A Heterogeneous Multimodal Graph Learning Framework for Recognizing User Emotions in Social Networks

Sree Bhattacharyya, Shuhua Yang, James Z. Wang
College of Information Sciences and Technology
The Pennsylvania State University, University Park
{*sfb6038, sky5341, jzw11*}*@psu.edu*

***Abstract*—The rapid expansion of social media platforms has provided unprecedented access to massive amounts of multimodal user-generated content. Comprehending user emotions can provide valuable insights for improving communication and understanding of human behaviors. Despite significant advancements in Affective Computing, the diverse factors influencing user emotions in *social networks* remain relatively understudied. Moreover, there is a notable lack of deep learning-based methods for predicting user emotions in social networks, which could be addressed by leveraging the extensive multimodal data available. This work presents a novel formulation of personalized emotion prediction in social networks based on heterogeneous graph learning. Building upon this formulation, we design *HMG-Emo*, a Heterogeneous Multimodal Graph Learning Framework that utilizes deep learning-based features for user emotion recognition. Additionally, we include a dynamic context fusion module in HMG-Emo that is capable of adaptively integrating the different modalities in social media data. Through extensive experiments, we demonstrate the effectiveness of HMG-Emo and verify the superiority of adopting a graph neural network-based approach, which outperforms existing baselines that use rich hand-crafted features. To the best of our knowledge, HMG-Emo is the first multimodal and deep-learning-based approach to predict personalized emotions within online social networks. Our work highlights the significance of exploiting advanced deep learning techniques for less-explored problems in Affective Computing.**

***Index Terms*—Emotion Recognition, Multimodal AI, Social Networks, Heterogeneous Graphs, Graph Attention**

I. Introduction

Advancements in deep learning and artificial intelligence over the last two decades have sparked significant research interests in methods for automatic emotion recognition that utilize multimodal data. Existing research in emotion recognition can be broadly categorized into evoked and expressed emotion recognition [2]. Evoked emotion refers to the emotions elicited in individuals when interacting with certain external stimuli, regardless of whether those emotions are explicitly expressed. Expressed emotion recognition, on the other hand, regards individuals as the source of specific emotional actions, conveying their feelings through facial expressions, verbal expressions, or body movements. When studying evoked emotion, researchers often focus solely on the external stimuli that prompt the corresponding emotional change. However, in real-life scenarios, evoked emotion can be a consequence of multiple factors that have a simultaneous impact. Understanding emotions within such complex contexts requires the integration of diverse information from multiple modalities or contexts, leading to the development of context-aware emotion recognition (CAER) tasks.

Fig. 1. Example from the IESN [1] dataset. The emotion labels are dependent on both the image contents and the user comments, showing the complex nature of emotion interpretation.

While most CAER methods for evoked emotion recognition focus on contextual information from the external stimuli, such as historical context in conversations [3], scene and object information in images [4], [5], or a fusion of multiple modalities from video data [6], there remains a gap in considering contexts from the individuals whose emotions are evoked. This information, however, is particularly critical in social networks, where user contexts play a pivotal role. In social media platforms, user-specific context can be extremely important, as individual users coming from various backgrounds constantly interact with visual content on the platform and actively get influenced by their connections with other users [7]. The challenge in emotion recognition is further amplified by the possibility of varied interpretations of the same media by different individuals, leading to personalized evoked emotions. For instance, consider the scenario depicted in Fig. 1. We can observe that the emotions evoked by the same image can vary among different users. It is only when contextual cues, such as user comments alongside the images, are taken

979-8-3315-1643-7/24/$31.00 ©2024 IEEE

into account that personalized emotion assessments become meaningful. Additionally, leveraging user-related information, such as social ties, can provide valuable context, as community dynamics often significantly influence individual emotions [8].

Investigating personalized emotions within social networks can potentially benefit numerous downstream tasks. To start with, fine-grained knowledge of user behaviors in social media has multifarious applications such as content recommendations, polarity detection, and content moderation [9]. Since emotions are often associated with opinion mining [10], [11], examining emotions within social media platforms can also be used to gauge user sentiment on global matters. Additionally, the evolving emotional states of users over time are important for identifying social media-related mental health issues, especially in unusual circumstances like the pandemic, which saw a sharp spike in digital engagement [12].

Several researchers utilized social media data to investigate affective information drawing on various modalities of multimedia content [13]. However, few approaches address the problem from a user-centric perspective, predicting personalized emotions evoked in users [1], [7], by including information directly associated with users as an influencing factor. Existing user-centric methods rely only on low-level hand-crafted features from different modalities and probabilistic or deterministic graph modeling for personalized emotion prediction. The power of graph-based deep learning has not been exploited adequately to address personalized emotion recognition for users in social networks. Therefore, in this work, we present the first deep graph learning-based framework for personalized user emotion recognition in social networks. We create a framework that allows features learned from different modalities to be combined adaptively in the form of a user-media graph, wherein the features are further refined through graph learning. We utilize the only public dataset for personalized emotion prediction [1] to test the effectiveness of our framework, as it includes information about users and the images they upload and interact with. We also include a comparison of our method with the existing baseline [1] that employs multi-task hypergraph learning. Our main **contributions** can be summarized as follows:

- We introduce a novel formulation for the problem as an edge classification task in a heterogeneous user-image graph, which enables the intuitive and easy use of advanced graph learning methods to approach personalized emotion recognition in social networks.
- We create **HMG-Emo**, a Heterogeneous, Multimodal, Graph Learning framework that utilizes an adapted Graph Attention Network [14] for emotion recognition. It accumulates information from multiple modalities simultaneously and includes a plug-and-play dynamic context combination module to attend adaptively to different modalities during the emotion classification task.
- With thorough experiments on the well-established public dataset in this domain [1], we verify that our framework outperforms the existing baselines for emotion classification. Furthermore, our method does not require high-level, hand-crafted features. We also demonstrate the effectiveness of using multiple contextual sources of information and the adaptive combination module.
- Through extensive ablation experiments, we further validate the robustness of different components in the proposed framework. We also generatively augment part of the benchmark dataset, to examine the importance of using multiple factors in emotion prediction.

II. Related Work

In this section, we highlight recent related work on context-aware and multimodal emotion recognition and introduce studies of emotion in the context of social networks.

A. Emotion Recognition: Context-Aware and Multimodal

CAER has seen the use of both unimodal and multimodal data for automatically recognizing emotions with a situated perspective. Among published work that focuses on using the *visual modality* singularly, recent approaches adopt the fusion of contextual information in various ways, such as through expanding the original emotion label space into a combined emotion-context matrix [15], using information from the full image along with other specific components like the human body or a segmentation map of an image [5]. More recent approaches use background knowledge in the form of spatial information [16], object and scene information [4], and human body features [17].

In the *modality of text*, applications such as emotion recognition in conversations (ERC) have seen the use of historical information from the conversation as context, using self and inter-speaker dependency to guide the task of identifying emotions [3]. Recently, several unique approaches have exploited context dependency in conversations. SACL [18] uses contextual negative samples for improved robustness in ERC, and another approach learns separate features based on whether a particular training instance should be considered context-dependent or independent to mitigate any impact from noisy contextual information [19]. Recent work has also included meta-information such as discourse structure in conversations [20] for emotion recognition.

For *video data*, a wide variety of contextual information has been used. This includes using features such as the appearance and motion of subjects in videos [21], face and gait information, along with background scene information and depth maps for video frames [6]. Other approaches have used dense optical flow [22], body language and movements [23] [24] [25] [26], or facial expressions [27]. The development of video-based context-aware datasets [28] has also propelled the growth of such methods. However, recent years have seen a marked shift to using all modalities in conjunction, to create an emotion recognition pipeline that resembles the human process of understanding emotions even more closely.

When emotion recognition uses *multiple modalities*, each modality could be considered to serve as complementary contextual information for the other modalities. Relatively

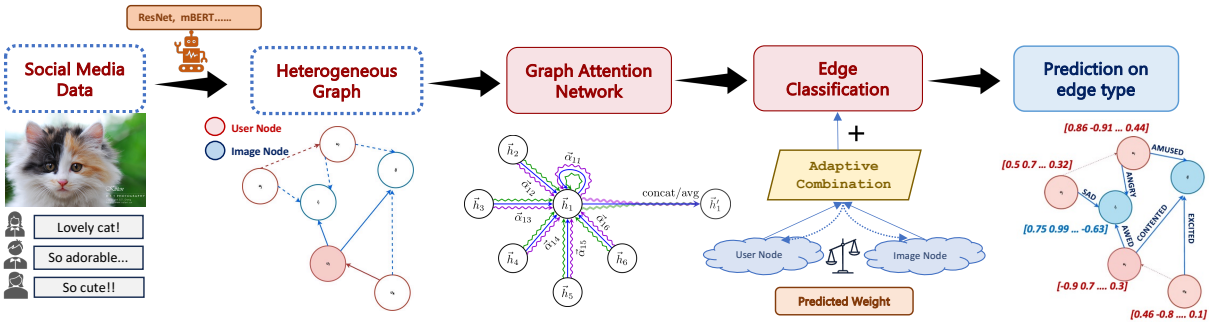

Fig. 2. An overview of the complete methodology pipeline.

earlier approaches in using multimodal features simultaneously include the introduction of a Tensor Fusion Network to learn intra-modality and inter-modality dynamics for inputs from three different modalities [29], and models with gating units storing speaker-specific, and multimodal, information in ERC [30]. More recent approaches have focused on the complexities of the interaction of multiple modalities. This includes the use of cross-modal context fusion networks [31], pretrained multimodal models for feature extraction [32], contrastive learning on multimodal features [33], and cross-modal attention-based methods [34].

B. Emotion Recognition in Social Networks

With social media being ubiquitous today, emotion recognition on social media content has received widespread attention in the field of affective computing. There have been popular datasets for emotion recognition that utilize social media platforms as data sources, including Twitter [13], [35], [36], Instagram and Flickr [37]. However, most of these datasets do not consider user context, such as user preferences and connections. Some approaches also attempt to examine emotions by combining multiple modalities such as text, images, and videos from social media [38], [39]. Two earlier approaches attempt to model user emotions holistically from a user-centered view, considering information including the user contacts, groups, and sequential order of users viewing different images in Flickr, along with studying the multimedia content [1], [7]. One work uses hypergraph-based learning, devising a learning algorithm that explores the correlation between the different user relations and images [1], while the other work uses statistical modeling to study the effect of social relations on user affect. Essentially, the use of deep learning for user emotion recognition remains under-explored. Further, utilizing deep graph learning for multimodal and personalized emotion recognition in social networks remains an open challenge.

III. Methodology

In this section, we give a detailed description of the proposed formulation of the personalized user emotion prediction problem and our framework. Fig. 2 provides a visual reference for the complete steps in the framework.

A. Problem Formulation

Our goal is to develop a framework to predict personalized emotional responses from users during their interaction with visual media in a social network. We aim to include user contexts alongside features of the visual stimuli they interact with. This requires a robust framework that can combine multiple factors seamlessly to make reliable predictions for evoked personalized emotions in users.

To formulate this problem, we consider two types of user interactions in a social network: interactions with other users, and interactions with the visual contents, namely, images. Therefore, we utilize two basic entities, the users and the images, to create a heterogeneous interaction graph that can model both interactions simultaneously. We define the heterogeneous graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ represents the set of nodes containing two subsets: $\mathcal{V}_{\text{user}}$ and $\mathcal{V}_{\text{image}}$. The set of edges $\mathcal{E}$ encompasses two types of edges, denoted as edge type *views* and edge type *connects*. Here, $\mathcal{E}_{\text{views}} \subseteq \mathcal{V}_{\text{user}} \times \mathcal{V}_{\text{image}}$ represents connections between users and the images they view on the platform, while $\mathcal{E}_{\text{connects}} \subseteq \mathcal{V}_{\text{user}} \times \mathcal{V}_{\text{user}}$ corresponds to edges connecting users who are contacts or connections of each other.

In our graph representation, the images are considered as independent nodes rather than features of the user nodes. We also embed additional information as node and edge features in $\mathcal{G}$. For example, the extracted image features can be included as the node features for all nodes of type $\mathcal{V}_{\text{image}}$. Likewise, the user comment i when viewing image j can be considered as a feature of the edge $e_{i,j}$, where $e_{i,j} \in \mathcal{E}_{\text{views}}$. For user i and image j, the evoked emotion ground-truth labels are considered as the supervision label for edge $e_{i,j}$, indicating that the evoked emotion comes from viewing image j. Thus, the task of predicting evoked emotions in users becomes a supervised edge classification task on a single type of edge in the graph ($\mathcal{E}_{\text{views}}$), which is done by aggregating information from various sources using the graph structure, node, and edge features. This formulation presents two advantages. Firstly, the users and images exist as independent nodes, allowing for the creation of user features pertaining directly to each user, as described in the next section. The separation of the entities to constitute two different types of interactions also helps in effectively increasing the number of instances to

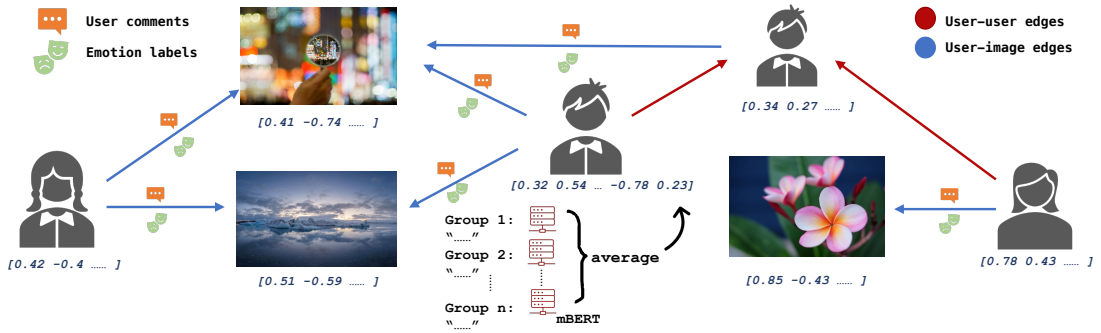

Fig. 3. The structure of the created heterogeneous graph. The feature generation process for a particular user is depicted in one of the user nodes.

use for modeling emotions. Secondly, this formulation makes advanced graph learning methods an intuitive choice to solve the problem.

B. Unimodal Processing of Contexts

We first process each source of contextual information to be included in HMG-Emo as follows:

- *Visual Context*: This refers to the actual visual media, or images, being viewed by the users on the social media platform. We use high-level features obtained using a pretrained Resnet-50 [40] model, moving beyond the low-level, hand-crafted features that have been employed for this purpose in the past. The features obtained are then treated as node attributes for all nodes of the type *image*.
- *User Context*: The user context aims to capture information directly related to the user. This is subdivided into two parts:
  - *User-Image Interaction Context*: We consider the comments left by users on images as the contextual information of a single user-image interaction. We include the comment information as edge features for edges connecting users and images.
  - *General User Context*: Besides incorporating the information of user connections implicitly through the heterogeneous graph structure, we also include information about different special interest groups joined by users on a social media platform. Such groups on photo-sharing websites are usually joined by like-minded users interested in sharing images aligned to a particular theme. We use the textual descriptions of each such special interest group and obtain their features. Then, for every user, we aggregate the features obtained for all such groups the user belongs to, to be used as a node feature for user nodes in the graph. Formally, consider $\mathcal{S}$ to be the set of all possible special interest groups present in the social media platform. Now, for user i, say u_i is a part of all groups in $\mathcal{S}_i$, where $\mathcal{S}_i \subseteq \mathcal{S}$. Along with that, we have a text description t, corresponding to each $s \in \mathcal{S}$. Thus, for user u_i, we calculate the group context feature by the following formulation:

$$g_i = \frac{1}{|\mathcal{S}_i|} \sum_{s \in \mathcal{S}_i} f(s) \,, \tag{1}$$

where f is the feature extractor for each textual description t for all $s \in \mathcal{S}$.

For both the *interaction-level context*, and *general user context*, we thus need to use a text-based feature extractor. We use zero-shot features learned from multilingual-BERT [41], to handle text in multiple languages. The features learned using comments are used as edge features for user-image edges and the features learned from group descriptions are used as node features for all user nodes. Wherever the comment texts are missing, we initialize the edge attribute with a uniform feature tensor.

The reason we rely on popular pre-trained models for zero-shot feature extraction is to ensure that the unimodal feature learning process can be as straightforward as possible, and not require any human supervision. As opposed to previously used methods, learning features from models in a zero-shot manner saves the requirement of domain knowledge for designing hand-crafted low-level features or spending time and resources on retraining models to obtain features, while also providing significant performance improvements in the downstream task of classifying evoked emotion. Fig. 3 shows the graph structure created, along with the process for creating node features for user nodes.

C. Graph Learning

1) Graph Attention Network: Starting with the heterogeneous graph, we include additional information as low-dimensional features learned in a zero-shot manner to add as node and edge attributes. To have an algorithm that aggregates all of this information and predicts edge labels for a single type of edge, we use a graph attention network backbone [14], adapted to suit the structure of our heterogeneous graph. Essentially, the network should assimilate information from the graph structure, node, and edge features, in multiple steps or hops, and update the representation learned for the nodes. Based on the low-dimensional node representations, a classification problem is formulated, that predicts the emotion label for an edge using the representations of its source and destination nodes. We ensure that the learning mechanism uses edge attributes only for user-image edges. Further, it is important to note that the model should aggregate information

over all types of nodes and edges while using supervision only for user-image edges that have ground truth emotion labels. Mathematically, considering a spatial view of the heterogeneous graph, for every layer of graph attention, the following is computed:

$$x_i' = \alpha_{i,i}\mathbf{\Theta}x_i + \sum_{j\in\mathcal{N}(i)} \alpha_{i,j}\mathbf{\Theta}x_j \ , \tag{2}$$

where x_i' represents the updated representation for node x_i and $\alpha_{i,i}$ and $\alpha_{i,j}$ are the attention coefficients for self and neighboring nodes of the node x_i [14]. $\mathbf{\Theta}$ represents the learnable weight matrix for the node interactions, and can be considered the general weight matrix that transforms nodes into high-level feature representations. Specifically, the attention coefficient $\alpha_{i,j}$ is calculated as

$$\frac{\exp\left(f\left(\mathbf{a}_s^\top\mathbf{\Theta}\mathbf{x}_i + \mathbf{a}_t^\top\mathbf{\Theta}\mathbf{x}_j + \mathbf{a}_e^\top\mathbf{\Theta}_e\mathbf{e}_{i,j}\right)\right)}{\sum_{k\in\mathcal{N}(i)\cup\{i\}}\exp\left(f\left(\mathbf{a}_s^\top\mathbf{\Theta}\mathbf{x}_i + \mathbf{a}_t^\top\mathbf{\Theta}\mathbf{x}_k + \mathbf{a}_e^\top\mathbf{\Theta}_e\mathbf{e}_{i,k}\right)\right)} \ , \tag{3}$$

where $f(\cdot)$ is a non-linearity like LeakyRelu, $\mathbf{a}$ is the weight matrix associated solely with the attention mechanism, with $\mathbf{a_s}, \mathbf{a_t}, \mathbf{a_e}$ denoting the weights for attention to self, neighbor and the edge connecting them, respectively. $\mathbf{\Theta}_e$ is the learnable weight matrix for transforming the edge attribute features, used only for user-image edges. In HMG-Emo, each Graph Attention layer is followed by Batch Normalization [42], and a non-linearity of ReLU [43].

Once the individual node representations are learned, they are combined using an adaptive mechanism and then subjected to a classification module. Precisely, beyond the unsupervised graph learning stage, we have the node features $x_i \in \mathcal{R}^{\mathcal{D}}$, where $\mathcal{D}$ is the final dimension for the node features. Then, the classification task becomes to find the mapping:

$$h : g(x_i, x_j) \rightarrow \mathcal{C} \ , \tag{4}$$

where $g(\cdot)$ represents a mechanism to combine the node features of the source and destination node, and $\mathcal{C}$ represents the class of emotion labels. Note that this is carried out only for edges of the *views* type. Our classification module consists of a simple feedforward network, consisting of 3 hidden layers, with non-linear activation in the form of ReLU [43], and Dropout [44] layers after each of them, except the last.

2) Adaptive Combination of Node Features: The node features learned using the graph network have to be fused in some form for the final classification to take place. Common choices from the literature include concatenation, addition, or a dot product. In this work, before combining the features finally through such a mechanism, we pass the features through an adaptive weight prediction module, similar to the attention mechanism [45]. We denote this adaptive combination method as AC. The module works in the form of a single-layered feedforward network, that predicts a single weight coefficient to be applied in a complementary manner to the node features of both the source and destination nodes of any edge being considered for classification. Mathematically, the weight β is calculated in the following way:

$$\beta = \exp(\text{LeakyReLU}(\mathbf{w}_u x_u + \mathbf{w}_i x_i)) \ , \tag{5}$$

TABLE I
STATISTICS OF THE CREATED GRAPH

User Nodes	Image Nodes	User-User Edges	User-Image Edges
108899	85157	1649058	197561

where $x_u \in \mathcal{V}_{user}$ and $x_i \in \mathcal{V}_{image}$, and $\mathbf{w}_u, \mathbf{w}_i$ are learnable scalar weights for the user and image node features respectively. We do not include the edge attributes to be adaptively combined as we initialize missing edge attributes uniformly. The randomly initialized feature tensors might not hold information reliable enough for the final classification step. Once the weight coefficient is obtained, the combination takes place as follows,

$$x_{\text{comb}} = c(\beta x_u, (1-\beta)x_i) \ . \tag{6}$$

Here, $c(\cdot)$ denotes the final combination mechanism.

IV. EXPERIMENTS AND RESULTS

A. Dataset Description

We use the Image-Emotion-Social-Net (IESN) dataset [1] to validate the effectiveness of HMG-Emo. As the sole publicly available dataset for personalized emotion prediction in social networks, IESN stands out by including *contexts about users* from the social media platform Flickr[1], alongside images and ground-truth emotion labels for emotion prediction. The dataset includes eight emotion classes, based on Mikel's model [46] - *Amusement*, *Anger*, *Awe*, *Contentment*, *Disgust*, *Excitement*, *Fear*, and *Sadness*, as well as an *Unknown* emotion category. We use the following diverse information from the dataset:

- *Actual Emotion of User-Image Interaction*: This part of the dataset provides details on users, images, and their interaction such as the timestamp, along with ground truth emotion labels representative of the *actual* emotion evoked in users by viewing the images. These ground truth labels are primarily derived from the user comments by calculating their average Valence, Arousal, and Dominance (VAD) using VAD norms [47].
- *User Group Information*: The dataset provides information about special interest groups on Flickr, where multiple users join in to share images related to a specific topic. This information includes descriptions of the group interests and a membership list.
- *User Information*: Additional user data such as the user contact lists are also available in this dataset.

As IESN [1] was introduced in 2016, some images and comments included in the dataset may no longer exist, or their parent user profiles may have been deleted. We remove such entries. A number of the user-image interactions also contain multiple related emotion labels or an emotion label of *Unknown*. We use only a single ground truth emotion label for a user-image edge and remove instances that are

[1]https://www.flickr.com

TABLE II
EMOTION CLASSIFICATION RESULTS

Method	F1	Precision	Recall
RMTHG (V) [1]	0.44 ± 0.07	0.36 ± 0.16	0.68 ± 0.12
RMTHG (Fusion) [1]	0.60 ± 0.1	0.50 ± 0.1	$\mathbf{0.72 \pm 0.1}$
HMG-Emo$-\{$T, AC$\}$	0.75 ± 0.004	0.92 ± 0.009	0.64 ± 0.001
HMG-Emo$-\{$AC$\}$	0.76 ± 0.003	0.94 ± 0.005	0.64 ± 0.002
HMG-Emo	$\mathbf{0.77 \pm 0.003}$	$\mathbf{0.97 \pm 0.007}$	0.65 ± 0.001

TABLE III
EMOTION CLASSIFICATION WITH COMMENTS GENERATED BY LLAVA

Method	F1	Precision	Recall
HMG-$\mathcal{T}$ comments only	0.30 ± 0.003	0.31 ± 0.006	0.34 ± 0.002
LLaVA comments only	0.36 ± 0.002	0.38 ± 0.003	0.38 ± 0.002

labeled *Unknown*. The comments are also presented only in the form of comment IDs, and the actual text is obtained directly from Flickr, using those IDs. Then, we construct the heterogeneous graph based on the user, image, contact list, and comment information. In case the actual comment text is missing, we initialize the corresponding edge attributes to be a uniform tensor. Table I provides an overview of the statistics of the created graph. Among the 197,561 user-image edges, a large number of edges lack the original comment texts, leading to only 34% comment features available extracted using mBERT [41]. Note that the number of user nodes depicted in Table I can refer to users as either image uploaders, image viewers, or both. Still, they do not necessarily have to be both simultaneously.

B. Experimental Setup

The implementations in all experiments are based on PyTorch Geometric [48]. The models are trained on an NVIDIA A40 GPU node with four GPU cores. Given the dataset's significant class imbalance, we report the weighted F1 score as the classification performance metric. We make an 80:10:10 train-validation-test split and report 3-fold cross-validated scores on the test data, along with the standard deviation across total runs. We use an output dimension size of 64 for the graph network in HMG-Emo. The models are trained for 20 epochs using the Adam optimizer [49], with a base learning rate of 0.005, and cosine annealing scheduler [50]. The batch size used for the experiments is 512. The embedding dimensions chosen for the initial features are 128 for the user, 128 for the image, and 256 for the comment features, chosen empirically. We use five layers of each graph neural network backbone, with the default choice for HMG-Emo being Graph Attention layers [14]. Convergence in training takes approximately 4 hours at the longest.

C. Emotion Prediction Experiments

1) With IESN and HMG-Emo: The primary results using HMG-Emo are presented in Table II. We compare our proposed method with the only baseline in this problem, introduced in [1], which adopted a Rolling Multi-task Hypergraph Learning framework (RMTHG). RMTHG considered multiple factors, such as image features, similarity between users in terms of the comments they make, the groups they join, as well as interaction records such as the timestamps of comments. RMTHG predicts the emotional state of the user when they view a particular image, based on all of these contextual pieces of information. It however considers image features and user information to be part of a single complex vertex in the hypergraph. The variant of RMTHG where only visual features are used is denoted using RMTHG (V). Similarly, for HMG-Emo, the variant *without* comment features or adaptive combination is denoted as HMG-Emo$-\{$T, AC$\}$, and the variant that uses all features - visual, user context, and comments, but not the adaptive combination, is denoted as HMG-Emo$-\{$AC$\}$. Our formulation of the problem achieves a significant boost in the performance of personalized emotion classification, with our complete framework achieving an improvement of 28% over the multimodal baseline, and 75% over the vision-only baseline. The use of multiple features within the graph network, along with an adaptive combination in the classification step further strengthens the model. HMG-Emo tends to have a very low rate of predicting false positives, leading to high precision, as opposed to the significantly high value of recall achieved by the RMTHG method. We speculate that the difference in precision and recall is due to the highly imbalanced nature of the data. The higher values of precision compared to recall signal better performance in the majority class, which can also be understood by the significantly lower number of samples present for the minority class. However, the precision and recall scores achieved by HMG-Emo can be considered substantially high enough for the model to be useful for downstream tasks.

2) Effect of Multimodal Features: As described early in Section IV, the ground truth emotion labels for actual emotions of users, in IESN [1], are obtained primarily from the comments left by the users. This may spark a question as to whether the comments can only be considered sufficient for predicting personalized emotions in users. To verify, we have also conducted an experiment using only the comment features from HMG-Emo. For a fair comparison in classification performance, we use the same feedforward neural network architecture as in the HMG-Emo classification module, described in Section III-C. The first row in Table III shows the results, where HMG-$\mathcal{T}$ denotes the comment features originally used in HMG-Emo as edge attributes. The classification performance is far below what is obtained using a graph-based approach, underlining the need for a multimodal approach.

However, we also consider that the original comment feature set HMG-$\mathcal{T}$ has only 34% of the feature representations learned from actual comment texts, as the actual texts for others were missing, leading to them being filled with randomly sampled feature tensors. To delve deeper and verify whether that is the reason for the poor classification perfor-

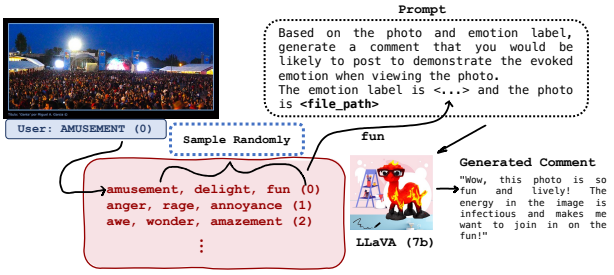

Fig. 4. The prompting method used for LLaVA

mance, we employ LLaVA (7b) [51] to generate actual texts for the missing comments. LLaVA is chosen because it is open-source, and relatively lightweight, making it easier to reproduce the results with minimal resource constraints. For every edge that is missing the original comment features, we use the corresponding images, and ground truth emotion labels to prompt LLaVA to generate comments that are representative of the ground truth emotion. The prompt thus depends solely on the image, and the associated evoked emotion label. This presents a challenge, as multiple edges exist between the same image, and different users, often having the same emotion label. This could lead to the same prompt being used for multiple users, leading to identical comments being generated for different users viewing the same image. To ensure that the comments generated by LLaVA for different users are unique, randomly sampled synonyms of the emotion label name are used in the prompts. The entire process is depicted in Fig. 4. We manually and qualitatively validate a subset of such generations from LLaVA and find them to be highly correlated with the corresponding emotion labels. The comments often contain keywords directly related to the ground truth emotion label, which makes them significantly richer than real-world comments, which may often contain unrelated or noisy information. The same feature extraction method of mBERT [41] is used on the comments generated using LLaVA, and emotion classification is carried out. As can be noted from Table III, even with rich features learned from comments that are highly related to the actual emotion label, the classification performance is largely below par. This stresses the need for using a framework that learns from multiple modalities in tandem, to achieve a holistic understanding of user emotions in social networks.

3) Graph Backbone Modification: We experiment with different backbone graph layers to understand the impact of using a specific graph layer in our model. We compare our model with the counterparts obtained when the Graph Attention layers in them are replaced by GraphSAGE [52] and GraphConv [53] layers. Both GraphSAGE and GraphConv do not however take into consideration the edge attributes by default. Thus, for a fair comparison, we compare them with HMG-Emo with the comment features removed. Note that this still includes the adaptive combination of user and image modalities, hence it is different from HMG $-\{$T, AC$\}$. From Table IV, we can observe that with our framework, there is not much variation in the emotion prediction performance, even when the backbone layers are changed, demonstrating that the improvement in performance does not only come from the choice of graph attention as the backbone.

TABLE IV
ABLATION STUDY FOR DIFFERENT GRAPH BACKBONE

Graph Backbone	F1	Precision	Recall
GAT	0.77 ± 0.002	0.98 ± 0.004	0.65 ± 0.0002
GraphSAGE	0.75 ± 0.004	0.92 ± 0.012	0.65 ± 0.001
GraphConv	0.76 ± 0.001	0.93 ± 0.006	0.66 ± 0.002

TABLE V
ABLATION STUDY FOR NEGATIVE SAMPLING IN TRAINING AND INFERENCE

Method	F1	Precision	Recall
W/ Negative Sampling	0.71 ± 0.002	0.80 ± 0.002	0.64 ± 0.002
W/o Negative Sampling	0.75 ± 0.004	0.92 ± 0.009	0.64 ± 0.001

4) Negative Sampling Strategies: Negative sampling within a graph involves generating potential edges that are not present in the original graph. In our framework, we study the impact of creating negative edges. With negative sampling, the model is trained to classify edges into the 8 emotion classes while also learning to discern some edges as non-existent. This strategy leads to a slight drop in the performance of the model due to the ambiguity of generating negative samples in a dynamic social network. In this case, some non-existent edges can potentially become true edges in the future, making the decision on the existence of edges more challenging. Table V compares the performance of our simplest model (HMG$-\{$T, AC$\}$) with and without negative sampling.

V. CONCLUSION

We present a graph and deep learning-based framework for personalized emotion prediction in social networks. We demonstrate significant gains in performance when utilizing deep features and multiple modalities simultaneously. In conclusion, several promising directions can be considered for future research. Firstly, the proposed framework in this paper can be further enhanced by incorporating additional contexts, such as geo-tags of images and timestamps of posting images or comments, alongside adopting more sophisticated feature extraction techniques. With our method being easily extensible to new modalities of information, the inclusion of diverse data can be seamless. Moreover, given the increasing popularity of Large Language Models (LLMs) and Graph-LLMs, there exists an opportunity to input data into Graph-LLMs for emotion prediction directly. Addressing the challenge of integrating multiple contexts into prompts for such LLMs remains an open area of inquiry. We hope that the insights from this study will inspire future research endeavors in this field.

VI. Ethical Impact Statement

Our research introduces a novel framework for personalized emotion prediction in social networks, leveraging graph and deep learning techniques. While our work contributes to the advancement of affective computing, we recognize the importance of addressing ethical considerations in this domain.

The proposed automatic pipeline and model raise concerns about algorithm transparency and accountability. We acknowledge the challenge of reproducing the same results in the experiments due to the complexity of deep learning models and the reliance on different hyperparameters. Therefore, we provide comprehensive details of the proposed method in this research and plan to make our codes for data collection and experiments publicly available upon acceptance. Additionally, we emphasize that it is necessary to carry out strict evaluation to ensure the reliability and fairness of our method. This includes conducting experiments over multiple runs and providing corresponding error estimates.

Furthermore, the utilization of social media data brings concerns regarding data privacy and user consent. We understand that it is pivotal to address user privacy rights and obtain explicit consent for the collection and analysis of sensitive user data. We confirm that we only use content that is already publicly available from social media and do not infringe upon the privacy of users on the social media platform.

In a supplementary experiment utilizing a Large Language Model (LLM) to generate missing data, we acknowledge that the data includes the biases of the training set for the particular model. We also note that data generated from the LLM for our purpose is not representative of the real-world data quality. Based on this observation, we refrain from training our framework on the augmented dataset to avoid potential misleading performance gains obtained from idealistic data.

Moreover, we note that although this research aims to predict user emotions in social networks for social good, we should not rule out the possibility of the research being used for a malicious purpose, such as manipulating content on social media based on user emotions. Therefore, we advocate for the responsible usage of our research and similar endeavors, emphasizing thorough risk analysis it can pose to users on social media and its usage only in well-charted scenarios that prioritize user welfare.

Acknowledgments

The work was supported in part by the National Science Foundation (NSF) under Grant No. 2234195. This work used cluster computers at the National Center for Supercomputing Applications through an allocation from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by NSF Grants Nos. 2138259, 2138286, 2138307, 2137603, and 2138296. We would also like to express our appreciation to the anonymous reviewers for their constructive feedback.

Some images are incorporated in figures for illustrative purposes and to support the conceptual discussions in this paper. All copyrights remain with their respective owners. The authors acknowledge and appreciate the work of the creators.

References

[1] S. Zhao, H. Yao, Y. Gao, R. Ji, W. Xie, X. Jiang, and T.-S. Chua, "Predicting personalized emotion perceptions of social images," in *Proceedings of the 24th ACM International Conference on Multimedia*, MM '16, (New York, NY, USA), p. 1385–1394, Association for Computing Machinery, 2016.

[2] J. Z. Wang, S. Zhao, C. Wu, R. B. Adams, M. G. Newman, T. Shafir, and R. Tsachor, "Unlocking the emotional world of visual media: An overview of the science, research, and impact of understanding emotion," *Proceedings of the IEEE*, vol. 111, no. 10, pp. 1236–1286, 2023.

[3] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, and A. Gelbukh, "DialogueGCN: A graph convolutional neural network for emotion recognition in conversation," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (K. Inui, J. Jiang, V. Ng, and X. Wan, eds.), (Hong Kong, China), pp. 154–164, Association for Computational Linguistics, Nov. 2019.

[4] S. Nagappan, J. Q. Tan, L.-K. Wong, and J. See, "Context-aware multi-stream networks for dimensional emotion prediction in images," in *2023 IEEE International Conference on Image Processing (ICIP)*, pp. 2480–2484, 2023.

[5] K. Peng, A. Roitberg, D. Schneider, M. Koulakis, K. Yang, and R. Stiefelhagen, "Affect-dml: Context-aware one-shot recognition of human affect using deep metric learning," in *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pp. 1–8, IEEE, 2021.

[6] T. Mittal, P. Guhan, U. Bhattacharya, R. Chandra, A. Bera, and D. Manocha, "Emoticon: Context-aware multimodal emotion recognition using Frege's principle," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14234–14243, 2020.

[7] Y. Yang, J. Jia, S. Zhang, B. Wu, Q. Chen, J. Li, C. Xing, and J. Tang, "How do your friends on social media disclose your emotions?," in *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, AAAI'14, p. 306–312, AAAI Press, 2014.

[8] S. Jin and R. Zafarani, "Emotions in social networks: Distributions, patterns, and models," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, p. 1907–1916, 2017.

[9] K. Ravi and V. Ravi, "A survey on opinion mining and sentiment analysis: tasks, approaches and applications," *Knowledge-based Systems*, vol. 89, pp. 14–46, 2015.

[10] J. Zhao, K. Liu, and L. Xu, "Book review: Sentiment analysis: Mining opinions, sentiments, and emotions by bing liu," *Computational Linguistics*, vol. 42, no. 3, pp. 595–598, 2016.

[11] A. Kumar, P. Dogra, and V. Dabas, "Emotion analysis of twitter using opinion mining," in *2015 Eighth International Conference on Contemporary Computing (IC3)*, pp. 285–290, IEEE, 2015.

[12] Y. Yang, K. Liu, S. Li, and M. Shu, "Social media activities, emotion regulation strategies, and their interactions on people's mental health in COVID-19 pandemic," *International Journal of Environmental Research and Public Health*, vol. 17, no. 23, p. 8931, 2020.

[13] J. Yang, M. Sun, and X. Sun, "Learning visual sentiment distributions via augmented conditional probability neural network," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, p. 224–230, 2017.

[14] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *International Conference on Learning Representations*, 2018.

[15] P. Balouchian and H. Foroosh, "Context-sensitive single-modality image emotion analysis: A unified architecture from dataset construction to cnn classification," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, pp. 1932–1936, 2018.

[16] M.-H. Hoang, S.-H. Kim, H.-J. Yang, and G.-S. Lee, "Context-aware emotion recognition based on visual relationship detection," *IEEE Access*, vol. 9, pp. 90465–90474, 2021.

[17] T. Khargonkar, S. Choudhary, S. Kumar, and B. R. KR, "SeLiNet: Sentiment enriched lightweight network for emotion recognition in images," in *2023 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 1–5, IEEE, 2023.

[18] D. Hu, Y. Bao, L. Wei, W. Zhou, and S. Hu, "Supervised adversarial contrastive learning for emotion recognition in conversations," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (Toronto, Canada), pp. 10835–10852, Association for Computational Linguistics, July 2023.

[19] G. Tu, B. Liang, R. Mao, M. Yang, and R. Xu, "Context or knowledge is not always necessary: A contrastive learning framework for emotion recognition in conversations," in *Findings of the Association for Computational Linguistics: ACL 2023* (A. Rogers, J. Boyd-Graber, and N. Okazaki, eds.), 2023.

[20] D. Zhang, F. Chen, and X. Chen, "Dualgats: Dual graph attention networks for emotion recognition in conversations," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7395–7408, 2023.

[21] Y. Fan, X. Lu, D. Li, and Y. Liu, "Video-based emotion recognition using CNN-RNN and C3D hybrid networks," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, ICMI '16, (New York, NY, USA), p. 445–450, Association for Computing Machinery, 2016.

[22] I. Pikoulis, P. P. Filntisis, and P. Maragos, "Leveraging semantic scene characteristics and multi-stream convolutional architectures in a contextual approach for video-based visual emotion recognition in the wild," in *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pp. 01–08, IEEE, 2021.

[23] P. P. Filntisis, N. Efthymiou, G. Potamianos, and P. Maragos, "Emotion understanding in videos through body, context, and visual-semantic embedding loss," in *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pp. 747–755, Springer, 2020.

[24] Y. Luo, J. Ye, R. B. Adams, J. Li, M. G. Newman, and J. Z. Wang, "ARBEE: Towards automated recognition of bodily expression of emotion in the wild," *International Journal of Computer Vision*, vol. 128, pp. 1–25, 2020.

[25] N. Dael, M. Mortillaro, and K. Scherer, "Emotion expression in body action and posture," *Emotion (Washington, D.C.)*, vol. 12, pp. 1085–101, 11 2011.

[26] H. Gunes and M. Piccardi, "A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior," in *18th International Conference on Pattern Recognition (ICPR'06)*, vol. 1, pp. 1148–1153, 2006.

[27] W. Zheng, J. Yu, R. Xia, and S. Wang, "A facial expression-aware multimodal multi-task learning framework for emotion recognition in multi-party conversations," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15445–15459, 2023.

[28] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza, "Context based emotion recognition using EMOTIC dataset," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 11, pp. 2755–2766, 2020.

[29] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (M. Palmer, R. Hwa, and S. Riedel, eds.), (Copenhagen, Denmark), pp. 1103–1114, Association for Computational Linguistics, Sept. 2017.

[30] D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L.-P. Morency, and R. Zimmermann, "Conversational memory network for emotion recognition in dyadic dialogue videos," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (M. Walker, H. Ji, and A. Stent, eds.), (New Orleans, Louisiana), pp. 2122–2132, Association for Computational Linguistics, June 2018.

[31] X. Zhang and Y. Li, "A cross-modality context fusion and semantic refinement network for emotion recognition in conversation," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13099–13110, 2023.

[32] D. Bose, R. Hebbar, K. Somandepalli, and S. Narayanan, "Contextually-rich human affect perception using multimodal scene information," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, IEEE, 2023.

[33] G. Hu, T.-E. Lin, Y. Zhao, G. Lu, Y. Wu, and Y. Li, "UniMSE: Towards unified multimodal sentiment analysis and emotion recognition," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 7837–7851, 2022.

[34] T. Shi and S.-L. Huang, "MultiEMO: An attention-based correlation-aware multimodal fusion framework for emotion recognition in conversations," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14752–14766, 2023.

[35] Q. You, J. Luo, H. Jin, and J. Yang, "Robust image sentiment analysis using progressively trained and domain transferred deep networks," in *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, p. 381–388, AAAI Press, 2015.

[36] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang, "Large-scale visual sentiment ontology and detectors using adjective noun pairs," in *Proceedings of the 21st ACM International Conference on Multimedia*, MM '13, (New York, NY, USA), p. 223–232, Association for Computing Machinery, 2013.

[37] M. Katsurai and S. Satoh, "Image sentiment analysis using latent correlations among visual, textual, and sentiment views," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2837–2841, 2016.

[38] A. Illendula and A. Sheth, "Multimodal emotion classification," in *Companion Proceedings of the 2019 World Wide Web Conference*, pp. 439–449, 2019.

[39] B. Xu, Y. Fu, Y.-G. Jiang, B. Li, and L. Sigal, "Heterogeneous knowledge transfer in video emotion recognition, attribution and summarization," *IEEE Transactions on Affective Computing*, vol. 9, no. 2, pp. 255–270, 2018.

[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

[41] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (J. Burstein, C. Doran, and T. Solorio, eds.), (Minneapolis, Minnesota), Association for Computational Linguistics, June 2019.

[42] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML '15, p. 448–456, JMLR.org, 2015.

[43] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning*, pp. 807–814, 2010.

[44] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[45] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS '17, (Red Hook, NY, USA), p. 6000–6010, Curran Associates Inc., 2017.

[46] J. A. Mikels, B. L. Fredrickson, G. R. Larkin, C. M. Lindberg, S. J. Maglio, and P. A. Reuter-Lorenz, "Emotional category data on images from the international affective picture system," *Behavior Research Methods*, vol. 37, pp. 626–630, 2005.

[47] A. B. Warriner, V. Kuperman, and M. Brysbaert, "Norms of valence, arousal, and dominance for 13,915 english lemmas," *Behavior Research Methods*, vol. 45, pp. 1191–1207, 2013.

[48] M. Fey and J. E. Lenssen, "Fast graph representation learning with PyTorch Geometric," in *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.

[49] D. P. Kingma and J. L. Ba, "Adam: Amethod for stochastic optimization," in *International Conference on Learning Representations*, pp. 1–15, 2014.

[50] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," in *International Conference on Learning Representations*, 2016.

[51] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 2023.

[52] W. L. Hamilton, R. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, (Red Hook, NY, USA), p. 1025–1035, Curran Associates Inc., 2017.

[53] C. Morris, M. Ritzert, M. Fey, W. L. Hamilton, J. E. Lenssen, G. Rattan, and M. Grohe, "Weisfeiler and Leman go neural: Higher-order graph neural networks," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, pp. 4602–4609, 2019.