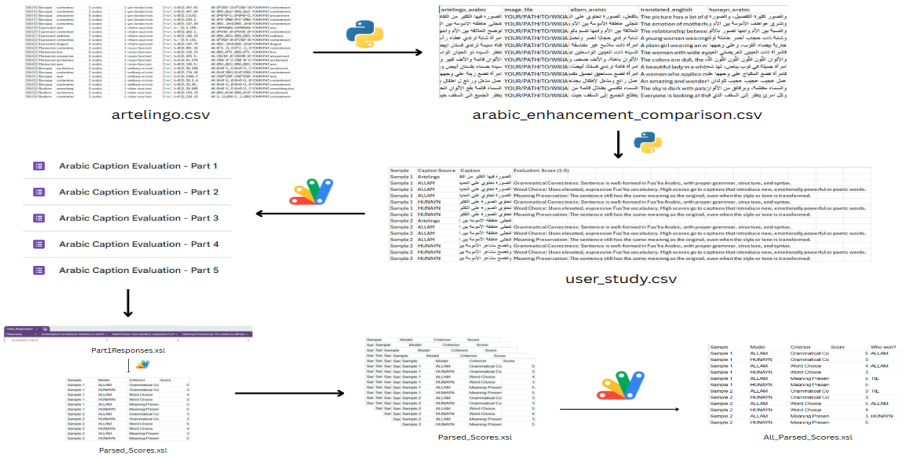

Figure 2: User study pipeline

- Word Choice: Uses elevated, expressive Fus'ha vocabulary. High scores go to captions that introduce new, emotionally powerful or poetic words.
- Meaning Preservation: The sentence still has the same meaning as the original, even when the style or tone is transformed.

100 samples - 50 samples from ALLaM's enhancement and 50 samples from Hunayn's enhancement have been evaluated by Arabic linguists. This criterion was carefully designed to specifically target which method has introduced new strong poetic words while still preserving meaning and correct syntax.

The results revealed that ALLaM was favored 65 times, Hunayn was favored 42 times, and were tied 43 times. These findings suggest that participants preferred the stylistic and grammatical quality of ALLaM's enhancements over those of Hunayn.

Since ALLaM was selected as the sole enhancement method for scaling, this chosen method was applied to 332,152 Arabic captions from the ArtELingo dataset, resulting in a fully enhanced, poetic, Arabic caption dataset.

3.4 Training BLIP

The captions enhanced by ALLaM, along with their corresponding image file paths from WikiArt, were used to train a BLIP model. In parallel, another BLIP model was trained using the original ArtELingo captions paired with the same images.

While preparing the 380,000+ image-caption pairs, a minor obstacle emerged:

some images in the WikiArt dataset were corrupted or missing. As a result, a filtering step was performed to discard any sample with an unreadable or absent image, reducing the dataset size from 386,411 to 332,152 samples (a 14

Both BLIP models were trained for 5 epochs. Training on the ArtELingo captions took 19 hours and 21 minutes, while training on the ALLaM-enhanced captions took 18 hours and 23 minutes. Figure 3 shows the training loss for both methods.

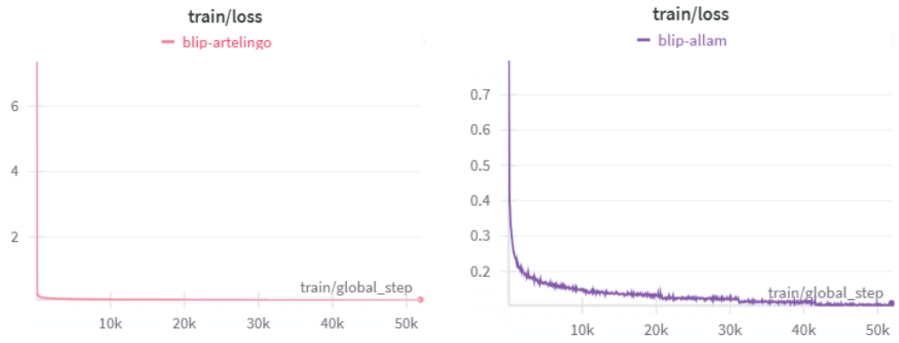

Figure 3: Recorded loss over 5 epochs for both BLIP training methods

4 Evaluation and Results

To evaluate the performance of the poetic Arabic image captioning system, both quantitative and qualitative assessments were conducted.

4.1 Quantitative Evaluation

Two standard metrics were employed to evaluate the 332,152 captions generated by each BLIP model: BLEU [8] and METEOR [9]. BLEU calculates a precision-based score reflecting n-gram overlap with the reference captions (the original ArtELingo dataset), while METEOR incorporates semantic similarity through stemming, synonym matching, and word order.

While ArtELingo captions achieve slightly higher scores under both metrics, the absolute values remain low. This result is expected in tasks focused on stylistic or poetic generation, where surface-level lexical overlap with references is inherently limited. Traditional metrics like BLEU and METEOR are designed for factual machine translation and fail to capture key qualities such as emotional tone, metaphorical expression, or rhetorical depth. Therefore, despite the numerical difference, these metrics offer limited insight into the effectiveness of

Metric	ALLaM	ArtELingo
BLEU	0.0025	**0.0096**
METEOR	0.0385	**0.0494**

Table 1: Quantitative evaluation comparing ALLaM and ArtELingo captions. Bolded values indicate higher performance.

stylistic captioning. In this context, qualitative evaluation remains the more reliable measure of success.

4.2 Qualitative Evaluation

Since quantitative evaluation was considered inadequate for assessing poetic captions, we next evaluate the captions qualitatively.

4.2.1 LLM-as-a-judge

To provide an objective evaluation, ChatGPT4o was employed as an automated judge to compare captions generated by BLIP models trained on either ALLaM-enhanced or original ArtELingo captions. Captions were evaluated based on the same criteria used in the human study: Grammatical Correctness, Word Choice, and Meaning Preservation, each rated on a 1–5 scale. For each image-caption pair, ChatGPT4o was prompted to assign scores and provide detailed justifications. The average score across the three criteria was computed, and the caption with the higher average received a point. The model with the most points was deemed the superior stylistic captioner.

Although ArtELingo contains over 29,000 Arabic image-caption pairs suitable for evaluation, a smaller, diverse subset was sampled to ensure stylistic variety and reduce token usage costs.

Figure 4 presents an example evaluated by ChatGPT4o. The caption generated by the ALLaM-trained BLIP model received scores of 5 for grammatical correctness, 5 for emotional word choice, and 4 for semantic meaning preservation.

From the 29,000 available ArtELingo samples, a stylistically diverse subset of art images was evaluated. Figure 4 shows one example where the ALLaM caption scored higher due to its expressive vocabulary ('حب', 'حنان') and coherent structure, while the ArtELingo version was judged repetitive and less emotionally rich.

Overall, ChatGPT4o favored ALLaM-enhanced captions 1,933 times versus 1,067 for ArtELingo, which is a 2:1 margin, highlighting ALLaM's superior stylistic and emotional expressiveness.

Figure 4: The Artist's Wife and Baby by David Bomberg

5 Conclusion

This work presents a stylistic captioning framework that transforms factual Arabic descriptions into poetic Fus'ha using two pipelines: direct enhancement via ALLaM and translation-based rewriting via Hunayn. Evaluations by both humans and LLMs showed consistent preference for ALLaM-enhanced captions due to their fluency and expressive richness.

Using ALLaM, 332,152 captions were generated and used to train a BLIP model. Despite low BLEU and METEOR scores, qualitative assessments confirmed the superiority of poetic captions in capturing artistic and emotional depth. This framework opens paths for culturally rich captioning and future work in style-guided generation, metaphor use, and classical Arabic forms.

Acknowledgements

We would like to thank the five individuals who took their time to evaluate the captions in the user study. We also greatly appreciate the mentorship and guidance gifted by Kilichbek Haydarov, whose feedback and constant support guided this research project to its success.

References

[1] Li, Junnan, et al. BLIP: Bootstrapped Language-Image Pretraining for Unified Vision-Language Understanding and Generation. ECCV 2022.

[2] Lin, Tsung-Yi, et al. Microsoft COCO: Common Objects in Context. ECCV 2014.

[3] Sharma, Piyush, et al. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-Text Dataset For Automatic Image Captioning. ACL 2018.

[4] Mohammed, Youssef, et al. ArtELingo: A Million Emotion Annotations of WikiArt with Emphasis on Diversity over Language and Culture. EMNLP 2022.

[5] WikiArt.org. (n.d.). Visual Art Encyclopedia. Retrieved from https://www.wikiart.org

[6] Bari, Saiful, et al. ALLaM: Large Language Models for Arabic and English.

[7] Almousa, Nasser, et al. Hunayn: Elevating Translation Beyond the Literal.

[8] Papineni, Kishore, et al. "BLEU: a method for automatic evaluation of machine translation." ACL 2002.

[9] Banerjee, Satanjeev, and Alon Lavie. "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments." ACL Workshop 2005.