Supplementary Material of MIP-GAF: A MLLM-annotated Benchmark for Most Important Person Localization and Group Context Understanding

S. Madan[1], S. Ghosh[2], L. R. Sookha[1], M.A. Ganaie[1], R. Subramanian[1,3], A. Dhall[4], T. Gedeon[2]
[1]IIT Ropar, [2] Curtin University, [3] University of Canberra, [4] Monash University
{surbhi.19csz0011,lownish.23csz0010,mudasir,s.ramanathan}@iitrpr.ac.in,
{shreya.ghosh,tom.gedeon}@curtin.edu.au, abhinav.dhall@monash.edu.au

Figure 1. *More Sample Images from the MIP-GAF Dataset.* Additional examples from the MIP-GAF dataset along with their corresponding MIPs. Row (1) represents negative emotion images, row (2) represents positive emotion images, and row (3) represents neutral emotion images. These examples illustrate the diversity and complexity of scenes, emphasizing the varying contexts in which MIPs must be identified.

1. MIP-GAF: Additional MIP Examples

We have presented additional example images from the MIP-GAF dataset (Figure 1) to illustrate the diversity in scenes and the complexity involved in identifying the MIP in real-world scenarios. These examples are categorized into three emotional settings: Positive, Negative, and Neutral (**Note:** Emotion information is not used anywhere in the paper but has been kept as meta-data for future work).

- For the Negative emotion images, we have a scene where a mother scolds her child, with the mother marked as the MIP since she is the primary caregiver. Another example is a protest scene featuring a central figure who could be the leader or activist, marked as the MIP. Additionally, there is a roadside scene showing two poor women selling vegetables, with the woman whose face is visible marked as the MIP. Further, we include a movie scene where the person holding a rifle is marked as the MIP, suggesting he could be the main character. Lastly, we present another march scene where a woman holding a banner is marked as the MIP, as her presence and the banner contribute significantly to the image's message.

- For the Positive emotion images, we have an image featuring a politically dominant personality, marked as the MIP due to being the President of the United States. Another example is a wedding scene where the groom sitting on the chair is marked as the MIP, with his bride standing beside him. We also show a stage performance scene where pop star and singer Ariana Grande is marked as the MIP. Additionally, there are group images where the person at the center is marked as the MIP for being the center of attention.

- For the Neutral emotion images, we have a rally scene where Indian Prime Minister Narendra Modi is marked as the MIP due to his political prominence. Another example is a protest scene against the war in Syria, where a girl giving an interview is marked as the MIP because the microphone is close to her, and she is reading something in front of a group of people. Additionally, there is a classroom scene where a teacher standing behind the students is marked as the MIP. We also show a business meeting setup where multiple people are looking at a central figure, marking that person as the MIP for being the center of attention.

2. Experiments and Results

2.1. More Details on MIP-CLIP: Stage 1

Given an input image I and a text description of MIP as L (Figure 2), the process begins by encoding the visual features (V_I) using a vision encoder [1] and the textual features (T_L) using a text encoder [4]:

$$V_I = ResNet50(I)$$
$$T_L = BERT(L)$$

Since we are working with visual features, the image features are down-sampled by a factor of k in both the width and height of the image:

$$Height = Height_I/k$$
$$Width = Widtht_I/k$$

Next, both the visual and textual embeddings are projected into a common latent space with dimension d termed C_d, using a projection layer.

To align the domain differences, we enhance the visual features using text features and the text features using visual features. This enhancement facilitates the classification process: enhancing text features with visual information helps in classification, while enhancing visual features with textual information aids in localization [2]. An affine transformation is then applied to both the visual (A_v) and text features (A_t) using the following equations:

$$A_t = SoftMax((VW_1^v) \otimes (TW_2^t)^T)/\sqrt{C_d}$$
$$A_v = SoftMax((TW_1^t) \otimes (VW_2^v)^T)/\sqrt{C_d}$$

Where, W_*^v and W_*^t are learnable parameters. These affine transformations are used to create the final vision and textual features as described by the following equations:

$$\text{Text}_{\text{Final}} = A_t^T \otimes (VW_3^v)$$
$$\text{Vision}_{\text{Final}} = \text{Re}(A_v^T \otimes (TW_3^t))$$

This stage is trained in a contrastive classification manner, where positive and negative descriptions associated with the image are classified. A response map is generated for each description (both positive and negative, denoted as R_p and R_N , and an image-level score (y_j) is computed for each description in a contrastive manner. The goal is to make $y_j = 1$ for positive image-description pairs and $y_j = 0$ for negative image-description pairs.

2.2. More Qualitative Evaluations

Figure 3 presents a performance comparison of various state-of-the-art models on the MIP-GAF, MS, and NCAA datasets. In the figure, the dotted line represents the predicted bounding box, while the solid line indicates the ground truth. The results reveal that our MIP-GAF dataset poses significant challenges, necessitating the development of more robust algorithms. Unlike other datasets, the MIP in MIP-GAF is rarely positioned at the center, and the largest

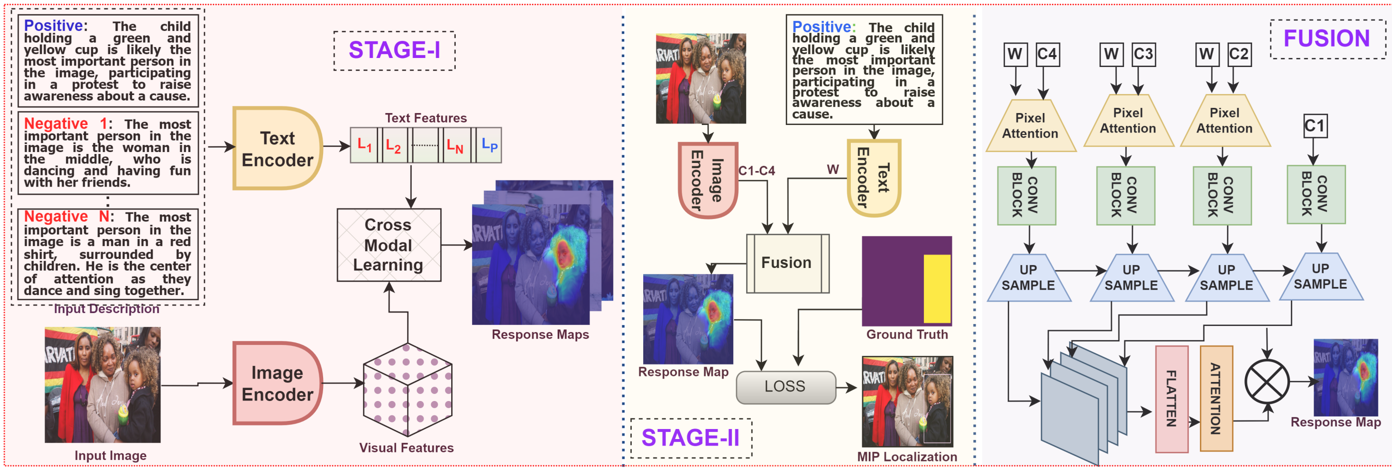

Figure 2. *Our proposed MIP-CLIP framework.* Stage 1: It learns to classify text inputs and uses positive expressions to locate the MIP on response maps. Stage 2: Trained image and text encoders generate feature maps, and a **fusion** model localizes MIP using response maps.

face is not always marked as the MIP. Additionally, the POINT model frequently struggles to accurately identify the MIP. In contrast, our MIP-CLIP method demonstrates superior performance by effectively utilizing contextual information and scene descriptions to locate the MIP. This highlights the complexity of our dataset and the difficulties inherent in identifying MIPs within real-world images.

3. Other Details

3.1. MLLM Failed Cases

We employ the KOSMOS-2 [3] model as a Multimodal Large Language Model (MLLM) to annotate MIP in the first round. However, the MLLM encountered considerable confusion in various complex scenarios, some of which are depicted in Figure 4 . Below are detailed examples of where the MLLM struggled:

1. **Multiple Individuals in Similar Attire:** The MLLM is unable to accurately identify MIP when there are several people in the image wearing similar formal attire or the same type of sash. This situation often occurs in professional meetings or celebratory events, where distinguishing between individuals can be difficult due to the uniformity of their attire or accessories.
2. **Meeting Setup with Focused Person Not Visible:** The model has difficulty when the image depicts a meeting setup where all participants are looking at a particular person, but that person is not visible in the frame. This creates ambiguity for the MLLM, as it relies on visual cues that are absent.
3. **Hands Prominently Positioned:** In images where hands are prominently positioned towards the camera, the MLLM gets confused. The focus on hands instead of faces or bodies disrupts the model's ability to correctly identify and annotate the MIP.
4. **Group Photos with Uniform Poses:** The model encounters issues in group photos where all participants are looking at the camera in the same way. This uniformity in pose and direction makes it challenging for the MLLM to single out and annotate the MIP accurately.

These examples clearly illustrate the limitations of the MLLM in handling complex visual contexts. Such scenarios underscore the necessity for human intervention to ensure accurate annotation and identification of MIP, highlighting the complementary role of human oversight in conjunction with automated models.

3.2. Face-API

Face-api.js [5] is a powerful, browser-based face recognition library built on TensorFlow.js, designed for seamless integration into web applications. It offers a comprehensive suite of features, including face detection for single and multiple faces, providing precise bounding box coordinates, and face recognition with matching capabilities against known faces using unique face descriptors. The library also supports facial landmarks detection with a 68-point model, aiding in face alignment, and can recognize various facial expressions and emotions in real-time video streams. Additionally, it predicts age and classifies gender of detected faces. Optimized for performance, Face-api.js utilizes WebGL for real-time processing and works efficiently across modern web browsers on desktops, laptops, tablets, and smartphones. It includes pre-trained models for different tasks and supports fine-tuning or training

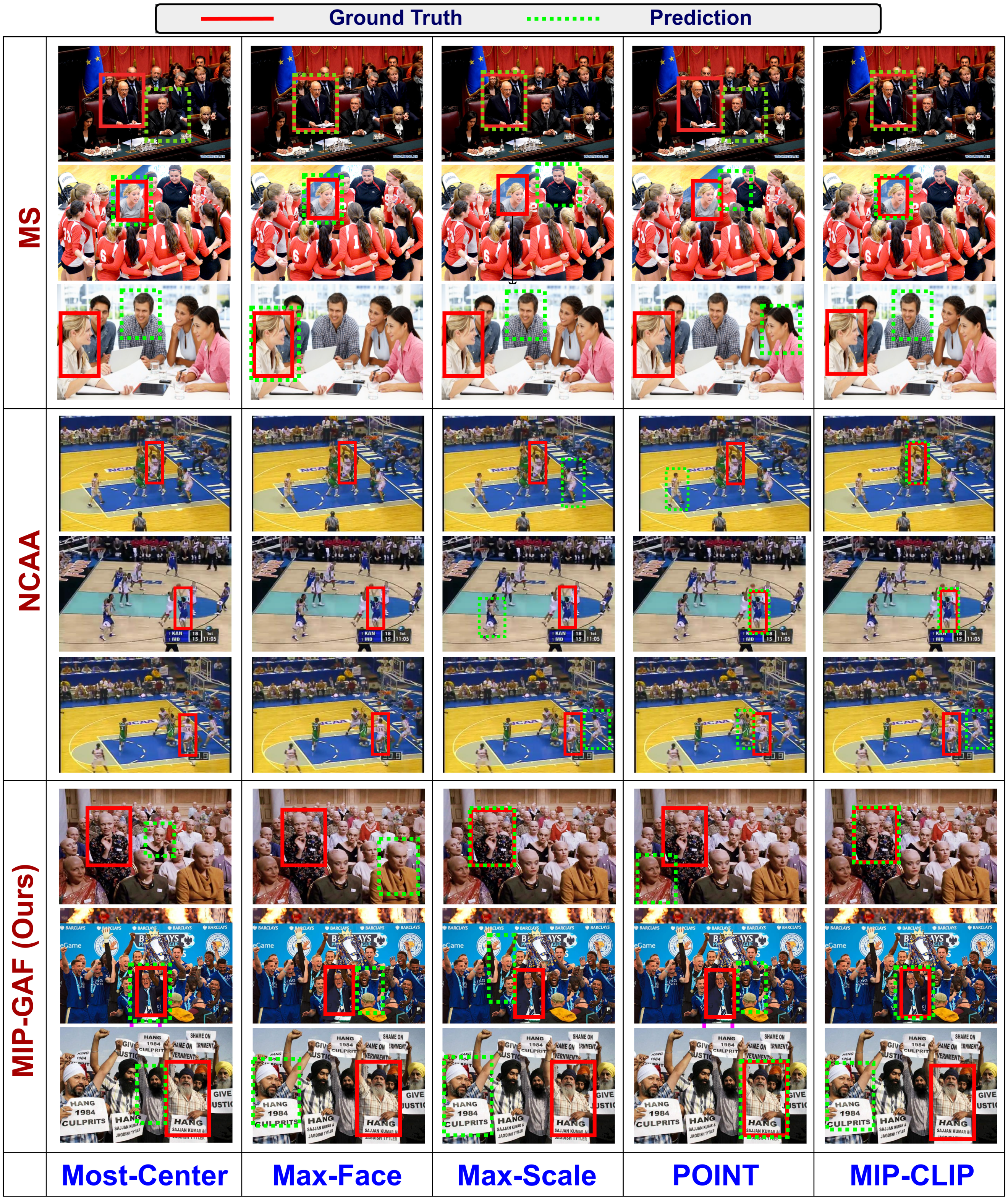

Figure 3. *Qualitative Analysis.* We compare the output of different off-the-shelf methods on MS, NCAA, and MIP-GAF datasets. Here, the dotted line(green) indicates the predicted bounding box and the solid line (red) bounding box indicates the ground truth.

Figure 4. *MLLM Failed Cases.* Instances where MLLM failed to annotate the MIP. These scenarios include: multiple individuals in similar attire or wearing the same sash, meeting setups where the focused person is not visible, hands prominently positioned, and group photos with uniform poses. Different colors of bounding boxes (bboxes) indicate that MLLM marked all of these bboxes as MIP in the given figures. The colors of the bboxes are random and hold no significance.

custom models using TensorFlow.js, ensuring customization and extensibility.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[2] Fang Liu, Yuhao Liu, Yuqiu Kong, Ke Xu, Lihe Zhang, Baocai Yin, Gerhard Hancke, and Rynson Lau. Referring image segmentation using text supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22124–22134, 2023. 2

[3] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 3

[4] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[5] Vladmandic. Vladmandic/face-api: Faceapi: Ai-powered face detection & rotation tracking, face description & recognition, age & gender & emotion prediction for browser and nodejs using tensorflow/js. 3