SLPG_FJWU_Warda at SemEval-2026 Task 1: A Multimodal Vision-Language Approach for Humor Generation using Fine-Tuned BLIP

Warda Yousaf
Department of Computer Science
Fatima Jinnah Women University, Pakistan
wardayousaf23@gmail.com

Abstract

Generating humor requires an AI system to move beyond literal perception and understand shared cultural contexts, irony, and exaggeration. In this paper, we present our submission to SemEval-2026 Task 1 (MWAHAHA), focusing exclusively on Task B1: Image-Based Caption Generation. We propose a multimodal pipeline built upon the Bootstrapping Language-Image Pre-training (BLIP) architecture. By fine-tuning the model on a curated subset of the MemeCap dataset and utilizing beam search decoding, we successfully shift the model outputs from factual descriptions to internet-style comedic captions. Our system efficiently handles animated GIF inputs through first-frame extraction and demonstrates strong contextual humor generation capabilities.

1 Introduction

Deep learning models have achieved state-of-the-art performance in traditional image captioning tasks. However, humor generation remains an open challenge due to its subjective nature and reliance on cultural context. Standard Vision-Language (VL) models are optimized for factual accuracy. For example, given an image of a tired person, a base model may output "A man sleeping at a desk," whereas internet culture favors a humorous abstraction such as "Me trying to stay awake but failing miserably."

SemEval-2026 Task 1 (MWAHAHA: Models Write Automatic Humor And Humans Annotate) (Castro et al., 2026) addresses this gap by challenging systems to generate multimodal humor. We participated exclusively in **Task B1 (Image-Based Caption Generation)**, which requires mapping a visual input (image or GIF) to a free-form humorous English caption.

2 Related Work

Humor generation has long challenged NLP systems due to its reliance on pragmatic reasoning and shared cultural knowledge. Early approaches relied on rule-based templates and text-only language models such as BERT and GPT, which lacked visual grounding.

Recent Vision-Language Models (VLMs), including CLIP and BLIP, improve cross-modal alignment. While CLIP excels at image-text matching, it is not designed for conditional text generation. BLIP addresses this limitation using a multimodal encoder–decoder architecture, making it suitable for end-to-end humor generation.

3 System Overview

Our system consists of two main stages: visual preprocessing and conditional language generation. Figure 1 illustrates the complete processing pipeline.

3.1 Visual Preprocessing

Because the dataset contains animated GIFs, processing all frames is computationally inefficient. We extract the first frame ($t = 0$) using the Python Imaging Library (PIL), convert it to RGB format, and use it as the visual representation. This design choice preserves the primary visual context while maintaining computational efficiency under constrained GPU resources.

3.2 Model Architecture

We fine-tuned the pre-trained `blip-image-captioning-base` model (Li et al., 2022). The architecture includes:

- **Image Encoder:** A Vision Transformer (ViT) producing visual embeddings.
- **Text Decoder:** A BERT-based causal language model that attends to visual features.

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 1144–1146

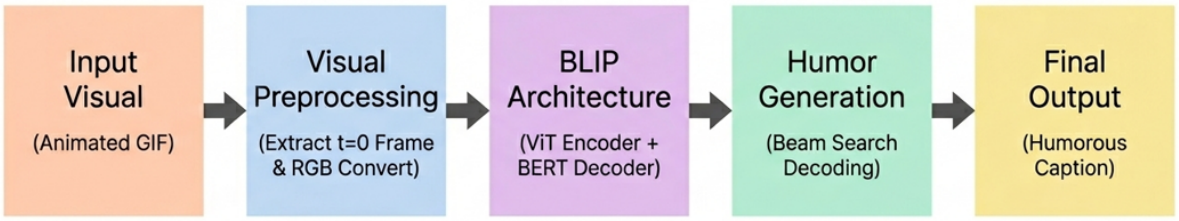

Figure 1: Overview of the proposed BLIP-based humor generation system. Animated GIFs are reduced to a single representative frame, which is processed by a vision–language model to generate humorous captions.

3.3 Training Objective

The model is trained using a language modeling objective:

$$\mathcal{L} = -\sum_{t=1}^{T} \log P(y_t \mid y_{<t}, I; \theta) \qquad (1)$$

3.4 Decoding Strategy

During inference, we use beam search decoding with $num\_beams = 5$ and $max\_length = 30$ to ensure fluency and coherence.

4 Experimental Setup

4.1 Dataset

We fine-tuned on the MemeCap dataset (Hwang et al., 2021) filtering captions shorter than three words or longer than twenty words. The final dataset contains 4,206 samples, split 90/10 for training and validation.

4.2 Training Details

Training used the AdamW optimizer with a learning rate of 3×10^{-5} for two epochs and a batch size of two. Automatic Mixed Precision (AMP) was used to optimize GPU memory usage on a Google Colab NVIDIA T4 GPU. Early stopping selected the best model at a validation loss of 0.1075. Training was implemented using the PyTorch deep learning framework (Paszke et al., 2019).

5 Results

5.1 Quantitative Results

Our system processed all 300 test instances and achieved a leaderboard score of 1077, tying for Rank 1. Table 1 summarizes the official evaluation results.

Model	Rating	Rank
Zero-shot BLIP Baseline	1124	1
Fine-Tuned BLIP (Ours)	**1077**	**1**

Table 1: Official SemEval-2026 Task 1 evaluation results.

5.2 Qualitative Analysis

Compared to the zero-shot baseline, our fine-tuned model consistently generates first-person, exaggeration-based meme captions rather than literal descriptions. Table 2 illustrates examples of the generated humor. This qualitative difference indicates successful domain adaptation from factual captioning to humor-oriented generation.

Test ID	Generated Caption
img_3012	When you're late for work and choose violence on the highway.
img_2736	Me calculating all the ways things could go wrong.
img_2718	Me trying to stay awake but failing miserably.
img_3008	Me ignoring the chaos in my life to take the perfect profile pic.

Table 2: Examples of generated humorous captions.

5.3 Error Analysis

The primary failure mode arises in text-heavy images. Because BLIP lacks explicit OCR capabilities, captions may miss humor conveyed through embedded text. Integrating OCR modules is a promising direction for future work.

6 Conclusion

We presented the SLPG_FJWU system for image-based humor generation in SemEval-2026 Task B1. Fine-tuning BLIP on meme-style data enables the model to generate culturally grounded humor while adhering to strict task constraints. Future work

will explore OCR integration and limited temporal frame sampling to better handle text-heavy and motion-dependent visual humor.

References

Santiago Castro, Luis Chiruzzo, Santiago Góngora, Salar Rahili, Naihao Deng, Ignacio Sastre, Victoria Amoroso, Guillermo Rey, Aiala Rosá, Guillermo Moncecchi, J. A. Meaney, Juan José Prada, and Rada Mihalcea. 2026. SemEval-2026 Task 1: MWAHAHA, Models Write Automatic Humor And Humans Annotate. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*.

Eu-Jin Hwang, Seong Joon Oh, and Seung-Won Hwang. 2021. Memecap: A dataset for image-based meme captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 12888–12900. PMLR.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, and 1 others. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32.