

## 1045 A BASELINES

1046 S2S [34], develops an image captioning model that uses a CNN to encode the  
1047 input image into a fixed-length feature vector. This vector is then used by a  
1048 Transformer architecture to generate captions, as detailed in Section III-C.  
1049 The generated captions are visualized on the input images by considering  
1050 the size of the images and the texts, placing the captions at the bottom of  
1051 the images to produce the final meme images, as described in Section III-D.

1052 Dank Learning [2] introduces a new meme generation system that pro-  
1053 duces a humorous and relevant caption for any given image. The system  
1054 can also generate content based on a user-defined label related to the meme  
1055 template. It uses a pretrained Inception-v3 network to return an image  
1056 embedding, which is fed to an attention-based deep-layer LSTM model to  
1057 create the caption, inspired by the Show and Tell Model.

1058 Transformer [30] is an end-to-end neural and probabilistic architecture  
1059 for meme generation. It consists of a meme template selection module to  
1060 identify a compatible image for the input sentence, forming part of the  
1061 meme generation process.

1062 MEMEIFY [33] utilizes the transformer-based GPT-2 architecture as its  
1063 base language generative model. It includes class information for different  
1064 memes by prepending a meme caption with its class name, assisting in  
1065 generating class-specific captions. The GPT-2's self-attention capabilities,  
1066 large-scale generative pre-training, and adaptability to multiple tasks enable  
1067 it to express humor in the text effectively.

1068 BLIP-2-7B [17] is a highly efficient and powerful vision-language pre-  
1069 trained model that achieves impressive performance on various tasks by  
1070 leveraging off-the-shelf frozen image encoders and language models, while  
1071 maintaining significantly fewer trainable parameters compared to existing  
1072 methods.

1073 MiniGPT-4-7B [39] is a powerful vision-language model that aligns a  
1074 frozen visual encoder with a frozen language model, achieving advanced  
1075 multi-modal capabilities such as detailed image description generation,  
1076 website creation from hand-drawn drafts, and other emerging capabilities  
1077 like story and poem writing, and teaching cooking based on food photos.

1078 InstructBLIP-7B [12] is a vision-language model that conducts systematic  
1079 and comprehensive vision-language instruction tuning based on pretrained  
1080 BLIP-2 models. It achieves state-of-the-art zero-shot performance on various  
1081 tasks, outperforming BLIP-2 and larger Flamingo models.

1082 LLaVA-7B [23] is a large multimodal model that integrates a vision  
1083 encoder and a language model. It is trained end-to-end using language-only  
1084 GPT-4 to generate multimodal language-image instruction-following data.

1085 LLaVA-1.5-7B [22] is a large multimodal model that demonstrates the  
1086 surprising power and efficiency of the fully-connected vision-language  
1087 cross-modal connector. By using CLIP-ViT-L-336px with an MLP projection  
1088 and incorporating academic-task-oriented VQA data with simple response  
1089 formatting prompts, it establishes stronger baselines and achieves state-of-  
1090 the-art performance across 11 benchmarks.

1091 Unified-IOXL-2B [25] is a unified model that can perform a wide range  
1092 of AI tasks encompassing computer vision, vision-and-language, and nat-  
1093 ural language processing tasks. It overcomes the challenges posed by the  
1094 heterogeneous inputs and outputs of each task by representing them as  
1095 sequences of discrete vocabulary tokens.

1096 Shikra-7B [9] is designed to handle spatial coordinate inputs and out-  
1097 puts in natural language, enabling referential dialogue and various vision-  
1098 language tasks. It features a simple architecture with a vision encoder,  
1099 alignment layer, and LLM, eliminating the need for extra vocabularies,  
1100 position encoder, or external plug-in models.

1101 Qwen-VL-Chat-7B [3] is part of the Qwen-VL series, a set of large-scale  
1102 vision-language models designed to perceive and understand both text and  
1103 images. It incorporates a visual receptor, input-output interface, 3-stage  
1104 training pipeline, and multilingual multimodal cleaned corpus to enhance  
1105 its visual capacity.

1103 GPT4v<sup>6</sup>, presented by OpenAI, is a model that can accept images as  
1104 inputs and generate captions, classifications, and analyses.

1105 VTfM [27] stands for Video-Text Fusion Model, where text and video  
1106 features are fused before being fed into the classifier.

1107 MSAM [27] stands for Multimodal Self Attention Model. It uses BERT to  
1108 obtain text encoding representations for each dialogue turn and C3D for  
1109 video encoding representation, given a sequence of dialogue turns.

1110 HKT [14] is the Humor Knowledge Enriched Transformer Model. First,  
1111 it creates unimodal representations of the punchline conditioned on the  
1112 context. Then, the humor-centric feature-enriched language and non-verbal  
1113 embedding undergo Bimodal Cross Attention layers to create multimodal  
1114 fusion.

## 1115 B LIMITATIONS AND ETHIC STATEMENT

1116 Despite the promising results, our study comes with certain limitations.  
1117 First and foremost, our XMECAP framework might not encapsulate the full  
1118 depth of humor nuances present in memes, as humor is often subjective  
1119 and multi-dimensional. Secondly, while the XMECAP framework excels in  
1120 handling image-text relationships, it might be susceptible to cultural biases  
1121 or overlook region-specific meme templates. Furthermore, we acknowledge  
1122 that the granularity of image-text relationship understanding can be further  
1123 improved, as meme comprehension isn't just about discrete elements but  
1124 also their implicit connotations. Addressing these limitations will be crucial  
1125 in crafting more holistic and universally applicable meme analysis tools.

1126 Ethical considerations are paramount in the realm of research and devel-  
1127 opment, especially when analyzing internet phenomena like memes, which  
1128 often reflect societal values and biases. In this study, we have made every  
1129 effort to ensure that our XMECAP and analysis respect the diverse cultures  
1130 and beliefs represented in the meme dataset. We acknowledge that memes  
1131 can be sensitive and are susceptible to being interpreted in multiple ways.  
1132 The intention of our XMECAP framework is purely academic and aims to  
1133 understand the interplay of images and text in meme caption generation,  
1134 without promoting or endorsing any particular sentiment or ideology. We  
1135 always prioritize ethical considerations and avoid propagating potentially  
1136 harmful or misleading content.

1137  
1138  
1139  
1140  
1141  
1142  
1143  
1144  
1145  
1146  
1147  
1148  
1149  
1150  
1151  
1152  
1153  
1154  
1155  
1156  
1157  
1158  
1159  
1160  
<sup>6</sup><https://chat.openai.com/>