Deep Multimodal Complementarity Learning

Daheng Wang, Tong Zhao, Wenhao Yu, Nitesh V. Chawla, *Fellow, IEEE*, and Meng Jiang, *Member, IEEE*

***Abstract*— Complementarity plays a significant role in the synergistic effect created by different components of a complex data object. Complementarity learning on multimodal data has fundamental challenges of representation learning because the complementarity exists along with multiple modalities and one or multiple items of each modality. Also, an appropriate metric is needed for measuring the complementarity in the representation space. Existing methods that rely on similarity-based metrics cannot adequately capture the complementarity. In this work, we propose a novel deep architecture for systematically learning the complementarity of components from multimodal multi-item data. The proposed model consists of three major modules: 1) unimodal aggregation for extracting the intramodal complementarity; 2) cross-modal fusion for extracting the intermodal complementarity at the modality level; and 3) interactive aggregation for extracting the intermodal complementarity at the item level. To quantify complementarity, we utilize the TUBE distance metric to measure the difference between the composited data object and its label in the representation space. Experiments on three real datasets show that our model outperforms the state-of-the-art by +6.8% of mean reciprocal rank (MRR) on object classification and +3.0% of MRR on hold-out item prediction. Qualitative analyses reveal that complementarity is significantly different from similarity.**

***Index Terms*— Complementarity modeling, deep learning, multimodal machine learning.**

I. Introduction

COMPLEMENTARITY describes the synergistic effect created by different components of a complex data object [1], [2]. The characteristics of being complementary to each other refer to the potential of stimulating synergistic interactions to create additional utilities by incorporating the target component [3]. For example, the profile image(s) and description text for a product displayed on e-commerce platforms are all critical, affecting its exposure and popularity level. Therefore, experienced sellers create a set of complementary images (e.g., different angles and occasions) and complementary textual descriptions (e.g., different features and specifications) instead of similar ones to maximize the probability of receiving wider public attention. Take a research project team as another example. Each researcher makes a certain amount of contributions to the work when considered alone. In addition, when two or more researchers have complementary expertise and skillsets, they can create additional value and improve the success rate for the team. Song *et al.* [4] and [5] and Lin *et al.* [6] and suggesting materials for effective learning [7], and discovering complementary medicine for multimorbidity [8].

Real-world data objects exhibit complex structural characteristics, which can be a fertile source for learning the complementarity information [9]–[11]. On one hand, data objects around us involve multiple perceivable modalities. Modality refers to the form in which the data object is presented [12], [13]. For example, a product description consists of modalities such as image and text; and, a project team has multiple modalities like researcher, engineer, and hardware resource. Such data is therefore defined as multimodal data when it contains multiple modalities [14], [15]. On the other hand, each modality shows a set structure composed of one or more items. For example, a word or phrase in the product description can be considered an item of the text modality. Likewise, a researcher (author) is an item of the corresponding modality in a project team (research paper). A multimodal multi-item data object may have multiple items of different modalities. In real world, it is often accompanied by label information. The label can be either a class label [16] (e.g., category of a product or venue of a paper) or a numerical label [17] (e.g., product sales or number of paper citations). Moreover, the label serves as the condition for measuring the complementarity of the object's items. For example, a product description's image and text items are highly complementary if they mutually qualify the category or increase the product's sales. Similarly, the items of a research paper (e.g., coauthors, references, keywords) are complementary toward publishing the paper at the target venue or receiving many citations. The target label is an indispensable part of the multimodal multi-item data object [18]. Overall, these structural characteristics of multimodal multi-item data provide an excellent opportunity to study complementarity.

Learning the complementarity of multimodal multi-item data faces fundamental challenges. First, the complementary relationships exist in three different kinds of interactions inside a multimodal multi-item data object, and they all need to be carefully considered during the learning process: 1) modality-level intermodal interactions, i.e., the complementary relationship among a variety of modalities; 2) item-level intramodal interactions, i.e., the complementary relationship among items of the same modality; and 3) item-level intermodal interactions, i.e., the complementary relationship among items across different modalities. How to systematically model the complementarity information in both the intra- and intermodal perspectives and how to jointly capture the intermodal complementarity at both the modality level and item level largely remain an open problem.

Manuscript received July 18, 2021; revised December 8, 2021 and March 10, 2022; accepted April 1, 2022. This work was supported in part by the NSF under Grant IIS-1849816, Grant CCF-1901059, Grant IIS-2119531, and Grant IIS-2146761. *(Corresponding author: Meng Jiang.)*

The authors are with the Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN 46556 USA (e-mail: dwang8@nd.edu; tzhao2@nd.edu; wyu1@nd.edu; nchawla@nd.edu; mjiang2@nd.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2022.3165180.

Digital Object Identifier 10.1109/TNNLS.2022.3165180

Second, the metric used for measuring complementarity in the representation space plays a crucial role in preserving such a relationship from multimodal multi-item data. Most existing work [6], [19] relied on the Bayesian personalized ranking (BPR) [20], which was originally proposed as a generic optimization criterion for personalized ranking from implicit feedback. It essentially measured the similarity between (positive and negative) pairs of items and organized the relative position of items in a returned user-specific ranking list. However, this similarity-based metric could not be adequately used for precisely modeling the complementarity. Although being complementary with others is likely to be similar to some extent, similarity does not fully indicate complementarity because the synergistic effect is created by the unique value of each component. One recent work by Wang *et al.* [21] modeled the success of project teaming composed of multiple team members. However, it did not consider the complex structure of multimodal multi-item data objects and ignored the conditional impact from the object's label. Correctly measuring the complementarity and fully considering the structure of multimodal multi-item data are needed for effective complementarity learning.

In this work, we propose a deep representation learning model, called multimodal and multi-item TUBE (M^{2}TUBE), for systematically learning the complementary relationships from multimodal multi-item data. M^{2}TUBE has three major modules: 1) unimodal aggregation; 2) cross-modal fusion; and 3) interactive aggregation. Specifically, the unimodal aggregation module uses items of the same modality to summarize the modality's representations. It has an attentive process that captures the intramodal complementarity at the item level. Then, the cross-modal fusion module combines the modalities' representations into the multimodal object's representations. This design allows the model to capture the intermodal complementarity at the modality level. Next, the interactive aggregation module enhances the unimodal aggregation by conditioning the modality representation compared with item representations of other modalities and the object's label. This interactive attention architecture enables the model to capture the intermodal complementarity at the item level.

To quantify the complementarity in the vector space, we utilize the metric called TUBE distance [22] to measure the difference between the composited data object and its label. In particular, the label is represented as a ray starting from a point in the space that stands for the minimum qualification of belonging to the class. Intuitively, the ε-TUBE distance assigns the same distance value ε for data objects' representation vectors around the label's ray. Recall that in our running example of the product description, images and texts are considered multimodal items, and the product's category is the data object's label. Fig. 1 visualizes two labels ("accessories" and "electronics") as red/blue rays and two multimodal multi-item data objects (product descriptions of headphone and backpack) as paths of the items' representation vectors. The dotted path visualizes the deep aggregation of the items' vectors as the addition of the vectors. The product categories' rays reflect their characteristics. If the product's representation vectors align close to their corresponding category's rays

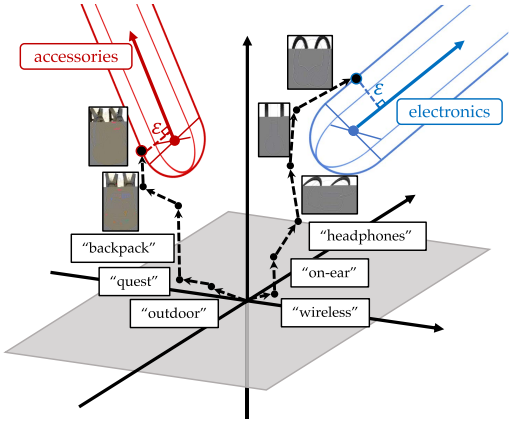


Fig. 1. Two data object examples of learning item complementarity in the vector space. A black dashed arrow denotes an item's representation. The object's vector is obtained via deep multimodal aggregations (simply illustrated as additions) of its items. Each item is a part of the aggregation instead of an independent vector. The red/blue dots stand for the "minimum" vectors of labels, and a label is represented as a ray starting from the dot. The tube-shaped region (not traditionally a sphere) shows the distance between the object's point and the label's ray, allowing the model to explore the possibility of going beyond the minimum.

in terms of TUBE distance, the product description items are complementary to each other to describe the product's characteristics. Based on the TUBE distance, we derive the pairwise item complementarity metric in the representation space conditioning on: 1) representations of the items of the pair in a data object; 2) representations of other items in the object; and 3) representation of the data object's label.

We use real datasets from three domains such as academic publication, social media, and e-commerce to evaluate the proposed model. Experiments demonstrate that M^{2}TUBE outperforms the state-of-the-art by $+6.8\%$ of mean reciprocal rank (MRR) on the task of data object classification and by $+3.0\%$ of MRR on the task of hold-out item prediction. Additionally, qualitative analyses reveal that the learned intra- and intermodal complementarities are significantly different from similarity.

Here we summarize the main contributions of this work.

1) We propose to model the complementarity of multimodal multi-item data from: 1) intramodal interactions of items; 2) intermodal interactions of modalities; and 3) intermodal interactions of items across modalities.
2) We design a novel model that consists of three deep modules to learn complementarity from data. We quantify complementarity in the vector space by TUBE distance.
3) We conduct extensive experiments on datasets from three domains to demonstrate the effectiveness of the proposed model. Also, we provide a suite of case studies to illustrate the captured complementarity.

II. Related Work

A. Complementarity Modeling

Complementarity has been studied in recommender systems to discriminate products between substitute and complement [23]–[26]. McAuley *et al.* [27] proposed to combine topic modeling and supervised link prediction for inferring

product graph representations. Wang *et al.* [28] incorporated path constraints in pairwise relational modeling and adopted additional relation-aware parameters to model multi-item relations. Hao *et al.* [29] proposed first to predict complementary product types and then predict the products of each type based on the distant supervision labels. These methods explicitly rely on product items' profile and category information. Complementarity has also been widely studied in the fashion domain as the compatibility for matching different items in an outfit. Cui *et al.* [30] modeled the relations of fashion items into a graph to learn the compatibility embedding. Lin *et al.* [6] proposed a two-stage model to capture the compatibility and user preference among a various number of fashion items for recommendation. The major drawback of these methods is that their latent spaces were built on BPR [20] that measured the relative similarity between item pairs in a user-specific ranking list. Being complementary with others is likely to be similar to some extent. However, similarity does not fully indicate complementarity.

B. Multimodal Learning

A task or dataset can be characterized as multimodal when it includes multiple modalities such as image, audio signal, and text [14], [31]. Conventional multimodal learning tasks such as audio-visual speech recognition have been studied for decades [32]–[34]. Zhu *et al.* [35] proposed to leverage group convolutions in the generator and progressively decreased the group numbers of the convolutions in the decoder for generating multimodal images. Peng *et al.* [36] proposed a self-guided word relation attention scheme and two question adaptive visual relation attention modules for VQA. There is a line of work exploring the opportunity of leveraging external data to improve multimodal learning. Kumar *et al.* [37] proposed a privacy-preserving method which establishes *ad hoc* social networks to augment speech intelligibility. Su *et al.* [38] proposed to use images for disambiguation in unsupervised neural machine translation. Recently, Huang *et al.* [39] proposed an approach utilizing the visual space as the approximate pivot to align the multilingual multimodal embedding space for unsupervised multimodal machine translation. We refer readers to a survey by Baltrušaitis *et al.* [14] on multimodal learning methods and their applications.

C. Multiple Instance Learning

Classic supervised learning discovers hidden relationships between data objects and labels. However, a class label is assigned to a bag (or set) of instances in multiple instance learning, and the instance-level label is not (or partially) known. Murphy *et al.* [40] proposed Janossy pooling expressing a permutation-invariant function as the average of a permutation-sensitive function applied to all rearrangement of the input sequence. By contrast, in our work, a class label is assigned to a multi-item data object where the label at modality or item level is not applicable. Some efforts simultaneously tackle multimodal learning and multiple instance learning. Meng *et al.* [41] presented a hierarchical sequence-attention model for learning set-of-sets embeddings, which essentially is a multimodal and multiple instance learning method. None of the work can model the complementarity among multiple modalities and one or multiple items of each modality.

III. Problem Definition

This section introduces basic concepts and formally defines the research problem.

Definition (Multimodal and Multi-Item Data Object): A multimodal multi-item data object X_i can be characterized by its features of multiple modalities, i.e., $X_i := \{X_i^1, \ldots, X_i^K\}$ where K is the number of modalities. And, each one of its modality X_i^k $(1 \leq k \leq K)$ is composed of a set of one or multiple items, i.e., $X_i^k := \{v_i^{k,1}, \ldots, v_i^{k,S_i^k}\}$ where S_i^k is the number of items in the kth modality of X_i.

For brevity, we refer to a multimodal multi-item data object as a multimodal object throughout this article. We use $\mathcal{V}^k = \bigcup_{i=1}^{N} X_i^k$ to denote the set of all items in the kth modality.

We denote the raw feature vector of an item $v^k \in \mathcal{V}^k$ by $\mathbf{v}^k \in \mathbb{R}^{D_k}$ where D_k is the dimensions of the kth modality. For example, we can use pre-trained word embeddings to represent the semantics of word tokens in the product description. In case of no prior knowledge is available, the one-hot encoding scheme can be used to indicate the identification of the item ($D_k = |\mathcal{V}^k|$). We use the matrix $\mathbf{X}_i^k := [\mathbf{v}_i^{k,1}, \ldots, \mathbf{v}_i^{k,S_i^k}] \in \mathbb{R}^{S_i^k \times D_k}$ to represent the feature vectors of all items in the kth modality of X_i. The jth row of $\mathbf{X}_i^k$ contains the feature vector of item $v_i^{k,j}$ $(1 \leq j \leq S_i^k)$. All vectors in this article are treated as row-vectors.

Each multimodal object X_i is also accompanied with a specific label $y_i \in \mathcal{T} := \{1, \ldots, T\}$ where T is the number of classes. For a product, its label can be the category it belongs to, and, for a research project, the label could be its venue, such as the target conference. Note that different from the classical multiclass classification setting where the label is often treated as a nominal variable for regression, the label y_i of the multimodal object X_i can also be seen as a generalized item complementary to other items/modalities of X_i in our case. We denote label y_i's initial features as $\mathbf{y}_i \in \mathbb{R}^{D_\mathcal{T}}$ where $D_\mathcal{T}$ is the dimensions for label. In analogous to items, if represented in the one-hot encoding scheme, the number of dimensions of the label's representation vector equals the number of classes, i.e., $D_\mathcal{T} = T$.

Now, we formally define the problem of learning complementarity from multimodal multi-item data.

Problem: Given a large dataset $\mathcal{D}$ of N multimodal objects with their paired labels, i.e., $\mathcal{D} := \{(X_1, y_1), \ldots, (X_N, y_N)\}$, we aim to learn: 1) an item embedding function $f_v(\mathcal{D}) : \{\mathcal{V}^k\}_{k=1}^K \rightarrow \mathbb{R}^{D_M}$ that maps each item $v^k \in \mathcal{V}^k$ into a D_M-dim hidden representation $\mathbf{h}^k$ and 2) a label embedding function $f_l(\mathcal{D}) : \mathcal{T} \rightarrow \mathbb{R}^{D_M}$ that maps each label y_i into a hidden representation $\mathbf{g}_i$, where $D_M \ll \min(\{|\mathcal{V}^k|\}_{k=1}^K)$ is the number of dimensions. The output item embeddings $\{\mathbf{h}^k \mid v^k \in \mathcal{V}^k\}_{k=1}^K$ and label embeddings $\{\mathbf{g}_i \mid y_i \in \mathcal{T}\}$ capture the complementarity of each multimodal object $X_i \in \mathcal{D}$ in both the intra and intermodal perspectives.

For obtaining the multimodal object X_i's hidden representation $\mathbf{p}_i$, existing methods rely on the linear dependence

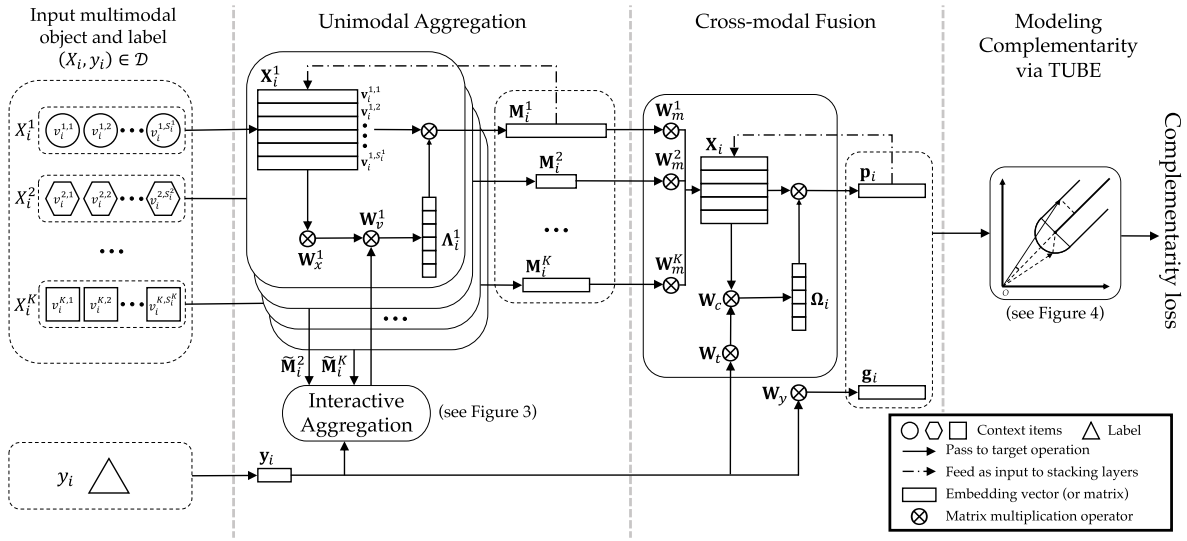


Fig. 2. Framework of M^{2}TUBE for learning complementarity from multimodal multi-item data: 1) unimodal aggregation module for generating representations for each modality individually; 2) cross-modal fusion architecture for summarizing modality embeddings into the multimodal object embedding; and 3) interactive aggregation module for considering interactions between items of different modalities and the condition of labels. The novel metric of TUBE distance [22] between the object and label embeddings is calculated for quantifying complementarity in the latent space.

assumption that $\mathbf{p}_i$ equals the weighted sum of its item embeddings [21], [22]. However, this oversimplified assumption no longer holds in our multimodal multi-item data setting since there is potentially rich complementarity along with multiple modalities and one or multiple items of each modality. In this work, we also want to find a deep nonlinear mapping f_o which can effectively transform item embeddings into the multimodal object's latent representations.

IV. Methodology

In this section, we present a novel approach M^{2}TUBE for modeling the complementarity from multimodal multi-item data in both the intra- and intermodal perspectives. A preliminary version of our work proposed an algorithm TUBE [22] for quantifying complementarity. This article makes one significant step to tell: 1) difference between complementarity and similarity and 2) model design of aggregations on items and modalities to learn complementarity from complex data objects of multimodal items. The proposed framework consists of three major components: 1) unimodal aggregation module for generating representations for each modality individually; 2) cross-modal fusion architecture for summarizing various modality embeddings into the holistic, multimodal object embedding; and 3) interactive aggregation module for considering interactions between items of different modalities and the condition of labels. We leverage the novel metric of TUBE distance [22] between the object and label embeddings as the training objective for extracting the complementarity information. Fig. 2 illustrates the overall framework.

A. Unimodal Aggregation

Given an input multimodal object $X_i \in \mathcal{D}$, we first aggregate its items' information inside each modality to uncover complementarity in the intramodal perspective. It is natural to assume that different modalities have variable sizes and each item is of varying importance grade, we propose an adapted self-attention method to generate a fixed-length embedding by attentively combining item features. Formally, for the kth modality of object X_i, we first calculate the importance values of its items

$$\mathbf{\Lambda}_i^k = \text{softmax}\big(\sigma\left(\mathbf{X}_i^k \cdot \mathbf{W}_x^k\right) \cdot \mathbf{W}_v^k\big) \tag{1}$$

where σ is the nonlinear function of tanh, $\mathbf{W}_x^k \in \mathbb{R}^{D_k \times D_x}$ is the parameter matrix that transforms item features into D_x-dim query vectors, which are further transformed by parameter matrix $\mathbf{W}_v^k \in \mathbb{R}^{D_x \times 1}$ into attention energies. Then, the softmax function squashes all unbounded energy values into attention weights $\mathbf{\Lambda}_i^k \in \mathbb{R}^{S_i^k \times 1}$ summing up to 1. The summarized modality embedding $\mathbf{M}_i^k$ can be generated by multiplying $\mathbf{\Lambda}_i^k$ with the item feature matrix $\mathbf{X}_i^k$

$$\mathbf{M}_i^k = {\mathbf{\Lambda}_i^k}^\top \cdot \mathbf{X}_i^k \tag{2}$$

where $\top$ is the matrix transpose operator and $\mathbf{M}_i^k \in \mathbb{R}^{1 \times D_k}$.

However, this transformation only provides us a single view of the multimodal object attending on its items inside each modality. To generalize into multiple views and stabilize the attention mechanism [42], we can easily generate multihead attention weights $\mathbf{\Lambda}_i^k \in \mathbb{R}^{S_i^k \times H}$ where H is the number of heads by adjusting the size of $\mathbf{W}_v^k \in \mathbb{R}^{D_x \times H}$ in (1). Thus, the result modality embedding matrix $\mathbf{M}_i^k \in \mathbb{R}^{H \times D_k}$ contains H summarized embedding vectors. We can further treat them as the input and stack L layers to extract a highly expressive summarizing modality vector

$$\mathbf{\Lambda}_i^{k^{(l)}} = \text{softmax}\Big(\sigma\left(\mathbf{M}_i^{k^{(l-1)}} \cdot \mathbf{W}_x^{k^{(l)}}\right) \cdot \mathbf{W}_v^{k^{(l)}}\Big) \tag{3}$$

$$\mathbf{M}_i^{k^{(l)}} = \left(\mathbf{\Lambda}_i^{k^{(l)}}\right)^\top \cdot \mathbf{M}_i^{k^{(l-1)}} \tag{4}$$

where the superscript (l) indicates the lth layer $(1 \le l \le L)$. The initial input includes all raw item feature vectors, i.e., $\mathbf{M}_i^{k^{(0)}} = \mathbf{X}_i^k$, and the parameter matrix at last layer is fixed at $\mathbf{W}_v^{k^{(L)}} \in \mathbb{R}^{D_x \times 1}$ so that the final output is a single modality

embedding vector $\mathbf{M}_i^{k(L)} \in \mathbb{R}^{1\times D_k}$ summarizing all information of X_i^k. For brevity, we omit the superscript (l) and use $\mathbf{M}_i^k$ to denote the summarized modality embedding of X_i^k after L stacking layers. In practice, we use $L = H = 3$.

Note that there exists some connections between the encoder layer of transformer [42] and our proposed unimodal aggregation and because they are stackable and contain multiple attention views. One critical difference is that the former method is originally designed for natural language with position embeddings enforcing sequence format, which cannot be directly applied to items in a modality of a set structure. In contrast, our method is carefully designed for aggregating a set of items in a modality with an easy extension on additional conditions from other modalities (see Section IV-C).

B. Cross-Modal Fusion

After aggregations on single modalities, we now have X_i's K modality embedding vectors, i.e., $\{\mathbf{M}_i^k\}_{k=1}^K$. Each one of these embeddings $\mathbf{M}_i^k \in \mathbb{R}^{D_k}$ captures the complementary relations of items inside the corresponding modality.

For modeling the intermodal complementarity information, the next step is to fuse all modality embeddings into a holistic latent representation of the multimodal object X_i. A straightforward way is treating $\{\mathbf{M}_i^k\}_{k=1}^K$ as the item embeddings of a "unified" modality, and again apply our proposed unimodal aggregation module (see Section IV-A). Here, we use matrix $\mathbf{X}_i \in \mathbb{R}^{K\times D_M}$ to denote X_i's modality embeddings transformed to the same size. Each row of $\mathbf{X}_i$ contains the mapped D_M-dim embedding through a linear layer $\sigma(\mathbf{M}_i^k \cdot \mathbf{W}_m^k)$ where $\mathbf{W}_m^k \in \mathbb{R}^{D_k\times D_M}$ is the parameter matrix. The final latent representation $\mathbf{p}_i \in \mathbb{R}^{1\times D_M}$ of X_i can be generated as

$$\mathbf{\Omega}_i = \text{softmax}(\mathbf{X}_i \cdot \mathbf{W}_c), \quad \mathbf{p}_i = \mathbf{\Omega}_i{}^\top \cdot \mathbf{X}_i \tag{5}$$

where $\mathbf{W}_c \in \mathbb{R}^{D_M\times 1}$ is the parameter matrix to transform $\mathbf{X}_i$ into the importance values of all modalities $\mathbf{\Omega}_i \in \mathbb{R}^{K\times 1}$. But this overlooks the critical impact from the label y_i on modalities. In other words, the significance of each modality should be conditioned on the multimodal object label y_i. Therefore, we inject the label information by transforming the label feature vector $\mathbf{y}_i$ to attend on all modalities of X_i

$$\mathbf{W}_c = \sigma\left(\mathbf{W}_t \cdot \mathbf{y}_i{}^\top\right) \tag{6}$$

where $\mathbf{W}_t \in \mathbb{R}^{D_M\times D_\mathcal{T}}$ is the parameter matrix to map $\mathbf{y}_i \in \mathbb{R}^{1\times D_\mathcal{T}}$ into a D_M-dim query over all modality embeddings. By substituting (6) into (5), we can fuse all modality embeddings into the final multimodal object embedding $\mathbf{p}_i$, conditioned on the multimodal object label. And, in analogous to the unimodal aggregation module [see (3) and (4)], we can achieve multiple views of cross-modal fusion by replacing the label embedding $\mathbf{y}_i$ in (6) with matrix $\hat{\mathbf{y}}_i \in \mathbb{R}^{H\times D_\mathcal{T}}$ stacking H row duplicates of $\mathbf{y}_i$. So, the multimodal object embedding $\mathbf{p}_i$ captures the intermodal complementarity between different modalities of X_i.

C. Interactive Aggregation

So far, we have attentively aggregated the information of items within each modality and fused their summarizing modality embeddings based on the multimodal object label. However, there are two limitations for capturing the rich complementarity information across different modalities. First, the cross-modal fusion only considers the intermodal complementarity on the upper modality level but ignores the interaction across modalities on the lower item level. There could be rich complementary interactions between items of different modalities. For example, in a research project with modalities of researchers, topics, methods, and datasets, the same researcher could take variable levels of importance for different topics. Second, the unimodal aggregations of items inside each modality are independent of each other, and the multimodal object label should be taken into consideration.

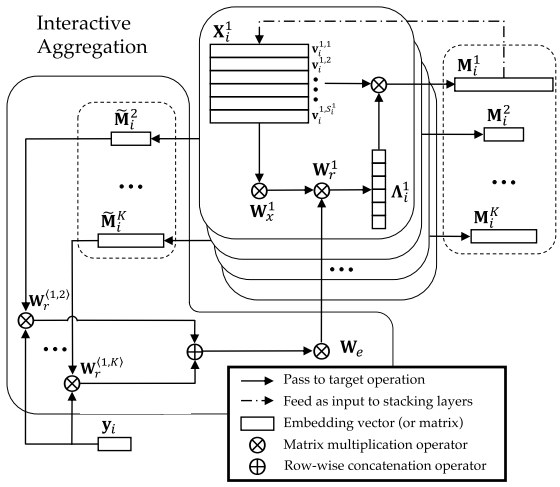


Fig. 3. Interactive aggregation module for generating modality embeddings fully considers the intermodal complementarity information on the item level, as well as the conditional impact of the object label.

These valuable intermodal complementarity information on the item level which conditions on the object label should also be carefully modeled. To this end, we propose an interactive aggregation module (see Fig. 3) to augment the power of M^{2}TUBE for capturing complementarity across modalities. It fully considers the interactions between items of different modalities and the conditional impact of the label when generating the summarized modality embedding. Particularly, for the kth modality of X_i, we first calculate a pseudo modality embedding $\widetilde{\mathbf{M}}_i^{k'} \in \mathbb{R}^{1\times D_{k'}}$ for each one of its other modality $\{k' \mid 1 \le k' \le K, k \ne k'\}$ by applying the mean pooling operator on their item embeddings

$$\widetilde{\mathbf{M}}_i^{k'} = \sum\nolimits_{v_i^{k',j}\in X_i^{k'}} \mathbf{v}_i^{k',j} \Big/ S_i^{k'}. \tag{7}$$

These pseudomodality embeddings are then transformed via a bilinear mapping together with the label feature vector

$$\widetilde{\mathbf{W}}_r^k = \Big\Vert_{k'=1,k'\ne k}^{K} \widetilde{\mathbf{M}}_i^{k'} \cdot \mathbf{W}_r^{\langle k,k'\rangle} \cdot \mathbf{y}_i{}^\top \tag{8}$$

where $\mathbf{W}_r^{\langle k,k'\rangle} \in \mathbb{R}^{D_{k'}\times D_\mathcal{T}}$ is the parameter matrix and $\Vert$ is the row-wise concatenation operator. The result matrix $\widetilde{\mathbf{W}}_r^k \in \mathbb{R}^{(K-1)\times 1}$ contains all interactive information from other modalities and the label. And, it is further transformed into the query vector $\mathbf{W}_r^k \in \mathbb{R}^{D_x\times 1}$ over all items of X_i^k

$$\mathbf{W}_r^k = \sigma\left(\mathbf{W}_e \cdot \widetilde{\mathbf{W}}_r^k\right) \tag{9}$$

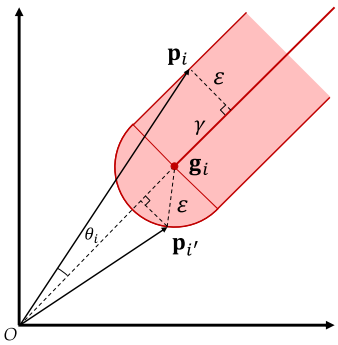


Fig. 4. Metric of TUBE distance for measuring the difference between representations of the multimodal object X_i and its label y_i. Solid vector $\mathbf{p}_i$ (and $\mathbf{p}_{i'}$) refers to the object embedding of X_i and vector $\mathbf{g}_i$ positions the label embedding of y_i. The γ-value demonstrates the achievement of X_i w.r.t. y_i, and the ε value reflects the TUBE distance.

where $\mathbf{W}_e \in \mathbb{R}^{D_x \times (K-1)}$ is the parameter matrix. Thus, we constructed an interactive query $\mathbf{W}_r^k$ capturing the complementary relations between X_i^k's items and other modalities', as well as the condition of label. By substituting (9) into (1) as the $\mathbf{W}_v^k$ matrix, we achieved the goal of modeling the intermodal complementarity on both the upper modality level and the lower item level.

D. Modeling Complementarity via TUBE Distance

Now, we have generated the multimodal object hidden representation $\mathbf{p}_i \in \mathbb{R}^{D_M}$. For calculating the complementarity of items in multimodal object X_i with respect to its label y_i, we leverage the metric of TUBE distance [22]. This method deviates from previous methods using the BPR objective, which constructs the latent space via the pairwise similarity property. We derive the complementarity metric for capturing the synergistic effects between items of multimodal object conditioned on the label based on the TUBE distance, which shapes like a test-tube in the latent space.

We transform the label features into a vector $\mathbf{g}_i \in \mathbb{R}^{D_M}$ of the same size as multimodal object embedding $\mathbf{p}_i$ through a liner layer $\mathbf{g}_i = \sigma(\mathbf{y}_i \cdot \mathbf{W}_y)$ where $\mathbf{W}_y \in \mathbb{R}^{D_\mathcal{T} \times D_M}$ is the parameter matrix. We define the relative achievement of multimodal object X_i with respect to its label y_i as follows:

$$\gamma(X_i \mid y_i) = \frac{\|\mathbf{p}_i\| \cos\theta_i}{\|\mathbf{g}_i\|} \tag{10}$$

where θ_i refers to the angle between the object X_i's embedding vector $\mathbf{p}_i$ and the label y_i's embedding $\mathbf{g}_i$ vector

$$\cos\theta_i = \frac{\mathbf{p}_i \cdot \mathbf{g}_i}{\|\mathbf{p}_i\| \|\mathbf{g}_i\|} \tag{11}$$

and $\gamma \in (-\infty, \infty)$. When the achievement value $\gamma \geq 1$, the projection of the object embedding $\mathbf{p}_i$ on the label's direction locates on the ray starting from $\mathbf{g}_i$ (see Fig. 4).

Next, we define the ε-region of multimodal object X_i with respect to its label y_i in the latent space as follows:

$$\varepsilon(X_i \mid y_i) = \begin{cases} \|\mathbf{p}_i\| \sin\theta_i, & \text{for } \gamma(X_i \mid y_i) \geq 1 \\ \|\mathbf{p}_i - \mathbf{g}_i\|, & \text{for } \gamma(X_i \mid y_i) < 1 \end{cases} \tag{12}$$

where $\varepsilon \in [0, +\infty)$. As shown in Fig. 4, the distance from an object embedding to the corresponding label embedding depends on the relative achievement $\gamma(X_i \mid y_i)$: 1) for a multimodal object of low achievement value $\gamma < 1$ (e.g., $\mathbf{p}_{i'}$ in the figure), the value of ε equals to the normal Euclidean distance between the object's embedding $\mathbf{p}_{i'}$ and label's embedding $\mathbf{g}_i$ and 2) for an object of high achievement value $\gamma \geq 1$ (e.g., $\mathbf{p}_i$ in the figure), the value of ε is defined as the distance from the multimodal object embedding $\mathbf{p}_i$ to the ray starting from the label's embedding vector $\mathbf{g}_i$.

Intuitively, given a fixed value of ε, each label y_i can determine a region in which any multimodal object embedding has a not-larger-than-ε distance to it, which is called the ε-region of the label y_i. The shape of ε-region in the embedding space looks like a test tube, as shown in Fig. 4.

We then normalize the value of ε as follows:

$$p(X_i \mid y_i) = \tanh\left(\frac{1}{\varepsilon(X_i \mid y_i)}\right) \in (0, 1]. \tag{13}$$

It transforms the value of ε into a $(0, 1]$-space for learning so that we can define $p(X_i \mid y_i) = 1$ when $\varepsilon(X_i \mid y_i) = 0$; and, $p(X_i \mid y_i) \to 0$ when $\varepsilon(X_i \mid y_i) \to +\infty$. Thus, the model can output a bounded scalar value p, which can be interpreted as the probability of object X_i being close to its label y_i in the latent space in terms of the TUBE distance.

Based on this normalized TUBE distance p, we derive the item pairwise complementarity as the extent that these two items together being in a specific multimodal object increase the p-value over either one of them being included. That is, given two items v_i and u_i of a multimodal object X_i (assuming $|X_i| \geq 3$) with label y_i, the complementarity between item v_i and item u_i of multimodal object X_i can be defined as

$$c(v_i, u_i \mid X_i) = p(X_i \mid y_i) - \max\{p(X_i \setminus \{v_i\} \mid y_i), p(X_i \setminus \{u_i\} \mid y_i)\} \tag{14}$$

where $X_i \setminus \{v_i\}$, or $X_i \setminus \{u_i\}$, means item v_i, or u_i, is excluded from the multimodal object X_i, respectively. In a trivial case when $X_i = \{v_i, u_i\}$ we can get $c(v_i, u_i \mid X_i) = p(X_i \mid y_i) - \max\{p(v_i \mid y_i), p(u_i \mid y_i)\}$. Then, the pairwise complementarity between items v_i, u_i can be taken as the average over the entire dataset of multimodal objects

$$c(v_i, u_i) = \frac{1}{|\mathcal{D}^{\{v_i, u_i\}}|} \sum_{X_i \in \mathcal{D}^{\{v_i, u_i\}}} c(v_i, u_i \mid X_i) \tag{15}$$

where $\mathcal{D}^{\{v_i, u_i\}}$ is the subset of multimodal objects in $\mathcal{D}$ which contains both item v_i and item u_i. By substituting (13) and (14) into (15), we can see this novel complementarity metric is calculated via the TUBE distance conditioning on: 1) representations of the pair of items; 2) representations of other items of the same multimodal multi-item data object; and 3) representation of the label.

E. Optimization and Negative Sampling

For learning the model and preserving the complementarity of all multimodal objects in the dataset $\mathcal{D}$, the optimization process aims at minimizing the following objective:

$$O = d(\hat{p}(\cdot \mid \cdot), p(\cdot \mid \cdot)) \tag{16}$$

where $d(\cdot, \cdot)$ is the distance between two distributions. We choose to use the KL-divergence of the observed and estimated distributions, and by replacing $d(\cdot, \cdot)$ with the KL-divergence, we get the following objective for optimization:

$$O = -\sum_{(X_i, y_i) \in \mathcal{D}} \hat{p}(X_i \mid y_i) \log p(X_i \mid y_i). \tag{17}$$

We adopt the asynchronous stochastic gradient algorithm (ASGD) [43] for optimizing the objective. In each step, the ASGD algorithm samples one positive example and t negative examples and updates the model parameters, where t is the rate of negative sampling ($t \geq 1$). We presented the objective functions for positive examples O^+ in (17). If a sampled multimodal object $X_{i'}$ and label $y_{i'}$ make a negative example $(X_{i'}, y_{i'})$ that does not exist in dataset $\mathcal{D}$, the objective O^- is

$$O^- = -\sum_{(X_{i'}, y_{i'}) \notin \mathcal{D}} \log \tanh(\varepsilon(X_{i'} \mid y_{i'})) \tag{18}$$

and the final optimization objective is to minimize the overall loss function $O' = O^+ + t \times O^-$. On one hand, the first term of the overall objective [see (17)] pushes the embedding $\mathbf{p}_i$ of multimodal object X_i and the embedding $\mathbf{g}_i$ of label y_i to have a smaller TUBE distance indicating the items in X_i are highly complementary to each other with respect to the label y_i; on the other hand, the second term of the overall objective [see (18)] pulls $\mathbf{p}_i$ of multimodal object X_i and $\mathbf{g}_i$ of label y_i to have a larger TUBE distance indicating items of the object are of low complementarity.

During negative sampling, the label of negative example $y_{i'}$ is sampled from the set of all labels $\mathcal{T}$, which is the same as positive examples. However, the negative multimodal object $X_{i'}$ does not necessarily come from the set of positive multimodal objects $\{X_i\}_{i=1}^N$. The space for sampling negative multimodal object contains the combination (with replacement) of all items $\{\mathcal{V}^k\}_{k=1}^K$ of arbitrary size.

In practice, we find three strategies that are useful.

1) *S1:* Keep the multimodal object fixed, i.e., $X_{i'} = X_i$, and randomly sample a different label $y_{i'} \neq y_i$ ($y_{i'} \in \mathcal{T}$).
2) *S2:* Keep the label fixed, i.e., $y_{i'} = y_i$, and randomly drop an item from a random modality of X_i, i.e., $X_{i'} = X_i \setminus \{v_i^{k',j'}\}$ ($k' \neq k, 1 \leq k' \leq K, 1 \leq j' \leq S_i^{k'}$).
3) *S3:* Keep the label fixed $y_{i'} = y_i$, and randomly sample a modal-distribution-constrained multimodal object $X_{i'}$ based on the negative sampling strategy of [21].

Intuitively, the S1 strategy indicates the positive multimodal object X_i should be tailored for the positive label y_i instead of any other label $y_{i'} \neq y_i$. The S2 strategy emphasizes that any item in the positive multimodal object $v_i^{k,j} \in X_i$ should be indispensable and critical for the complementarity based on the label. The S3 strategy contrasts the positive multimodal object X_i with a randomly assembled multimodal object $X_{i'}$ of the same modal size distribution.

V. Experiments

In this section, we evaluate the effectiveness of M^{2}TUBE against competitive baselines on three real large multimodal datasets. We aim at answering these research questions.

TABLE I
STATISTICS ON THREE MULTIMODAL MULTI-ITEM DATASETS

Dataset	#objects N	#labels T	#mods. K	modalities	#items $\lvert\mathcal{V}^k\rvert$
$\mathcal{D}_\text{p}$	806,211	2,000	3	author keyword reference	103,460 9,962 254,402
$\mathcal{D}_\text{t}$	2,293,560	18	3	word hashtag URL	164,898 47,233 58,398
$\mathcal{D}_\text{e}$	798,368	29	2	image (MBR) word	1,836,246 72,647

1) *RQ1:* How does the proposed method perform compared against the state-of-the-art methods for complementarity learning on multimodal multi-item data?
2) *RQ2:* How do the unimodal aggregation, cross-modal fusion, and interactive aggregation modules of the proposed framework affect the overall performance?
3) *RQ3:* Is the metric of TUBE distance leveraged by the proposed model more effective for quantifying complementarity in latent space compared with other metrics?
4) *RQ4:* What are some concrete examples of the learned complementarity and their differences with similarity?
5) *RQ5:* What are the recommended setting of hyper-parameters for applying the model in practical cases?

A. Datasets

We conduct experiments on three large multimodal datasets from different domains. Statistics are given in Table I.

1) Academic Publication Dataset $\mathcal{D}_p$: We collected 1.3M academic papers published in 13 081 venues from the Microsoft Academic Graph. We built a dataset by limiting the number of venues to 2000. The publication venue is treated as the label of the paper. On average, each paper has 2.8 authors, 5.4 keywords, and 9.2 references (17.4 items in total).

2) Social Media Dataset $\mathcal{D}_t$: We extracted 37M tweets from the public COVID-19-TweetIDs dataset. After preprocessing, we built a dataset of 2 293 560 tweets and 270 529 items of 3 modalities: 1) word; 2) hashtag; and 3) URL. Each tweet has at least one word, one hashtag, and an arbitrary number of URLs. For tweets, we use the number of their retweets in logarithmic scale (and round it) as class label. This can be interpreted as the popularity level. On average, each tweet has 4.7 words, 2.4 hashtags, and 0.8 URLs (7.9 items in total).

3) E-Commerce Dataset $\mathcal{D}_e$: We utilized the dataset of KDD Cup 2020 Challenges for Modern E-Commerce Platform. For each product image, the pre-extracted features and locations of one or more minimum bounding rectangles (MBRs) are provided, along with the classification category of the detected object in the MBR. For assigning the label of product category, we first link each non-phrase with its most frequent object category and then take the majority vote of object category from all non-phrases and MBRs. On average, each product has 2.3 MBRs and 2.4 words (4.7 items in total).

B. Experimental Settings

1) Baseline Methods: We compare M^{2}TUBE against the state-of-the-art models that handle either or both of the

TABLE II
MODEL'S CAPABILITY ON: 1) MULTIMODAL DATA; 2) MULTI-ITEM DATA; AND 3) COMPLEMENTARITY. (◐: PARTIAL SUPPORT, ●: FULL SUPPORT)

Models	Multimodal	Multi-item	*Complementarity*
ATTNMIL	-	●	-
HATS	●	●	-
OUTFITNET	-	●	●
PMSC	-	◐	●
GP-BPR	◐	-	●
CTO-NET	◐	●	●
LEARNSUC	◐	●	●
M^{2}TUBE	●	●	●

multimodal and multi-item data: ATTNMIL [44], HATS [41], and OUTFITNET [6]. Also, we consider these methods that are specifically designed for modeling complementarity information with partial support on multimodal or multi-item data: PMSC [28], GP-BPR [5], CTO-NET [45], and LEARNSUC [21]. A summary of each model's capability is shown in Table II. We follow the recommended setup guideline for all baseline methods whenever possible. For models that cannot fully handle multimodal data, we merge items into a unified modality (two modalities for GP-BPR); and, we only retain the first item of each modality (and drop others) for models that cannot handle multi-item data. We cast the co-occurrence of items inside or across modalities as multirelations for PMSC, and we ignore the preference factor of GP-BPR since the user-item interactions are out of the scope of this work. We implemented CTO-NET following the given guidance.

We set $L = H = 3$ for both the unimodal aggregation and cross-modal fusion modules, and use the default S3 negative sampling strategy for M^{2}TUBE and its variants. When applicable, we use the same embedding dimensions of 256 for all items, modalities, multimodal objects, and object labels. For fair comparisons, we use the same random split of 80%/10%/10% for training/validation/test at each round across methods and report the average performance of five runs. We set a constant learning rate of 1e-4 for the optimizer ASGD with zero momentum factor and weight decay for optimizing the model's parameters. All models are trained for a maximum of 100 epochs with early stopping patience of 5, and the saved best model on the validation set is used for evaluation on the test dataset.

2) Evaluation Protocols: We evaluate the proposed method and baselines through a suite of experiments on two tasks: (T1) label prediction and (T2) hold-out item recommendation. For task T1, we are given a multimodal object $X_i \in \mathcal{D}$, and we aim to predict its correct label y_i from $\mathcal{T}$. And, for T2, we are given a pair of multimodal object and label but with an arbitrary item masked out, i.e., $(X_i \setminus \{v_i^{k,j}\}, y_i)$ where $v_i^{k,j} \in X_i$, and we aim to recover the masked item $v_i^{k,j}$ from modality $\mathcal{V}^k$. We use a fully-connected layer with sigmoid nonlinearity as the predictive model for all methods and both tasks. And, we adopt two sets of evaluation metrics.

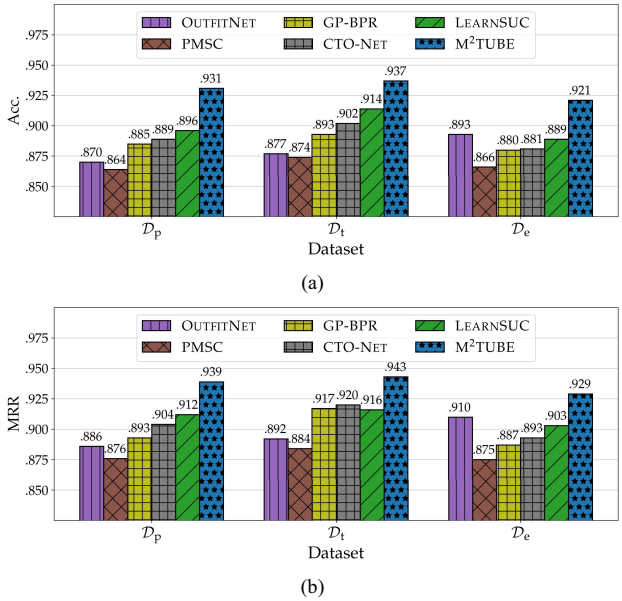


Fig. 5. Performance of M^{2}TUBE and baselines on task T2 for recovering a masked item (author for $\mathcal{D}_p$, hashtag for $\mathcal{D}_t$, and word for $\mathcal{D}_e$) on three datasets in terms of (a) Acc. and (b) MRR.

1) *Accuracy (Acc.) and Hit@k:* There is one true label for T1 (masked item for T2). These two metrics check whether the top-K predictions can find the ground truth ($K = 1$ for Acc.). Higher values indicate better performance.
2) *MRR and Harmonic Mean of Ranks (HMR) [21]:* These two ranking-based metrics check whether the method ranks the ground truth at the top of the returned list. A higher MRR value or a lower HMR value indicates better model performance.

For calculating MRR and HMR, it is time-consuming to enumerate through the candidate space during inference. We adopt the conventional strategy in practice of truncating the returned list to include 10 000 randomly sampled candidates plus the ground truth if the number of candidates is larger.

C. Overall Performance (RQ1)

The performance of M^{2}TUBE and baselines on task T1 for predicting the object label is presented in Table III, and the results on T2 of hold-out item recovery is shown in Fig. 5.

First, we can see baselines ATTNMIL and HATS which do not consider any form of complementarity information of the multimodal object perform inferior to other baselines. HATS generally outperforms ATTNMIL on T1 across metrics and datasets (except for Acc. and Hit Ratio at Top 20 (Hit@20) on $\mathcal{D}_e$) because it can capture both of the multimodal and multi-item composition information. We exclude their results on T2 from Fig. 5 due to their performance being lower than 0.850 of Acc. and MRR. Moreover, there is a large margin between the performance of HATS and any other method that is able to capture complementarity (e.g., -10.1% of MRR relatively over PMSC on $\mathcal{D}_t$). This observation verifies the importance of modeling the complementarity for object label prediction and hold-out item recovery given the multimodal object.

TABLE III

PERFORMANCE OF THE PROPOSED M^{2}TUBE AND BASELINES ON TASK T1 FOR PREDICTING THE LABEL GIVEN THE MULTIMODAL OBJECT, IN TERMS OF ACC., HIT@20, MRR, AND HMR, ON THREE DATASETS. HIGHER VALUES OF ACC., HIT@20, AND MRR INDICATE BETTER MODEL PERFORMANCE. FOR HMR, LOWER VALUES ARE BETTER. BOLD AND UNDERLINE HIGHLIGHT THE BEST AND THE SECOND BEST VALUES

Method	$\mathcal{D}_\text{p}$				$\mathcal{D}_\text{t}$				$\mathcal{D}_\text{e}$			
	Acc.	Hit@20	MRR	HMR	Acc.	Hit@20	MRR	HMR	Acc.	Hit@20	MRR	HMR
ATTNMIL	.584	.653	.601	158.9	.568	.634	.573	67.9	.725	.735	.702	51.4
HATS	.664	.704	.645	133.8	.603	.671	.639	78.1	.714	.732	.717	49.9
OUTFITNET	.837	.864	.880	12.1	.824	.844	.831	15.0	.870	.892	.879	5.7
PMSC	.722	.754	.733	48.8	.673	.707	.711	41.4	.765	.781	.771	32.4
GP-BPR	.813	.821	.824	21.6	.742	.759	.750	39.7	.823	.844	.815	16.9
CTO-NET	.827	.835	.831	17.8	.781	.787	.793	28.6	.849	.873	.881	10.3
LEARNSUC	.844	.886	.873	11.6	.856	.889	**.895**	13.9	.873	.887	.881	5.7
M^{2}TUBE	**.922**	**.950**	**.932**	**8.9**	**.896**	**.910**	.892	**10.4**	**.930**	**.945**	**.936**	**4.2**

TABLE IV

PERFORMANCE OF THE PROPOSED M^{2}TUBE AND ITS VARIANTS ON TASK T1 FOR PREDICTING THE LABEL GIVEN THE MULTIMODAL OBJECT, IN TERMS OF ACC., HIT@20, MRR, AND HMR, ON THREE DATASETS. HIGHER VALUES OF ACC., HIT@20, AND MRR INDICATE BETTER MODEL PERFORMANCE. FOR HMR, LOWER VALUES ARE BETTER. BOLD AND UNDERLINE HIGHLIGHT THE BEST AND THE SECOND BEST VALUES

Method	$\mathcal{D}_\text{p}$				$\mathcal{D}_\text{t}$				$\mathcal{D}_\text{e}$			
	Acc.	Hit@20	MRR	HMR	Acc.	Hit@20	MRR	HMR	Acc.	Hit@20	MRR	HMR
M^{2}TUBE-intra(m)	.834	.865	.872	12.8	.822	.840	.828	15.5	.872	.879	.882	6.4
M^{2}TUBE-inter(m)	.828	.859	.870	13.4	.830	.859	.841	14.8	.879	.889	.887	5.3
M^{2}TUBE-indp	.912	.942	.921	9.23	.878	.896	.880	11.8	.919	.936	**.942**	**4.2**
M^{2}TUBE-BPR	.846	.880	.890	11.7	.847	.874	.879	14.2	.880	.890	.888	5.5
M^{2}TUBE-SUC	.885	.912	.905	11.0	.862	.887	.884	13.3	.886	.906	.908	5.1
M^{2}TUBE	**.922**	**.950**	**.932**	**8.96**	**.896**	**.910**	**.892**	**10.4**	**.930**	**.945**	.936	**4.2**

Secondly, among baselines OUTFITNET, PMSC, GP-BPR, and CTO-NET that are designed for complementarity modeling but can only partially handle the multimodal data or multi-item data, OUTFITNET generally has a better performance. Compared with PMSC's partial support on multi-item data (by constructing an item graph), OUTFITNET effectively considers the complementary relation between multiple items utilizing an attention layer, thus can score an MRR of 0.880 for task T1 on $\mathcal{D}_p$ (+20.1% relatively over PMSC). GP-BPR does not support multi-item data but can partially handle multimodal data by capturing the complementarity within a fixed bi-modalities formulation. It underperforms OUTFITNET on task T1 across datasets and metrics, but can sometimes generate better performance on T2 (e.g., an MRR of 0.917 on $\mathcal{D}_t$ which is +2.8% relatively over OUTFITNET). CTO-NET is the state-of-the-art method for modeling compatibility by formulating items graph to support multi-item data and can partially handle multimodal data by assigning weight values. It is based on a disentangled graph learning scheme and performs on par with GP-BPR. It underperforms OUTFITNET on task T1 and can generate slightly better performance on T2 (e.g., an MRR of 0.920 on $\mathcal{D}_t$ which is +3.1% relatively over OUTFITNET). This observation indicates the existence of complementarity information between multiple items of the same modality and complementary relations between items across modalities. So, models must learn the complementarity information in both the intra- and intermodal perspectives.

Thirdly, the best baseline method LEARNSUC can mainly outperform all other baselines, which can be probably attributed to two major reasons. On the one hand, it is the only baseline that supports complementarity learning on multimodal and multi-item data. Although it is limited to consider multimodalities by manually assigning modality weights. On the other hand, unlike other baselines capturing complementarity via BPR, it models the behavior success as a proxy of the complementary information inside the behavior of a multitype itemset structure. It can score MRRs of 0.895 for T1 on $\mathcal{D}_t$ (which is the global best value) and 0.912 for T2 on $\mathcal{D}_p$ (+2.1% relatively over GP-BPR). This observation justifies the effectiveness of complementarity learning from the complex multimodal and multi-item data and emphasizes the necessity of constructing the latent space of complementarity via an appropriate metric other than similarity-based BPR.

M^{2}TUBE achieves the best performance on both tasks across datasets and metrics (except MRR for T1 on $\mathcal{D}_t$). By utilizing the unimodal aggregations, cross-modal fusion, and interactive aggregations, it fully considers intra- and intermodal complementary relation between multimodal items. It scores MRRs of 0.936 for task T1 on $\mathcal{D}_e$ and 0.943 for task T2 on $\mathcal{D}_t$ (+z6.5% and +5.7% relatively over OUTFITNET). Moreover, by regulating the latent space of complementarity via the metric of TUBE distance [22] conditioning on the label, M^{2}TUBE is highly effective in preserving the captured multimodal complementarity in the latent representations. So, it can score +6.2% higher MRR for T1 on $\mathcal{D}_e$, and +2.9% higher MRR for T2 on $\mathcal{D}_t$, relatively over LEARNSUC. This validates our goal for M^{2}TUBE to model the complementarity in both the intra- and intermodal perspectives.

D. Ablation Studies (RQ2)

We dive deeper into the underlying rationale of M^{2}TUBE's effectiveness by examining the contribution of each one of its components. In particular, we set up controlled experiments on: 1) the unimodal aggregation module (see Section IV-A); 2) the cross-modal fusion module (see Section IV-B); and 3) the interactive aggregation module (see Section IV-C), by building ablated versions of the model and compare M^{2}TUBE against them. Specifically, we test on the following model versions.

1) *M^{2}TUBE-Intra(m):* We remove the unimodal aggregation module and apply mean pooling on item embeddings of each modality, i.e., $\mathbf{M}_i^k = \text{mean}(\mathbf{X}_i^k)$. This model only considers complementarity across modalities.
2) *M^{2}TUBE-Inter(m):* We replace the cross-modal fusion module with a mean pooling operator on summarizing modality embeddings, i.e., $\mathbf{p}_i = \text{mean}(\mathbf{X}_i)$. This version ignores the intermodal complementarity and only considers it in the intramodal perspective.
3) *M^{2}TUBE-Indp:* We disable the interactive aggregation module so this version does not model the intermodal complementarity on the item level.

We focus on task T1 of object label prediction and present the results of M^{2}TUBE and its various versions in Table IV.

First, it is clear to see that the variants of M^{2}TUBE-intra(m) and M^{2}TUBE-inter(m) underperform other model variants. Intuitively, removing the unimodal aggregation module is essentially squashing each modality into an item and casting the multimodal object into a single modality. Similarly, the removal of the cross-modal fusion module can be seen as averaging out all modalities into a unified modality. As a result, both methods failed in "losslessly" extracting the complementarity from the multimodal object. And, they can only perform on par with the baseline method OUTFITNET (see Table III) which learns complementarity solely from multi-item data. M^{2}TUBE-inter(m) performs slightly better than M^{2}TUBE-intra(m) (on $\mathcal{D}_t$ and $\mathcal{D}_e$) probably because there are less number of labels and the intramodal complementarity captured by the the former model plays a more important role. Nevertheless, both these variants are incompetent for fully modeling the complementarity of a multimodal object. In contrast, by leveraging the modules for unimodal aggregation and cross-modal fusion, M^{2}TUBE shows noticeable improvements. This observation demonstrates the effectiveness of the unimodal aggregation module and the cross-modal fusion module in M^{2}TUBE for learning the multimodal object.

Second, we see M^{2}TUBE-indp can produce very competitive performance compared with the former two variants. It scores an MRR of 0.942 on $\mathcal{D}_e$ (the global best) which is +6.8% and +6.2% relatively over M^{2}TUBE-intra(m) and M^{2}TUBE-inter(m). In fact, it can generally outperform the best baseline LEARNSUC (see Table III) except for MRR on $\mathcal{D}_t$, although it does not model the intermodal complementarity of items. In contrast, the proposed model M^{2}TUBE consistently improves the performance on $\mathcal{D}_p$ and $\mathcal{D}_t$ by utilizing the interactive aggregation module. It can produce MRRs of 0.932 and 0.892 on two datasets which are +1.2% and +1.4% relatively over M^{2}TUBE-indp. We also note the improvements of M^{2}TUBE on $\mathcal{D}_e$ is not as stable as expected: it scores the exact value of HMR and slightly lower MRR. Because there are only two modalities and the items of the image modality, i.e., bounding boxes, shows weaker connections to noun phrases of the word modality, the intermodal complementarity on the item level may not be salient. Nevertheless, we verify the contribution of complementary interactions between items across modalities. M^{2}TUBE shows effectiveness in extracting the complementarity on both the modality and item levels.

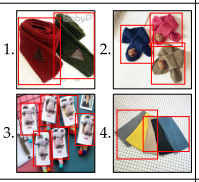
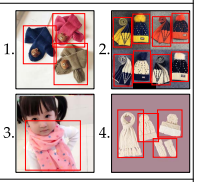
$\mathcal{D}$	Query	Similarity	Complementarity
$\mathcal{D}_p$	Dr. Jure Leskovec	1. Caroline Lo 2. Jaewon Yang 3. Seth Myers 4. Justin Cheng 5. Ashton Anderson 6. Mary McGlohon 7. Gregory Kossinets 8. Siddharth Suri	1. Eric Horvitz 2. Jon Kleinberg 3. Christos Faloutsos 4. Susan Dumais 5. Samuel Madden 6. Carlos Guestrin 7. Daniel Jurafsky 8. Robert West
$\mathcal{D}_t$	#FlattenTheCurve	1. #Covid19 2. #CoronavirusOutbreak 3. #BattlingCovid19 4. #CoronavirusUpdate 5. #StopCovid19 6. #CoronaAlert 7. #PublicHealth 8. #Pandemic	1. #Coronavirus 2. #SocialDistancing 3. #Covid19 4. #StayAtHome 5. #BattlingCovid19 6. #WearAMask 7. #MentalHealth 8. #ProtectOurHeroes
$\mathcal{D}_e$	"Korean-style fashion girl scarf"		
$\mathcal{D}_e$		1. "shoes" 2. "sneakers" 3. "children" 4. "running" 5. "workout" 6. "toddler" 7. "shock-proof" 8. "kids"	1. "sneakers" 2. "comfortable" 3. "shoes" 4. "children" 5. "breathable" 6. "nonslip" 7. "protect" 8. "lightweight"

Fig. 6. Case study on the complementarity learned by M^{2}TUBE. Top similar and complementary results are returned according to the query. First two queries show learned intra-modal complementarity in $\mathcal{D}_p$ and $\mathcal{D}_t$. And, last two queries show the learned intermodal complementarity in $\mathcal{D}_e$.

E. Metrics of Measuring Complementarity (RQ3)

This section further considers the factor of different metrics for quantifying complementarity in the latent space. We replace the metric of TUBE distance for measuring complementarity (see Section IV-D) in M^{2}TUBE with other traditionally used metrics and build the following model variants.

1) *M^{2}TUBE-BPR:* We replace the metric of TUBE distance (see Section IV-D) with the conventionally used BPR [20] for measuring the complementarity information in the latent embedding space.
2) *M^{2}TUBE-SUC:* We replace the metric of TUBE distance (see Section IV-D) with the unconditioned behavior success metric LEARNSUC [21] for measuring complementarity.

We test the effectiveness of these variants equipped with different metrics, and the results are also presented in Table IV. We can see that M^{2}TUBE-SUC consistently outperforms

M^{2}TUBE-BPR across datasets. This observation can be easily explained because BPR is a similarity-based metric that pulls similar embeddings closer in the hidden space. While being similar is likely to indicate being complementary, fully modeling the complementarity also needs to consider the uniqueness of each item, as well as the condition of label [22]. M^{2}TUBE-SUC utilizes the distance metric of LEARNSUC [21] learning the unconditional composition of the multimodal object and produces a better performance. However, it has limited capability in handling multimodalities and label conditions and performs on par with the best baseline method LEARNSUC (see Table III). In contrast, the proposed M^{2}TUBE leverages the metric of TUBE distance [22] in the latent space for quantifying the complementarity, which enhances the metric of LEARNSUC by explicitly conditioning the multimodal object on its label. In this way, M^{2}TUBE outperforms all its ablated variants, and it scores MRRs of 0.932 and 0.936 on $\mathcal{D}_p$ and $\mathcal{D}_e$ which are $+3.0\%$ and $+3.1\%$ relatively over M^{2}TUBE-SUC. This observation justifies the importance of measuring complementarity via TUBE.

F. Qualitative Analysis (RQ4)

In Fig. 6, we provide data examples for illustrating the characteristics of complementarity learned by the M^{2}TUBE, and highlight its differences with similarity in both the intra- and intermodal perspectives. First, our preliminary work in [22] gives a comprehensive qualitative analysis on the intra-modal complementarity on academic data. Second, we similarly illustrate the learned intra-modal complementarity on social media data by using the hashtag #FlattenTheCurve as a query to look up its most similar and complementary hashtags. We can see most of the top eight similar hashtags are popular and general ones that are likely to be included in any random covid-related tweet: 1) #Covid19, #CoronavirusOutbreak, #CoronavirusUpdate, and #CoronaAlert are overused hashtags with limited discriminative power and 2) #BattlingCovid19 and #StopCovid19 serve more like the general covid morale slogans. By contrast, we can perceive much more information from the top eight complementary hashtags covering specific measures implemented for the goal of #FlattenTheCurve: 1) #StayAtHome and #MentalHealth are related to the urgent lockdown across countries and cities; 2) #SocialDistancing and #WearAMask reflect the specific countermeasures recommended by the government; and 3) #ProtectOurHeroes shows people's sympathy and support on the front line workers. A good choice of complementary hashtags to be used in a tweet can effectively convey valuable content and potentially receive wider public attention.

Third, we demonstrate the learned intermodal complementarity at the modality level by using a short query sentence of words to find the most similar and complementary images in $\mathcal{D}_e$ since each image naturally contains multiple items of bounding boxes (red rectangles in the figure). We can see that each most similar image of the query "Korean-style fashion girl scarf" includes multiple scarfs of identical style but different colors. By contrast, the second and the fourth images ranked by the complementarity of the query include products that are indeed complementary to the scarf, e.g., paired hats in the second image and gloves in the fourth image. The third complementary image is a try-on of the query scarf. This verifies that the proposed model can effectively capture intermodal complementarity at the modality level.

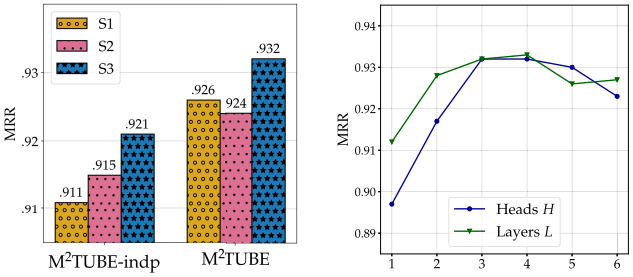


Fig. 7. Parameter sensitivity of the proposed M^{2}TUBE on (a) negative sampling strategy and (b) number of attention heads H and layers L.

Finally, for demonstrating the learned intermodal complementarity at the item level, we use a bounding box from a product image depicting a pair of kids' shoes in $\mathcal{D}_e$ as a query to look up the most similar and complementary items of the other modality, i.e., word tokens. We can see all the top eight words ranked by the similarity metric generally describe the identity of the query product: 1) "shoes," "sneakers," "running," and "workout" refer to the basic functionality of sports shoes and 2) "children," "toddler," and "kids" indicate the product is designed for children. By contrast, the top eight complementary words cover more aspects of the product characteristic besides its functionality: 1) "comfortable" and "breathable" are valuable properties of sports shoes that most users are willing to pay for and 2) "nonslip," "protect," "lightweight" are product features that parents especially pay attention to when buying shoes for their kids. We conclude that M^{2}TUBE is highly effective in learning the complementarity in both the intra- and intermodal perspectives.

G. Parameter Sensitivity (RQ5)

We test the sensitivity of M^{2}TUBE on hyper-parameters: 1) negative sampling strategies and 2) number of attention heads H and layers L. In Fig. 7(a), we report the MRR of M^{2}TUBE and its variant on T1 in D_p. We can see the S3 strategy of sampling modal-distribution-constrained multimodal object produces better performance; S1 works betters for M^{2}TUBE, and S2 is better for the model variant. In Fig. 7(b), we report the MRR of M^{2}TUBE on T1 by varying the values of H and L. We observe that M^{2}TUBE can achieve optimal performance when H and L is at 3 or 4. Setting these values smaller limit the model's expressiveness, and larger values will introduce additional noise. We recommend setting H and L within a reasonable range of $[2, 5]$ in practice.

VI. CONCLUSION

In this work, we proposed to learn complementarity from complex multimodal data. We designed a novel approach to model the complementarity of: 1) intramodal interactions of items; 2) intermodal interactions of modalities; and

3) intermodal interactions of items across different modalities. Our approach has three deep modules: unimodal aggregation, cross-modal fusion, and interactive aggregation. It uses the novel metric of TUBE distance to quantify the complementarity. Extensive experiments demonstrated the effectiveness of the proposed model. Future directions include: 1) examining the decay factor of complementarity when modalities are of large sizes and 2) leveraging free external data to align modalities.

REFERENCES

[1] P. L. Williams and R. D. Beer, "Nonnegative decomposition of multivariate information," 2010, *arXiv:1004.2515*.

[2] N. Timme, W. Alford, B. Flecker, and J. M. Beggs, "Synergy, redundancy, and multivariate information measures: An experimentalist's perspective," *J. Comput. Neurosci.*, vol. 36, no. 2, pp. 119–140, Apr. 2014.

[3] V. Griffith and C. Koch, "Quantifying synergistic mutual information," in *Guided Self-Organization: Inception*. Berlin, Germany: Springer, 2014, pp. 159–190.

[4] X. Song, F. Feng, J. Liu, Z. Li, L. Nie, and J. Ma, "NeuroStylist: Neural compatibility modeling for clothing matching," in *Proc. ACM Int. Conf. Multimedia*, Oct. 2017, pp. 753–761.

[5] X. Song, X. Han, Y. Li, J. Chen, X.-S. Xu, and L. Nie, "GP-BPR: Personalized compatibility modeling for clothing matching," in *Proc. ACM Int. Conf. Multimedia*, Oct. 2019, pp. 320–328.

[6] Y. Lin, M. Moosaei, and H. Yang, "OutfitNet: Fashion outfit recommendation with attention-based multiple instance learning," in *Proc. Web Conf.*, Apr. 2020, pp. 77–87.

[7] D. Kim, K. Saito, K. Saenko, S. Sclaroff, and B. Plummer, "Mule: Multimodal universal language embedding," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 11254–11261.

[8] M. J. Cardoso *et al.*, Eds., *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction With MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings*, vol. 10553. Springer, 2017.

[9] D. H. Park *et al.*, "Multimodal explanations: Justifying decisions and pointing to the evidence," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8779–8788.

[10] L. Ye, M. Rochan, Z. Liu, X. Zhang, and Y. Wang, "Referring segmentation in images and videos with cross-modal self-attention network," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jan. 26, 2021, doi: 10.1109/TPAMI.2021.3054384.

[11] W. Yu, M. Yu, T. Zhao, and M. Jiang, "Identifying referential intention with heterogeneous contexts," in *Proc. Web Conf.*, Apr. 2020, pp. 962–972.

[12] R. Felix, B. G. V. Kumar, I. Reid, and G. Carneiro, "Multi-modal cycle-consistent generalized zero-shot learning," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 21–37.

[13] A. Furnari and G. Farinella, "What would you expect? Anticipating egocentric actions with rolling-unrolling LSTMs and modality attention," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6252–6261.

[14] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019.

[15] P. Gao *et al.*, "Dynamic fusion with intra- and inter-modality attention flow for visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6639–6648.

[16] Y. Huang, J. Wang, and L. Wang, "Few-shot image and sentence matching via aligned cross-modal memory," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jan. 18, 2021, doi: 10.1109/TPAMI.2021.3052490.

[17] K. Wang, M. Bansal, and J.-M. Frahm, "Retweet wars: Tweet popularity prediction via dynamic multimodal regression," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 1842–1851.

[18] X. Xu, K. Lin, Y. Yang, A. Hanjalic, and H. T. Shen, "Joint feature synthesis and embedding: Adversarial cross-modal retrieval revisited," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Dec. 17, 2020, doi: 10.1109/TPAMI.2020.3045530.

[19] R. Yin, K. Li, J. Lu, and G. Zhang, "Enhancing fashion recommendation with visual compatibility relationship," in *Proc. Web Conf.*, 2019, pp. 3434–3440.

[20] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme, "BPR: Bayesian personalized ranking from implicit feedback," 2012, *arXiv:1205.2618*.

[21] D. Wang, M. Jiang, Q. Zeng, Z. Eberhart, and N. V. Chawla, "Multi-type itemset embedding for learning behavior success," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2018, pp. 2397–2406.

[22] D. Wang, T. Jiang, N. V. Chawla, and M. Jiang, "TUBE: Embedding behavior outcomes for predicting success," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2019, pp. 1682–1690.

[23] Y. Zhang, H. Lu, W. Niu, and J. Caverlee, "Quality-aware neural complementary item recommendation," in *Proc. ACM Conf. Recommender Syst.*, Sep. 2018, pp. 77–85.

[24] J. Bai *et al.*, "Personalized bundle list recommendation," in *Proc. Web Conf.*, May 2019, pp. 60–71.

[25] D. Wang *et al.*, "Calendar graph neural networks for modeling time structures in spatiotemporal user behaviors," in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2020, pp. 2581–2589.

[26] D. Wang, Q. Zeng, N. V. Chawla, and M. Jiang, "Modeling complementarity in behavior data with multi-type itemset embedding," *ACM Trans. Intell. Syst. Technol.*, vol. 12, no. 4, pp. 1–25, Jun. 2021.

[27] J. McAuley, R. Pandey, and J. Leskovec, "Inferring networks of substitutable and complementary products," in *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2015, pp. 785–794.

[28] Z. Wang, Z. Jiang, Z. Ren, J. Tang, and D. Yin, "A path-constrained framework for discriminating substitutable and complementary products in e-commerce," in *Proc. 11th ACM Int. Conf. Web Search Data Mining*, Feb. 2018, pp. 619–627.

[29] J. Hao *et al.*, "P-companion: A principled framework for diversified complementary product recommendation," in *Proc. 29th ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2020, pp. 2517–2524.

[30] Z. Cui, Z. Li, S. Wu, X.-Y. Zhang, and L. Wang, "Dressing as a whole: Outfit compatibility learning based on node-wise graph neural networks," in *Proc. Web Conf.*, May 2019, pp. 307–317.

[31] X. Deng and P. L. Dragotti, "Deep convolutional neural network for multi-modal image restoration and fusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3333–3348, Oct. 2021.

[32] D. Wang, P. Cui, M. Ou, and W. Zhu, "Deep multimodal hashing with orthogonal regularization," in *Proc. Int. Joint Conf. Artif. Intell.*, 2015, pp. 1–7.

[33] Y. Yang, D.-C. Zhan, X.-R. Sheng, and Y. Jiang, "Semi-supervised multi-modal learning with incomplete modalities," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 2998–3004.

[34] R. Gao and K. Grauman, "VisualVoice: Audio-visual speech separation with cross-modal consistency," 2021, *arXiv:2101.03149*.

[35] Z. Zhu, Z. Xu, A. You, and X. Bai, "Semantically multi-modal image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5467–5476.

[36] L. Peng, Y. Yang, Z. Wang, Z. Huang, and H. T. Shen, "MRA-Net: Improving VQA via multi-modal relation attention network," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 318–329, Jan. 2022.

[37] A. Kumar, T. Braud, L. H. Lee, and P. Hui, "Theophany: Multimodal speech augmentation in instantaneous privacy channels," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 2056–2064.

[38] Y. Su, K. Fan, N. Bach, C.-C.-J. Kuo, and F. Huang, "Unsupervised multi-modal neural machine translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10482–10491.

[39] P.-Y. Huang, J. Hu, X. Chang, and A. Hauptmann, "Unsupervised multi-modal neural machine translation with pseudo visual pivoting," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 8226–8237.

[40] R. L. Murphy, B. Srinivasan, V. Rao, and B. Ribeiro, "Janossy pooling: Learning deep permutation-invariant functions for variable-size inputs," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–21.

[41] C. Meng, J. Yang, B. Ribeiro, and J. Neville, "HATS: A hierarchical sequence-attention framework for inductive set-of-sets embeddings," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2019, pp. 783–792.

[42] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1–11.

[43] B. Recht, C. Re, S. Wright, and F. Niu, "Hogwild!: A lock-free approach to parallelizing stochastic gradient descent," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 1–9.

[44] M. Ilse, J. M. Tomczak, and M. Welling, "Attention-based deep multiple instance learning," 2018, *arXiv:1802.04712*.

[45] N. Zheng, X. Song, Q. Niu, X. Dong, Y. Zhan, and L. Nie, "Collocation and try-on network: Whether an outfit is compatible," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 309–317.