國立臺灣大學管理學院資訊管理學研究所

碩士論文

Department of Information Management
College of Management
National Taiwan University

Master’s Thesis

可調適廣告迷因標題生成模型

Adaptable Advertising Meme Caption Generation Model

葉家妤

Jia-Yu Yeh

指導教授：李瑞庭 博士

Advisor: Anthony J. T. Lee, Ph.D.

中華民國 114 年 7 月

July 2025

doi:10.6342/NTU202501318

謝辭

兩年的研究生生涯即將畫下句點，至今還是不敢相信我成功讓迷因成為我從台大資管所畢業的基石，用最有趣的方式立下我人生中重要的里程碑之一。無論是在反覆調整並確認模型結果時，看到好笑迷因的救贖感，或是在碰到困難時，開始把各種迷因砸到與朋友的聊天室裡，迷因成功地為每個研究中所遭遇的困境添上了一絲笑意。除了迷因，我也很幸運的獲得了許多人的幫助與支持，才能一路笑著走過種種波折，完成這段在笑與淚之間反覆穿梭的研究生旅程。

首先，最感謝的是指導教授李瑞庭老師，沒有老師一開始的支持，我不會有以迷因作為論文題目的勇氣，也不會對論文中千奇百怪的複雜式子有如此深入的了解。也感謝老師耐心引導我掌握學術研究的方法與脈絡，在我迷失方向時，點出了我所忽略的細節與尚未發現的可能性，感謝老師這兩年來全心的陪伴與指引，讓我的研究生涯得以踏實而豐富結束。也感謝兩位口試委員吳怡瑾教授和戴敏育教授給予的肯定及寶貴的建議，讓我的研究更加充實與完整。

接著，要感謝風雨同舟的 Tony Lab 的成員，也就是 Tony Lab 三條魚中的另外兩條－芳瑜和青妤，相似的特殊作息讓我們總是能在神奇的時間點分享討論彼此碰上的問題與困境，用相同的步調與歡樂的氣氛一起在這片學術海洋裡乘浪而行。也要謝謝學長姐們在各方面的幫助，我也很慶幸自己能成爲 Tony Lab 的一員。同時也要感謝並肩同行的 CH Lab 成員和其他碩班的朋友們，一起為我的研究生日常增添色彩，還能解決研討課程結束後 Tony Lab 無法不透過猜拳選定午餐，但又猜不出結果的慘況。最後，要感謝我的媽媽，沒有她的支持我無法傾盡全力一心專注於我的學業，甚至在大年初一的晚上開車載我去實驗室解決突發狀況，總是在我需要的時候默默給予支持。謝謝媽媽一直以來的陪伴與包容，讓我能一步步走過這段不凡的旅程。

這段旅程比我預想的更熱鬧、更混亂，也更多笑聲，不只是因為研究與迷因相關，更因為身邊總有人陪我一起笑著跌倒、爬起來、再繼續前進。謝謝所有在這兩年裡陪伴與支持我的人，這場把迷因變成論文、把笑聲變成模組的瘋狂冒險才能走到這裡。如今，這場研究旅程走到了尾聲，那些關於迷因、模型與笑聲的記憶，也悄悄被 adapter 微調進我人生的主模型裡，準備好遷移到人生的下一個任務裡。願我們都能保有對生活的幽默感，在未來的每一個任務中都順利又平安。

葉家妤 謹識

於國立臺灣大學資訊管理研究所

中華民國 114 年 7 月

doi:10.6342/NTU202501318

論文摘要

論文題目：可調適廣告迷因標題生成模型

作者：葉家妤

指導教授：李瑞庭　博士

越來越多品牌利用迷因行銷，讓他們的品牌更貼近目標客戶，引起他們的共鳴。許多迷因標題生成模型利用特定的模板產生迷因標題，但這限制創作的自由度，無法產生具原創性且有影響力的迷因標題。因此，在本研究中，我們提出一個可調適廣告迷因標題生成模型 GAMC 幫某個品牌的貼文產生迷因標題，我們提出的模型包含五個模組，首先，我們透過視覺特徵提取模組和情感、情緒與幽默特徵擷取模組，從圖片中提取視覺與情緒相關特徵；接著，我們運用共注意力模組學習不同模態特徵間的關係；然後，我們利用大型語言模型生成迷因標題，並利用主資料集訓練模型以增加生成標題的幽默感，其中主資料集包含許多的迷因資料；最後，我們利用品牌資料與可適應模組微調已訓練好的模型，讓生成的迷因標題更加契合品牌形象。實驗結果顯示，我們提出的模型在幽默度、友善性及流暢度等評分指標上均優於比較模型。我們的廣告迷因標題生成模型，可幫助品牌展現其幽默感，提升品牌形象與曝光度，讓它們的貼文更具病毒式擴散能力。

關鍵字：迷因標題生成模型、大型語言模型、可調適模組、注意力機制

doi:10.6342/NTU202501318

Thesis Abstract

Adaptable Advertising Meme Caption Generation Model
By Jia-Yu Yeh
DEPARTMENT OF INFORMATION MANAGEMENT
NATIONAL TAIWAN UNIVERSITY
July 2025
ADVISOR: Anthony J. T. Lee, Ph.D.

Many companies have used meme marketing to make them more relatable and approachable to their target audience. Many meme caption generation models use custom meme templates to generate meme captions; however, they can only generate meme captions on the pre-trained classes (or topics). These constraints on creative freedom can significantly hinder the ability to produce original and impactful meme captions. Therefore, in this study, we propose an adaptable model to Generate Advertising Meme Captions, called GAMC, for the posts of a brand. The proposed model contains five modules: the visual feature extraction module, the emotion-sentiment-humor (ESH) module, the co-attention module, caption generation module, and adaptation module. First, we apply the visual feature extraction module to extract the visual features and the ESH module to derive the emotion, sentiment, and humor features from the photo of the input post. Next, we use the co-attention module to learn the inter-relationships between features of different modalities. Fourth, we employ the Large Language Model (LLM) to generate the meme caption for the input post in the caption generation module, and train the proposed model by the main dataset to increase the sense of humor of generated captions, where the main dataset contains a large number of meme captions. Finally, we adapt the trained model to the brand by using the adapters and the dataset collected from the brand's posts to fine-tune the trained model in the adaptation module. The experimental results show that the proposed model outperforms the compared models in terms of humor, benign, and fluency scores. Our model can help businesses reveal their humorous side, enhance their brand image, promote effective communication with their customers, and spark positive word-of-mouth.

Keywords: meme caption generation model, large language model, adapter, attention mechanism

doi:10.6342/NTU202501318

Table of Contents

doi:10.6342/NTU202501318

List of Figures

doi:10.6342/NTU202501318

List of Tables

doi:10.6342/NTU202501318

Chapter 1 Introduction

Humor is one of the greatest communication tools, many studies have shown the infectious ability in humor-induced communication (Eisend 2022, Weber and Quiring 2019). Meme marketing is a marketing strategy that uses humor to connect with more people. Businesses often use this strategy to enhance their image, promote effective customer communication, and spark positive word-of-mouth. Many companies have used meme marketing to make them more relatable and approachable to their target audience such as Zoom, Google, Netflix on X (formerly Twitter), Instagram, and also the fast-food industry, including Sonic Drive-In, and McDonald's. In fact, meme campaigns achieve nearly ten times more reach than standard marketing graphics, with 60% organic engagement on Facebook and Instagram, compared to only about 5% for traditional marketing graphics.[1] Another example of click-through rates, meme marketing campaigns have shown 14% higher than those for email marketing,[2] and more than 60% of consumers tend to purchase from a business using meme advertising[3].

Memes allow businesses to reveal their human side through humor and give them a free content marketing tool. Additionally, businesses can quickly adjust their strategy through immediate customers' feedback and responses to the meme. According to Malodia et al.'s experiment (2022), what makes the meme go viral that helps firms increase their brand recall and consumer brand engagement are content-related factors (i.e., humor, relevance, iconicity, and shareability), customer-related factors (i.e., escapism, social and content gratifications), and media-related factors (i.e., seeding and distribution). For example, Zoom once posted a viral meme that compared individuals dressed casually out for the wedding but dressed professionally on their first day work-from-home (WFH), which connects with the audience's personal WFH experiences, uses popular phrases WFH, makes it more comprehensibly, reminds people the humor

---

[1] Memes: A Digital Marketing Tool for Every Industry, https://www.forbes.com/sites/forbescommunicationscouncil/2018/08/10/memes-a-digital-marketing-tool-for-every-industry/?sh=5e4511ba2664#:~:text=According%20to%20Mittal,just%20no%20comparison.%E2%80%9D

[2] Memes Statistics by Country, Devices, Users, Industry and Trends, https://www.enterpriseappstoday.com/stats/memes-statistics.html#google_vignette:~:text=Meme%20campaigns%20achieve%2014%25%20higher%20click%2Dthrough%20rates%20(CTR)%20than%20email%20marketing.

[3] TOP MEME STATISTICS IN 2023, https://www.amraandelma.com/meme-statistics/#:~:text=STATISTICS%20(EDITORS%20CHOICE)-,Over%2060%25%20of%20people%20say%20they%20would%20be%20more%20likely%20to%20buy%20from%20a%20company%20that%20uses%20memes%20in%20their%20marketing,-The%20click%2Dthrough

doi:10.6342/NTU202501318

WFH scenario that they might dress formally only on top where the camera can capture and casually off camera, fulfills the content-related factors. Also, temporarily escaping from the misery of COVID-19, sharing the meme with their friends and co-workers to express their self-identity and seek validation from them, and reminding them of Zoom's central role in remote work settings, fits the customer-related factors. All the meme needs are the media-related factors, a perfect release time, a media type that can easily spread memes like X or Instagram, and influencers to share the meme to blow the hot air balloon up and high.

With memes, brand equity, brand exposure and corporate awareness can be improved more efficiently at a lower cost. Also, engagement on posts and social media users can receive direct and positive impact. Moreover, the effects of Internet memes on post engagements are larger with original memes than often seen meme templates (Yang and Hayashi 2021). Therefore, it is essential and desirable to provide a tool to assist businesses in creating their original viral memes to become a "fashion creator."

Many image captioning methods (J. Li et al. 2022, 2023, R. Li et al. 2023, L. Liu et al. 2023, Mokady et al. 2021, Zhang et al. 2023) have been proposed to generate the caption for a given image. However, these models are designed to annotate the content of an image (i.e. objects, relationships between objects, etc.), not to generate meme captions for business-related posts. The chemistry between the text and the image, along with originality, is the key of creating advertising meme captions that not only hit the mark with humor but also leave a lasting impression and arouse the desire to share. Also, some models (Peirson and Tolunay 2018, Vyalla and Udandarao 2020, Wang and Lee 2024) have been presented for generating meme captions. However, Dank Learning (Peirson and Tolunay 2018) and MemeCraft (Wang and Lee 2024) can only generate meme captions on the pre-specified subjects in existing meme templates. Although Memeify (Vyalla and Udandarao 2020) can produce meme captions with custom meme templates, it can only generate meme captions on the pre-trained classes (or topics). These constraints on creative freedom can significantly hinder the ability to produce original and impactful meme captions.

Therefore, in this study, we propose an adaptable model to Generate Advertising Meme Captions, called GAMC, for the posts of a brand. The proposed model contains five modules: the visual feature extraction module, the emotion-sentiment-humor (ESH) module, the co-attention module, caption generation module, and adaptation module.

doi:10.6342/NTU202501318

First, the visual feature extraction module extracts the visual features from the photo of the input post. Second, the ESH module employs MiniGPT4 (Zhu et al. 2023) to derive the emotion, sentiment, and humor features from the input photo. Third, the co-attention module learns the inter-relationships between visual and ESH features from two different modalities and integrates the learned features into the fused features. Fourth, the caption generation module uses the embedding layer of the Large Language Model (LLM) (Almazrouei et al. 2023) to derive the textual features from the caption of the input post, concatenates the fused and textual features, and employ the LLM to generate the meme caption for the input post. We train the proposed model by the main dataset, OxfordTVG-HIC (R. Li et al. 2023), to increase the sense of humor of generated captions, where the main dataset contains a large number of meme captions. Finally, the adaptation module adapts the trained model to the brand by using the dataset collected from the posts of the brand to fine-tune the trained model.

The contributions of this study are summarized as follows.

- We propose a novel adaptable advertising meme caption generation model that generates captions for brands and can be easily adapted to various brands or industries, overcoming limitations of template-based or class-restricted methods.
- The proposed model innovatively incorporates emotion, sentiment, and humor features extracted from input images using MiniGPT-4 to enhance caption creativity and emotional resonance.
- Our model aligns universal humor generation with brand-specific tones by employing a two-stage training strategy. It begins with pretraining on a main dataset to enhance creativity and humor perception, and continues with fine-tuning on brand-specific adapted datasets using LoRA, BitFit and funny score tuning adapters.
- Experimental results demonstrate that our model outperforms baselines in humor and fluency. Also, the results of human evaluation indicate that the generated captions have strong potential to enhance brand recall and deepen consumer brand engagement.
- Our proposed model can help businesses reveal their humorous side, enhance their brand image, promote effective communication with their customers, and spark positive word-of-mouth.

doi:10.6342/NTU202501318

- The brand elements and platform-specific cultural cues can be easily incorporated into the meme captions generated by our model, such as hashtags, to resonate with audiences and align with online humor trends.
- The generated meme captions can also support multi-platform event promotion by including event-related hashtags, effectively balancing humor with accurate promotional information. This makes the model particularly suitable for businesses new to meme culture that aim to strengthen their brand image.

The rest of this thesis is organized as follows. We survey the related work in Chapter 2. Next, we present our proposed model in Chapter 3 and conduct the performance evaluation in Chapter 4. Last, we discuss the concluding remarks and future work in Chapter 5.

doi:10.6342/NTU202501318

Chapter 2 Related Work

Our study is related to three research streams, namely meme caption generation, image captioning, and adapter. We will survey these three research streams in the following sections.

2.1 Meme Caption Generation

Dawkins (1976) introduced the concept of memes to denote the cultural units that spread virally. However, with the proliferation of the Internet, memes have become a significant cultural phenomenon in online communication. Moreover, Pech (2003) argued that balancing the use of memes can foster innovation, thereby impacting the company's profitability, which is called “meme marketing.”

Peirson and Tolunay (2018) proposed a model, called Dank Learning, which uses the Long Short-Term Memory model (LSTM) (Hochreiter and Schmidhuber 1997) to generate a humorous meme caption for a photo. Vyalla and Udandarao (2020) developed a model, called Memeify, by employing the Generative Pre-trained Transformer 2 (GPT-2) (Radford et al. 2019) to generate meme captions based on custom images or pre-trained classes. Wang and Lee (2024) presented a model, called MemeCraft, to produce humorous meme captions with the conditions given by users. They showed that MemeCraft outperformed Dank Learning in terms of authenticity and hilarity scores. However, Dank Learning and MemeCraft can only generate meme captions on the pre-specified subjects in existing meme templates. Although Memeify can produce meme captions with custom meme templates, it can only generate meme captions on the pre-trained classes (or topics). Instead of creating original memes, they can only reuse the pre-trained meme templates or classes. Therefore, they may not meet the advertising demand of businesses, since reusing existing meme templates is unlikely to become the “fashion creator.” Unlike these template-based models, our model generates meme captions entirely based on user-provided photos without using any templates or predefined formats.

doi:10.6342/NTU202501318

2.2 Image Captioning

Many image captioning methods (J. Li et al. 2022, 2023, Mokady et al. 2021) adopt the Contrastive Language-Image Pre-Training encoder (CLIP) (Radford et al. 2021) to generate the caption for an image. ClipCap (Mokady et al. 2021) applies the fine-tuned mapping network to produce a prefix and feeds it to a language model to generate the caption for an image. Blip (J. Li et al. 2022) proposes a multimodal mixture of encoder-decoder architecture to generate the caption for a given image. Blip-2 (J. Li et al. 2023) exploits the trainable Q-Former module (Zhao et al. 2022), containing three training processes, image-text matching, image-grounded text generation, and image-text contrastive learning, to bridge the gap between the CLIP and the LLM. Bootstrapping Interactive Image–Text Alignment model (BITA) (Yang et al. 2024) is similar to Blip-2 but employs only image–text contrastive learning (ITC) to constrain the Interactive Fourier Transformer (IFT) module. Despite using fewer training processes and parameters, it surpasses Blip-2 in performance. However, the CLIP-based models may struggle to capture the emotions and the relationships among objects in the images (Radford et al. 2021).

In addition, some transformer-based models adopt the encoder-decoder framework (L. Liu et al. 2023, Zhang et al. 2023) to generate the caption for an image. SwinCaption (L. Liu et al. 2023) employs Swin Transformer (Liu et al. 2021) as the encoder and LSTM as the decoder. The Deep Fusion Transformer (DFT) (Zhang et al. 2023) utilizes the Co-Transformer mechanism (Lu et al. 2019) to fuse the region features derived from the pre-trained CNNs, such as ResNet-101 (He et al. 2015) and VGG-16 (Simonyan and Zisserman 2015), and the grid features derived from R-CNN (Ren et al. 2017), then feeds them into the proposed Cross on Cross Attention module to capture the relation between context and photo features. Scene Graph Guiding Captioning (SGGC) (Chen et al. 2021) and ReFormer (Yang et al. 2022) extend the transformer framework by incorporating the scene graph to capture the relationships between objects, and enable their models to generate captions based on both semantic relationships and visual features. Comprehending and Ordering Semantics Networks (COS-Net) (Y. Li et al. 2022) utilizes the CLIP to retrieve the relevant sentences from the training sentence dataset and then uses the retrieved sentences to generate a caption. Vision transformer-based image captioning model (ViTCAP) (Fang et al. 2022) introduces a concept token network to predict the concept of a token to enrich

doi:10.6342/NTU202501318

the semantic information for the image captioning task. However, these models are designed to annotate the content of an image, not to generate meme captions for businesses' posts. The captions generated by these models focus on describing objects and object relationships on the image. The chemistry between the text and the image, along with originality, is the key of creating advertising meme captions that not only hit the mark with humor but also leave a lasting impression and arouse the desire to share.

2.3 Adapter

Many studies (Devlin et al. 2019, Liu et al. 2019) have shown the effectiveness of fine-tuning pre-trained language models on the NLP tasks, and so do the adapters on adapting language models (He et al. 2021, Yan et al. 2023). Hu et al. (2021) proposed the Low-Rank Adaptation (LoRA), an adapter implemented by a bottleneck architecture including a down-projection and an up-projection. The LoRA runs parallelly with the frozen linear layers of queries and values, and greatly reduces the number of trainable parameters for the pre-trained language model. Zaken et al. (2022) designed the BitFit adapter by freezing most of the network parameters and only fine-tuning the bias terms.

By integrating BitFit and LoRA into our model, we can fine-tune a small number of parameters in the pre-trained meme generator, and enhance transfer learning for better generalization to various brands and industries.

doi:10.6342/NTU202501318

Chapter 3 The Proposed Framework

In this chapter, we propose an adaptable model to Generate Advertising Meme Captions, called GAMC, for the posts of a brand in Figure 1, where the first four modules are displayed by bold boxes: the visual feature extraction module in the bottom left corner, the emotion-sentiment-humor (ESH) module at the bottom center, the co-attention module in the top left, and the caption generation module on the right; and the adaptation module is applied to the modules located within the orange-shaded area. The proposed model contains five modules: the visual feature extraction module, the emotion-sentiment-humor (ESH) module, the co-attention module, caption generation module, and adaptation module. First, we extract the visual features from the photo of the input post in the visual feature extraction module. Second, we employ MiniGPT4 (Zhu et al. 2023) to derive the emotion, sentiment, and humor features from the input photo in the ESH module. Third, we use the co-attention module to learn the inter-relationships between visual and ESH features from two different modalities, and integrate the learned features into the fused features. Fourth, we exploit the embedding layer of the Large Language Model (LLM) (Almazrouei et al. 2023) to derive the textual features from the caption of the input post, concatenate the fused and textual features, and employ the LLM to generate the meme caption for the input post in the caption generation module. We train the proposed model by using the main dataset, OxfordTVG-HIC dataset (R. Li et al. 2023), to learn how to generate humorous captions, where the main dataset contains a large number of meme captions. The OxfordTVG-HIC dataset features a wide range of emotional and semantic diversity resulting in out-of-context examples that are particularly ideal for generating humorous captions. Finally, we adapt the trained model to the brand by using the dataset collected from the posts of the brand to fine-tune the trained model in the adaptation module. In addition, we fine-tune the model with funny score to reinforce the sense of humor of generated captions.

doi:10.6342/NTU202501318

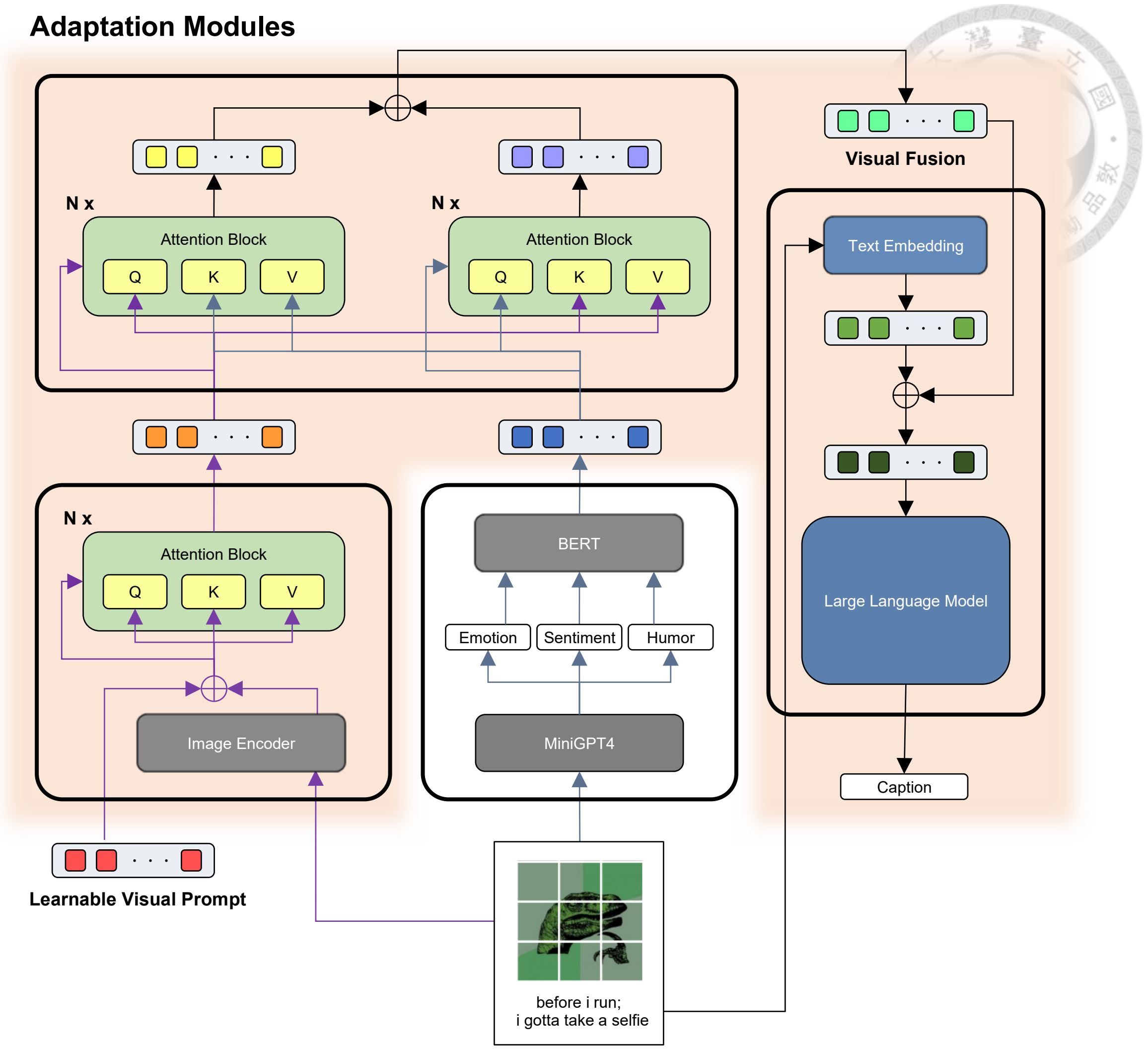


Figure 1. The GAMC Model

3.1 Visual Feature Extraction Module

Swin Transformer (Ze Liu et al. 2022) has been successfully applied to many vision tasks, including semantic segmentation (Hatamizadeh et al. 2022, He et al. 2022), object detection (Giroux et al. 2023, Jung et al. 2022, Z. Liu et al. 2022), facial emotion recognition (Bousaid et al. 2022, Han et al. 2023, Xie and Zhao 2023), and emotion extraction from body posture and gestures (Ninh et al. 2023). We apply the image encoder, Swin Transformer, to extract visual features from the photo of the input post by Eq. (1), where $SWIN$ denotes Swin Transformer, P_0 denotes the photo of the input post. In addition, we initialize a learnable visual prompt as the input, where $LVP \in \mathbb{R}^{n_p \times d}$ denotes the learnable visual prompt, n_p denotes the number of patches, d denotes the dimensionality of the feature extracted from each patch in Swin Transformer, and $P_1 \in \mathbb{R}^{2n_p \times d}$ denotes the visual features. The learnable visual

doi:10.6342/NTU202501318

prompt is considered as the model parameters used to extract the most informative features of the photo.

$$P_1 = Concat(SWIN(P_0), LVP) \quad (1)$$

Next, we employ the attention block to consider the relationships among the visual features. The attention block contains the multi-head attention mechanism (Vaswani et al. 2023), followed by a residual network, a layer normalization layer, a two-layer feed-forward network and a residual network. Specifically, we apply the multi-head attention mechanism (Vaswani et al. 2023) to the extracted visual features in Eqs. (2)-(7), where $P_2 \in \mathbb{R}^{2n_p \times d}$ denotes the visual features, output by the multi-head attention mechanism, $LayerNorm$ denotes the layer normalization layer, $Attention$ denotes the self-attention mechanism, $Q_1 \in \mathbb{R}^{d \times d_k}, K_1 \in \mathbb{R}^{d \times d_k}$ and $V_1 \in \mathbb{R}^{d \times d_k}$ respectively denote the query, key and value matrices in the multi-head attention mechanism, $W_i^Q \in \mathbb{R}^{d \times d_k}$, $W_K \in \mathbb{R}^{d \times d_k}$ and $W_i^V \in \mathbb{R}^{d \times d_k}$ are the learnable parameter matrices, $Q_2 \in \mathbb{R}^{n_v \times d_k}$, $K_2 \in \mathbb{R}^{n_v \times d_k}$ and $V_2 \in \mathbb{R}^{n_v \times d_v}$ respectively denote the query, key and value matrices in the self-attention mechanism, ${K_1}^T$ denotes the transpose of K_1, n_v denotes the number of visual features, $d_k = n_v/h$, $d_v = n_v/h$ denotes the dimensionality of the output of each head, h denotes the number of heads, and $W_Q \in \mathbb{R}^{d \times d_k}$, $W_K \in \mathbb{R}^{d \times d_k}$, $W_V \in \mathbb{R}^{d \times d_v}$ are the learnable parameter matrices. We also implement residual connection and layer normalization in our module. To more specific, we apply the multi-head attention mechanism with layer normalization in the query matrix to the visual features in Eqs. (2)-(4), and the residual connection with layer normalization and two-layer feed-forward network in Eqs. (5)-(7). The multi-head attention mechanism in Eqs. (3)-(4) computes the relevance of each visual feature in the input sequence to every other visual feature, and allows the model prioritize different elements in the input visual features and the learnable visual prompt, which helps capture relationships among the features extracted from the photo and the prompt. As for the two-layer feed-forward network, the residual connection and the layer normalization in Eqs. (5)-(7), they apply non-linear transformations to the output of the attention mechanism, which enables the model to learn richer representations. $P_3 \in \mathbb{R}^{2n_p \times d}$ denotes the visual features after the residual connection and layer normalization, $P_4 \in \mathbb{R}^{n_p \times d}$ denotes the attended visual features, FFN denotes a two-layer feed-forward network, $W_1 \in \mathbb{R}^{n_v \times d_f}$ and $W_2 \in \mathbb{R}^{d_f \times n_v}$ are the learnable parameter matrices, d_f denotes the dimension of the hidden layer in the two-layer

doi:10.6342/NTU202501318

feed-forward network, and $b_1 \in \mathbb{R}^{d_f}$ and $b_2 \in \mathbb{R}^{n_v}$ are the biases. Note that the visual feature extraction module adopts the design based on ClipCap (Mokady et al. 2021), extracting only the visual prompt in P_4 for caption generation.

$$P_2 = MultiHead(LayerNorm(P_1), P_1, P_1) \tag{2}$$

$$MultiHead(Q_1, K_1, V_1) = Concat(head_1, \ldots, head_h)W^O,$$
$$\text{where } head_i = Attention\left(Q_1 W_i^Q, K_1 W_i^K, V_1 W_i^V\right) \tag{3}$$

$$Attention(Q_2, K_2, V_2) = Softmax\left(\frac{Q_2 {K_2}^T}{\sqrt{d_k}}\right) V_2 \tag{4}$$

$$P_3 = LayerNorm(P_1 + P_2) \tag{5}$$

$$P_4 = FFN(P_3) + (P_1 + P_2) \tag{6}$$

$$FFN(x) = max(0, max(0, xW_1 + b_1)\, W_2 + b_2) \tag{7}$$

3.2 ESH Module

Meme is an emotion-related culture, most of the meme can be recreate with different kinds of stories, but with the same feeling and intension given from the photo. Thus, to enrich the visual features with useful information for meme caption generation, we derive the emotion, sentiment, humor descriptions from the photo by MiniGPT-4 (Zhu et al. 2023).

Next, we apply the pre-trained Bidirectional Encoder Representations from Transformers (BERT) model (Devlin et al. 2019) to extract the ESH features in Eqs. (8)-(10), where T_E, T_S and T_H respectively denote the emotion, sentiment and humor descriptions derived by MiniGPT-4, $P_E \in \mathbb{R}^{n_{ESH} \times d}, P_S \in \mathbb{R}^{n_{ESH} \times d}$ and $P_H \in \mathbb{R}^{n_{ESH} \times d}$ denote the extracted textual features, n_{ESH} denotes the numbers of tokens in the emotion, sentiment and humor descriptions, and d denotes the dimensionality of the feature extracted from each token, $P_{ESH} \in \mathbb{R}^{n_{ESH} \times d}$ denotes the ESH features, and $W^{ESH} \in \mathbb{R}^{3n_{ESH} \times n_{ESH}}$ denotes the learnable parameter matrix.

$$P_E = BERT(T_E) \tag{8}$$

$$P_S = BERT(T_S) \tag{9}$$

$$P_H = BERT(T_H) \tag{10}$$

$$P_{ESH} = Concat(P_E, P_S, P_H)W^{ESH} \tag{11}$$

doi:10.6342/NTU202501318

3.3 Co-Attention Module

Also, to consider the inter-relationships between the visual features and ESH features, we apply Co-Attention Transformer (Lu et al. 2019) to the attended visual features in Eqs. (9)-(15), where $P_7 \in \mathbb{R}^{n_p \times d}$ denotes the visual-attending-ESH features, output by Co-Attention Transformer, $P_{10} \in \mathbb{R}^{n_p \times d}$ denotes the ESH-attending-visual features, output by Co-Attention Transformer, $MultiHead$ denotes the multi-head attention mechanism (Vaswani et al. 2023), $P_{11} \in \mathbb{R}^{n_p \times d}$ denotes the integrated visual and ESH features, and $W^{CO} \in \mathbb{R}^{2n_p * n_p}$ denotes the learnable parameter matrix. Co-Attention Transformer comprises a multi-head attention mechanism in Eqs. (9) and (12) and a feed-forward network, each augmented by residual connection and layer normalization in Eqs. (10)-(11) and Eqs. (13)-(14). The visual-attending-ESH features are derived by generating the query from the attended visual features, and the key and value from the ESH features. On the other hand, the ESH-attending-visual features are derived by generating the query from the ESH features, and the key and value from the attended visual features.

Using both visual-attending-ESH features and ESH-attending-visual features in the co-attention module allows the model to capture the complementary information from both modalities. By visual-attending-ESH features, the model can attend to the parts of a description that best describe a particular region of the photo, enhancing the alignment between visual content and its ESH descriptions. On the other hand, ESH-attending-visual features can identify visual elements in the photo that correspond to specific words or phrases in the ESH descriptions. These approaches lead the model to a deeper, more accurate understanding of the relationships between the photo and ESH descriptions, enhancing the model's performance in meme caption generation.

$$P_5 = MultiHead(LayerNorm(P_4), P_{ESH}, P_{ESH}) \tag{9}$$

$$P_6 = LayerNorm(P_4 + P_5) \tag{10}$$

$$P_7 = FFN(P_6) + (P_4 + P_5) \tag{11}$$

$$P_8 = MultiHead(LayerNorm(P_{ESH}), P_4, P_4) \tag{12}$$

$$P_9 = LayerNorm(P_{ESH} + P_8) \tag{13}$$

$$P_{10} = FFN(P_9) + (P_{ESH} + P_8) \tag{14}$$

$$P_{11} = Concat(P_7, P_{10})W^{CO} \tag{15}$$

doi:10.6342/NTU202501318

3.4 Caption Generation Module

Next, we employ the Large Language Model (LLM) (Almazrouei et al. 2023) to generate the meme caption for the input post by concatenating the integrated visual and ESH features and the textual features extracted from the caption of the input post, and passing it the LLM, so that the LLM can generate the meme caption base on the world-knowledge from the origin language model and the caption of the input post. Specifically, we transform the caption of the input post into the LLM understandable features through the embedding layer of the LLM in Eq. (16), where T_0 denotes the caption of the input post, and $T_1 \in \mathbb{R}^{n_t \times d}$ denotes the textual features extracted from the embedding layer of the LLM. By concatenating the textual features and the integrated visual and ESH features, we sent it through the LLM to generate a meme caption in Eq. (17), where O denotes the generated meme caption, and LLM denotes the Large Language Model.

$$T_1 = Embed(T_0) \tag{16}$$

$$O = LLM\big(Concat(P_{11}, T_1)\big) \tag{17}$$

The loss function of the caption generation is defined in Eq. (18), where $LogSoftmax$ denotes a Softmax followed by a logarithm, and $NLLLoss$ denotes the negative log likelihood loss.

$$\mathcal{L}_c = NLLLoss(LogSoftmax(O), T_0) \tag{18}$$

3.5 Adaptation Module

Many studies (He et al. 2021, Hu et al. 2021, Zaken et al. 2022, W. Liu et al. 2023, Yan et al. 2023) show that fine-tuning with a small subset of the parameters in the model is an efficient adaptation strategy, and retains high model quality on transfer learning models. To fit the proposed model to the brand, we integrate three adapters into our model, namely Low-Rank Approximation (LoRA) (Hu et al. 2021), BitFit (Zaken et al. 2022), and the funny score tuning, as shown in Figure 2.

doi:10.6342/NTU202501318

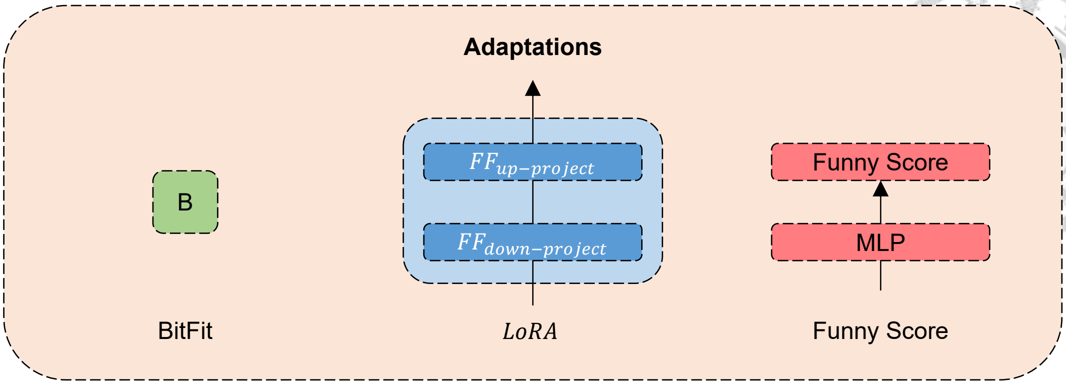


Figure 2. Three Adapters

3.5.1 Low-Rank Approximation and BitFit

The Low-Rank Approximation (LoRA) adapter, denoted by the blue boxes in the adapted attention block, as shown in Figure 3, comprises a down-projection and an up-projection in Eq. (19) and the multi-head attention network in Eq. (20), where W^L_{down} denotes the learnable parameter matrix of the down-projection, W^L_{up} denotes the learnable parameter matrix of the up-projection, and b_A denotes the bias.

The BitFit adapter is a localized and fast fine-tuning of the pre-trained transformers without additional modules. The BitFit, denoted by the green boxes in Figure 3, freezes the original weights and fine-tunes the biases in the adaptation.

The LoRA corporates with the BitFit to fine-tine the model. Specifically, in the attention mechanism in Eqs. (3) and (4), we replace the query $Q_1W_i^Q$ with $Q_1W_i^Q + LoRA(Q_1) + b_i^Q$, the key $K_1W_i^K$ with $K_1W_i^K + b_i^K$, and the value $V_1W_i^V$ with $V_1W_i^V + LoRA(V_1) + b_i^V$ in Eq. (20), where b_i^Q, b_i^K, and b_i^V are the adaptive biases. We freeze the components marked red and fine-tune the components marked blue or green during the adaptation in Eqs. (20)-(22). Both LoRA and BitFit adapters are applied within the attention block in Figure 1. The adapters enable efficient adaptation of the attention mechanism while minimizing the number of trainable parameters.

$$LoRA(x) = x\Delta W = (xW^L_{down})W^L_{up} \tag{19}$$

$$\begin{aligned} head_i = Attention(&Q_1W_i^Q + LoRA(Q_1) + b_i^Q, K_1W_i^K + b_i^K, \\ &V_1W_i^V + LoRA(V_1) + b_i^V) \end{aligned} \tag{20}$$

$$Attention(Q_2, K_2, V_2) = Softmax\left(\frac{Q_2K_2{}^T}{\sqrt{d_k}} + b_A\right)V_2 \tag{21}$$

$$FFN(x) = max(0, max(0, xW_1 + b_1)\,W_2 + b_2) \tag{22}$$

doi:10.6342/NTU202501318

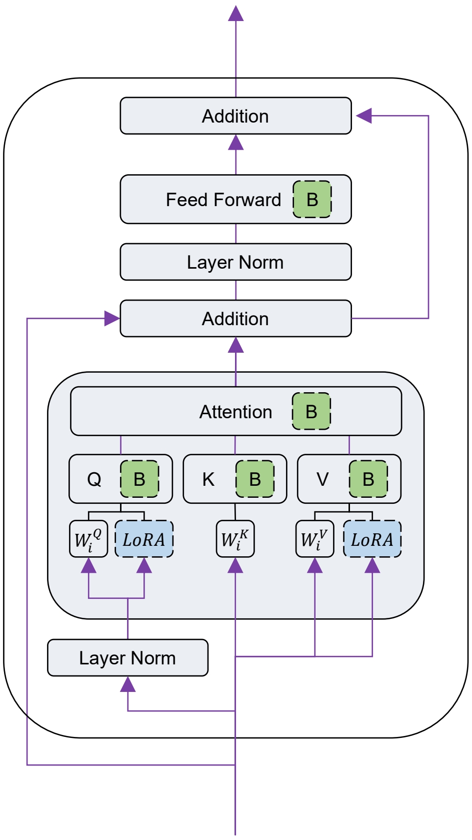


Figure 3. The Adapted Attention Block

3.5.2 Funny Score Tuning

The adapted dataset is collected from the posts of the brand, which contains fewer humorous captions than the main dataset, the OxfordTVG-HIC dataset. To address the lack of humor and the reduced parameters in adaptation, we apply the funny score tuning adapter to enhance the model's ability to recognize and generate humorous captions. We take the concatenation of integrated visual and ESH features, and textual features as input through a two-layer multilayer perceptron network in Eq. (23), where $W_{F1} \in \mathbb{R}^{d \times 1}$ and $W_{F2} \in \mathbb{R}^{(n_v+n_t) \times 1}$ are the learnable parameter matrices, d_f denotes the dimension of the hidden layer in the two-layer feed-forward network, $b_{F1} \in \mathbb{R}$ and $b_{F2} \in \mathbb{R}$ are the biases, and S is the sigmoid function to predict the funny score. We use the binary cross entropy loss on the funny score tuning in Eq. (24), where FC_0 denotes the ground-truth funny score. Finally, we combine the two loss functions in Eq. (25), where α is the hyperparameter. The funny score tuning adapter is integrated into the caption generation module, as illustrated in Figure 4, where it guides the model toward producing humor-aware captions by supervising the generation process with an additional objective.

doi:10.6342/NTU202501318

$$FC = S(max(0, Concat(P_{11}, T_1)W_{F1} + b_{F1})\, W_{F2} + b_{F2}) \quad (23)$$

$$\mathcal{L}_f = \mathcal{L}(FC, FC_0)$$

$$= \frac{1}{M}\sum_{m=1}^{M} -\left[FC_{0,m} \cdot log(FC_m) + (1 - FC_{0,m}) \cdot log(1 - FC_m)\right] \quad (24)$$

$$\mathcal{L} = \mathcal{L}_c + \alpha\mathcal{L}_f \quad (25)$$

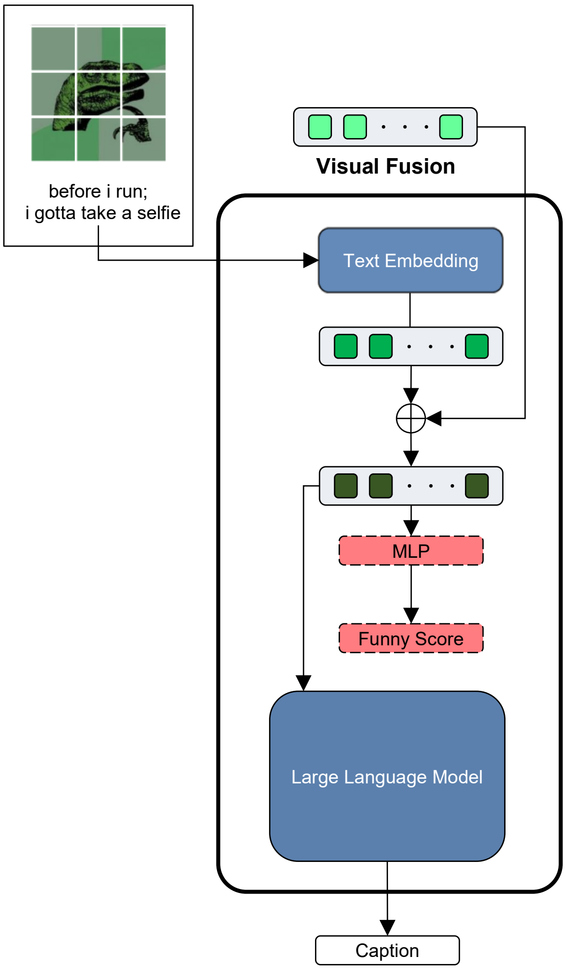


Figure 4. Funny Score Tuning in Caption Generation Module

The training process for the proposed model begins with the OxfordTVG-HIC dataset, ensuring that it captures the broad understanding of humor from the visual-textual features. Then, the proposed model adapts with a brand-specific dataset. During the adaptation phase, most of the model's parameters are frozen, and only the parameters of the adapters are fine-tuned. This approach ensures that the generator preserves the knowledge acquired from the OxfordTVG-HIC dataset and adjusts the model to the specific domain, to achieve optimal performance on generating advertising meme captions for a specific brand.

doi:10.6342/NTU202501318

Chapter 4 Experimental Results

We evaluate the performance of our framework in the following sections. We present an overview of the dataset and the evaluation metrics in Section 4.1, performance evaluation in Section 4.2, ablation study in Section 4.3, human evaluation in Section 4.4, and the meme caption examples in Section 4.5.

4.1 Dataset and Evaluation Metrics

We use the OxfordTVG-HIC dataset (R. Li et al. 2023) as the main dataset, and the posts collected from the Sonic Drive-In and McDonald's as the adapted datasets. The OxfordTVG-HIC dataset[4] contains 3,265,261 photo-caption pairs with 116,649 unique images. Sonic Drive-In and McDonald's adapted datasets are collected from Instagram, having 1,522 unique photo-caption pairs in the Sonic Drive-In adapted dataset[5] and 787 unique photo-caption pairs in the McDonald's adapted dataset[6], where each photo-caption pair is classified as humorous or not by the annotators.

Following (R. Li et al. 2023), we use four metrics namely, humor score, benign score (Muennighoff 2020), fluency score (Damodaran 1989) and diversity score (Radford et al. 2021), to evaluate the performance.

- **Humor score:**

  Base on the humor evaluation model (R. Li et al. 2023), we develop a refined version with several key improvements. The humor evaluation model (R. Li et al. 2023) utilizes ResNet50 (He et al. 2015) as image encoder, extracting features from its final layer, GPT-2 (Radford et al. 2019) serves as the text encoder, and a linear classification layer to predict whether the input post is humorous or not. Since this model cannot effectively predict humorous posts collected from Instagram, it is not suitable for humor score evaluation. To resolve such a problem, we develop a humor evaluation model, called HumorEva, by extracting features from the penultimate layer of ResNet50 which provides more refined features, and employing the contrastive learning module to improve the model performance.

---

4 The OxfordTVG-HIC dataset, https://drive.google.com/drive/folders/1BDuUcMeaWrFD8TwgHLhFPkuAwmoHaVNQ

5 Instagram - SONIC Drive-In (@sonicdrivein), https://www.instagram.com/sonicdrivein/

6 Instagram - McDonald's Switzerland (@mcdonalds_switzerland), https://www.instagram.com/mcdonalds_switzerland/

doi:10.6342/NTU202501318

Our humor evaluation model is trained with 3,800 photo-caption pairs, containing 1,900 photo-caption pairs from the OxfordTVG-HIC dataset, 1,300 photo-caption pairs from the Sonic Drive-In adapted dataset, and 600 photo-caption pairs from the McDonald's adapted dataset as shown in Table 1. It is tested with 475 photo-caption pairs, containing 325 photo-caption pairs from the Sonic Drive-In adapted dataset, and 150 photo-caption pairs from the McDonald's adapted dataset. Thus, our model can effectively learn the caption structures of the adapted datasets while simultaneously capturing the widest range of humor styles. We use the focal loss to deal with imbalanced data distribution in the dataset and the InfoNCE loss for the contrastive learning. Our humor evaluation model achieves the accuracy of 74.8% and the F1 score of 0.721 in the Sonic Drive-In and the accuracy of 94% and the F1 score of 0.943 in the McDonald's, and outperforms the origin humor evaluation model presented in R. Li et al. (2023), as shown in Tables 2 and 3.

Table 1. Data Composition of the Humor Evaluation Model

	OxfordTVG-HIC	Sonic Drive-In	McDonald's	Total
Train data	1,900	1,300	600	3,800
Test data	-	325	150	475

Table 2. Performance of the Humor Evaluation Model in Sonic Drive-In Dataset

	R. Li et al.	HumorEva
Accuracy	0.283	**0.748**
Precision	0.576	**0.749**
Recall	0.283	**0.748**
F1 score	0.219	**0.721**

Table 3. Performance of the Humor Evaluation Model in McDonald's Dataset

	R. Li et al.	HumorEva
Accuracy	0.407	**0.940**
Precision	0.681	**0.949**
Recall	0.407	**0.940**
F1 score	0.361	**0.943**

doi:10.6342/NTU202501318

- **Benign score:**

  Based on the benign violation theory (McGraw and Warren 2010), effective humor requires a high level of benign intent while avoiding offensive content to the audience. We measure the benign score of an photo-caption pair by the output probability of the Villo model (Muennighoff 2020), which estimates the likelihood that the input photo-caption pair is hateful. The benign score is defined as the complement of the predicted hateful probability, i.e., $1 - \mathrm{P(hateful)}$, where the prediction is obtained from the vision–language model, Villo.

- **Fluency score:**

  We measure the fluency score of generated captions using Parrot (Damodaran 2021), a T5-based language model with an evaluation component built on a fine-tuned BERT classifier. This classifier outputs a probability between 0 and 1 indicating the likelihood that a caption is fluent. Compared to traditional metrics like n-gram overlap or perplexity, this approach leverages contextual understanding and self-attention to capture grammatical structure, word order, and semantic coherence, including subtle phrasing issues that simpler metrics may overlook. This enables Parrot to provide a more human-aligned assessment of caption fluency.

- **Diversity score:**

  We use the pre-trained CLIP model (Radford et al. 2021) as a tool to evaluate semantic diversity. Specifically, we compute the cosine similarity between the semantic features of each generated caption pair to quantify variation. This approach leverages CLIP's strong cross-modal semantic understanding to detect subtle differences in meaning beyond surface-level textual variation. For each caption, we exclude self-comparison and identify the most similar caption based on cosine similarity. We then compute the average of the maximum similarity values to represent the typical semantic closeness among the captions in Eq. (26), where MS_j denotes the maximum similarity value for the j-th caption and R denotes the total number of captions being evaluated. When captions are semantically similar and exhibit low diversity, the mean maximum similarity approaches 1, resulting in a diversity score near 0. Conversely, when captions are more semantically distinct and diverse, the mean similarity is lower, yielding a higher diversity score approaching 1.

doi:10.6342/NTU202501318

$$diversity = \sqrt{1 - \left(\frac{1}{R}\sum_{j=1}^{R} MS_j\right)} \tag{26}$$

Table 4. Data Composition of the Main and Adapted Datasets

	OxfordTVG-HIC	Sonic Drive-In	McDonald's
Train data	2,400	224	224
Test data	600	56	56

Due to the hardware limitation, we are unable to train the model using the full OxfordTVG-HIC dataset. We select 3,000 photo-caption pairs, and divide the photo-caption pairs into 80% for training and 20% for testing. Each text from the photo-caption pairs is augmented into 300 diverse variants using a masked language model based on RoBERTa (Liu et al. 2019), which predicts masked tokens in context to generate semantically coherent alternatives. Also, we select 280 photo-caption pairs from each adapted dataset and divide the selected photo-caption pairs into 80% for training and 20% for testing as shown in Table 4.

The model is implemented in Python 3.11, utilizing the Pytorch 2.6.0 version with a NVIDIA GeForce RTX 4090 GPU with 22GB memory. We use the AdamW optimizer with the learning rate 0.00002, and the batch size of 20. Also, we set the hyper-parameters α to 10. We adopt the Falcon (Almazrouei et al. 2023) as the pre-trained Large Language Model. Our model is trained in two stages. In the first stage, we train the model by the OxfordTVG-HIC dataset and use the negative log likelihood loss (Eq. 18) for caption generation. In the second stage, the trained model is fine-tuned by the adapted datasets, where the loss function contains the negative log likelihood loss and the binary cross-entropy loss of funny scores (Eq. 25).

4.2 Performance Evaluation

We compare our proposed model with the state-of-the-art image captioning model ClipCap (Mokady et al. 2021) and BITA (Yang et al. 2024) in Tables 5 and 6. GAMC scores the highest in humor, indicating its advantage in generating humorous captions in both adapted datasets. By using the adaptation module, GAMC can maintain more humor elements when generating meme captions. Even though different datasets might cause slightly different performance, BITA's humor score is relatively low in

doi:10.6342/NTU202501318

Sonic Drive-In adapted dataset but higher in McDonald's adapted dataset. ClipCap performs better than BITA in terms of humor score. All models show relatively stable performance on benigh scores, scoring 0.91 or higher, indicating their effectiveness in avoiding harmful content. GAMC has a slightly higher benign score than BITA and ClipCap, suggesting that its generated content may be less controversial. GAMC outperforms the compared models in terms of humor scores, showing the capacity of creating humorous captions.

Table 5. Performance Evaluation in Sonic Drive-In Dataset

	Humor	Benign	Fluency	Diversity
BITA	0.309	0.929	0.922	0.496
ClipCap	0.265	0.923	0.836	**0.609**
GAMC	**0.600**	**0.933**	**0.945**	0.450

Table 6. Performance Evaluation in McDonald's Dataset

	Humor	Benign	Fluency	Diversity
BITA	0.322	0.919	0.913	0.461
ClipCap	0.195	0.926	0.905	**0.529**
GAMC	**0.507**	**0.928**	**0.944**	0.401

GAMC leads in fluency scores, meaning its generated captions are most readable, while ClipCap performs the worst in both adapted datasets. However, ClipCap achieves the highest diversity score, then followed by BITA, indicating their content offers more variation than GAMC. This phenomenon can be explained with the concept of spectrum. Most of the humor texts in the OxfordTVG-HIC dataset feature simpler and more complete sentences, which also aligns with Malodia et al.'s experiment (2022) on viral memes, emphasizing the necessity of iconicity as a key characteristic. In contrast, the texts from Instagram often include additional information such as links, numerous hashtags and account tagging. Therefore, when the generated captions lean toward the humor side of the spectrum (OxfordTVG-HIC dataset), their concise nature leads to lower diversity scores with shorter sentences. Conversely, when the generated captions shift toward the informational side (the adapted datasets from Instagram), longer sentences result in higher diversity scores, but at the cost of a greater risk of

doi:10.6342/NTU202501318

lower fluency. Since the additional information on Instagram are usually the names or the slogans of the event and the partners or customers the brands are engaging with. It is usually not present in the image itself and is predetermined before they needed an Instagram post to do the promotion. GAMC tends to strike a better balance by incorporating enough information from adapted datasets while maintaining the humor-driven quality found in the OxfordTVG-HIC dataset. Therefore, GAMC outperforms BITA and ClipCap.

4.3 Ablation Study

Tables 7 and 8 illustrate the performance of the variants of our model. There are three modules that can be adapted namely, the visual feature extraction module, the co-attention module, and the Large Language Model (LLM) module, where ✓ denotes that the module is adapted with LoRA and BitFit, and ✗ denotes that the module is not adapted. We conduct experiments on all possible combinations, resulting in eight different variants of the model.

By comparing the first and third rows as well as the sixth and eighth rows in Tables 7 and 8, we observe that diversity scores generally increase when the co-attention module adaptation is excluded, indicating that this module may reinforce structural consistency but at the expense of variations in generated captions. As for the benign scores, they remain consistently high across all variations, showing that the adapted modules effectively mitigate potentially harmful content.

From the results in the last four rows in Tables 7 and 8, the humor and fluency scores have shown the adaptation in different modules successfully maintain the humor from the main dataset, also proceed a higher fluency score with the sacrifice of the diversity score, slightly shift to the humor side of the spectrum. However, in both datasets, the full adaptation (visual + co-attention + LLM), shows in the first row of Tables 7 and 8, does not necessarily yield the highest humor scores, suggesting that certain module combinations better optimize humor in caption generation. The combination of visual + co-attention adaptation (in the second row of Tables 7 and 8) yields the best humor performance in both datasets. This suggests that this adaptation setup is optimal for generating humor-driven captions. In addition, the presence of humor tends to enhance fluency in most cases, reinforcing the idea that humor-oriented captions benefit both readability and engagement.

doi:10.6342/NTU202501318

Table 7. Performance of Variant Adapted Modules in Sonic Drive-In Dataset

Adapted Modules			Humor	Benign	Fluency	Diversity
Visual	Co-A	LLM				
✓	✓	✓	0.456	0.928	0.896	0.484
✓	✓	✗	**0.600**	0.933	**0.945**	0.450
✓	✗	✓	0.388	0.930	0.869	**0.517**
✗	✓	✓	0.442	0.928	0.930	0.492
✓	✗	✗	0.382	0.926	0.922	0.489
✗	✓	✗	0.389	0.930	0.933	0.504
✗	✗	✓	0.445	0.933	0.910	0.475
✗	✗	✗	0.431	**0.936**	0.840	0.515

Table 8. Performance of Variant Adapted Modules in McDonald's Dataset

Adapted Modules			Humor	Benign	Fluency	Diversity
Visual	Co-A	LLM				
✓	✓	✓	0.304	0.922	0.896	0.475
✓	✓	✗	**0.507**	0.928	**0.944**	0.401
✓	✗	✓	0.310	0.928	0.904	0.509
✗	✓	✓	0.231	0.933	0.908	0.557
✓	✗	✗	0.346	0.928	0.920	0.531
✗	✓	✗	0.262	0.925	0.893	0.537
✗	✗	✓	0.328	**0.936**	0.869	0.536
✗	✗	✗	0.256	0.930	0.677	**0.558**

Tables 9 and 10 illustrate the performance of the variants of the adaptation setting in our model, where FCT denotes the funny score tuning, ✓ denotes that the adapter is applied and ✗ denotes that the adapter is not applied. There are three adapters, LoRA, BitFit, and funny score tuning, funny score tuning in our model. We conduct experiments on all possible combinations, resulting in eight different variants of the adaptation setting on adapters in our model.

doi:10.6342/NTU202501318

Table 9. Performance of Variant Adapters in Sonic Drive-In Dataset

Adapted Modules			Humor	Benign	Fluency	Diversity
LoRA	BitFit	FCT				
✓	✓	✓	**0.600**	0.933	**0.945**	0.450
✓	✓	✗	0.446	0.925	0.867	0.467
✓	✗	✓	0.459	0.934	0.901	0.492
✗	✓	✓	0.465	0.924	0.908	0.483
✓	✗	✗	0.390	0.923	0.830	0.487
✗	✓	✗	0.463	0.930	0.922	0.441
✗	✗	✓	0.431	**0.936**	0.840	**0.515**
✗	✗	✗	0.357	0.922	0.623	0.494

Table 10. Performance of Variant Adapters in McDonald's Dataset

Adapted Modules			Humor	Benign	Fluency	Diversity
LoRA	BitFit	FCT				
✓	✓	✓	**0.507**	0.928	**0.944**	0.401
✓	✓	✗	0.274	**0.932**	0.913	0.522
✓	✗	✓	0.322	0.926	0.919	0.469
✗	✓	✓	0.410	0.923	0.900	0.440
✓	✗	✗	0.298	0.922	0.891	0.491
✗	✓	✗	0.352	0.923	0.939	0.459
✗	✗	✓	0.256	0.930	0.677	**0.558**
✗	✗	✗	0.246	0.921	0.630	0.486

The benign scores in Tables 9 and 10 remain stable across all variants, demonstrating that neither LoRA nor BitFit adaptations, nor the application of funny score tuning, introduce harmful or controversial elements.

When the funny score tuning is activated without LoRA or BitFit, the model achieves the highest diversity score but the lowest fluency score among the variations with the funny score tuning enabled. In this configuration, the generated captions are more information-oriented, incorporating a wider range of details and variations. However, this increased diversity comes at the expense of fluency, as the addition information such as hashtags and account tagging may not be seen fluence as regular

doi:10.6342/NTU202501318

sentences. In contrast, across all variations where the funny score tuning is not applied, the model generally struggles to capture humorous elements effectively. As a result, the model exhibits a learning bias toward content annotation rather than humor generation, favoring literal image annotation over the generation of humor-oriented meme captions. This often results in more descriptive captions, such as those in the McDonald's dataset. These captions are relatively inoffensive, as indicated by higher benign scores, but tend to be less humorous, corresponding to lower humor scores. These straightforward captions are easier for the model to learn, leading to relatively higher fluency and greater diversity. The fluency scores tend to be higher when either LoRA or BitFit or funny score tuning is enabled, suggesting their positive impact on sentence structure and readability. Especially, the BitFit-only adaptation shows a more significant improvement in fluency in McDonald's dataset compared to Sonic Drive-In. This suggests that BitFit contributes more effectively to sentence coherence and readability in McDonald's captions, possibly due to differences in caption structure or linguistic patterns. The highest humor score is achieved when LoRA, BitFit and funny score tuning are enabled, in the first row of Tables 9 and 10, highlighting their combined effectiveness in humor optimization.

To balance humor, fluency, and diversity, adaptation strategies should prioritize the visual feature extraction module and the co-attention module for humor-driven captions, while adjusting LoRA, BitFit, and funny score tuning usage to optimize fluency and variation. The trade-off between fluency and diversity highlights the importance of tuning adaptation methods based on dataset characteristics and caption objectives. Ultimately, the most balanced model for meme caption generation is the visual feature extraction module and the co-attention module adaptation (excluding LLM adaptation), which consistently achieves the highest humor scores across both datasets. This adaptation setup maximizes humor while maintaining fluency, making it an optimal choice for humor-driven caption generation. Meanwhile, fluency is best enhanced when LoRA and BitFit adaptations are combined with funny score tuning, reinforcing the positive impact of these adapters on sentence coherence.

doi:10.6342/NTU202501318

4.4 Human Evaluation

To evaluate the effectiveness of GAMC, ClipCap, and the original post, we conduct a human evaluation to assess the quality of the captions. The study involved 30 participants, with 76.7% being women and 23.3% men. All participants were 25 years old or younger, and 80% reported using Instagram daily, while 96.7% were familiar with memes. Since our goal is for the target audience of meme marketing to perceive branded memetic posts as effective, the participants are general social media users with an understanding of meme culture.

The questionnaire consists of two evaluation sets, each corresponding to a different brand, Sonic Drive-In and McDonald's. For each evaluation set, we present five original photos from the Instagram-adapted dataset, accompanied by captions generated by GAMC and ClipCap, as well as the original caption from the Instagram post. After reviewing these photo-caption pairs, participants are asked to answer five questions related to the five evaluation metrics to assess the effectiveness of the captions.

Based on Malodia et al.'s experiment (2022), we use five metrics to evaluate the meme virality. Four of these are content-related—humor, relevance, iconicity, and shareability while the fifth represents a customer-related factor—customer gratification. To assess the effectiveness of photo-caption pairs, we employ a five-point Likert scale (Likert, 1932), ranging from "strongly disagree" (1) to "strongly agree" (5).

- **Relevance:**

  The extent to which content aligns with the target audience by being topical, familiar, relatable, contemporary, and noncontroversial, ensuring maximum engagement.
- **Iconicity:**

  The degree to which content is clear and well-structured, using simple, complete sentences, with a writing style that aligns with the context and incorporates situational descriptions to enhance relatability.
- **Humor:**

  The extent to which an appropriate humor strategy enhances meme virality.
- **Shareability:**

  The degree to which recipients are willing to further distribute the content.

doi:10.6342/NTU202501318

- **Customer Gratification:**

  The extent to which individuals engage with memes for escape, social connection, and self-expression. Specifically, this evaluation considers how memes provide humor, shared experiences, and nihilistic perspectives, enabling users to connect, seek validation, and express their identity, while also serving as a medium for sharing and receiving informative content.

To evaluate the effectiveness of the meme caption, the participants are required to familiarize with the five metrics and their definitions, along with the provided information about the brand. Then, they are asked to rate the photo-caption pairs of the original post and those generated by each mode. The ratings for each metric across models and the original post are averaged based on the number of evaluation sets and participants. That is, the quality of the captions in the photo-caption pairs is determined through the average ratings.

(a)	(b)	(c)
		
GT: happy world fashion day!	**GT:** the bell rings and the evening is over.	**GT:** deck your doggo out in SONIC swag for their next trip to the Drive-In. Our Wag Shop is now OPEN with toys, costumes, and more. 🐶
GAMC: i'm going to ask you a few questions about your food and drink preferences.	**GAMC:** we've got a lot of different people who are trying to get through this crisis.	**GAMC:** the only thing i know for sure is that the can of tuna in a sandwich was made with real meat.
ClipCap: do you know the feeling of a new year? i wish you all a wonderful new year.	**ClipCap:** it's about to get spicy: come and try our new spicy mcchicken jalapeño today and show us your #spiceface🌶	**ClipCap:** it doesn't matter if you don't have it.

Figure 5. Meme Caption Examples

doi:10.6342/NTU202501318

Tables 11 and 12 present the results of human evaluation for Sonic Drive-In Meme and McDonald's meme captions. GAMC significantly outperforms both ClipCap and the original post in most cases. However, in the McDonald's meme captions, although GAMC receives a higher iconicity score, the score compared to the original post is not statistically significant. This is because the McDonald's original posts sometimes use extremely short sentences simply to celebrate or describe discounts, without additional humorous elements. As a result, the sentences tend to be relatively brief, as seen in Figure 5 (a), making it difficult to achieve a humor-driven post for meme marketing, but received a relatively high score in iconicity. The shareability score of GAMC in the two evaluation sets is significantly greater than the score of ClipCap. However, despite achieving a higher score than the original post, the score is not statistically significant better in McDonald's evaluation set. Since people are generally less likely to share purely promotional posts, McDonald's original advertisement image, such as Figure 5 (b), may have lowered the willingness to share, leading to a lower score and ultimately reducing the difference in shareability. On the other hand, the customer gratification score of GAMC in Sonic Drive-In meme captions shows a notable advantage over ClipCap. While it also achieves a higher score than the original post, but not statistically significant in both evaluation sets. GAMC is designed to transform purely promotional Instagram posts into content with the promotion that allows audiences to escape from the reality of life into the realm of humor. In addition, if shared, these posts contribute to social gratification. For example, in Figure 5 (c), GAMC incorporates humor by presenting a food-obsessed dog sharing its thoughts on pet-related activities. This integration reduces the overall promotional content density while enhancing escapism gratification through humor, ultimately resulting in a lower overall score. Conversely, the original post contains high promotional content, simply describing what was available in the Wag Shop, with relatively low escapism gratification, which also contributed to a lower overall score. As for ClipCap, its depiction of pet-related activities includes a quirky, somewhat absurd notion, that even those without pets could participate, as if the idea is simply "why not?". While this carries a hint of humor, it is not as strong as GAMC's. Compared to the original post, the promotional content is entirely absent, leading to the lowest score among the three. The generally low scores make it difficult to identify statistically significant differences in the results.

doi:10.6342/NTU202501318

Table 11. Human Evaluation of Meme Captions in Sonic Drive-In

Metrics	GAMC	ClipCap	GT
Relevance	3.547	3.060*	3.060*
Iconicity	3.767	3.267*	3.333*
Humor	3.460	2.687**	2.720*
Shareability	3.120	2.467*	2.633*
Customer gratification	2.813	2.293*	2.493

NOTE: * and ** indicate that the results of the GAMC are significantly better than those of the compared models (t-test, * indicates $p < 0.05$, and ** indicates $p<0.001$).

Table 12. Human Evaluation of Meme Captions in McDonald's

Metrics	GAMC	ClipCap	GT
Relevance	3.700	3.240*	3.320*
Iconicity	3.700	3.353*	3.400
Humor	3.500	2.873*	3.047*
Shareability	3.053	2.547*	2.700
Customer gratification	2.773	2.440	2.520

NOTE: * and ** indicate that the results of the GAMC are significantly better than those of the compared models (t-test, * indicates $p < 0.05$, and ** indicates $p<0.001$).

GAMC significantly outperforms both compared methods and the original Instagram post in relevance and humor. This demonstrates its potential to improve brand recall and brand engagement, making promotional posts more entertaining and shareable. However, the lack of statistical significance in its improvements, challenges in differentiation, and the trade-offs between humor and content density, along with overall low scores that blur meaningful differences, ultimately constrain its advantage.

doi:10.6342/NTU202501318

4.5 Meme Caption Examples

In this section, we present some captions generated by each model, where GT denotes the ground-truth caption. Figure 6 presents the caption examples with Sonic Drive-In. Take the first photo on the left in Figure 6 as the example, the photo shows a Sonic corndog mascot taking a picture with a little girl and a handwriting-style words "Happy Day!" edited on the photo, the ground truth caption has nothing related with humor or meme, since it simply describes the emotions of the edited words, then provides the information about the upcoming event, which is not provided in the given photo. In contrast, GAMC not only reads the "happy" emotion through the edited words, but also notices the shadow on the face of the Sonic corndog mascot costume, which can be interpreted as a Yandere-like characteristic, a characteristic originates from Japan's ACG culture (Anime, Comic, Game) and is favored by certain audiences. This characteristic is often defined by extreme obsession or love, sometimes escalating into irrational behavior or even violence. With the caption generated by GAMC, we can observe the connection between "Happy Day!" and "the best thing" in the caption. In addition, the Sonic corndog mascot appears to be asking what the best thing to do in 2019 is, with a threatening, Yandere-like expression. The implication is that there is only one acceptable answer "with me, the Sonic corndog", which creates humor by contrasting the cheerful emotion with the underlying tension. The caption generated by ClipCap clearly illustrates the previously mentioned issue. Without adaptation, it hardly absorbs elements from Instagram, causing the caption to lean toward the information-side of the spectrum. Not only does it produce longer sentences, making its structure more distinct from other captions and potentially leading to a higher diversity score, but the inclusion of hashtags affects fluency and readability of the caption. What's worse, the generated hashtags may not accurately align with the brand's intend of promoting its upcoming event, potentially sending the wrong message and diluting the intended marketing impact. However, ClipCap successfully captures the happy emotion conveyed in the image. BITA's generated caption repeats "what's this" three times, indicating that BITA successfully captured the hand details of the corndog mascot. However, it fails to recognize the emotional context present in the image. Thus, the caption generated by GAMC has higher relevance to the image and contains stronger humor elements, making it the most suitable meme caption among all the generated options.

doi:10.6342/NTU202501318

GT: This little bit was so excited to meet Mr. Coney at SONIC Summer Movie Night at OKC's Myriad Gardens. Tonight, "The Sandlot", next week "Forrest Gump."	**GT:** congrats to our grand prize #doubledogdare winners! @lparrinelli @jusserv @tia_lorraine pls email your name and mailing	**GT:** how many corn dogs you'll have is still a toss-up. drop by sonic on august 31st and enjoy 50¢ corn dogs all day.
GAMC: we all know that this is the best thing to do in 2019.	**GAMC:** when i was 16 and my mom told me to go get a dog, he said.	**GAMC:** this is the perfect recipe for a healthy, delicious and satisfying breakfast.
ClipCap: it' s a good day and a miss… what' s your favorite moment of the year? #msfhc #seasonshowslife #msfhc2 #calendardaylooks_like_this #calendarbreak #calendar_looks_like_this #calendar_looks_like_	**ClipCap:** we' re going to make a new sonic cup of food.	**ClipCap:** this is what you' re told.
BITA: `` what 's this? " `` what 's this? " `` what 's this? "	**BITA:** 🚨🚨🚨🚨🚨🚨🚨	**BITA:** it' s halloween, and this is the corn dog you' re looking for

(a)

GT: drown your tastebuds in rich, creamy 1/2 price shakes after 8pm at sonic for the next five days.#sonicdrivein #shakebreak #shakes #📅 #🕗	**GT:** wake up, it's the first of the month
GAMC: for # shakespeare' s birthday, we were thinking of a lil bit more serious.	**GAMC:** i'm not a dog person but i love these little dogs! and they' re so cute, it makes me want to be one too.
ClipCap: it' s time for a change.	**ClipCap:** when you wake up feeling like you belong in the middle of nowhere.
BITA: if you' re a fan of the slushies you' re going to love this one	**BITA:** when you know you' re in the dog house you know you' re in the dog house you

(b)

Figure 6. Meme Caption Examples with Sonic Drive-In

doi:10.6342/NTU202501318

Figure 7 displays the caption generated by each model with McDonald’s. Taking the first photo on the left in Figure 7 as an example, the photo portrays someone experiencing McDonald's app for the first time, specifically trying out the Order & Pay service. The app displays the tracking map of their meal, also they use a nervous dog meme on the photo to express their anticipation and excitement. The relationship between the photo and the ground truth text begins with the text, then completed by the photo. Together, they illustrate a contrast between two types of people, those who are familiar with using the service, Order & Pay, and those who are trying it for the first time, feeling both nervous and excited. For this contrast to be effectively conveyed, the alignment between the caption and the photo must be highly cohesive. Instead of emphasizing the contrast, the caption generated by GAMC focuses on the dramatic emotions embedded in the photo. The caption precisely captures the emotional tone intended by the dog meme, authentically presenting the inner monologue of the people, using Order & Pay for the first time. By breaking down the caption, we can see how it effectively achieves our goal. It not only integrates the brand-related elements but also captures the complex emotional layers within the photo, demonstrating a refined understanding of meme construction. First, even though the photo only contains a black dog and a smartphone, the first part of the caption introduces an element not present in the photo, a burger. This demonstrates that GAMC has indeed learned brand-related concepts from the adapted dataset, enabling it to generate contextually rich content. Second, with the support of the ESH module, GAMC is able to perceive and reflect two distinct emotions from the photo. One is a sense of nervousness, which is clearly conveyed through the dog meme in the caption. The other, though more subtle, is a feeling of eager anticipation for the meal. Rather than merely describing what is literally visible, GAMC uses the photo as a foundation, skillfully integrating brand-related insights from the adapted dataset to creatively craft a cohesive and contextually relevant meme. The caption from ClipCap emphasizes that ClipCap might have recognized the smartphone or classified the photo as app-related. It attempts to create humor by exaggerating the app’s capabilities, but the actual humor remains subtle and lacks the emotional engagement that could strengthen its memetic impact. On the contrary, BITA only focuses on the emotions, possibly because the emotional aspect is highly prominent in the photo, making it the main concept BITA identifies. As a result, BITA generates captions based on the discovered emotional theme, while overlooking other

doi:10.6342/NTU202501318

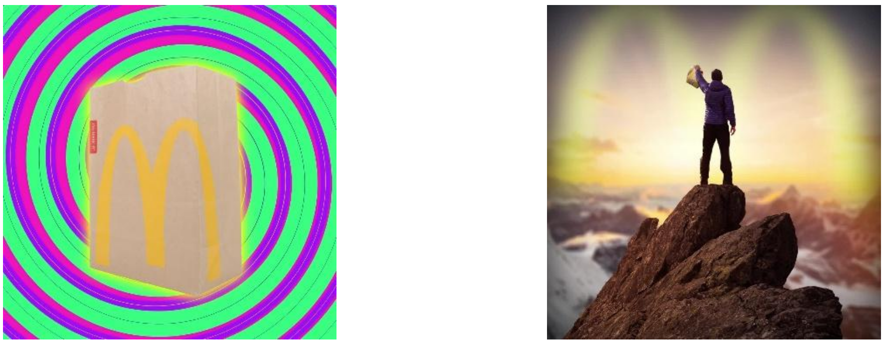
photo details and ignoring the knowledge from the adapted dataset. Thus, GAMC excels in meme captioning by effectively integrating brand-related elements, recognizing complex emotional layers, and constructing contextually rich and humorous captions. Unlike ClipCap, which emphasizes app-related aspects but lacks emotional engagement, or BITA, which solely focuses on detecting emotions while overlooking other details, GAMC strikes a balance between emotion, context, and brand relevance. By leveraging adaptations from the adapted dataset and the ESH module, GAMC enhances photo-driven storytelling, precisely reflecting emotions, incorporating missing but relevant brand details, and transforming a static photo into a compelling meme.

In summary, the adapted dataset effectively compensates for the lack of brand knowledge in the main dataset, allowing the model to generate humorous meme captions based on photo elements and emotions. For Sonic Drive-In, the adapted dataset mainly consists of standard advertisements, enabling the model to integrate products with humor. For example, in the left image of Figure 6 (b), the model identifies the product "shakes" and introduce the idea of a "discount for Shakespeare’s birthday" through wordplay, adding a touch of humor to the photo-caption pairs which originally only conveyed a straightforward promotional discount. Likewise, in the middle of Figure 6 (a) and the right of Figure 6 (b), the adapted model recognizes the products “hotdog” and “corndog,” and generates dog-related jokes. These captions are created by combining humor from the main dataset with product knowledge from the Sonic Drive-In adapted dataset, resulting in meme captions suitable for Instagram posts promoting Sonic Drive-In. On the other hand, the McDonald's adapted dataset not only consists of promotional product information, which allows the model to generate burger-related posts when processing the left image in Figure 7 (a). In addition, since McDonald's Instagram posts reflects its people-oriented brand awareness, often conveying themes such as kids, family, friendship, and happiness, the adapted model incorporates elements of warmth and togetherness. For instance, with the left photo of Figure 7 (b), the model generates a caption related to giving McDonald's as a gift for a friend, further reinforcing the brand’s emphasis on sharing joyful experiences through food and fostering social connections among friends and family. Thus, adaptation successfully enables the model to learn brand-specific characteristics from the adapted dataset, allowing it to generate humorous meme captions that align with each brand’s original brand personality and visual elements.

doi:10.6342/NTU202501318

GT: there are those who use order&pay, and there are the others...	**GT:** only people using order&pay in the app can relate.	**GT:** i know you would never decline that, would you?? 📞 #shrimps #shrimpscalling
GAMC: the burger is ready! get on the way... and enjoy your meal.	**GAMC:** i love you! " go for it.	**GAMC:** we are trying to figure out what the best app for your brain is.
ClipCap: they are just as talented as you are.	**ClipCap:** gut feeling like you left out of mcdonald 's month? thank you to our l.	**ClipCap:** this is the moment you 're not afraid to try your hand at a menu.
BITA: you 're going to meet a friend who 's going to take you to a place where you 'll	**BITA:** are you ready for some love? 💕 do n't forget to turn on your favorite playlist and shop	**BITA:** if you 're a hiccup, you 'll taste the smell of the mcflurry in the

(a)



GT: no need to wait for the sun to go down. enjoy our famous sunset offers in the mcdonald's app every day as of 17h! #sunsetoffers	**GT:** our take-away bag. proud sponsor of having fun at home. who agrees? 😜 ♡
GAMC: here's the best gift for a friend who always comes with you, even when they don t want to go out.	**GAMC:** the reason you' re here is because your body was designed to do it.
ClipCap: let it be real, we 've got something for everyone.	**ClipCap:** this is the first time we 've ordered a mcnuggets® collection for lunch.
BITA: when it comes to dating, who 's the first one you 're going after? i 'm going	**BITA:** it 's not always easy to let go of your favorite Mcdonald 's, but at least you know

(b)

Figure 7. Meme Caption Examples with McDonald's

doi:10.6342/NTU202501318

Chapter 5 Conclusions and Future Work

We propose an adaptable framework to generate an advertising meme caption for the posts of a brand. The proposed framework first employs the visual feature extraction module to extract the visual features and the ESH module to extract the features of emotion, sentiment, and humor from the photo of the input post. Third, it utilizes the co-attention module to learn the inter-relationships between visual and ESH features. Fourth, it exploits the Large Language Model (LLM) to generate the meme caption for the input post and the main dataset to train the model for increasing the sense of humor of generated captions. Last, it adopts the adapters, LoRA and BitFit, and the funny score tuning, to fine-tune the trained model to generate advertising meme captions for the posts of the brand.

The experimental results show that the proposed model outperforms the compared models in terms of humor, benign, and fluency scores, and has a better balance by incorporating enough information from adapted datasets while maintaining the humor-driven qualities found in the OxfordTVG-HIC dataset. Furthermore, by utilizing adapters, the generated captions successfully lean toward the humor side of the spectrum while still integrating relevant information from Instagram, resulting in captions that are both entertaining and strategically aligned with the platform's branding style.

In addition, our model outperforms the variant models in terms of humor score, demonstrating its strength in generating engaging and entertaining captions. However, enabling all the modules simultaneously does not produce the best results. Instead, the adaptation with the visual feature extraction module with the co-attention adaptation achieves the highest humor score while maintaining fluency and diversity, striking the optimal balance for meme caption generation. As for the composition of the adapters, enabling all adapters together yields the best overall performance.

Furthermore, we conduct a survey to assess the effectiveness of the captions in the photo-caption pairs based on five metrics: relevance, iconicity, humor, shareability and customer gratification. The results demonstrates that our model significantly outperforms the compared model in all five metrics and surpasses the original Instagram post in relevance and humor, highlighting its potential to enhance brand recall and brand engagement.

doi:10.6342/NTU202501318

The proposed model has the following theoretical implications. First, we propose an adaptable model that generates advertising meme captions resonated with a specific brand while remaining the ability extended to other brands and industries. Second, our approach leverages MiniGPT4 to extract key features, such as emotion, sentiment, and humor from input images and integrates them through a co-attention module that bridges multiple modalities. Third, we train the model on a comprehensive humor dataset to instill a rich sense of wit and creativity in the generated captions. Then, we fine-tune the pre-trained model on an adapted dataset collected from the specific brand's Instagram account, ensuring that the final captions capture both the universal appeal of humor and the brand's particular tone. Last, the two-stage training strategy combined with our novel multimodal integration framework sets a new benchmark for creative advertising and opens promising avenues for future innovations in digital marketing and creative content generation.

In addition, the proposed model offers the following managerial and practical value for businesses, particularly in digital marketing and brand communication. By crafting emotionally engaging and contextually meaningful meme captions, our model enables businesses to present a more humorous, relatable brand personality. This can help strengthen emotional connections with consumers, making brand messages more memorable and impactful. The ability of our model goes beyond surface-level image description and creatively integrates brand-related elements, such as the products of the brand, and some platform-specific cultural cues from Instagram, such as the usage of the hashtags, which allows it to generate content that aligns closely with audience expectations and online humor trends. This not only enhances brand storytelling but also improves click-through rates, content shareability, and overall engagement on social media platforms. When brands are preparing to promote events across multiple platforms, these events typically already have predefined names and basic content. After generating captions with our model, brands can seamlessly integrate event names in hashtag format, achieving both the humor required for meme marketing and the necessary promotional context. This allows brands to maintain their event identity while leveraging meme virality, making our model the ideal partner for businesses unfamiliar with meme culture but eager to enhance their brand image with meme.

In the future, we could develop an integrated model that takes the name or the purpose of the event as a prompt, generating potentially useful slogans as background

doi:10.6342/NTU202501318

knowledge for meme captioning. By combining the generated meme with hashtagging the generated slogan, we can create a more complete and cohesive Instagram post that enhances engagement, relevance, and branding. Moreover, memes are a form of cultural trend that constantly evolves across regions, languages, and time. What resonates today in one context may not be as effective tomorrow or in a different cultural setting. Therefore, it is crucial for the training data to remain up-to-date and culturally adaptive. To keep pace with these shifts, future work could involve incorporating real-time meme trends and regional variations through continuous data collection and fine-tuning, ensuring that the model maintains both cultural relevance and humor effectiveness in dynamic social media environments.

doi:10.6342/NTU202501318

References

Almazrouei E, Alobeidli H, Alshamsi A, Cappelli A, Cojocaru R, Debbah M, Penedo G et al. (2023) The falcon series of open language models. arXiv:2311.16867. https://doi.org/10.48550/arXiv.2311.16867

Bousaid R, El Hajji M, Es-Saady Y (2022) Facial expression recognition using a hybrid ViT-CNN aggregator. *Proceedings of the International Conference on Business Intelligence.* 61-70. https://doi.org/10.1007/978-3-031-06458-6_5

Chen H, Wang Y, Yang X, Li J (2021) Captioning transformer with scene graph guiding. *Proceedings of the IEEE International Conference on Image Processing*. 2538-2542. https://doi.org/10.1109/ICIP42928.2021.9506193

Damodaran P (2021) Parrot: Paraphrase generation for nlu. GitHub Open Source.

Dawkins R (1976) *The Selfish Gene*. Oxford: Oxford University Press, New York, USA.

Devlin J, Chang M W, Lee K, Toutanova K (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.* 4171-4186. https://doi.org/10.18653/v1/N19-1423

Eisend M (2022) The influence of humor in advertising: Explaining the effects of humor in two-sided messages. *Psychology & Marketing* 39(5):962-973. https://doi.org/10.1002/mar.21634

Fang Z, Wang J, Hu X, Liang L, Gan Z, Wang L, Yang Y, Liu Z (2022) Injecting semantic concepts into end-to-end image captioning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 17988-17998. https://doi.org/10.1109/CVPR52688.2022.01748

Giroux J, Bouchard M, Laganière R (2023) T-FFTRadNet: Object detection with Swin vision transformers from raw ADC radar signals. *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*. 4032-4041. https://doi.org/10.1109/ICCVW60793.2023.00435

Han S, Chang H, Shi Z, Hu S (2023) Facial expression recognition algorithm based on Swin transformer. *Proceedings of the 9th International Conference on Systems and Informatics.* 1-6. https://doi.org/10.1109/ICSAI61474.2023.10423327

Hatamizadeh A, Nath V, Tang Y, Yang D, Roth HR, Xu D (2022) Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images.

doi:10.6342/NTU202501318

Proceedings of the International MICCAI Brainlesion Workshop. 272-284. https://doi.org/10.1007/978-3-031-08999-2_22

He K, Zhang X, Ren S, Sun J (2015) Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 770-778. https://doi.org/10.48550/arXiv.1512.03385

He R, Liu L, Ye H, Tan Q, Ding B, Cheng L, Low J, Bing L, Si L (2021) On the effectiveness of adapter-based tuning for pretrained language model adaptation. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing.* 2208-2222. https://doi.org/10.18653/v1/2021.acl-long.172

He X, Zhou Y, Zhao J, Zhang D, Yao R, Xue Y (2022) Swin transformer embedding UNet for remote sensing image semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing* 60:1-15. https://doi.org/10.1109/TGRS.2022.3144165

Hochreiter S, Schmidhuber J (1997) Long short-term memory. *Neural Computation* 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, Wang L, Chen W (2021) LoRA: Low-rank adaptation of large language models. arXiv.2106.09685. https://doi.org/10.48550/arXiv.2106.09685

Jung D, Shim S, Choo C, Hwang D, Nah Y, Oh S (2022) A preliminary result of food object detection using Swin transformer. *Proceedings of the 8th International Conference on Computer Technology Applications*. 183-187. https://doi.org/10.1145/3543712.3543731

Li J, Li D, Savarese S, Hoi S (2023) Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *Proceedings of the International Conference on Machine Learning.* 19730-19742. https://doi.org/10.48550/arXiv.2301.12597

Li J, Li D, Xiong C, Hoi S (2022) Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *Proceedings of the International Conference on Machine Learning.* 12888-12900. https://doi.org/10.48550/arXiv.2201.12086

Li R, Sun S, Elhoseiny M, Torr P (2023) OxfordTVG-HIC: Can machine make humorous captions from images? *Proceedings of the IEEE/CVF International*

doi:10.6342/NTU202501318

Conference on Computer Vision. 20293-20303. https://doi.org/10.1109/ICCV51070.2023.01856

Li Y, Pan Y, Yao T, Mei T (2022) Comprehending and ordering semantics for image captioning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 17969-17978. https://doi.org/10.1109/CVPR52688.2022.01746

Likert R (1932) A technique for the measurement of attitudes. *Archives of psychology.* 22(140):55.

Liu L, Jiao Y, Li X, Li J, Wang H, Cao X (2023) Swin transformer-based image captioning with feature enhancement and multi-stage fusion. *Proceedings of the 19th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery.* 1-7. https://doi.org/10.1109/ICNC-FSKD59587.2023.10281090

Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019) RoBERTa: A robustly optimized BERT pretraining approach. arXiv.1907.11692. https://doi.org/10.48550/arXiv.1907.11692

Liu Z, Hu H, Lin Y, Yao Z, Xie Z, Wei Y, Ning J, Cao J, Zheng Z, Dong L, Wei F, Guo B (2022) Swin transformer v2: Scaling up capacity and resolution. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 11999-12009. https://doi.org/10.1109/CVPR52688.2022.01170

Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B (2021) Swin transformer: Hierarchical vision transformer using shifted windows. *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 10012-10022. https://doi.org/10.1109/ICCV48922.2021.00986

Liu Z, Tan Y, He Q, Xiao Y (2021) SwinNet: Swin transformer drives edge-aware RGB-D and RGB-T salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology* 32(7):4486-4497. https://doi.org/10.1109/TCSVT.2021.3127149

Lu J, Batra D, Parikh D, Lee S (2019) ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in Neural Information Processing Systems* 32. https://doi.org/10.48550/arXiv.1908.02265

doi:10.6342/NTU202501318

Malodia S, Dhir A, Bilgihan A, Sinha P, Tikoo T (2022) Meme marketing: How can marketers drive better engagement using viral memes? *Psychology & Marketing* 39(9):1775-1801. https://doi.org/10.1002/mar.21702

McGraw AP, Warren C (2010) Benign violations: Making immoral behavior funny. *Psychological Science*. 21(8):1141-1149. https://doi.org/10.1177/0956797610376073

Mokady R, Hertz A, Bermano A H (2021) ClipCap: Clip prefix for image captioning. arXiv.2111.09734. https://doi.org/10.48550/arXiv.2111.09734

Muennighoff N (2020) Vilio: State-of-the-art visio-linguistic models applied to hateful memes. arXiv:2012.07788. https://doi.org/10.48550/arXiv.2012.07788

Ninh QB, Nguyen HC, Huynh T, Tran MT, Le TN (2023) Multi-branch network for imagery emotion prediction. *Proceedings of the 12th International Symposium on Information and Communication Technology.* 371-378. https://doi.org/10.1145/3628797.3628954

Pech RJ (2003) Memetics and innovation: Profit through balanced meme management. *European Journal of Innovation Management* 6(2):111-117.

Peirson VAL, Tolunay EM (2018) Dank learning: Generating memes using deep neural networks. arXiv.1806.04510. https://doi.org/10.48550/arXiv.1806.04510

Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G, Sutskever I (2021) Learning transferable visual models from natural language supervision. *Proceedings of the International Conference on Machine Learning.* 8748-8763. https://doi.org/10.48550/arXiv.2103.00020

Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I (2019) Language models are unsupervised multitask learners. *OpenAI Blog* 1(8):9.

Ren S, He K, Girshick R, Sun J (2017) Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39(6):1137-1149. https://doi.org/10.1109/TPAMI.2016.2577031

Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556. https://doi.org/10.48550/arXiv.1409.1556

Vaswani A (2017) Attention is all you need. arXiv.1706.03762. https://doi.org/10.48550/arXiv.1706.03762

doi:10.6342/NTU202501318

Vyalla SR, Udandarao V (2020) Memeify: A large-scale meme generation system. *Proceedings of the 7th ACM India Joint International Conference on Data Science and Management of Data*. 307-311. https://doi.org/10.1145/3371158.3371403

Wang H, Lee RKW (2024) MemeCraft: Contextual and stance-driven multimodal meme generation. *Proceedings of the ACM on Web Conference*. 4642-4652. https://doi.org/10.1145/3589334.3648151

Weber M, Quiring O (2019) Is it really that funny? Laughter, emotional contagion, and heuristic processing during shared media use. *Media Psychology* 22(2):173-195. https://doi.org/10.1080/15213269.2017.1302342

Xie Z, Zhao C (2023) Micro-expression recognition based on dual-branch Swin transformer network. *Proceedings of the International Conference on Intelligent Computing*. 544-554. https://doi.org/10.1007/978-981-99-4742-3_45

Yan Y, Xue K, Shi X, Ye Q, Liu J, Ruan T (2023) AF adapter: Continual pretraining for building Chinese biomedical language model. *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine*. 953-957. https://doi.org/10.1109/BIBM58861.2023.10385733

Yang C, Li Z, Zhang L (2024) Bootstrapping interactive image–text alignment for remote sensing image captioning. *IEEE Transactions on Geoscience and Remote Sensing*. 62:1-12. https://doi.org/10.1109/TGRS.2024.3359316

Yang X, Hayashi T (2021) Exploring the effects of internet memes in social media marketing through A/B testing. *Proceedings of the IEEE 23rd Conference on Business Informatics*. 97-106. https://doi.org/10.1109/CBI52690.2021.10060

Yang X, Liu Y, Wang X (2022) ReFormer: The relational transformer for image captioning. *Proceedings of the 30th ACM International Conference on Multimedia*. 5398-5406. https://doi.org/10.1145/3503161.3548409

Zaken EB, Goldberg Y, Ravfogel S (2022) BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*. 1-9. https://doi.org/10.18653/v1/2022.acl-short.1

Zhang J, Xie Y, Ding W, Wang Z (2023) Cross on cross attention: Deep fusion transformer for image captioning. *IEEE Transactions on Circuits and Systems*

doi:10.6342/NTU202501318

for Video Technology 33(8):4257-4268. https://doi.org/10.1109/TCSVT.2023.3243725

Zhao Y, Cong G, Shi J, Miao C (2022) QueryFormer: A tree transformer model for query plan representation. *VLDB Endowment* 15(8): 1658-1670. https://doi.org/10.14778/3529337.3529349

Zhu D, Jun C, Xiaoqian S, Xiang L, Mohamed E (2023) MiniGPT-4: Enhancing vision-language understanding with advanced large language models. arXiv.2304.10592. https://doi.org/10.48550/arXiv.2304.10592

doi:10.6342/NTU202501318

Appendix A

Table A1 shows the architectural configurations of the baseline models and the proposed model.

Table A1. Comparison of the Baselines and the Proposed Models

Model	ClipCap		BITA		GAMC	
Image Size	224×224		224×224		256×256	
Visual Encoder	CLIP ViT-B/32		CLIP ViT-L/14		swinv2-tiny-patch4-window8-256	
	Size	Base	Size	Large	Size	Tiny
	Transformer	Standard	Transformer	Standard	Transformer	SWIN
	TF Layers	12	TF Layers	24	TF Layers	2+2+6+2
	Patch Size	32×32	Patch Size	14×14	Patch Size	4×4
	Hidden Dim	768	Hidden Dim	768	Hidden Dim	768
	Trainable	Frozen	Trainable	Frozen	Trainable	Frozen
Mapping Module	lightweight transformer-based mapping network		Interactive Fourier Transformer		Visual Feature Extraction + ESH + Co-Attention Modules	
Text Decoder	GPT2		opt-2.7b		Falcon3-1B-Base	
	Model Type	LM	Model Type	LLM	Model Type	LLM
	Model Size	137M	Model Size	2.7B	Model Size	1.67B
	Trainable	Frozen	Trainable	Frozen	Trainable	Frozen
First Stage	Language Model Alignment via Mapping Network		Image–Text Alignment via Contrastive Representation Learning		Meme Caption Generation	
Second Stage	✗		Visual Feature-Guided Language Generative Learning		Brand adaptation via LoRA, BitFit and funny score tuning	

doi:10.6342/NTU202501318

Appendix B

In the ESH Module, we leverage MiniGPT-4 to extract the emotion, sentiment, and humor descriptions that enrich visual features with useful information for meme caption generation. As a multimodal vision–language model (VLM), MiniGPT-4 is capable of generating structured textual responses based on both images and textual prompts. We input the image along with the following prompt: "Analyze the given image and break down its emotional, sentimental, and humorous aspects by listing their key elements in simple words. Answer 'None' if none exist." Then the generated text is processed through an automated pipeline.

First, we split the generated plain text by line breaks. Second, we identify the aspect keywords, "emotion," "sentiment," or "humor", to categorize the following content accordingly. Third, we determine the structure of the text output. According to our prompt design, the generated text typically falls into three recognizable structural patterns. In the first structure, as shown in Figure A1 (a), each aspect keyword is followed by a list in which individual elements are introduced using the star sign (*). The second structure, illustrated in Figure A1 (b), presents the key elements under each aspect as a numbered list. In the third structure, shown in Figure A1 (c), the aspect keyword appears together with its corresponding description within the same paragraph, without separation or bulleting. To determine the structure of the generated text, we examine each paragraph sequentially. Specifically, we assess whether the following paragraph introduces a new aspect keyword, which signals the start of a different section. In addition, we check whether the current line begins with a star symbol (*) or a numeric prefix, both of which indicate a list format. These two cues together allow us to classify the structural pattern and apply the appropriate parsing strategy for extracting each aspect description. In Figures A1 (a) and (b), we eliminate the list markers, either the star symbol (*) or the leading numbers, then concatenate the individual lines into a single unified description. In Figure A1 (c), no additional processing is required, the entire paragraph serves as the final description.

However, due to the randomness and length constraints of MiniGPT-4, the generation may terminate unexpectedly, resulting in missing aspect descriptions. For instance, as shown in Figure A1 (b), the output may only include emotion and sentiment descriptions while omitting humor description. To address this issue, we repeat the extraction process up to five times and dynamically adjust the prompt so that

doi:10.6342/NTU202501318

MiniGPT-4 focuses only on aspects that remain unextracted. After five attempts, if an aspect still cannot be extracted, a fallback mechanism automatically assigns 'None' as the final description. This pipeline ensures robust and reproducible extraction of emotional, sentimental, and humorous signals from visual content across different photo–caption pair datasets.

	Input Image	MiniGPT-4 Output
(a)		The image shows a woman standing in front of the words "blindfold outfit challenge". The woman is wearing a red and black striped hoodie, black leggings, and black boots, and is holding up her hands to show off her outfit. The background of the image is a white wall with a red square on it. Here are the key elements of the image's emotion, sentiment, and humor aspects: Emotion: * The woman's smile suggests happiness and excitement. * The pose of the woman, with her hands held up to show off her outfit, conveys a sense of pride and confidence. Sentiment: * The image is playful and lighthearted, with the woman's outfit being unusual and unexpected. * The challenge theme in the background adds a sense of fun and adventure. Humor: * The woman's outfit is unexpected and amusing, creating a sense of surprise or amusement for the viewer. * The idea of a blindfold outfit challenge is humorous, as it adds an element of unpredictability and challenge to the image. Overall, the image has a playful and lighthearted feel to it, with the woman's smile and outfit adding to the mood and sentiment.
(b)		The image provided is a cartoon of a minion holding a sign that says "Swap to help us find the stolen smiles". The key elements of emotion in this image are: 1. Happiness - The minion is smiling and holding a sign that says "Swap to help us find the stolen smiles", indicating happiness. 2. Excitement - The minion's body language and facial expression suggest excitement and enthusiasm for finding the stolen smiles. 3. Determination - The minion's grip on the sign and his posture convey a sense of determination to find the stolen smiles. The sentiment of this image is positive, as the minion is trying to help find something that has been taken. The sentiment can be described as: 1. Helpful - The minion is actively trying to help find the stolen smiles, which conveys a sense of helpfulness. 2. Kind - The minion's smile and friendly demeanor suggest kindness towards the person who has lost the smiles. 3. Optimistic - The image conveys an optimistic tone, as the minion believes he can help find the stolen smiles.
(c)		The image shows a group of people taking selfies with their mouths open and eyes closed. There are various food items such as burgers, fries, chicken wings, and drinks in the foreground. In the background, there is a park with trees, grass, and flowers. The emotion expressed in the image is happiness and excitement. The people in the image are smiling and laughing, and they are enjoying their meal. The food items also contribute to the happy atmosphere, as they are all delicious and satisfying. The sentiment in the image is positive. The people in the image are engaged in a fun activity and seem to be enjoying themselves. The food items are also positive, as they are all tasty and satisfying. The humor aspect of the image is that the people are taking selfies while eating and enjoying their meal. It is humorous because it highlights the fact that food is something that brings people together and makes them happy.

Figure A1. Illustration of MiniGPT-4 Text Generation Structures

doi:10.6342/NTU202501318