

## A BASELINES

S2S [34], develops an image captioning model that uses a CNN to encode the input image into a fixed-length feature vector. This vector is then used by a Transformer architecture to generate captions, as detailed in Section III-C. The generated captions are visualized on the input images by considering the size of the images and the texts, placing the captions at the bottom of the images to produce the final meme images, as described in Section III-D.

Dank Learning [2] introduces a new meme generation system that produces a humorous and relevant caption for any given image. The system can also generate content based on a user-defined label related to the meme template. It uses a pretrained Inception-v3 network to return an image embedding, which is fed to an attention-based deep-layer LSTM model to create the caption, inspired by the Show and Tell Model.

Transformer [30] is an end-to-end neural and probabilistic architecture for meme generation. It consists of a meme template selection module to identify a compatible image for the input sentence, forming part of the meme generation process.

MEMEIFY [33] utilizes the transformer-based GPT-2 architecture as its base language generative model. It includes class information for different memes by prepending a meme caption with its class name, assisting in generating class-specific captions. The GPT-2's self-attention capabilities, large-scale generative pre-training, and adaptability to multiple tasks enable it to express humor in the text effectively.

BLIP-2-7B [17] is a highly efficient and powerful vision-language pre-trained model that achieves impressive performance on various tasks by leveraging off-the-shelf frozen image encoders and language models, while maintaining significantly fewer trainable parameters compared to existing methods.

MiniGPT-4-7B [39] is a powerful vision-language model that aligns a frozen visual encoder with a frozen language model, achieving advanced multi-modal capabilities such as detailed image description generation, website creation from hand-drawn drafts, and other emerging capabilities like story and poem writing, and teaching cooking based on food photos.

InstructBLIP-7B [12] is a vision-language model that conducts systematic and comprehensive vision-language instruction tuning based on pretrained BLIP-2 models. It achieves state-of-the-art zero-shot performance on various tasks, outperforming BLIP-2 and larger Flamingo models.

LlVA-7B [23] is a large multimodal model that integrates a vision encoder and a language model. It is trained end-to-end using language-only GPT-4 to generate multimodal language-image instruction-following data.

LlVA-1.5-7B [22] is a large multimodal model that demonstrates the surprising power and efficiency of the fully-connected vision-language cross-modal connector. By using CLIP-ViT-L-336px with an MLP projection and incorporating academic-task-oriented VQA data with simple response formatting prompts, it establishes stronger baselines and achieves state-of-the-art performance across 11 benchmarks.

Unified-IOXL-2B [25] is a unified model that can perform a wide range of AI tasks encompassing computer vision, vision-and-language, and natural language processing tasks. It overcomes the challenges posed by the heterogeneous inputs and outputs of each task by representing them as sequences of discrete vocabulary tokens.

Shikra-7B [9] is designed to handle spatial coordinate inputs and outputs in natural language, enabling referential dialogue and various vision-language tasks. It features a simple architecture with a vision encoder, alignment layer, and LLM, eliminating the need for extra vocabularies, position encoder, or external plug-in models.

Qwen-VL-Chat-7B [3] is part of the Qwen-VL series, a set of large-scale vision-language models designed to perceive and understand both text and images. It incorporates a visual receptor, input-output interface, 3-stage training pipeline, and multilingual multimodal cleaned corpus to enhance its visual capacity.

GPT4v<sup>6</sup>, presented by OpenAI, is a model that can accept images as inputs and generate captions, classifications, and analyses.

VTfM [27] stands for Video-Text Fusion Model, where text and video features are fused before being fed into the classifier.

MSAM [27] stands for Multimodal Self Attention Model. It uses BERT to obtain text encoding representations for each dialogue turn and C3D for video encoding representation, given a sequence of dialogue turns.

HKT [14] is the Humor Knowledge Enriched Transformer Model. First, it creates unimodal representations of the punchline conditioned on the context. Then, the humor-centric feature-enriched language and non-verbal embedding undergo Bimodal Cross Attention layers to create multimodal fusion.

## B LIMITATIONS AND ETHIC STATEMENT

Despite the promising results, our study comes with certain limitations. First and foremost, our XMECAP framework might not encapsulate the full depth of humor nuances present in memes, as humor is often subjective and multi-dimensional. Secondly, while the XMECAP framework excels in handling image-text relationships, it might be susceptible to cultural biases or overlook region-specific meme templates. Furthermore, we acknowledge that the granularity of image-text relationship understanding can be further improved, as meme comprehension isn't just about discrete elements but also their implicit connotations. Addressing these limitations will be crucial in crafting more holistic and universally applicable meme analysis tools.

Ethical considerations are paramount in the realm of research and development, especially when analyzing internet phenomena like memes, which often reflect societal values and biases. In this study, we have made every effort to ensure that our XMECAP and analysis respect the diverse cultures and beliefs represented in the meme dataset. We acknowledge that memes can be sensitive and are susceptible to being interpreted in multiple ways. The intention of our XMECAP framework is purely academic and aims to understand the interplay of images and text in meme caption generation, without promoting or endorsing any particular sentiment or ideology. We always prioritize ethical considerations and avoid propagating potentially harmful or misleading content.

<sup>6</sup><https://chat.openai.com/>