Breaking Isolation: Multimodal Graph Fusion for Multimedia Recommendation by Edge-wise Modulation

Feiyu Chen
Sichuan Artificial Intelligence Research Institute
Yibin, China
chenfeiyu@uestc.edu.cn

Junjie Wang
Waseda University
Tokyo, Japan
wjj1020181822@toki.waseda.jp

Yinwei Wei
National University of Singapore
Singapore
weiyinwei@hotmail.com

Hai-Tao Zheng*
Shenzhen International Graduate School, Tsinghua University
Shenzhen, China
zheng.haitao@sz.tsinghua.edu.cn

Jie Shao*
University of Electronic Science and Technology of China
Chengdu, China
shaojie@uestc.edu.cn

ABSTRACT

In a multimedia recommender system, rich multimodal dynamics of user-item interactions are worth availing ourselves of and have been facilitated by Graph Convolutional Networks (GCNs). Yet, the typical way of conducting multimodal fusion with GCN-based models is either through *graph mergence* fusion that delivers insufficient inter-modal dynamics, or through *node alignment* fusion that brings in noises which potentially harm multimodal modelling. Unlike existing works, we propose EgoGCN, a structure that seeks to enhance multimodal learning of user-item interactions. At its core is a simple yet effective fusion operation dubbed EdGe-wise mOdulation (EGO) fusion. EGO fusion adaptively distils edge-wise multimodal information and learns to modulate each unimodal node under the supervision of other modalities. It breaks isolated unimodal propagations, allows the most informative inter-modal messages to spread, whilst preserving intra-modal processing. We present a hard modulation and a soft modulation to fully investigate the multimodal dynamics behind. Experiments on two real-world datasets show that EgoGCN comfortably beats prior methods.

CCS CONCEPTS

• **Information systems** → **Multimedia information systems**.

KEYWORDS

multimedia recommendation; graph fusion; multimodal dynamics

ACM Reference Format:
Feiyu Chen, Junjie Wang, Yinwei Wei, Hai-Tao Zheng, and Jie Shao. 2022. Breaking Isolation: Multimodal Graph Fusion for Multimedia Recommendation by Edge-wise Modulation. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3503161.3548399

*Corresponding authors: Hai-Tao Zheng and Jie Shao.

MM '22, October 10–14, 2022, Lisboa, Portugal

ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3548399

1 INTRODUCTION

Multimedia recommendation has been a core service in many multimedia sharing platforms to curate customized streams of content tailored to each user. Typically, a recommender system aims to predict whether a user will interact with (e.g., click, rate) an item [11, 22, 27, 28]. Multimedia content contains multimodal data streams such as video, text and music, hence leads to complex multimodal relations between users and items, a.k.a., multimodal dynamics [7]. There exist two main characteristics regarding multimodal dynamics behind user-item interactions:

i) **Variation**: User preferences may vary widely across modalities and items, and how each user reacts to different modalities of an item as well as different items in the same modality are complex and implicit. For instance, a user may prefer item I_1 to item I_2 at first sight due to the more exquisite visual content (i.e., intra-modal dynamics), but is disappointed by the poor sound tracks of item I_1 (i.e., inter-modal dynamics), and thus eventually chooses item I_2.
ii) **Interrelation**: User preference is a joint contribution by all modalities, and the preferences towards different modalities may interrelate. For instance, in the above case, the positive impression towards the visual content of item I_1 is affected and neglected due to another modality.

Nevertheless, the multimodal dynamics behind user-item interactions are totally implicit. Therefore, it is essential to learn informative representations of users and items which capture rich and complex multimodal dynamics of user-item interactions.

Early works on multimedia recommendation tend to employ neural collaborative filtering models [5, 10] which focus more on descriptive multimodal features but somewhat underestimate user-item interactions. More recently, therefore, researchers usually resort to building on Graph Convolutional Networks (GCNs) [15] in which high-order user-item associations can be effectively modelled, yielding remarkable improvements [6, 20, 26].

Many efforts have been dedicated to modelling multimodal dynamics in GCN-based multimedia recommendation through graph fusion [13]. Most existing works utilize *graph mergence* fusion,

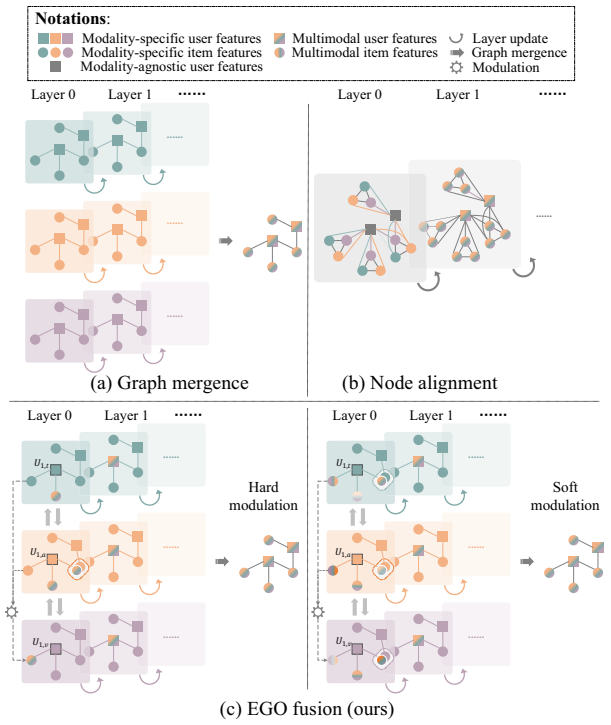


Figure 1: A conceptual overview of existing fusion methods and our proposed EGO fusion (where U_1 is the central node).

as shown in Figure 1(a), which merges isolated unimodal sub-graphs into a single one through operations such as concatenation [24, 26] and attention [30]. However, under such circumstances, since unimodal sub-graphs propagate independently, the learned high-order interactions are prone to underrate latent interrelation among modalities and deliver insufficient multimodal dynamics. Another line of research for multimodal graph fusion has used *node alignment* fusion, as shown in Figure 1(b), which constructs a huge heterogeneous graph where each modality of an item is regarded as a node, and connected with other modalities of the same item as well as connected with the user nodes [3]. Despite that more multimodal messages can thus flow within such graphs, it may introduce redundant noisy information that potentially harms multimodal modelling. In addition, its application is limited to those data at a small scale as it greatly enlarges edge quantity and memory usage. For a large dataset, this paradigm leads to an extremely huge graph with training process that normal machines can barely hold.

To address the limitations, in this work we propose **Edg**e-wise Multimodal M**o**dulated **G**raph **C**onvolutional **N**etwork (EgoGCN), which aims to simultaneously model *variation* and *interrelation* among modalities, and partially ease the existence of noises. It contains three components: multimodal propagation, ID embedding propagation and a prediction layer. First, the multimodal propagation module maintains modality-specific user-item sub-graphs to model *variation* among modalities, similar to the graph mergence method, but does not likewise conduct isolated propagation. Instead, at its core is a self-adaptive and effective fusion operation dubbed **Ed**ge-wise m**O**dulation (EGO) fusion, which aims at modelling *interrelation* among modalities, and is performed once at the very beginning of all graph operations. Concretely, as shown in Figure 1(c), for a central user (item) node, EGO fusion learns an edge-wise multimodal modulator to accordingly modulate features of each neighbouring item (user) node, by the messages from the other modalities of the same item (user) instance. This modulation enables each node to aggregate inter-modal messages from its neighbours, so as to deliver the most informative inter-modal messages across sub-graphs and break isolated unimodal propagations. This ensures that during separate intra-modal propagations, informative inter-modal messages can simultaneously flow within sub-graphs, delivering richer multimodal dynamics. Regarding modulator learning and modulation operation, we propose two mechanisms: 1) hard modulation with importance-aware modulators (EGO_{hard}) and 2) soft modulation with influence-driven modulators (EGO_{soft}). Basically, for a central node, the EGO_{hard} method learns a fuse-or-retain operation for each neighbouring node, and consequently fuses *a portion of* neighbouring nodes conditioned on the importance of user-item interactions. On the other hand, EGO_{soft} conducts fusion for *all* neighbouring nodes, with each fusion guided by the influence intensity of other modalities. In either mechanism, the modulators for each instance vary among modalities, and hence EGO fusion is capable of learning different fusion patterns for different modalities conditioned on the features of each modality. The second component is an ID embedding propagation layer that refines semantic-free collaborative embeddings of users and items through simplified graph convolution operation [11]. For simplicity, it does not borrow any information from the first multimodal module for guidance. Finally, a prediction layer takes multimodal embeddings and ID embeddings as input to perform predictions.

The overall intuition behind is that breaking isolated unimodal propagations allows the most informative inter-modal messages to spread, as well as preserves intra-modal processing. The core idea of EGO fusion is to spread a proper amount of multimodal information, partially ease the existence of noises and thus reinforce fine-grained multimodal embeddings. Our contributions are summarized as follows:

- We present EgoGCN, a structure for multimedia recommendation which enhances multimodal learning of user-item interactions as a supplement to collaborative signals.
- We develop a simple yet effective graph fusion approach EGO fusion that adaptively distils edge-wise multimodal information and modulates unimodal node features in sub-graphs, resulting in more sufficient multimodal processing. We propose two modulation mechanisms to fully investigate multimodal relations between users and items.
- Extensive experiments on two real-world multimedia recommendation datasets show that EgoGCN comfortably beats previous methods in three metrics.

2 RELATED WORK

2.1 GCN-Based Multimedia Recommendation.

Research in GCN-based multimedia recommendation can be coarsely classified into graph mergence and node alignment methods, based on how they fuse. The graph mergence method combines unimodal

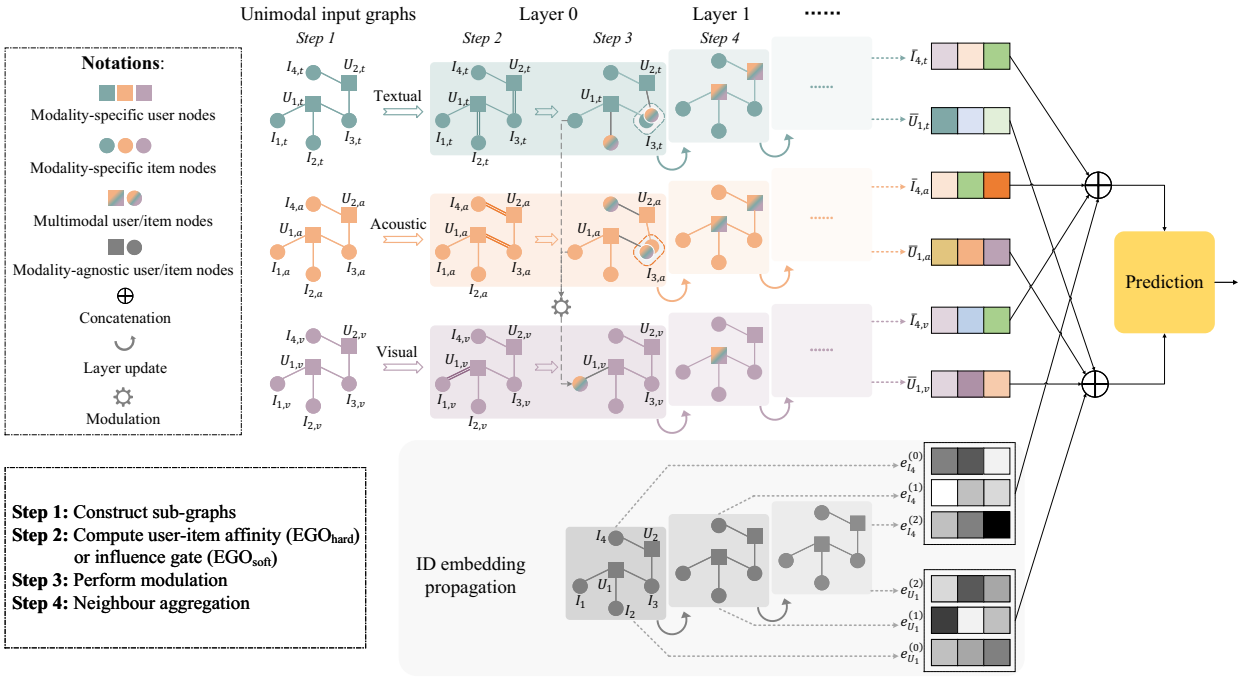


Figure 2: The overall structure of EgoGCN. EGO fusion is depicted with hard modulation where user nodes are central nodes.

sub-graphs after isolated processing, which is a dominant approach. For instance, MMGCN [26] learned user preference to each modality of an item by building modality-specific graphs. HUIGN [24] further modelled intra-level and inter-level relations among layers in each unimodal graph. GRCN [25] explored the influence of implicit feedback and integrated a soft pruning mechanism into the multimodal graph learning. LATTICE [30] explicitly considered relations among items by injecting item-item affinities into the merged multimodal graph. This fusion paradigm is easily achieved but is not skilled at capturing latent interrelation among modalities, incurring a bottleneck of performance improvement. In parallel, node alignment methods such as HHFAN [3] construct a huge heterogeneous graph that may bring in noisy information and enlarge memory burden.

Unlike prior works, we step further by mining richer multimodal dynamics that supplement collaborative signals of user-item interactions greatly.

2.2 Multimodal Fusion

The literature on multimodal fusion is vast, which includes its application to many research domains. With classic neural networks, multimodal fusion is typically classified into early fusion, late fusion, and a hybrid of them [2, 17, 18]. There are also some recent works that explore adaptive fusion methods [4, 23]. Joining the success of graph neural networks, several attempts are made to transfer multimodal learning onto graphs and accommodate the particularity of graphs. However, due to the difference of underlying propagation rules, it is not feasible to simply adopt existing advanced fusion strategies in graph neural networks. The typical way of multimodal graph fusion is to construct isolated sub-graphs [26] or a huge heterogeneous graph [14, 29], as discussed in Section 2.1. More advanced graph fusion mechanisms are worth exploring.

Distinct from the above works, our proposed EGO fusion learns a multimodal modulator to accordingly modulate unimodal node features, enables the most informative inter-modal messages to spread as well as preserves intra-modal processing, and better assists multimodal learning.

3 METHODOLOGY

We now formally describe the multimedia recommendation task setup. Let $\mathbf{A} \in \mathbb{R}^{P \times Q}$ represent the user-item interaction matrix, where P and Q are the numbers of users $\mathcal{U}$ and items $\mathcal{I}$ respectively. A nonzero entry $A_{u,i} = 1$ indicates that user $u \in \mathcal{U}$ has interacted with item $i \in \mathcal{I}$; otherwise $A_{u,i} = 0$. In addition, multimedia content information of each item is available, including textual (t), visual (v) and acoustic (a) modalities. We begin by constructing a user-item bipartite graph $\mathcal{G}$ where a node represents a user or item while an edge refers to an interaction.

The structure of the proposed EgoGCN is shown in Figure 2. In general, EgoGCN contains three modules: i) the multimodal propagation module that learns modality-specific fusion patterns and multimodal dynamics of user-item interactions; ii) the ID embedding propagation module that refines collaborative signals of users and items; and iii) a prediction layer that performs final prediction.

3.1 Multimodal Propagation

The main idea of the multimodal propagation module is to dig into latent multimodal dynamics of high-order user-item associations. We take two characteristics of multimodal dynamics behind user-item interactions into account: *variation* and *interrelation* of

multimodal user preferences. Variation results from the varying user preferences across modalities and items, while interrelation implies that user preferences towards different modalities may interrelate. Therefore, we leverage the multimodal propagation module to simultaneously model both characteristics.

Concretely, we maintain one unimodal user-item graph $\mathcal{G}_m$ for each modality $m \in \mathcal{M} = \{t, v, a\}$, where each item node contains its corresponding unimodal features i_m. At the first iteration, we define a trainable embedding vector $u_{m,(0)} \in \mathbb{R}^{D_m}$ as the preference representation of user u in modality m. In addition, $i_{m,(0)} \in \mathbb{R}^{D_m}$ denotes the unimodal item features projected into the same embedding space of users. The design of parallel sub-graphs is able to model the variation among modalities, but the widely used isolated unimodal propagation is inadequate for modelling the interrelation. To this end, we propose **E**d**G**e-wise m**O**dulation (EGO) fusion that enables the most informative inter-modal information to spread whilst preserving intra-modal processing, so that the interrelation can be modelled along with the variation.

3.1.1 EGO Fusion. Generally, the aim of EGO fusion is to break the isolation, spread a proper amount of inter-modal messages that is most beneficial to the model, and partially ease the existence of noises. Since every node in the graph aggregates messages from its neighbouring nodes, we propose to enable each node to aggregate inter-modal messages from its neighbours as well. We hence formulate the key problem as: *for a central node, to mine and pass the most informative inter-modal messages of its neighbouring nodes.* Therefore, we propose to modulate unimodal neighbouring features under the supervision of other modalities, so as to spread informative inter-modal messages. Towards effective and efficient message passing across modalities, we propose two modulation mechanisms that mine inter-modal messages from two different perspectives:

i) EGO$_{hard}$: hard modulation with importance-aware modulators. This mechanism mines full inter-modal messages of a portion of neighbouring nodes that are important;
ii) EGO$_{soft}$: soft modulation with influence-driven modulators. This mechanism mines influential inter-modal messages of all neighbouring nodes.

1. Hard Modulation (EGO$_{hard}$). Intuitively, the preference representation of a user is biased to those neighbouring item nodes that he/she has a strong preference for. This is known as *affinity* between users and items that reflects how close they are correlated [25]. Our key insight here is that more closely-correlated user-item pairs may provide better guidance of learning how user preference is affected. It thus motivates us to mine and leverage those inter-modal messages from closely-correlated user-item pairs. Technically, for a central node, we explore to learn a fuse-or-retain operation for each neighbour, and consequently fuse a portion of neighbours conditioned on the importance of user-item affinity.

Specifically, at the first iteration, the bidirectional unimodal affinities between user u and item i can be inferred from the inner product:

$$s^m_{i\to u} = \frac{exp({i_{m,(0)}}^T u_{m,(0)})}{\sum_{j\in\mathcal{N}_u} exp({j_{m,(0)}}^T u_{m,(0)})}, \qquad s^m_{u\to i} = \frac{exp({u_{m,(0)}}^T i_{m,(0)})}{\sum_{v\in\mathcal{N}_i} exp({v_{m,(0)}}^T i_{m,(0)})}, \tag{1}$$

where $s^m_{i\to u}, s^m_{u\to i} \in [0,1]$, and $\mathcal{N}_x$ is the neighbouring nodes of node x. A larger affinity score $s^m_{i\to u}$ suggests that content $i_{m,(0)}$ provides more important contribution for modelling the user preference in modality m, and vice versa. As discussed earlier, an important inference we make here is that the inter-modal messages behind content $i_{m,(0)}$ which has a larger $s^m_{i\to u}$ are *more likely to be essential for the model to learn how preference of user u is affected.* Hence, in this mechanism, we propose to perform a hard modulation that learns a fuse-or-retain operation for each neighbouring node controlled by the affinity scores. Formally, for a central node u_m (i_m), we modulate its neighbouring node $i_{m,(0)}$ ($u_{m,(0)}$) to obtain multimodality-aware neighbouring features $\hat{i}_{m,(0),u}$ ($\hat{u}_{m,(0),i}$):

$$\hat{i}_{m,(0),u} = \begin{cases} \frac{1}{|\mathcal{M}|}\sum_{m'\in\mathcal{M}} i_{m',(0)}, \text{ if } s^m_{i\to u} > \epsilon; \\ i_{m,(0)}, \text{ else;} \end{cases} \tag{2a}$$

$$\hat{u}_{m,(0),i} = \begin{cases} \frac{1}{|\mathcal{M}|}\sum_{m'\in\mathcal{M}} u_{m',(0)}, \text{ if } s^m_{u\to i} > \epsilon; \\ u_{m,(0)}, \text{ else;} \end{cases} \tag{2b}$$

where $\epsilon \in [0,1]$ is a hyper-parameter denoting importance threshold and m' denotes any modality in the database. The third subscripts in $\hat{i}_{m,(0),u}$ and $\hat{u}_{m,(0),i}$ correspond to the name *edge-wise*. The hat symbol indicates the awareness of multimodal information. The fusion operation is conducted through average. One can also use concatenation or weighted sum to fuse features, while we empirically observe that a simple average can deliver good performance with minimal cost.

One may have concerns that Equation 2 seems to ignore the fact that an important unimodal neighbouring instance does not necessarily imply the high importance of other modalities to the same central node. Please kindly note that Equation 2 does not intent to suggest the high importance of other modalities, but the *informativeness of inter-modal messages* behind this neighbour. Viewed dynamically, through this mechanism, the model learns itself to adaptively adjust the correlation between users and items by gradient decent, so as to achieve positive message passing across modalities.

2. Soft Modulation (EGO$_{soft}$). To fully investigate multimodal relations between users and items, we further propose a soft modulation mechanism that distils multimodal information from another perspective. Different from the hard modulation, EGO$_{soft}$ mines inter-modal messages of all neighbouring nodes but controls the messages that pass through. Let $i_{m,(0)}$ denote the features of a neighbouring node of user u in modality m. In this mechanism, in order to modulate $i_{m,(0)}$, we first learn a modality-specific influence gate $g^{h\to m}_{i\to u}$ for the corresponding features $i_{h,(0)}$ in modality h, to model the influence intensity of $i_{h,(0)}$ on $i_{m,(0)}$. The user-based process is symmetric. Formally,

$$g^{h\to m}_{i\to u} = \sigma\left(W^{h\to m}_{i\to u} i_{h,(0)} + b^{h\to m}_{i\to u}\right), \tag{3a}$$

$$g^{h\to m}_{u\to i} = \sigma\left(W^{h\to m}_{u\to i} u_{h,(0)} + b^{h\to m}_{u\to i}\right), \tag{3b}$$

where $g^{h\to m}_{i\to u}, g^{h\to m}_{u\to i} \in [0,1]$. $h \in \mathcal{M}, h \neq m$. $W^{h\to m}_{i\to u}$ and $W^{h\to m}_{u\to i}$ are weight vectors, and $b^{h\to m}_{i\to u}$ and $b^{h\to m}_{u\to i}$ are scalar biases. $\sigma(\cdot)$ is the sigmoid function. Next, for a central node u_m (i_m), we modulate

its neighbour $i_{m,(0)}$ ($u_{m,(0)}$) to obtain multimodal neighbouring features $\hat{i}_{m,(0),u}$ ($\hat{u}_{m,(0),i}$), guided by the influence gates:

$$\hat{i}_{m,(0),u} = \frac{1}{|\mathcal{M}|}\left(i_{m,(0)} + \sum_{h\neq m} g_{i\to u}^{h\to m} i_{h,(0)}\right), \tag{4a}$$

$$\hat{u}_{m,(0),i} = \frac{1}{|\mathcal{M}|}\left(u_{m,(0)} + \sum_{h\neq m} g_{u\to i}^{h\to m} u_{h,(0)}\right). \tag{4b}$$

In this fashion, the messages flowing from modality h to modality m is adaptively controlled by the influence gates, to ensure that only the most influential and informative messages from modality h can pass through. It can be noticed that the learned influence gates are both modality-specific (through superscripts) and edge-wise (through subscripts), which enables the model to mine inter-modal messages in a fine-grained level.

An example of EGO_{hard} on an item-based view is depicted at layer 0 in Figure 2. The user-based process is symmetric. Please kindly note that EGO fusion does not change the neighbouring nodes themselves, but only *what attends to the aggregation of the central node*, hence the name *edge-wise*. Take the case in Figure 2 as an example. In a nutshell, edge-wise means that although both central user nodes U_1 and U_2 have interacted with item I_3, the modulated item features that are aggregated to U_1 and U_2 may be different, conditioned on the respective situations. This provides more flexibility for modelling richer multimodal dynamics.

3.1.2 Neighbour Aggregation. After modulating neighbouring nodes and delivering informative multimodal messages, we now pass the messages to the central nodes and to local neighbourhood. For this purpose, we reformulate the neighbour routing mechanism [16, 25] to propagate multimodal embeddings. Specifically, we first recompute and update the affinities based on the refreshed features to reflect the modified correlation of the refreshed features:

$$\begin{aligned} \hat{s}_{i\to u}^{m} &= \frac{exp((\hat{i}_{m,(0),u})^T u_{m,(0)})}{\sum_{j\in\mathcal{N}_u} exp((\hat{j}_{m,(0)})^T u_{m,(0)})}, \\ \hat{s}_{u\to i}^{m} &= \frac{exp((\hat{u}_{m,(0),i})^T i_{m,(0)})}{\sum_{v\in\mathcal{N}_i} exp((\hat{v}_{m,(0)})^T i_{m,(0)})}. \end{aligned} \tag{5}$$

Next, we pass the messages to the central nodes by adopting a weighted sum aggregator:

$$\begin{aligned} \hat{u}_{m,(1)} &= \sum_{i\in\mathcal{N}_u} \hat{s}_{i\to u}^{m} \hat{i}_{m,(0),u}, \\ \hat{i}_{m,(1)} &= \sum_{u\in\mathcal{N}_i} \hat{s}_{u\to i}^{m} \hat{u}_{m,(0),i}, \end{aligned} \tag{6}$$

where $\hat{u}_{m,(1)}$ and $\hat{i}_{m,(1)}$ are the latent user preference and item features at layer 1 in modality m, which are multimodality-aware. We observe experimentally that including the representations of central nodes at layer 0 (a.k.a self-loops) has no positive effect on the performance.

We now gradually spread the essential multimodal messages over. It is observed empirically that further performing EGO fusion at subsequent layers is of little benefit to the results. The main reason we suggest is that it might further fuse those features that have already been fused, and thus introduce noises. Yet, the key idea of EGO fusion is to spread a proper amount of multimodal information and partially ease the existence of noises. Hence, for subsequent layers, we conduct regular node aggregation and graph update to propagate embeddings, following procedures in [25]. Recursively,

$$\begin{cases} \hat{u}_{m,(k+1)} = \hat{u}_{m,(k)} + \sum_{i\in\mathcal{N}_u} \hat{s}_{i\to u}^{m} \hat{i}_{m,(k)}, \\ \hat{s}_{i\to u}^{m} = \frac{exp((\hat{i}_{m,(k)})^T \hat{u}_{m,(k)})}{\sum_{j\in\mathcal{N}_u} exp((\hat{j}_{m,(k)})^T \hat{u}_{m,(k)})}, \end{cases} \tag{7}$$

$$\begin{cases} \hat{i}_{m,(k+1)} = \hat{i}_{m,(k)} + \sum_{u\in\mathcal{N}_i} \hat{s}_{u\to i}^{m} \hat{u}_{m,(k)}, \\ \hat{s}_{u\to i}^{m} = \frac{exp((\hat{u}_{m,(k)})^T \hat{i}_{m,(k)})}{\sum_{v\in\mathcal{N}_i} exp((\hat{v}_{m,(k)})^T \hat{i}_{m,(k)})}, \end{cases} \tag{8}$$

where $k \in \{1, 2, ..., K\}$. By this means the high-order user and item embeddings are gradually refined based on the outputs of previous layer. Lastly, we get the outputs of the final layer as representations of the multimodal user preference and item content, respectively:

$$\overline{u}_m = \hat{u}_{m,(K)}, \overline{i}_m = \hat{i}_{m,(K)}. \tag{9}$$

3.2 ID Embedding Propagation

We follow previous works [24, 25] to further maintain a branch that learns semantic-free collaborative embeddings of users and items. We assign an initial trainable ID embedding $e_u^{(0)} \in \mathbb{R}^{D_{id}}$ ($e_i^{(0)} \in \mathbb{R}^{D_{id}}$) for each user u (item i) and utilize the simplified graph operation from LightGCN [11] to aggregate messages:

$$\begin{aligned} e_u^{(l+1)} &= \sum_{i\in\mathcal{N}_u} \frac{1}{\sqrt{|\mathcal{N}_u|}\sqrt{|\mathcal{N}_i|}} e_i^{(l)}, \\ e_i^{(l+1)} &= \sum_{u\in\mathcal{N}_i} \frac{1}{\sqrt{|\mathcal{N}_i|}\sqrt{|\mathcal{N}_u|}} e_u^{(l)}, \end{aligned} \tag{10}$$

in which $e_u^{(l)}$ ($e_i^{(l)}$) is the ID embedding of user u (item i) at the l-th layer and the normalization term $\frac{1}{\sqrt{|\mathcal{N}_u|}\sqrt{|\mathcal{N}_i|}}$ follows the standard design in GCN [15]. By stacking L layers, each node receives the collaborative signals from L-hop neighbours and we combine the representations at each layer to obtain the final ID embeddings:

$$e_u = \sum_{l=0}^{L} e_u^{(l)}, e_i = \sum_{l=0}^{L} e_i^{(l)}. \tag{11}$$

As one may have noticed, we do not borrow any auxiliary information from the first multimodal propagation module, unlike some prior works such as GRCN [25]. Such decoupling design promotes simplicity and improves performance, as will be discussed in Section 5.2. Additionally, as the main purpose of this work is to explore multimodal dynamics rather than collaborative signals, we use a light and valid module to propagate ID embeddings, and avoid complicating EgoGCN unnecessarily. More advanced modules with heavier structures may further boost up the performance.

3.3 Prediction Layer

The prediction layer takes as input the concatenated multimodal features and ID embeddings to perform the final prediction, which contain both the high-order multimodal dynamics and collaborative signals of user-item interaction. Mathematically,

$$\begin{aligned} e_u^* &= e_u \oplus \overline{u}_t \oplus \overline{u}_v \oplus \overline{u}_a, \\ e_i^* &= e_i \oplus \overline{i}_t \oplus \overline{i}_v \oplus \overline{i}_a, \end{aligned} \tag{12}$$

Table 1: The statistics of datasets. The dimensions of visual, acoustic and textual features are denoted by V, A, and T, respectively.

Dataset	#Interactions	#Items	#Users	Sparsity	V	A	T
Movielens	1,239,508	5,986	55,485	99.63%	2048	128	100
Tiktok	726,065	76,085	36,656	99.97%	128	128	128

where $\oplus$ is concatenation operation. Lastly, the output of the model is obtained by the inner product of user and item representations:

$$\hat{y}_{u,i} = {e_u^*}^T e_i^*. \tag{13}$$

Here, $\hat{y}_{u,i}$ is the predicted affinity between user u and target item i.

3.4 Training Objective

Following prior works, we use Bayesian Personalized Ranking (BPR) [19] to conduct the pair-wise ranking, which encourages the score of an observed entry to be higher than its unobserved counterparts:

$$\mathcal{L} = \sum_{(u,i,j)\in\mathcal{T}} -ln\sigma(\hat{y}_{u,i} - \hat{y}_{u,j}) + \lambda\|\theta\|_2, \tag{14}$$

where $\mathcal{T} = \{(u,i,j)|A_{u,i} = 1, A_{u,j} = 0\}$ is a triplet containing a user u, an observed item i and an unobserved item j. $\sigma(\cdot)$ is the sigmoid function. λ is the L2-regularizer weight and θ denotes the trainable parameters in the model.

4 EXPERIMENTAL SETUP

4.1 Datasets

We compare the performance of EgoGCN against prior works on two real-world multimedia recommendation datasets, and employ text, video and audio modalities, following settings in previous works [24–26]. The statistics of datasets are summarized in Table 1.

Movielens[1] is a widely used dataset for the recommendation task and has been extended by researchers to multimodal scenario. Specifically, raw data are first obtained by collecting descriptions of movies from Movielens-10M and crawling the corresponding trailers from YouTube. Textual features are then extracted from descriptions by Sentence2Vector [1]. For visual modality, key frames are first extracted from the crawled videos and then fed into a pre-trained ResNet50 [9] to obtain visual features. For acoustic modality, the features are obtained with VGGish [12], following the soundtrack separation procedure using FFmpeg software.

Tiktok[2] is a dataset released by TikTok, a popular multimedia sharing platform. It contains user interactions with short videos and extracted multimodal features. No raw data is published. In particular, the textual features are extracted from the captions uploaded by users.

On both datasets, we follow the standard data split protocol as in previous works [24–26], and split the historical interactions of each user by an 8:1:1 ratio to obtain the training, validation, and testing sets. Note that previous works conduct experiments on Kwai[3] as well. However, our method is not applicable to the Kwai dataset, in which only visual modality is available.

[1]https://movielens.org
[2]https://www.tiktok.com
[3]https://www.kwai.com/

Table 2: Details of hyper-parameters in our experiments.

Dataset	Batch	Optimizer	D^m	D^{id}	K	L	ϵ
Movielens	2048	Adam (lr=2e-4)	64	64	4	3	0.4
Tiktok	1024	Adam (lr=3e-4)	64	64	2	2	0.6

4.2 Baselines

We contrast the performance of EgoGCN against the following GCN-based models:

- **GraphSAGE** [8] is an inductive model for computing node embeddings, which generates embeddings by sampling and aggregating features from a node's local neighbourhood.
- **NGCF** [22] explicitly learns the collaborative signals between users and items, deepening the use of graph convolutional networks with high-hop neighbours.
- **DisenGCN** [16] disentangles the latent factors between nodes by introducing a neighbourhood routing mechanism which relies on disentangling information from a local range.
- **GAT** [21] learns different weights to different nodes within a neighbourhood by using attention mechanism, so as to enhance the learned embeddings.
- **MMGCN** [26] models modality-specific user preference and fuses them as the representation of user to score their affinities to the item content features.
- **HUIGN** [24] models intra-level and inter-level associations among layers in each unimodal graph.
- **LightGCN** [11] removes feature transformation and nonlinear activation from standard GCNs to construct a light structure for collaborative filtering.
- **GRCN** [25] refines the user-item bipartite sub-graphs in different modalities and accordingly re-expresses the user and item to optimize the prediction of their interactions.

Please kindly note that the compared methods all utilize graph mergence fusion. We thus provide the comparison against both fusion paradigms under a fair setting in Section 5.2. In addition, since several methods above (i.e., GraphSAGE, NGCF, DisenGCN, GAT and LightGCN) are not originally designed under multimodal scenario, for a fair comparison, we concatenate multimodal features as the input node embeddings, with early fusion.

4.3 Settings and Evaluation Metrics

Our proposed model is implemented using PyTorch and torch-geometric packages. The networks are trained on a machine with 1 NVIDIA GeForce RTX 3090. On the training set, negative sampling is used to create training triplets. During validation and testing, for each user, we regard the items with which the user has no interaction as the negative samples. The interactions of user-item pairs are scored and ranked in a descending order. We follow dominant evaluation protocols to use precision@10, recall@10 and Normalized Discounted Cumulative Gain (NDCG)@10 as the metrics to measure the performance. During training, we employ early stopping if recall@10 on the validation set does not increase for 20 successive epochs. We test K in the range from 1 to 4, L from 1 to 3, and ϵ in the range of $\{0.3, 0.4, 0.5, 0.6, 0.7\}$. For EGO_{hard}, best results are achieved by $\{K = 4, L = 3, \epsilon = 0.4\}$ on Movielens and

Table 3: Comparison with previous methods on Movielens and Tiktok. ◇ from [25]; ⊘ from [24]; □ from our reimplementation.

Methods	Movielens			Tiktok		
	Precision	Recall	NDCG	Precision	Recall	NDCG
GraphSAGE◇ [8]	0.0496	0.1984	0.2136	0.0128	0.0631	0.0606
NGCF◇ [22]	0.0547	0.2196	0.2342	0.0135	0.0780	0.0661
DisenGCN◇ [16]	0.0555	0.2222	0.2401	0.0145	0.0760	0.0639
GAT◇ [21]	0.0569	0.2307	0.2434	0.0166	0.0891	0.0802
MMGCN◇ [26]	0.0581	0.2345	0.2517	0.0144	0.0808	0.0674
HUIGN⊘ [24]	0.0619	0.2522	0.2677	0.0164	0.0884	0.0769
LightGCN□ [11]	0.0621	0.2538	0.2701	0.0193	0.0944	0.0826
GRCN◇ [25]	0.0639	0.2569	0.2754	0.0195	0.1048	0.0938
EgoGCN-hard (ours)	**0.0679**	**0.2745**	**0.2945**	0.0214	0.1183	0.1028
EgoGCN-soft (ours)	0.0676	0.2734	0.2929	**0.0219**	**0.1199**	**0.1046**
Relative improvement	6.26%	6.85%	6.94%	12.31%	14.41%	10.33%

Table 4: Ablation studies of EgoGCN. GM denotes graph mergence and NA denotes node alignment.

No.	Methods	Movielens			Tiktok		
		Precision	Recall	NDCG	Precision	Recall	NDCG
1	w/o multimodal info.	0.0641	0.2614	0.2793	0.0187	0.1032	0.0901
2	ID embeddings with soft pruning	0.0672	0.2646	0.2867	0.0209	0.1102	0.0989
3	EgoGCN-GM	0.0671	0.2715	0.2912	0.0208	0.1139	0.0995
4	EgoGCN-NA	0.0672	0.2726	0.2910	0.0203	0.1111	0.0977
	EgoGCN-hard	0.0679	0.2745	0.2945	0.0214	0.1183	0.1028
	EgoGCN-soft	0.0676	0.2734	0.2929	0.0219	0.1199	0.1046

$\{K = 2, L = 2, \epsilon = 0.6\}$ on Tiktok. For EGO_{soft}, the best-performing values for K and L are the same. Full details of hyper-parameters for both datasets are shown in Table 2.

5 RESULTS AND ANALYSIS

5.1 Comparison with State-of-the-Arts

Table 3 presents the performance of our EgoGCN with a wide range of state-of-the-art methods. We can notice that on Movielens, the hard modulation outperforms soft modulation, and on Tiktok the opposite is the case. However, with either modulation mechanism, the proposed EgoGCN comfortably beats previous methods in all metrics on both datasets, which demonstrates its effectiveness. Specifically, EgoGCN outperforms the state-of-the-art GRCN by a good margin. In addition, it can be seen that the performance of LightGCN, which is used as a constituent of our model, is even inferior to GRCN. The main reason we suggest is that LightGCN is designed for semantic-free collaborative filtering and is not skilled at capturing multimodal semantics in multimedia data.

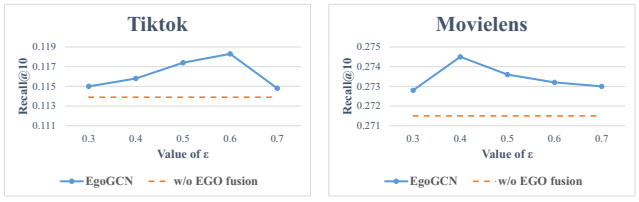
5.2 Ablation Studies

To gain insights to the constituents of our model, we study some configurations of EgoGCN, and present key results in Table 4. More visualization examples are presented in appendix.

Effect of multimodal propagation. We first explore the effect of multimedia semantic information by removing the multimodal propagation module and directly obtaining predictions based on ID embeddings only, as variant 1. A sharp decrease can be observed in all metrics, which validates the usefulness of taking advantage of multimodal dynamics of user-item interactions. Moreover, it outperforms the multimodal version of LightGCN in Table 3, which confirms that LightGCN is more powerful in semantic-free scenario.

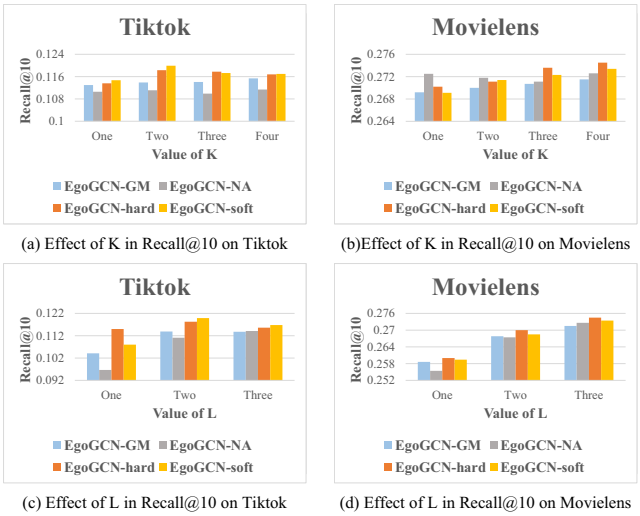
On ID embeddings. The ID embeddings in our model are decoupled from multimodal propagation. We contrast this setting against the soft pruning method in GRCN [25], which employs multimodal affinity scores as the normalization terms for propagating ID embeddings. Variant 2 shows the results using soft pruning and a decline of performance can be observed. This indicates the effectiveness of the simpler decoupling design.

Effect of EGO fusion. EGO fusion is the core of our method. We hence conduct experiments to compare it against other fusion methods. Variants 3 and 4 report the results of EgoGCN with graph mergence (EgoGCN-GM) and node alignment (EgoGCN-NA), respectively. EgoGCN-GM is implemented by removing EGO fusion and leaving behind isolated unimodal propagations which exactly employ graph mergence fusion. EgoGCN-NA is implemented by replacing modality-specific sub-graphs with a huge heterogeneous graph, and we use unified node aggregation protocol in Equation 7 or Equation 8 for all nodes. Variants 3 and 4 report the results



(a) Effect of ε in Recall@10 on Tiktok (b) Effect of ε in Recall@10 on Movielens

Figure 3: Performance study of ϵ.



(a) Effect of K in Recall@10 on Tiktok (b)Effect of K in Recall@10 on Movielens

(c) Effect of L in Recall@10 on Tiktok (d) Effect of L in Recall@10 on Movielens

Figure 4: Results of EgoGCN and two variants (EgoGCN-GM and EgoGCN-NA) at different layers. In (a) and (b), effects of K are tested by fixing L as in our best-performing models. In (c) and (d), effects of L are tested by fixing K as in our best-performing models.

with the same K and L as in our best-performing models. In terms of the quantitative results, we can observe a consistent drop in performance in all metrics with either method. We suggest that it is because graph mergence tends to underrate inter-modal relations and node alignment is prone to introduce noises, so they both deliver insufficient multimodal processing. This verifies the effectiveness of EGO fusion, empirically showing that spreading essential multimodal messages can facilitate multimodal dynamics. We provide more quantitative comparison among three fusion paradigms in Section 5.4.

5.3 Discussions on ϵ

In EGO$_{hard}$, a hyper-parameter ϵ is introduced as the importance threshold to balance between inter-modal and intra-modal processing. A smaller ϵ prioritizes inter-modal messages over intra-modal ones as it allows more instances to fuse. We hence empirically investigate the effect of ϵ on both datasets. We fix other hyper-parameters, test ϵ in the range of $\{0.3, 0.4, 0.5, 0.6, 0.7\}$ and present the results of recall@10 in Figure 3. Values of precision@10 and NDCG@10 show similar trends and thus are omitted.

On Tiktok, at first, the performance steadily improves as ϵ increases and peaks at ϵ=0.6, followed by a sharp decrease. On Movielens, the best performance is obtained at ϵ=0.4, and further increasing the threshold has a negative impact on the results. Nevertheless, it can be seen from Figure 3 that in all cases our method outperforms the variant that does not employ EGO fusion. This again, in our view, implies the usefulness of taking advantage of multimodal dynamics, as spreading different amounts of information can, to different extent, benefit multimodal processing.

5.4 Discussions on Graph Layers

Figure 4 presents the recall@10 results of EgoGCN and two variants (EgoGCN-GM and EgoGCN-NA) at different layers in multimodal propagation (K) and ID embedding propagation (L). The main observations are as follows.

Effects of K: From Figure 4(a) we can see that on Tiktok our method consistently outperforms the other two fusion methods regardless of the values of K. EgoGCN-NA is inferior to EgoGCN-GM. From Figure 4(b) we can observe that on Movielens EgoGCN-NA beats EgoGCN-GM in most cases, and even outperforms EgoGCN with a small K. However, stacking more layers boosts up the performance of EgoGCN and delivers best results, as the multimodal messages spread farther.

Effects of L: It is noteworthy from Figures 4(c) and 4(d) that, with the same multimodal propagation layers, EgoGCN-NA depends heavily on the numbers of ID embedding layers L on both datasets. This indicates that EgoGCN-NA benefits much more from semantic-free ID embeddings instead of multimodal processing than the other two. On Movielens, EgoGCN-GM and our EgoGCN show similar trends w.r.t. the value of L while EgoGCN-GM is more sensitive to ID embeddings on Tiktok.

In most cases, our proposed EgoGCN-hard and EgoGCN-soft outperform the other two fusion paradigms, which shows the effectiveness of our proposed EGO fusion.

6 CONCLUSIONS

We present a new structure EgoGCN that is equipped with a simple yet effective graph fusion operation, to achieve a more sufficient modelling of multimodal dynamics for multimedia recommendation. Our fusion method, EGO fusion, adaptively distils edge-wise multimodal messages and allows the most essential inter-modal information to better guide the fusion. Experiments show that our approach achieves new state-of-the-art results, suggesting that multimodal processing is a promising avenue of investigation for multimedia recommendation. We plan to explore advanced fusion mechanism for situations with missing modalities in future.

ACKNOWLEDGMENTS

This research is supported by National Natural Science Foundation of China (No. 61832001 and No. 6201101015), Beijing Academy of Artificial Intelligence (BAAI), Natural Science Foundation of Guangdong Province (No. 2021A1515012640), Shenzhen Fundamental Research Program (No. JCYJ20210324120012033 and JCYJ20190813165003837), and Overseas Cooperation Research Fund of Tsinghua Shenzhen International Graduate School (No. HW2021008).

REFERENCES

[1] Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A Simple but Tough-to-Beat Baseline for Sentence Embeddings. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

[2] Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* 41, 2 (2019), 423–443.

[3] Desheng Cai, Shengsheng Qian, Quan Fang, and Changsheng Xu. 2022. Heterogeneous Hierarchical Feature Aggregation Network for Personalized Micro-video Recommendation. *IEEE Trans. Multim.* 24 (2022), 805–818.

[4] Feiyu Chen, Zhengxiao Sun, Deqiang Ouyang, Xueliang Liu, and Jie Shao. 2021. Learning What and When to Drop: Adaptive Multimodal and Contextual Dynamics for Emotion Recognition in Conversation. In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*. 1064–1073.

[5] Jingyuan Chen, Hanwang Zhang, Xiangnan He, Liqiang Nie, Wei Liu, and Tat-Seng Chua. 2017. Attentive Collaborative Filtering: Multimedia Recommendation with Item- and Component-Level Attention. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*. 335–344.

[6] Xiaolin Chen, Xuemeng Song, Guozhen Peng, Shanshan Feng, and Liqiang Nie. 2021. Adversarial-Enhanced Hybrid Graph Network for User Identity Linkage. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*. 1084–1093.

[7] Yashar Deldjoo, Markus Schedl, Balázs Hidasi, Yinwei Wei, and Xiangnan He. 2022. Multimedia Recommender Systems: Algorithms and Challenges. In *Recommender Systems Handbook*. Springer, 973–1014.

[8] William L. Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive Representation Learning on Large Graphs. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. 1024–1034.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. 770–778.

[10] Ruining He and Julian J. McAuley. 2016. VBPR: Visual Bayesian Personalized Ranking from Implicit Feedback. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*. 144–150.

[11] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. 2020. LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*. 639–648.

[12] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin W. Wilson. 2017. CNN architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*. 131–135.

[13] Andreas Holzinger, Bernd Malle, Anna Saranti, and Bastian Pfeifer. 2021. Towards multi-modal causability with Graph Neural Networks enabling information fusion for explainable AI. *Inf. Fusion* 71 (2021), 28–37.

[14] Jingwen Hu, Yuchen Liu, Jinming Zhao, and Qin Jin. 2021. MMGCN: Multimodal Fusion via Deep Graph Convolution Network for Emotion Recognition in Conversation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*. 5666–5675.

[15] Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

[16] Jianxin Ma, Peng Cui, Kun Kuang, Xin Wang, and Wenwu Zhu. 2019. Disentangled Graph Convolutional Networks. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*. 4212–4221.

[17] Sijie Mai, Haifeng Hu, and Songlong Xing. 2019. Divide, Conquer and Combine: Hierarchical Feature Fusion Network with Local and Global Perspectives for Multimodal Affective Computing. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*. 481–492.

[18] Wasifur Rahman, Md. Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Mohammed E. Hoque. 2020. Integrating Multimodal Information in Large Pretrained Transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*. 2359–2369.

[19] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian Personalized Ranking from Implicit Feedback. In *UAI 2009, Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, Montreal, QC, Canada, June 18-21, 2009*. 452–461.

[20] Rui Sun, Xuezhi Cao, Yan Zhao, Junchen Wan, Kun Zhou, Fuzheng Zhang, Zhongyuan Wang, and Kai Zheng. 2020. Multi-modal Knowledge Graphs for Recommender Systems. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*. 1405–1414.

[21] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.

[22] Xiang Wang, Xiangnan He, Meng Wang, Fuli Feng, and Tat-Seng Chua. 2019. Neural Graph Collaborative Filtering. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*. 165–174.

[23] Yikai Wang, Wenbing Huang, Fuchun Sun, Tingyang Xu, Yu Rong, and Junzhou Huang. 2020. Deep Multimodal Fusion by Channel Exchanging. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

[24] Yinwei Wei, Xiang Wang, Xiangnan He, Liqiang Nie, Yong Rui, and Tat-Seng Chua. 2022. Hierarchical User Intent Graph Network for Multimedia Recommendation. *IEEE Trans. Multim.* 24 (2022), 2701–2712.

[25] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, and Tat-Seng Chua. 2020. Graph-Refined Convolutional Network for Multimedia Recommendation with Implicit Feedback. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*. 3541–3549.

[26] Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. 2019. MMGCN: Multi-modal Graph Convolution Network for Personalized Recommendation of Micro-video. In *Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019*. 1437–1445.

[27] Jiancan Wu, Xiang Wang, Fuli Feng, Xiangnan He, Liang Chen, Jianxun Lian, and Xing Xie. 2021. Self-supervised Graph Learning for Recommendation. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*. 726–735.

[28] Lianghao Xia, Yong Xu, Chao Huang, Peng Dai, and Liefeng Bo. 2021. Graph Meta Network for Multi-Behavior Recommendation. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*. 757–766.

[29] Jianing Yang, Yongxin Wang, Ruitao Yi, Yuying Zhu, Azaan Rehman, Amir Zadeh, Soujanya Poria, and Louis-Philippe Morency. 2021. MTAG: Modal-Temporal Attention Graph for Unaligned Human Multimodal Language Sequences. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*. 1009–1021.

[30] Jinghao Zhang, Yanqiao Zhu, Qiang Liu, Shu Wu, Shuhui Wang, and Liang Wang. 2021. Mining Latent Structures for Multimedia Recommendation. In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*. 3872–3880.

A VISUALIZATION

In Figure 5, we qualitatively compare the samples derived from GRCN and EgoGCN (with hard modulation). We randomly select several user-item pairs from Movielens and visualize the concatenated embeddings in Equation 12 before and after one propagation of the networks. The lines between stars and points can reflect the distances of user-item embeddings. Shorter lines imply closer user-item pairs. As can be seen, despite that both methods can aggregate relevant items, with GRCN, star-point pairs after one propagation still have long lines (e.g., ID 21 and ID 39). However, the clusters learned by our EgoGCN are more discernible and user-item pairs are closer. It verifies the good capability of our method to capture user-item associations.

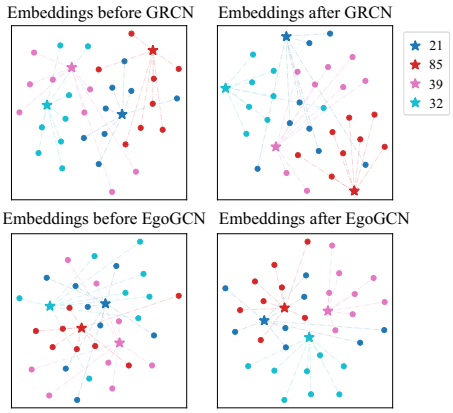


Figure 5: Visualization on learned t-SNE transformed embeddings. Each star denotes a user from Movielens, and the points in the same colour are the relevant items.