ABARUAH at SemEval-2026 Task 1: Leveraging High-Resolution VLMs and Reasoning LLMs for Multimodal Humor Generation

Arup Baruah
Dept. of CSE, Assam Don Bosco University
arup.baruah@gmail.com, arup.baruah@dbuniversity.ac.in

Abstract

This paper describes the systems developed for *"SemEval 2026 Task 1: Humor Generation"*. This shared task covered both unimodal text constraints and multimodal GIF-based humor generation. The proposed approach used a two-stage pipeline consisting of a Multimodal Grounding stage to extract semantic descriptions from GIFs and a Humor Synthesis stage to generate the final humorous output. The Qwen2-VL and Qwen3-8B models were used for these respective stages. The system achieved competitive Elo-like ratings of 1009, 973, and 914 for Subtasks A, B1, and B2, respectively, demonstrating its ability to address diverse humorous constraints. The system was ranked 4th in overall standings for Subtasks A and B1.

1 Introduction

Humor has been an important part of human interaction since distant past. It helps people connect with each other and lighten up situations. Humor is also complex because it often depends on culture, language, and creative thinking. Research in computational linguistics has tried to teach computers to generate jokes and humorous content. Although modern AI models have improved a lot, understanding and creating humor is still challenging for machines.

The *"SemEval 2026 Shared Task 1: Humor Generation"* (Castro et al., 2026) was organized to advance the state of the art in computational humor generation. The competition was structured into two primary subtasks. *Subtask A* focused on unimodal humor generation under specific constraints: *Word Inclusion*, which required the integration of two mandatory words without morphological variation, and *News Headline*, which tasked models with generating humorous punchlines inspired by a given news headline. *Subtask B* was a multimodal challenge involving GIF-based humor. This was further divided into two problem sets: *Subtask B1*, required the generation of a humorous caption for a given GIF, and *Subtask B2*, which required the completion of a partial sentence with humorous content based on the given GIF.

2 Related Work

A considerable amount of work has been performed in the domain of computational humor generation. Amin and Burghardt (2020) provides a comprehensive overview of the different approaches that has been used for computation humor generation.

The General Theory of Verbal Humor (GTVH) (Attardo and Raskin, 1991) suggested six hierarchical Knowledge Resources (KRs) that provide a structured framework to determine the strategies required for autonomous humor generation. The six parameters were script opposition, logical mechanism, situation, target, narrative strategy, and language. One of the earliest contributions to computational humor generation was the JAPE (Joke Analysis and Production Engine) system, developed to generate question-answer riddles (Binsted and Ritchie, 1994). This work developed a formal model of the underlying mechanisms in riddles. The rules were then implemented in a program. Valitutti et al. (2016) generated jokes by substituting words in short texts such as instant messages. The rules followed while substituting the words were that the new word should be orthographically or phonetically similar to the replaced word, the new word should be a taboo word, the new word should be placed at the end of the text or should be statistically consistent with the neighboring words. Yu et al. (2018) used a seq-2-seq model to generate puns. The model was able to generate without being explicitly trained on a pun dataset. To generate puns, a language model trained on a general corpus was used to generate sentences containing a target word.

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 3536–3543

Recent work have used Large Language Models (LLM) to generate humor. Vikhorev et al. (2024) fine-tuned a Llama-3.1-8B model to generate humor for English and Russian language. Zhong et al. (2024) mentions that Chain-of-Thought (CoT) while being effective for LLMs to solve logical reasoning tasks may not be conducive for creative problem-solving. Instead, Leap-of-Thought (LoT) may be more conducive for such tasks. The study developed a creative Leap-of-Thought (CLoT) method to enable LLMs to generate humor for a given text, image or both. Tikhonov and Shtykovskiy (2024) used the GPT-4 LLM to generate a generate a humor policy prompt by analyzing jokes from a dataset. The LLM was also used to generate associations given a topic. The associations and humor policy prompt was then utilized by the LLM to generate jokes for the seed topic.

3 Dataset

As outlined in Section 1, the shared task comprised three distinct problem statements: Subtask A, Subtask B1, and Subtask B2. For Subtask A, the evaluation dataset included 275 news headlines and 25 word pairs. A representative example of the word inclusion task involved the word pair *[drill, pumpkin]*, while a typical instance of the news headline task utilized the headline: *'How Guinness World Records began 7 decades ago'*. The datasets for Subtasks B1 and B2 each consisted of 300 GIF images each. The source URLs for the images were provided in the dataset.

A statistical analysis of the Subtask B1 corpus (Table 1) revealed a diverse range of visual formats. The GIFs had a mean aspect ratio of 1.32 ($\sigma = 94.55$), with a mean resolution of 449x368 pixels ($\sigma_w = 94.55$, $\sigma_h = 120.39$). Temporally, the dataset exhibited significant variance. The median duration was 2.62 seconds and the 95th percentile was 10.01 seconds. The average native frame rate was 15.14 FPS ($\sigma = 6.8$).

The Subtask B2 corpus exhibited a broader distribution of temporal and spatial features (Table 2). The mean resolution was 462x362 pixels ($\sigma_w = 81.94$, $\sigma_h = 110.16$). The temporal data revealed that the maximum duration was 969.6 seconds, with the 95th percentile being 13.59 seconds. The maximim total frames and native FPS was 808 and 50 respectively. The mean frame rate was 15.39 ($\sigma = 7.45$).

4 Methodology

This section describes the models, the system architecture, and the experimental setup used in this study.

4.1 Model

This study used a Vision Language Model (VLM) to understand the GIFs provided in the dataset and a Large Language Model (LLM) to generate the humor. The VLM and LLM used in this study are described below.

4.2 Vision Language Model

This study used Qwen2-VL (Wang et al., 2024) as the VLM to process the GIFs and generate descriptions for them. Unlike fixed resolution models, Qwen2-VL implements the Naive Dynamic Resolution feature which enables it to handle varying resolution and aspect ratio without resizing or cropping the images. As discussed in Section 3, the GIFs in the provided dataset were of varying resolution and aspect ratio [Mean resolution of 449x368 ($\sigma_w = 95$, $\sigma_h = 120$) and mean aspect ratio of $1.32(\sigma = 0.4)$ for subtask B1 dataset]. Thus, this feature of Qwen2-VL facilitated the processing of the GIFs without losing information. Qwen2-VL also introduced the Multimodal Rotary Position Embeddings (M-RoPE) feature. This feature decomposes positional information into temporal and spatial components, allowing the model to accurately resolve both the identity of objects within the GIF and their movement over time. This study used the Qwen2-VL-7B version of the model which consisted of a Vision Encoder with 675M parameters and an LLM with 7.6B parameters. The purpose of the Visual Encoder is to understand the GIF and generate the visual tokens. The LLM used the visual tokens and performed the reasoning task. Qwen2-VL's LLM was used to generate a description of the GIFs.

4.3 Large Language Model

This study used the Qwen3-8B instruction-tuned LLM (Yang et al., 2025) to perform the satirical reasoning and generate humorous captions for the GIFs. The model has a total of 8.2 billion parameters including 6.95 billion non-embedding parameters and was trained on a massive 36 trillion tokens. This model was specifically selected for its advanced reasoning capabilities and 'specialized thinking' modes, which enable it to perform com-

Statistic	Width	Height	Aspect Ratio	Total Frames	Native FPS	Total Duration (ms)
mean	449.02	368.14	1.32	49.65	15.14	3896.57
std	94.55	120.39	0.40	42.62	6.80	6151.44
min	156.00	132.00	0.34	2.00	0.00	0.00
50%	480.00	360.00	1.22	38.00	14.29	2620.00
95%	500.00	500.00	1.78	128.05	25	10013.5
max	1000.00	1000.00	2.40	433.00	33.33	83250.00

Table 1: Statistics of GIF Images in Subtask B1 Dataset

Statistic	Width	Height	Aspect Ratio	Total Frames	Native FPS	Total Duration (ms)
mean	462.67	362.23	1.38	55.39	15.39	7190.93
std	81.94	110.16	0.42	65.70	7.45	56034.68
min	160.00	118.00	0.40	2.00	0.00	0.00
50%	480.00	348.50	1.33	37.00	14.29	2475.00
95%	570.25	500.00	1.87	156.25	33.33	13589.00
max	800.00	720.00	2.94	808.00	50.00	969600.00

Table 2: Statistics of GIF Images in Subtask B2 Dataset

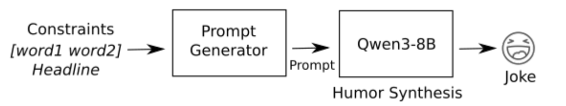

Figure 1: System Architecture for Subtask A

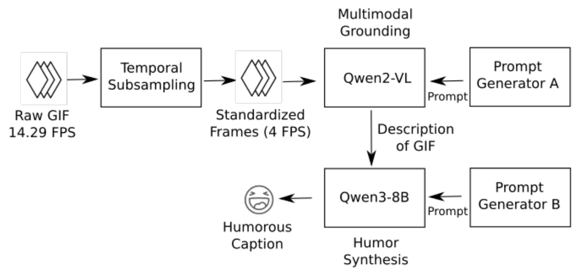

Figure 2: System Architecture for Subtask B

plex satirical reasoning required for high-quality humor. The choice of Qwen3-8B is justified by its strong empirical results in benchmarks relevant to humor generation. In the Creative Writing v3 task, Qwen3-8B achieves a score of 64.5 which is higher than the 52.8 reported for Llama-3.1-8B-Instruct (Yang et al., 2025). Gemma-3-12B-Instruct performed better in this task with a score of 79.9. However, Qwen3-8B has a smaller memory footprint and thus makes it ideal for resource-constrained environments. Qwen3-8B also outperformed Llama-3.1-8B-Instruct and occasionally the larger Gemma-3-12B-Instruct model in Math and Text Reasoning making it a robust choice for the complex logic required in humor synthesis.

4.4 System Architecture

Figure 1 shows the system architecture used for Subtask A. As can be seen, Qwen3-8B model was used to generate the joke. The constraints for generating the joke were embedded in a prompt. Listings 1 and 2 (see Appendix A) shows the structure of the prompt used for the word inclusion and the news headline tasks of Subtask A.

Figure 2 shows the system architecture used for subtasks B1 and B2. As discussed in Section 3, the GIFs had varying frame rates [Mean FPS of 15 (σ = 6.8) for subtask B1 dataset]. As such, the frame rate for the GIFs were standardized to 4.0 FPS before processing them. This conversion to 4.0 FPS was done by utilizing the native preprocessing pipeline of Qwen2-VL by setting the *fps* parameter to 4.0 within the multimodal input dictionary. A two stage pipeline was used to process the GIFs and generate humorous captions for them. The first stage, Multimodal Grounding, was used to perform the visual analysis of the GIFs and generate a description of them. The Qwen2-VL model was used in this stage. To optimally use the VRAM capabilities of the hardware (NVIDIA GeForce RTX 3080) the maximum pixel constraint was set to 768x28x28. As discussed in Section 3, this threshold accommodated 99% of the GIFs in the dataset. Thus, during the visual analysis phase, the Qwen2-VL Visual Encoder generated a maximum of 768 visual tokens. Qwen2-VL LLM then processed these tokens to generate a description of the GIFs. Listing 3 (see Appendix A) shows the structure of the prompt used in the Multimodal Grounding phase for Subtask B1 and B2.

The description of the GIFs produced by the

Multimodal Grounding stage was then provided as input to the Humor Synthesis stage. Based on the GIF descriptions, this stage generated the humorous caption for the GIF (for subtask B1) or complete a given text with humorous content (for subtask B2). The Qwen3-8B model was used in this stage. Listings 4 and 5 (see Appendix A) show the structure of the prompt used for Subtask B1 and B2 respectively. In both prompts, the attributes for initial state, final fail, expectation, and reality were programmatically populated using the descriptive output generated by Qwen2-VL during the preceding Multimodal Grounding phase.

4.5 Experimental Setup

The experiments in this study were conducted on an NVIDIA GeForce RTX 3080 GPU (16 GB VRAM). To address memory constraints, 4-bit QLoRA quantization (Dettmers et al., 2023) was used for model loading. The GIF frame rate was standardized to 4.0 FPS, with a maximum pixels limit for visual analysis was set to 768x28x28. For Subtask A, the temperature was set to 0.6 for the word-constrained task and 0.8 for the news headline task. A maximum token limit of 2048 tokens were used for the reasoning ('thinking') step. For Subtask B, a temperature of 0.01 was used for generating initial visual descriptions (Multimodal Grounding Stage), while a temperature of 0.8 was applied for the Humor Synthesis Stage. The maximum reasoning budget for the humor generation phase was set to 1024 tokens for Subtask B1 and 2048 tokens for Subtask B2.

5 Results and Discussion

An Elo-like rating system was used to evaluate the systems. In this scoring system, a pair of humor generated by two different systems was compared. The comparison was done by humans (Internet users and paid annotators from Prolific). In this scoring system, the score of the winning system increases and that of the losing system decreases. If a system beats a stronger opponent, the score increases significantly. All teams usually start with a baseline score of 1000. This type of rating system is used to calculate the relative skills of chess players.

The results declared for this shared task included the ratings for each system, the 95% confidence interval, and the rank. The scores obtained by proposed systems are listed in Tables 3 to 5. As can be seen from Table 3, the proposed system achieved an Elo-like rating of 1009 for Subtask A. The 95% Confidence Interval (CI) of [985, 1041] indicates a 95% certainty that the model's true performance level falls within this range. The proposed system secured the 4th rank in the overall standings, as the 95% CI of three other systems was higher than the model's. There were a total of 31 participants in this subtask. The baseline system emerged as the best performing model for this subtask, obtaining a rating of 1081 with a 95% CI of [1045, 1110].

System	Rank	Rating	95% CI
This study	4	1009	[985, 1041]
Baseline (Best System)	1	1081	[1045, 1110]

Table 3: Results for Subtask A

System	Rank	Rating	95% CI
This study	4	973	[938, 1018]
Baseline	1	1124	[1084, 1164]
Best System	1	1140	[1099, 1180]

Table 4: Results for Subtask B1

Table 4 shows that that proposed system for subtask B1 obtained a rating of 973 and a 95% CI of [938, 1018]. The model was ranked 4th in the overall ranking. There were 11 participants in the subtask. The best system obtained a rating of 1140 and a 95% CI of [1099, 1180].

Table 5 shows that the proposed system for subtask B2, obtained a rating of 914 and a 95% CI of [870, 950]. The model was ranked 7th in the overall standing. The best system in this subtask obtained a rating of 1065 and a 95% CI of [1032, 1102].

Listings 6 and 7 provide examples of some high quality generation and some sub-optimal generation for Subtask A Work Inclusion and News Headline tasks. As can be seen in the listings, the humor generated while including the words 'spray' and 'fridge' successfully captured a significant irony gap by subverting the expected outcome with a humorous reality. In contrast, the output generated while including the words 'move' and 'towel' lacks this subversion, resulting in a literal pun where the irony gap is absent. A similar trend is observed

System	Rank	Rating	95% CI
This study	7	914	[870, 950]
Baseline	1	1022	[991, 1060]
Best System	1	1065	[1032, 1102]

Table 5: Results for Subtask B2

in the News Headline task; here, Qwen3-8B performed very well in leveraging world knowledge to generate sophisticated satire for the 'Bush/Cheney' and 'Brazil/COP30' headlines. However, for the 'Ottawa goldfish' headline, the model seems to just restate the given headline without introducing any satirical irony.

An analysis of the GIF descriptions generated by Qwen2-VL reveals that the VLM occasionally failed to provide accurate semantic grounding. This prevented the Qwen3-8B model from generating high-quality humor during the humor generation stage. Furthermore, the explicit prompting for 'subtle physical cues' (such as sweating, trembling, or eye-rolling) appears to have introduced a hallucination bias. The VLM frequently suggested the presence of these reactions even when they were absent from the visual data, leading the humor generation model to base its satirical logic on non-existent physical states.

6 Conclusion

This paper presented a two-stage framework for the shared task "SemEval 2026 Task 1: Humor Generation". The study demonstrated the effectiveness of decoupling the task of humor generation into Visual Grounding via Qwen2-VL and Humorous Synthesis via Qwen3-8B. The proposed methodology achieved a competitive Elo-like ratings for Subtasn A and B1. The models obtained the 4th rank for these two subtasks. The performance shift from Subtask B1 (973) to B2 (914) suggests that sentence completion requires a deeper level of semantic alignment with visual context than captioning alone.

References

Miriam Amin and Manuel Burghardt. 2020. A survey on approaches to computational humor generation. In *Proceedings of the 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 29–41.

Salvatore Attardo and Victor Raskin. 1991. Script theory revis (it) ed: Joke similarity and joke representation model.

Kim Binsted and Graeme Ritchie. 1994. *An implemented model of punning riddles*. University of Edinburgh, Department of Artificial Intelligence.

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)*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115.

Alexey Tikhonov and Pavel Shtykovskiy. 2024. Humor mechanics: Advancing humor generation with multistep reasoning. *arXiv preprint arXiv:2405.07280*.

Alessandro Valitutti, Antoine Doucet, Jukka M Toivanen, and Hannu Toivonen. 2016. Computational generation and dissection of lexical replacement humor. *Natural Language Engineering*, 22(5):727–749.

Dmitry Vikhorev, Daria Galimzianova, Svetlana Gorovaia, Elizaveta Zhemchuzhina, and Ivan P Yamshchikov. 2024. Cleancomedy: Creating friendly humor through generative techniques. *arXiv preprint arXiv:2412.09203*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Zhiwei Yu, Jiwei Tan, and Xiaojun Wan. 2018. A neural approach to pun generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1650–1660.

Shanshan Zhong, Zhongzhan Huang, Shanghua Gao, Wushao Wen, Liang Lin, Marinka Zitnik, and Pan Zhou. 2024. Let's think outside the box: Exploring leap-of-thought in large language models with creative humor generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13246–13257.

A Appendix A: Prompt Templates and Sample Outputs

```
[SYSTEM INSTRUCTION]
Goal: Win the SemEval 2026 MWAHAHA Humor Generation competition.
Target: Top rank in human-evaluated 1-on-1 battles.
Strategy:
1. Strict Constraint: You MUST use the words [Word1] and [Word2] exactly as provided. Do not change
    tense or plurality.
2. Creativity: Avoid cliches. Use subversion, irony, and sharp observation.
3. Tone: Professional satirist. Sharp, cynical, and concise. (e.g., The Onion or Private Eye).
Format: Output ONLY the final joke.

[USER QUERY]
/think Create a joke: Current Task
Input Words: [drill, tomato]
Output Joke:
```

Listing 1: Prompt Template for Subtask A (Word Inclusion Task)

```
[SYSTEM INSTRUCTION]
Goal: Create a satirical punchline for a news headline.
Strategy:
1. Find the "Irony Gap" between the headline and reality. Use wit and sharp social commentary.
2. Find the most absurd angle. Do not explain the irony; just state it.
Tone: Sharp, cynical, and concise.
Format: Output ONLY the joke.

[USER QUERY]
/think Create a joke: Current Task
Headline: Carra: Liverpool must sign in January - even after summer spending spree
Joke:
```

Listing 2: Prompt Template for Subtask A (News Headline Task)

```
[INPUT CONFIGURATION]
Type: Video/GIF
File: img_3012.gif
FPS: 4.0
Max Pixels: 602,112

[TEXT INSTRUCTION]
Provide a highly detailed analysis for a comedy writer:
1. What is the character's initial state vs. their final fail?
2. Describe subtle physical cues (sweating, trembling, eye-rolling).
3. What is the specific 'expectation vs. reality' moment?
```

Listing 3: Prompt Template for Multimodal Grounding Phase of Subtask B1 and B2

```
[SYSTEM INSTRUCTION]
You are a cynical satire writer.
Your goal is to write a sharp, ironic caption for a GIF (STRICTLY under 20 words).
Use the 'Initial State vs. Final Fail' provided to craft a witty joke.

[USER QUERY]
/think Analyze the contrast between expectation and reality: Visual Analysis:

### Analysis for a Comedy Writer
1. Initial State vs. Final Fail
- Initial State: [This part is constructed based on the output generated by the VLM for this GIF] The
     character appears to be in a state of excitement or anticipation, possibly preparing for an
    event or activity. This is suggested by the character's focused expression and the presence of a
     steering wheel, indicating they might be driving or about to drive.
- Final Fail: [This part is constructed based on the output generated by the VLM for this GIF] The
    character's expression shifts to one of distress or disappointment, with sweat and tears
    indicating a sudden realization of failure or a negative outcome. This suggests a dramatic
    change from the initial state of excitement to a moment of failure or disappointment.

3. Expectation vs. Reality Moment
```

```
- Expectation: [This part is constructed based on the output generated by the VLM for this GIF] The
    character likely expected to have a successful or enjoyable experience, possibly related to
    driving or participating in an activity.
- Reality: [This part is constructed based on the output generated by the VLM for this GIF] The
    character's reaction suggests that the outcome was far from what was expected, leading to a
    moment of disappointment or failure. This expectation vs. reality moment is highlighted by the
    character's sudden change in expression and physical cues.

Task: Write a satirical caption (max 20 words).
```

Listing 4: Prompt Template for Humor Synthesis Phase of Subtask B1

```
[SYSTEM INSTRUCTION]
You are a witty satire writer.
You will be given a visual analysis of a GIF and the start of a caption.
Your task is to COMPLETE the caption in a way that is humorous and ironic.
STRICT RULE: The total completion must be under 20 words.

[USER QUERY]
/think Connect the prefix 'When your office prank goes too far and everyone is like ______' to the
    visual irony of the fail:

Visual Analysis:
### Analysis for a Comedy Writer
1. Initial State vs. Final Fail
- Initial State: [This part is constructed based on the output generated by the VLM for this GIF] The
     character is a man standing in front of a level 13 puzzle in a game. He appears to be focused
    and determined, possibly preparing to solve the puzzle.
- Final Fail: [This part is constructed based on the output generated by the VLM for this GIF] The
    character is a zombie, and he is running through a bathroom. He is holding a toothbrush and
    appears to be in a state of panic or confusion. He falls to the ground, and the scene ends with
    him lying on the floor.

2. Subtle Physical Cues
- Initial State: [This part is constructed based on the output generated by the VLM for this GIF] The
     man is standing upright, with his body slightly turned towards the puzzle. His facial
    expression is one of concentration, with his eyes focused on the puzzle. He is holding a
    toothbrush, which is a subtle cue that he might be preparing to solve the puzzle.
- Final Fail: [This part is constructed based on the output generated by the VLM for this GIF] The
    zombie is running with a toothbrush in his hand, which is a clear indication of his
    transformation. His body is in a state of disarray, with his limbs flailing and his face
    contorted in a state of panic. The fact that he is holding a toothbrush is a humorous contrast
    to his previous state.

3. Expectation vs. Reality Moment
- Expectation: [This part is constructed based on the output generated by the VLM for this GIF] The
    player is expected to solve the puzzle and progress to the next level.
- Reality: [This part is constructed based on the output generated by the VLM for this GIF] The
    player fails the puzzle and transforms into a zombie, which is a humorous twist.

Caption Start: When your office prank goes too far and everyone is like ______
Task: Complete the caption humorously.
```

Listing 5: Prompt Template for Humor Synthesis Phase of Subtask B2

```
--- SUB-OPTIMAL GENERATIONS ---

Words to include: move, towel
Humor generated: I told the towel to move, but it just dried up.

Words to include: shake, chair
Humor generated: The chair was so uncomfortable, I had to shake it out of my hands - turns out, it
    was a haunted seat with a shaky personality.

--- HIGH-QUALITY GENERATIONS ---

Words to include: blend, shoes
Humor generated: My shoes are a blend of comfort and style, but apparently, that's just a fancy way
    of saying they're both uncomfortable and unflattering - especially when the blend starts to rust
     and emit a strong odor of regret.
```

```
Words to include: spray, fridge
Humor generated: I sprayed my fridge with a new air freshener, but now it’s just a cloud of regret
    and expired milk.
```

Listing 6: Sample Outputs for Subtask A (Word Inclusion)

```
--- SUB-OPTIMAL GENERATIONS ---
Given headline: No place for Mostert distraction in South Africa team.
Humor generated: South Africa’s team finally found a use for Mostert: as the distraction in the
    jockstrap.

Given headline: ‘Mind-blowing’ population of feral goldfish in Ottawa pond to be euthanized.
Humor generated: Ottawa’s feral goldfish population so large, they’ve outgrown the pond-and the city’
    s budget for fish food.

--- HIGH-QUALITY GENERATIONS ---
Given headline: Dick Cheney funeral: George W. Bush delivers eulogy.
Humor generated: Because the only thing they ever agreed on was who to blame for the Iraq War’s ‘
    casualties’.

Given headline: Host Brazil pushes for progress on big issues at COP30 and some see cause for
    optimism.
Humor generated: Brazil hosts COP30 with a side of Amazonian bonfires-because progress is just a tree
     away.
```

Listing 7: Sample Outputs for Subtask A (News Headline)