



# M3PT: A Multi-Modal Model for POI Tagging

Jingsong Yang  
School of Data Science, Shanghai Key  
Laboratory of Data Science,  
Fudan University  
Shanghai, China  
jsyang21@m.fudan.edu.cn

Guanzhou Han  
Alibaba Group  
Hangzhou, China  
gavin.hgz@alibaba-inc.com

Deqing Yang\*  
School of Data Science, Shanghai Key  
Laboratory of Data Science,  
Fudan University  
Shanghai, China  
yangdeqing@fudan.edu.cn

Jingping Liu\*  
School of Information Science and  
Engineering, East China University of  
Science and Technology  
Shanghai, China  
jingpingliu@ecust.edu.cn

Yanghua Xiao  
School of Computer Science,  
Shanghai Key Laboratory of Data  
Science, Fudan University  
Shanghai, China  
shawyh@fudan.edu.cn

Xiang Xu  
Alibaba Group  
Hangzhou, China  
alexander.xx@alibaba-inc.com

Baohua Wu  
Alibaba Group  
Hangzhou, China  
zhengmao.wbh@alibaba-inc.com

Shenghua Ni  
Alibaba Group  
Hangzhou, China  
shenghua.nish@alibaba-inc.com

## ABSTRACT

POI tagging aims to annotate a point of interest (POI) with some informative tags, which facilitates many services related to POIs, including search, recommendation, and so on. Most of the existing solutions neglect the significance of POI images and seldom fuse the textual and visual features of POIs, resulting in suboptimal tagging performance. In this paper, we propose a novel **Multi-Modal Model for POI Tagging**, namely **M3PT**, which achieves enhanced POI tagging through fusing the target POI's textual and visual features, and the precise matching between the multi-modal representations. Specifically, we first devise a domain-adaptive image encoder (DIE) to obtain the image embeddings aligned to their gold tags' semantics. Then, in M3PT's text-image fusion module (TIF), the textual and visual representations are fully fused into the POIs' content embeddings for the subsequent matching. In addition, we adopt a contrastive learning strategy to further bridge the gap between the representations of different modalities. To evaluate the tagging models' performance, we have constructed two high-quality POI tagging datasets from the real-world business scenario of *Ali Fliggy*. Upon the datasets, we conducted the extensive experiments to demonstrate our model's advantage over the baselines of uni-modality and multi-modality, and verify the effectiveness of important components in M3PT, including DIE, TIF and the contrastive learning strategy.

\*The two corresponding authors have the same contribution.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](https://permissions.acm.org).

KDD '23, August 6–10, 2023, Long Beach, CA, USA

© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-0103-0/23/08...\$15.00

<https://doi.org/10.1145/3580305.3599862>

## CCS CONCEPTS

• **Information systems** → **Content ranking**; **Learning to rank**.

## KEYWORDS

point of interest, POI tagging, multi-modality

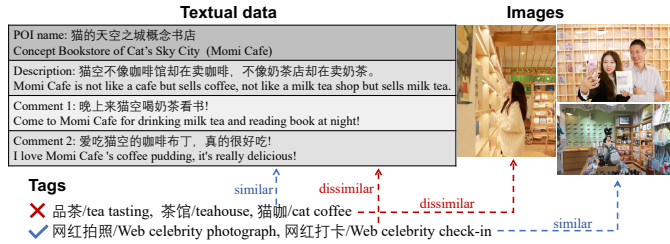
## ACM Reference Format:

Jingsong Yang, Guanzhou Han, Deqing Yang, Jingping Liu, Yanghua Xiao, Xiang Xu, Baohua Wu, and Shenghua Ni. 2023. M3PT: A Multi-Modal Model for POI Tagging. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23)*, August 6–10, 2023, Long Beach, CA, USA. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3580305.3599862>

## 1 INTRODUCTION

A point of interest (POI) is a location that someones may feel helpful or interesting, including a park, restaurant, shop, museum, and so on. In the last decades, various services related to POIs have become very popular on Web. POI tagging, i.e., annotating POIs with some informative tags (labels), which not only help users better understand POIs' characteristics, but also are useful to discover the relatedness or similarities between different POIs. As a result, POI tagging can facilitate many downstream applications, such as POI search [9] and recommendation [18, 41].

Many previous solutions achieve POI tagging based on **textual data** but only have limited performance. These solutions generally extract some features from the textual data relevant to a given POI, to infer the probability that each tag can be used to annotate the POI. The textual data mainly includes users' check-in logs [19, 20], POI taxonomy [39] and descriptions [21]. However, considering only textual data for POI tagging may suffer from the problems of false positive (FP) and false negative (FN). FP refers to the fact that it might annotate the POI with the tags that are semantically related to the textual data but incorrect in fact. FN refers to the fact



**Figure 1: A toy example to demonstrate the effects of incorporating images into POI tagging.**

that it might overlook the tags that are semantically dissimilar to the textual data but correct in fact.

We have found that there is sufficient **visual data** (such as images) related to POIs on many real-world platforms, which is in fact an essential supplement of the textual data, to solve the aforementioned problems. For example, Fig. 1 displays not only the textual data of a specific POI (Momi Cafe) including its name, description and user comments, but also some images posted in the comments. As shown in the figure, the FP and FN problems are alleviated or solved if the images are considered besides the textual data. Specifically, the POI was wrongly labeled with ‘品茶/tea tasting’ due to the comment ‘晚上来猫空喝奶茶看书!/Come to Momi Xafe for drinking milk tea and reading book at night!’. While the images do not display the scene of tea tasting, hence helping the model correct such FP problem. Besides, the incorrect tag ‘茶馆/teahouse’, and ‘猫咖/cat coffee’ could be filtered out since their semantics are not related to the images. For FN, the correct tag ‘网红拍照/Web celebrity photograph’ and ‘网红打卡/Web celebrity check-in’ would be inferred as they have similar semantics to the images, although they are less semantically similar to the texts. The significance of POI images on discovering correct tags inspires us to leverage the textual and visual data simultaneously to achieve effective POI tagging.

A straightforward approach of leveraging images is to transfer an existing multi-modal model (e.g., CLIP [32] and BLIP [23]). However, this approach’s effect is limited for the following reasons, especially given that the tagging task we focus on in this paper is specific to a real-world tour scenario of *Ali Fliggy*<sup>1</sup>. 1) There is a gap between the textual and visual representations in general, which would likely lead to the mismatching between the POI and the gold tags. 2) Leveraging the existing image pre-trained models (such as ViT [7]) might result in unsatisfactory representations of POI images, since our task is a domain-specific tagging task.

To address these problems, in this paper we propose a novel **Multi-Modal Model for POI Tagging**, namely **M3PT**, which is built based on a matching framework. Our model achieves enhanced tagging performance through the full fusion of the given POI’s textual and visual features, and the precise matching between the POI’s multi-modal representation and the candidate tag’s representation. Specifically, in M3PT’s feature encoding module, we devise a **Domain-adaptive Image Encoder**, namely **DIE**, to obtain the embeddings of POI images aligned to their gold tags’ semantics. Then, we design a **Text-Image Fusion** module, namely **TIF**, to fuse the

textual and visual features of the POI. In TIF, we construct a clustering layer followed by an attention layer to distill the significant features for generating the POI’s content embedding. Furthermore, we adopt a contrastive learning strategy to refine the embeddings of different modalities, and thus the cross-modality gap is bridged. In addition, the matching-based framework enables our M3PT to conveniently achieve the precise matching for a new tag, which is beyond the premise in traditional multi-label classification, i.e., the classification towards the close set of predefined labels.

Our contributions in this paper are summarized as follows.

1. To the best of our knowledge, this is the first work to exploit a multi-modal model incorporating the textual and visual semantics to achieve the POI tagging on a real-world business platform. To this end, we propose a novel POI tagging model M3PT based on a multi-modal matching framework, to achieve the precise matching between POIs and tags.

2. We specially devise a domain-adaptive image encoder DIE in our model to obtain the optimal embedding for each input image, which is aligned to the semantics of the image’s gold tags. The image embeddings generated by DIE are better adaptive to the requirements of the real-world scenario, resulting in enhanced tagging performance.

3. We have constructed two high-quality POI tagging datasets from the real-world tour scenario of *Ali Fliggy*, upon which we evaluate the models’s tagging performance. Our extensive experiment results not only demonstrate our M3PT’s advantage over the previous models of uni-modality and multi-modality, but also justify the effectiveness of the important components in M3PT.

## 2 RELATED WORK

### 2.1 POI Tagging Model

Existing POI tagging solutions can be divided into three groups. The first group uses users’ check-in data of POIs as inputs, and extracts discriminative features to predict location labels. Krumm et al. [19, 20] presented a set of manually designed features that are extracted from the public place logs and individual visits. Some other methods [14] leverage the features of user check-in activities and other behavior data to train the generative probabilistic models to infer POI tags. The authors in [39, 42, 45, 52] discussed more comprehensive check-in data including POI unique identifiers, user unique identifiers, the number, time and duration of check-ins, the latitude/longitude of user positions, as well as user demographic information. Label annotation for POIs was first studied in [45], which introduces a collective classification method for feature extraction. The authors in [52] studied how to select the most relevant features for POI tagging. Yang et al. [42] proposed an updatable sketching technique to learn compact sketches from user activity streams and then used a KNN classifier to infer POI tags. Wang et al. [39] proposed a graph embedding method to learn POI embeddings from POI temporal bipartite graphs, which are then used by an SVM classifier of POIs labels.

The second group includes the methods proposed in [13, 21, 51], which also use more fine-grained information of POIs besides user check-in data. For example, [13] includes general tags that may be related to categories and other information, e.g., "godzilla". [51] is the first to use POI name and address tokens, that are obtained by

<sup>1</sup>It is a famous Chinese tour platform and its URL is <https://www.fliggy.com>.

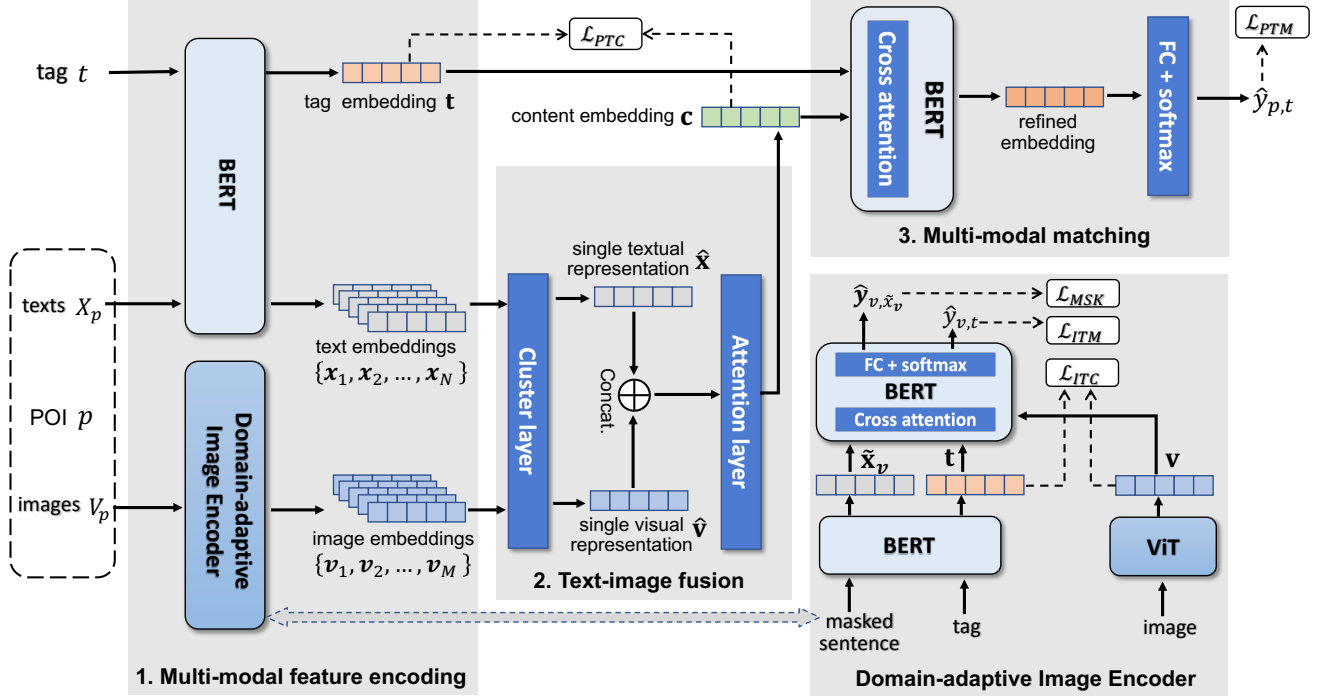


Figure 2: Our proposed M3PT consists of three modules. In the feature encoding module with a domain-adaptive image encoder (DIE), the textual and visual data of the given POI, and the candidate tag are encoded into the feature embeddings, respectively. Then, in the text-image fusion module (TIF), the text and image feature embeddings are fused into the POI’s content embedding. At last, the final probability is computed in the multi-modal matching module based on the matching between the content embedding and tag embedding.

the pre-training on a domain and language-specific corpus. Lagos et al. [21] found that POIs have several distinctive properties, including notably multiscript names, geospatial identities, and temporally defined context. Thus they proposed an approach to complete POI semantic tags in a crowdsourced database automatically.

Similar to our work, the third group leverages image tagging, of which many solutions use ViT [7] as the backbone to accomplish multiple image classification. Although many algorithms have been proposed for automatic image annotation [1, 4, 11], image tag refinement is treated as an independent problem, and has become an exciting issue [8, 26, 29, 46].

## 2.2 Vision-language Pre-trained Model

Vision-language pre-training (VLP) aims to improve the performance of downstream vision and language tasks by pre-training models on large-scale image-text pairs. The pre-training tasks of VLP models mainly include image-text contrastive learning and language modeling (LM) based tasks. VirTex [5], ICMLM [36], and ConVIRT [48] have demonstrated the potential of Transformer-based LM, masked LM, and contrastive LM on learning image representations from the text. CLIP[32], BLIP [23], FILIP [44], ALIGNm [17] and UNIMO [25] mainly make use of cross-modal contrastive learning which aligns the textual and visual information into a unified semantic space. VisualBERT [24], UNITER [3], and M6 [27] employ LM-like objectives, including both masked LM (e.g., Masked Language/Region Modeling) and auto-regressive LM (e.g., image

captioning, text-grounded image generation). In addition, some methods (such as BriVL [16]) rely on a pre-trained object detection model, such as Faster-RCNN [34], to extract regional image features offline. It requires extra labeled bounding-box data and makes the approach less scalable. Recent efforts such as SOHO [15] and SimVLM [40] try to alleviate this burden via visual dictionary or PrefixLM [33].

## 3 OVERVIEW

### 3.1 Problem Formalization

Given a POI  $p$ , the task objective of our model is to filter out some appropriate tags from all candidate tags, to characterize  $p$ . Expressly, in our scenario of Ali Fliggy, some textual contexts and images related to  $p$  are also provided to the model to achieve the tagging task better. The textual contexts include the name, category, description, and user comments of  $p$ . The images include  $p$ ’s main picture displayed on the platform and the pictures posted in user comments. Formally, we denote all textual contexts of  $p$  as  $X_p = \{x_1, x_2, \dots, x_N\}$ , and all related images as  $V_p = \{v_1, v_2, \dots, v_M\}$ . The model should compute the following matching score (probability) between  $p$  and a candidate tag  $t$ ,

$$\hat{y}_{p,t} = M3PT(X_p, V_p, t),$$

where  $M3PT()$  represents all operations in our model. With  $\hat{y}_{p,t}$  and a well-chosen threshold  $\pi$ , the model predicts whether  $t$  should be used to annotate  $p$ .

### 3.2 Solution Framework

As shown in Fig. 2, our proposed M3PT achieving POI tagging pipeline can be divided into the following three phases.

**1. Feature Encoding.** The candidate tag  $t$ , the textual contexts, and images of the given POI  $p$  are encoded into the feature embeddings in this first phase. The relevant operations are conducted in the feature encoding module of our model, which consists of a text encoder and an image encoder. Specifically, we take BERT<sub>base</sub> [6] as the text encoder to encode  $t$  into a tag embedding and encode  $X_p$  into a set of text feature embeddings of  $p$  simultaneously. For the image feature encoding of  $p$ , we adopt a new image encoder DIE, which is specific to our tagging task in Fliggy's domain, as we mentioned in Section . We will introduce the details of the feature encoding module, including DIE, in Section 4.1.

**2. Text-image Fusion.** The operations in the second phase are conducted in the TIF module of our model. The TIF mainly consists of two layers. The first layer is a clustering layer, through which  $p$ 's text embeddings and image embeddings obtained from the first phase are aggregated into a single textual representation and a single visual representation of  $p$ , respectively. Then, these two representations are concatenated and fed into the attention layer of TIF to generate  $p$ 's content embedding for the POI-tag matching in the subsequent phase. As we emphasized in Section , the clustering and attention operations in TIF are proposed to distill the significant features to represent  $p$ 's content, whose details will be introduced in Section 4.2.

**3. Multi-modal Matching.** To conduct the operations in the third phase, we build a multi-modal matching module fed with  $p$ 's content embedding and  $t$ 's tag embedding generated in the previous phases to output the final score  $\hat{y}_{p,t}$ . The main body of this module is also a BERT<sub>base</sub> encoder, whose details will be introduced in Section 4.3.

## 4 METHODOLOGY

### 4.1 Multi-modal Feature Encoding

In the feature encoding module of our model, the multi-modal features of  $p$  (including textual and visual features) and  $t$ 's textual features are first encoded into embeddings, respectively.

Specifically,  $p$ 's textual inputs  $\{x_1, x_2, \dots, x_N\}$  are fed into a text encoder constructed based on BERT<sub>base</sub> to generate a set of  $p$ 's text embeddings. The text encoder is initialized with the first 6 layers of BERT<sub>base</sub>. Formally, for an input text  $x_i (1 \leq i \leq N)$ , we get its feature embedding as

$$\mathbf{x}_i = EN_T(x_i) \in \mathbb{R}^D, \quad (1)$$

where  $EN_T()$  denotes the encoding operations in the text encoder, and  $D$  is the embedding's dimension.

Simultaneously,  $t$  is also fed into the same text encoder to obtain its tag embedding as

$$\mathbf{t} = EN_T(t) \in \mathbb{R}^D. \quad (2)$$

**4.1.1 Domain-adaptive Image Encoder.** Given that our tagging task is specific to the real-world scenario of Fliggy, we propose a new image encoder DIE to encode the input images of  $p$ , i.e.,  $\{v_1, v_2, \dots, v_M\}$ , into a set of image feature embeddings, instead of using an existing image pre-trained model [7, 24]. To adapt to

the particular goal of our POI tagging task, the embedding of an image in our model should be learned concerning the tags with the same semantics as it, i.e., its gold tags. It indicates that the learned embedding of an image should help the model recognize its gold tags. To this end, unlike the traditional image encoders such as ViT [7], which only receive images as input, our DIE takes an image and the text having the same semantics to constitute an input sample. Since DIE is also a pre-trained encoder, we propose a pretraining mask task.

**Mask Learning.** Specifically, given an image  $v$ , we first compose a sentence to indicate  $v$ 's semantics, denoted as  $x_v$ , which includes a gold tag of  $v$ , denoted as  $t_v$ . Then, we convert  $x_v$  into  $\tilde{x}_v$  by replacing one token in  $t_v$  with a special token [MASK]. For example, if an image  $v$  and its annotated tag  $t_v = \text{'cup'}$  are obtained, we have

$$x_v = \text{"This is a cup"}, \quad \tilde{x}_v = \text{"This is a [MASK]}."$$

Accordingly, we take  $(v, \tilde{x}_v)$  as the input of the mask learning task. As displayed in Fig. 2,  $i$  and  $\tilde{x}_v$  are fed into a ViT encoder and a BERT<sub>base</sub> in DIE, to generate  $v$ 's embedding  $\mathbf{v} \in \mathbb{R}^D$  and  $\tilde{x}_v$ 's embedding  $\tilde{\mathbf{x}}_v \in \mathbb{R}^D$ , respectively. Then, based on these two embeddings, DIE computes the following probability distribution indicating which token is on the position of [MASK] as,

$$\hat{\mathbf{y}}_{v, \tilde{x}_v} = \text{Pre}(\mathbf{v}, \tilde{\mathbf{x}}_v), \quad (3)$$

where  $\text{Pre}()$  represents all calculations in the prediction layer of DIE. Specifically,  $\hat{\mathbf{y}}_{v, \tilde{x}_v}$  is a vector of  $U$  dimensions and  $U$  is the size of token vocabulary. Each entry of  $\hat{\mathbf{y}}_{v, \tilde{x}_v}$  is the probability that the corresponding token is on the position of [MASK]. To compute this probability distribution more precisely,  $\mathbf{v}$  is first refined in terms of  $\tilde{\mathbf{x}}_v$  through the cross-attention operations as [37]. Then  $\hat{\mathbf{y}}_{v, \tilde{x}_v}$  is computed based on the refined  $\mathbf{v}$  through a fully-connected layer and softmax operation. At last, a token with the largest probability in  $\hat{\mathbf{y}}_{v, \tilde{x}_v}$  is predicted as on the position of [MASK].

To pretrain DIE, we formulate the mask learning's loss as

$$\mathcal{L}_{MSK} = \sum_{(v, u) \in \mathcal{N}_{MSK}} CE(\mathbf{y}_{v, \tilde{x}_v}, \hat{\mathbf{y}}_{v, \tilde{x}_v}), \quad (4)$$

where  $CE()$  is the cross-entropy function and  $\mathcal{N}_{MSK}$  is the training set.  $\mathbf{y}_{v, \tilde{x}_v} \in \mathbb{R}^U$  is the indicator vector, in which the entries of the real tokens are 1 and the rest entries are 0.

**Contrastive Learning.** In addition, to better align  $v$ 's embedding to its gold tags' embeddings, we further consider a contrastive learning loss. We first define  $v$ 's similarity score to each paired tag  $t$  as

$$s(v, t) = \frac{\exp(\mathbf{v}^\top \mathbf{t} / \tau)}{\sum_{t' \in \mathcal{T}_v} \exp(\mathbf{v}^\top \mathbf{t}' / \tau)}, \quad (5)$$

where  $t'$  is any one tag in  $v$ 's paired tag set  $\mathcal{T}_v$  including its gold tag  $t_v$ , and  $\mathbf{t} \in \mathbb{R}^D$  is  $t$ 's embedding obtained from the BERT<sub>base</sub> encoder. In addition,  $\tau$  is the temperature used to maintain the balance between the alignment and consistency of the contrasted samples [38]. To train the model better in contrastive learning, we tend to select the incorrect tags but semantically similar (having close embeddings) to the image as hard negative samples. Accordingly, a small  $\tau$  makes the model focus on discriminating the hard negative samples, resulting in better performance.

Similarly, a tag  $t$ 's similarity score to each paired image  $v$  is

$$s(t, v) = \frac{\exp(\mathbf{t}^\top \mathbf{v} / \tau)}{\sum_{v' \in \mathcal{V}_t} \exp(\mathbf{t}^\top \mathbf{v}' / \tau)}, \quad (6)$$

where  $v'$  is any one image in  $t$ 's paired image set  $\mathcal{V}_t$  including the gold images with the same semantics as  $t$ . Then, the contrastive learning loss is

$$\mathcal{L}_{ITC} = \frac{1}{2} \left\{ \sum_{(v, t) \in \mathcal{N}_{ITC}} CE(y_{v, t}, s(v, t)) + \sum_{(t, v) \in \mathcal{N}_{ITC}} CE(y_{t, v}, s(t, v)) \right\}, \quad (7)$$

where  $y_{v, t}$  and  $y_{t, v}$  are both the real label indicating the matching between  $v$  and  $t$ .

*Image-tag Matching Learning.* Furthermore, to better achieve the matching between images and tags, we append an image-tag matching loss for DIE's pretraining. We use  $\hat{y}_{v, t}$  to denote the matching probability between  $v$  and  $t$ . As the operations in Equation 3,  $\mathbf{v}$  is first refined in terms of  $\mathbf{t}$  through cross-attention and then used to compute  $\hat{y}_{v, t}$ . Then, the matching loss is

$$\mathcal{L}_{ITM} = \sum_{(v, t) \in \mathcal{N}_{ITM}} CE(y_{v, t}, \hat{y}_{v, t}). \quad (8)$$

After all, the overall loss of DIE's pre-training is<sup>2</sup>

$$\mathcal{L}_{DIE} = \mathcal{L}_{MSK} + \mathcal{L}_{ITC} + \mathcal{L}_{ITM}. \quad (9)$$

With the pre-trained DIE, the feature embedding of an image  $v_i (1 \leq i \leq M)$  is obtained by

$$\mathbf{v}_i = DIE(v_i) \in \mathbb{R}^D, \quad (10)$$

where  $DIE()$  denotes the encoding operations in DIE.

## 4.2 Text-image Fusion

The operations in this phase are conducted in the TIF module, where the textual and visual embeddings of  $p$  are fused into a condensed representation, namely content embedding, to achieve the subsequent matching conveniently.

*4.2.1 Clustering Layer.* The first step in TIF is to aggregate the multiple embeddings from either modality into a single embedding. To this end, we first build a clustering layer in TIF to perform the following operations.

For an image  $v$ 's feature embedding  $\mathbf{v} = [v_1, v_2, \dots, v_D]$  obtained by Equation 10, each  $v_i (1 \leq i \leq D)$  can be regarded as a frame-level descriptor of  $v$ . We used a clustering algorithm the same as NeXtVLAD [28] to cluster the  $d$  frame-level descriptors of  $v$  into  $K$  groups. Suppose the  $k$ -th cluster's centroid is  $c_k (1 \leq k \leq K)$ , we refine  $v_i$  in terms of  $c_k$  as

$$v_i^k = \alpha_k(\mathbf{v})(v_i - c_k), \quad (11)$$

where  $\alpha_k(\mathbf{v})$  is a function measuring  $\mathbf{v}$ 's proximity to the  $k$ -th cluster. We adopt a fully-connected layer with softmax activation to obtain  $\alpha_k(\mathbf{v})$  as

$$\alpha_k(\mathbf{v}) = \frac{e^{\mathbf{w}_k^\top \mathbf{v} + b_k}}{\sum_{k'=1}^K e^{\mathbf{w}_{k'}^\top \mathbf{v} + b_{k'}}}, \quad (12)$$

<sup>2</sup>The same coefficient for each sub-loss is the best choice based on our tuning studies.

where  $\mathbf{w} \in \mathbb{R}^D$  and  $b$  are both trainable parameters. Furthermore, there are  $K$  groups of  $\mathbf{w}, b$  in total.

Then, to aggregate the refined frame-level descriptors of all  $M$  image embeddings, we get the  $i$ -th aggregated descriptor in terms of the  $k$ -th cluster through the following sum pooling,

$$\hat{v}_i^k = \sum_{j=1}^M v_i^{j, k}, \quad 1 \leq i \leq D, 1 \leq k \leq K, \quad (13)$$

where  $v_i^{j, k}$  is the  $i$ -th refined frame-level descriptor of the  $j$ -th image embedding obtained by Equation 11. Accordingly, we have  $D \times K$  aggregated descriptors in total, which are then reduced into a single visual representation of  $H$  dimensions, denoted as  $\hat{\mathbf{v}} \in \mathbb{R}^H$ , through a fully-connected layer.

To aggregate the  $N$  text embeddings generated in the previous phase into a single textual representation, denoted as  $\hat{\mathbf{x}} \in \mathbb{R}^H$ , we adopt the same operations introduced above for generating the single visual representation.

*4.2.2 Attention Layer.* Next, we input the concatenation of  $\hat{\mathbf{x}}$  and  $\hat{\mathbf{v}}$  into the attention layer in TIF. This layer reassigns the feature weights in the input through the attention and reshape operations and outputs an embedding of  $D$  dimensions. Thus, the more significant features are highlighted in the output, which is just  $p$ 's content embedding, denoted as  $\mathbf{c} \in \mathbb{R}^D$ . Overall,  $\mathbf{c}$  condenses the significant multi-modal features of  $p$ .

## 4.3 Mutil-modal Matching

The objective of the multi-modal matching module in the third phase is to achieve the precise matching between  $p$  and  $t$ , based on  $p$ 's content embedding  $\mathbf{c}$  and  $t$ 's tag embedding  $\mathbf{t}$ .

In fact, the architecture of this module is similar to the prediction layer in DIE since both of their objectives are to achieve the matching between two objects. Specifically, the candidate tag's embedding  $\mathbf{t}$  is refined in terms of  $p$ 's content embedding  $\mathbf{c}$  through the cross-attention operations and then used to compute  $\hat{y}_{p, t}$  through a fully-connected layer and softmax operations. As last, M3PT predicts that  $t$  is a correct tag of  $p$  if  $\hat{y}_{p, t} > \pi$ .

## 4.4 Model Training

Similar to Equation 8, the main loss of our model training is the following POI-tag matching loss,

$$\mathcal{L}_{PTM} = \sum_{(p, t) \in \mathcal{N}_{PTM}} CE(y_{p, t}, \hat{y}_{p, t}), \quad (14)$$

where  $\mathcal{N}_{PTM}$  is the training set consisting of POI-tag pairs.

To improve the alignment of  $p$ 's content embedding and  $t$ 's tag embedding, we also append a contrastive learning loss the same as Equation 5~7 for our model training, except that  $\mathbf{v}$  is replaced with  $\mathbf{c}$  when computing  $s(p, t)$  and  $s(t, p)$ . Thus, this contrastive learning loss is

$$\mathcal{L}_{PTC} = \frac{1}{2} \left\{ \sum_{(p, t) \in \mathcal{N}_{PTC}} CE(y_{p, t}, s(p, t)) + \sum_{(p, t) \in \mathcal{N}_{PTC}} CE(y_{t, p}, s(t, p)) \right\}. \quad (15)$$

And the overall loss of M3PT's training is

$$\mathcal{L} = \mathcal{L}_{PTM} + \alpha \mathcal{L}_{PTC}, \quad (16)$$

where  $\alpha$  is a controlling parameter.

**Table 1: Statistics of our datasets.**

Dataset name	POI #	tag #	POI-tag #	tag # per POI	image # per POI	avg. text # per POI
MPTD1	63,415	354	197,254	3	64	126
MPTD2	6,415	286	27,486	4	8	16

## 5 EXPERIMENTS

### 5.1 Dataset Construction

We have constructed the following two datasets for our evaluation experiments, of which the detailed statistics are listed in Table 1.

**MPTD1:** The first dataset, named MPTD1 (Multi-modal POI Tagging Dataset), was constructed directly from the real-world tour scenario of Ali Fliggy, including more than 60,000 POIs, 354 unique tags, and more than 190,000 POI-tag pairs. For each POI in this dataset, its related textual contexts (texts in short) include its full name, introduction, categories, and user comments collected from Fliggy’s website. Each POI’s related images include the main picture shown on the top of its introduction page, as well as the pictures posted in its user comments. We first collected the original tags for each POI in some ways, including basic rules, manual selections, semantic-based algorithms, etc. When sufficient original tags had been collected, the experts verified and refined them based on their actual effects on Fliggy’s tour platform. At last, the reserved tags were identified as the gold (ground-truth) tags for these POIs. All POIs were divided into the training set, validation set, and test set according to 8:1:1. In each set, besides the POIs, their gold tags, texts, and images were also included.

**MPTD2:** Although the texts and images of each POI in MPTD1 are sufficient, the model training is time-consuming if we feed all of them into the model. In addition, most of them are not semantically related to the POI’s tags. Thus, to compare all models’ capabilities of leveraging a POI’s textual and visual features that are semantically related to its gold tags more efficiently, we further constructed a more concise dataset MPTD2. We randomly selected about one-tenth of the POIs in MPTD1, together with their tags, texts, and images. For each selected POI, we recruited some volunteers to check its texts and images and only retain those with similar semantics to its tags. For example, only the comments that directly mention the tags or are highly semantically related to the tags were retained. Similarly, only the images verified as semantically related to the POI’s tags were retained. As a result, the texts and images retained in MPTD2 can be directly leveraged as pieces of evidence to judge the POI’s matching to its tags. The ratio of the training set, validation set, and test set in MPTD2 is also 8:1:1.

### 5.2 Experimental Setup

**5.2.1 Baselines.** We compared our M3PT with the following models in our experiments, which can be categorized into two groups of uni-modal models and one group of multi-modal models.

The baselines in the first group only leverage the textual features of POIs to achieve POI tagging. In these models, a given POI’s texts are fed into their encoders to generate the textual embedding of the POI, which is used as the POI’s content embedding. This group includes **BERT** [6], **ALBERT** [22] and **ERNIE** [49]. The first two are both classic pre-trained language models. ERNIE incorporates

knowledge maps into the pretraining task to improve its representation capability, which has been successfully employed in the previous multi-label classification task.

The baselines in the second group are also uni-modal models, which only leverage the POIs’ images to achieve POI tagging. Specifically, only the images of a given POI are input into the models to generate the visual feature embedding used as its content embedding. This group includes the ResNet-based models including **ResNet-101** [12], **ResNet101(ASL)** [2] and **TResNet** [35]. All of them have been widely used to encode images in the computer vision (CV) community. In addition, we also considered **ViT** [7] and **ViT-Q2L** [30]. Inspired by the success of Transformer [37] in natural language processing (NLP), these two models were built based on the architecture of self-attention and cross-attention, showing their advantages over the image encoders based on convolutional neural networks (CNNs).

As M3PT, the baselines in the third group are both multi-modal models, including **M3TR** [50] and **TAILOR** [47]. M3TR employs the vision Transformer and builds a cross-modal attention module and a label-guided augmentation module to achieve multi-modal multi-label classification better. TAILOR first encodes the uni-modal features of texts and images and then fuses them into the cross-modal features as the encoder inputs to model the relationship between the modality and each label.

**5.2.2 Evaluation Metrics.** Our POI tagging task is just a multi-label classification if we regard the set of all tags as the label set of the POI classification. So we used the following metrics to evaluate all compared models’ performance in our experiments. We first considered the label-based classification metric **M-P** (Macro Precision), **M-R** (Macro Recall) and **M-F1** (Macro F1) [43]. When computing their scores, we first took a tag along with all POIs annotated by it as one sample, and reported the average score of all samples. Second, we also considered the object-based classification metrics, including **P-e** (Precision-exam), **R-e** (Recall-exam), and **F1-e** (F1-exam) [10]. When computing three metrics’ scores, we took a POI along with all of its tags as one sample, and reported the average score of all samples. In addition, we also considered **HLS** (Hamming Loss) [10], which is used to measure the misclassification samples on a single label (tag). Thus, a more miniature HLS score indicates the model’s better performance.

**5.2.3 Hyperparameter Settings.** We adopted AdamW [31] as the optimizer. Moreover, we set the initial learning rate to 1e-4 for pre-training. With steps increasing, we decreased the learning rate linearly to 1e-5. The settings of some crucial hyperparameters in

**Table 2: Hyperparameter settings of M3PT.**

Notation	Value	Description
$\tau_1$	0.12	temperature in $\mathcal{L}_{PTC}$
$\tau_2$	0.08	temperature in $\mathcal{L}_{ITC}$
$\pi$	0.5	threshold of prediction probability
$D$	768	embedding dimension
$K$	64	cluster number in TIF
$H$	414	textual/visual representation dimension in TIF
$\alpha$	0.5	controlling parameter in loss $\mathcal{L}$ (Equation 16)



**Table 3: Performance comparison results show M3PT’s advantage over the baselines on the two datasets, where the best and second-best scores in each group are in bold and underlined, respectively.**

Model	MPTD2							MPTD1						
	M-F	M-R	M-F1	P-e	R-e	F1-e	HLS	M-F	M-R	M-F1	P-e	R-e	F1-e	HLS
BERT	74.67	26.65	39.26	75.72	29.51	42.40	0.0835	45.25	12.18	19.46	54.33	20.34	29.59	0.1285
ALBERT	83.31	<u>29.53</u>	44.16	85.14	34.26	48.85	<u>0.0627</u>	54.23	<u>13.39</u>	21.35	64.55	23.90	34.02	<u>0.0610</u>
ERNIE	<u>84.16</u>	29.42	<u>45.56</u>	<u>86.63</u>	<u>36.06</u>	<u>50.93</u>	0.0683	<u>54.87</u>	13.29	<u>21.93</u>	<u>65.26</u>	<u>23.97</u>	<u>35.06</u>	0.0857
M3PT(text)	<b>86.34</b>	<b>32.57</b>	<b>47.11</b>	<b>88.38</b>	<b>38.63</b>	<b>53.72</b>	<b>0.0413</b>	<b>56.42</b>	<b>15.37</b>	<b>23.65</b>	<b>67.04</b>	<b>25.51</b>	<b>36.95</b>	<b>0.0512</b>
Improv.%	2.59	10.29	3.40	2.02	7.13	5.48	34.13	2.82	14.79	7.84	2.73	6.42	5.39	16.07
ResNet101	14.24	<u>9.40</u>	11.32	18.15	10.68	13.44	0.1141	10.21	7.87	8.81	13.60	9.48	11.17	0.1364
ResNet101(ASL)	14.37	9.36	11.36	<u>18.97</u>	10.07	13.12	0.1163	10.04	8.07	8.94	13.51	9.22	10.96	0.1328
TResNet	14.73	8.91	11.07	17.16	10.50	13.02	<b>0.0915</b>	10.80	<u>8.19</u>	9.31	13.60	10.13	11.61	<b>0.1243</b>
ViT	15.14	9.27	11.74	18.74	11.26	<u>14.06</u>	0.1149	10.14	<b>8.27</b>	9.11	13.06	9.61	11.07	0.1470
ViT(Q2L)	<u>15.26</u>	9.39	<u>12.06</u>	17.63	<b>11.89</b>	13.69	0.0952	<u>11.28</u>	8.14	<u>9.45</u>	<u>14.40</u>	<b>10.63</b>	<u>12.23</u>	0.1315
M3PT(image)	<b>15.85</b>	<b>10.95</b>	<b>12.93</b>	<b>19.34</b>	<u>11.47</u>	<b>14.21</b>	<u>0.0924</u>	<b>12.32</b>	7.94	<b>9.63</b>	<b>15.22</b>	<u>10.42</u>	<b>12.37</b>	<u>0.1293</u>
Improv.%	3.87	16.61	7.21	1.95	-3.53	1.07	-0.98	9.22	-3.99	1.90	5.69	-1.98	1.14	-4.78
M3TR	86.49	31.64	46.52	<u>88.42</u>	<u>39.03</u>	<u>54.15</u>	0.0620	57.41	15.34	24.29	<u>70.92</u>	26.57	38.65	0.0784
TAILOR	<u>87.72</u>	32.16	46.83	87.34	37.80	52.76	0.0427	57.92	16.26	25.39	69.45	27.66	39.56	<u>0.0512</u>
M3PT	<b>89.72</b>	<b>35.41</b>	<b>50.45</b>	<b>92.45</b>	<b>43.25</b>	<b>58.93</b>	<b>0.0242</b>	<b>59.51</b>	<b>17.56</b>	<b>27.11</b>	<b>72.15</b>	<b>28.50</b>	<b>40.85</b>	<b>0.0307</b>
Improv.%	2.28	10.11	7.73	4.56	10.81	8.83	43.33	2.75	8.00	6.77	1.73	3.04	3.26	40.04

our model are shown in Table 2, which were decided based on our parameter tuning studies. Specifically, we will display the results of tuning  $\tau$  and  $\pi$  afterward.

To reproduce our experiment results conveniently, we have published our model’s source code on <https://github.com/DeqingYang/M3PT>.

### 5.3 Main Results

The following performance scores of all compared models are reported as the average results of three runnings to alleviate the bias of single running. Table 3 lists the overall tagging performance of all models on MPTD1 and MPTD2 datasets, where the best and second-best scores in each group are in bold and underlined, respectively. In particular, to verify our M3PT’s capability of leveraging uni-modal features, we further proposed two ablated variants of M3PT. Wherein M3PT(text) only leverages the textual features, i.e., it only inputs the text embeddings into the cluster layer of TIF and directly uses the single textual representation  $\hat{x}$  as the POI’s content embedding  $c$ . Similarly, M3PT(image) only inputs image embeddings into TIF to generate the content embedding. M3PT’s performance improvement ratio relative to the best baseline (underlined) is also listed in each group. The results demonstrate that M3PT performs the best on most metrics. Moreover, we also have the following observations and analyses.

1. Compared with the first group of textual modality and the third group of multi-modality, M3PT is not the best on some metrics. Through the investigation of the two datasets, we found that the images of a POI are semantically related to only about one-fifth of the POI’s tags on average. It implies that M3PT can not exert its advantage sufficiently when only leveraging these images to discover the correct tags.

2. Compared with MPTD1, M3PT’s performance improvement over the baselines on MPTD2 is more apparent. It is because the POIs’ texts and images in MPTD2 are more related to the tags,

helping M3PT discover the matching across different modalities. Consequently, our model can achieve more precise POI tagging on MPTD2. It implies that our model is better at exploiting the high-quality dataset.

### 5.4 Detailed Analysis

**5.4.1 Ablation Studies.** We have conducted some ablation studies to justify the effects of the essential modules and strategy we designed in M3PT, including incorporating POI images, TIF, and the contrastive learning of POI-tag (PTC). Table 4 lists the results of different ablated variants of M3PT on MPTD2 and MPTD1, respectively, where the variant without image is just M3PT(text) in Table 3. In addition, in the variant without TIF, the single textual representation  $\hat{x}$  is just the sum of  $N$  text embeddings. Similarly, the single visual representation  $\hat{v}$  is just the sum of  $M$  image embeddings. Then, these two representations are concatenated and reshaped into  $p$ ’s content embedding  $c$ .

Besides the metric scores, the performance drop ratios of all variants relative to M3PT are also listed in Table 4. The results show that incorporating either one component of PTC, TIF, and image is helpful to M3PT’s performance improvement. Compared with TIF and image, PTC is more helpful for M3PT to achieve precise POI tagging since the performance drop of the variant without PTC is the most obvious.

**5.4.2 Impact of Domain-adaptive Image Encoder.** As mentioned before, the DIE specially designed in M3PT is an image encoder more adaptive to the requirements of the real-world scenario of Fliggy, resulting in M3PT’s enhanced performance. To verify it, we further propose some variants of M3PT by replacing the DIE with some previous image encoders, including ViT [7], CLIP [32] and BLIP [23].

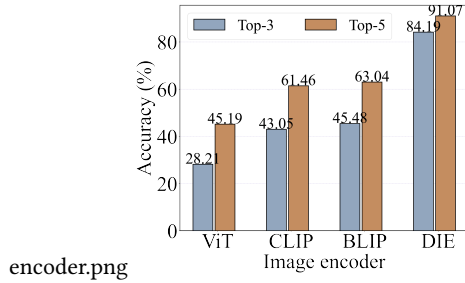
Due to space limitation, we only report the compared models’ scores of mAP, M-F1, and F1-e in Table 5 since they are the most representative metric of the ranking metrics, label-based classification

**Table 4: Ablation study results on the two datasets.**

MPTD2																
PTC	TIF	Image	M-P	Drop%	M-R	Drop%	M-F1	Drop%	P-e	Drop%	R-e	Drop%	F1-e	Drop%	HLS	Drop%
✗	✗	✗	80.71	10.04	26.03	26.49	39.36	21.98	81.1	12.28	36.06	16.62	49.92	15.29	0.0641	164.88
✗	✗	✓	83.28	7.18	28.39	19.82	42.43	15.90	83.05	10.17	36.9	14.68	51.19	13.13	0.0630	160.33
✗	✓	✓	85.91	4.25	30.9	12.74	45.45	9.91	86.55	6.38	37.04	14.36	51.87	11.98	0.0547	126.03
✓	✓	✗	86.34	3.77	32.57	8.02	47.11	6.62	88.38	4.40	38.63	10.68	53.72	8.84	0.0413	70.66
✓	✗	✓	87.81	2.13	32.36	8.61	47.25	6.34	89.16	3.56	38.05	12.02	53.37	9.43	0.0432	78.51
✓	✓	✓	<b>89.72</b>	-	<b>35.41</b>	-	<b>50.45</b>	-	<b>92.45</b>	-	<b>43.25</b>	-	<b>58.93</b>	-	<b>0.0242</b>	-
MPTD1																
PTC	TIF	Image	M-P	Drop%	M-R	Drop%	M-F1	Drop%	P-e	Drop%	R-e	Drop%	F1-e	Drop%	HLS	Drop%
✗	✗	✗	49.67	16.54	13.81	21.36	21.61	20.29	56.13	22.20	20.7	27.37	30.07	26.39	0.0809	163.52
✗	✗	✓	52.42	11.91	13.37	23.86	21.93	19.11	62.15	13.86	22.91	19.61	32.2	21.18	0.0753	145.28
✗	✓	✓	54.62	8.22	14.41	17.94	22.8	15.90	65.26	9.55	24.05	15.61	35.67	12.68	0.0773	151.79
✓	✓	✗	56.42	5.19	15.37	12.47	23.65	12.76	67.04	7.08	25.51	10.49	36.95	9.55	0.0542	76.55
✓	✗	✓	57.79	2.89	16.34	6.95	25.58	5.64	69.55	3.60	25.9	9.12	37.02	9.38	0.0504	64.17
✓	✓	✓	<b>59.51</b>	-	<b>17.56</b>	-	<b>27.11</b>	-	<b>72.15</b>	-	<b>28.5</b>	-	<b>40.85</b>	-	<b>0.0307</b>	-

**Table 5: Performance comparisons of adopting different image encoder in M3PT.**

Image encoder	MPTD2			MPTD1		
	mAP	M-F1	F1-e	mAP	M-F1	F1-e
ViT	81.60	47.16	<u>56.93</u>	66.20	<u>26.07</u>	35.62
CLIP	81.36	46.93	55.40	66.94	25.70	36.15
BLIP	<u>82.41</u>	<u>47.31</u>	55.03	<u>67.78</u>	25.62	<u>37.49</u>
DIE	<b>84.32</b>	<b>50.45</b>	<b>58.93</b>	<b>69.14</b>	<b>27.11</b>	<b>40.85</b>

**Figure 3: The precision of top-3/5 tags for the POIs on on MPTD2 predicted by the M3PTs with different image encoders.**

metrics, and object-based classification metrics. The results show that, with DIE, our M3PT performs better on both classification and ranking the correct tags. In addition, to concretely exhibit DIE's advantage in ranking the gold (correct) tags, in Fig. 3, we display the precision of top-3/5 tags predicted for the POIs on MPTD2. Obviously, all models' precision of top-5 is higher than that of top-3. Nevertheless, DIE's precision gap between top-3 and top-5 is much narrower than the other image encoders. Furthermore, in the table of Fig. 4, we list the top-3 tags predicted by the models for two POIs from MPTD2, where the incorrect tags are marked red. It also justifies that the M3PT with DIE can predict a bigger probability for the correct tags than the baselines.

POI	Image encoder	Tag1	Tag2	Tag3
圣亚海洋世界 Shengya Ocean World	ViT	湖泊 lake	滑雪 ski	企鹅 penguin
	CLIP	海豚 dolphin	湖泊 lake	海豚表演 dolphin show
	BLIP	白鲸 beluga whale	海豚 dolphin	海景 seascape
	DIE	海豚 dolphin	海洋馆 oceanarium	隧道 tunnel
青芝坞古镇 Qingzhiwu Ancient Town	ViT	酒吧 bar	乡村 countryside	小吃 snack
	CLIP	茶馆 teahouse	植物园 botanical garden	赏梅花 enjoy plum blossom
	BLIP	故居 former residence	品茶 tea tasting	乡村 countryside
	DIE	民宿 family hotel	乡村 countryside	茶庄 tea house

**Figure 4: Top-3 tags predicted by the models for two POIs on MPTD2 (better viewed in color).**

**5.4.3 Performance on Ranking Correct Tags.** We also evaluated M3PT's performance on the three ranking metrics. However, due to space limitation, we only compared it with the baseline performing the best in each group according to Table 3. The comparison results are listed in Table 6, showing that our model still has an advantage in ranking the correct tags on higher positions.

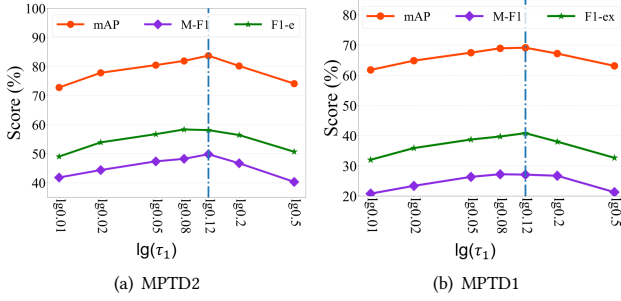
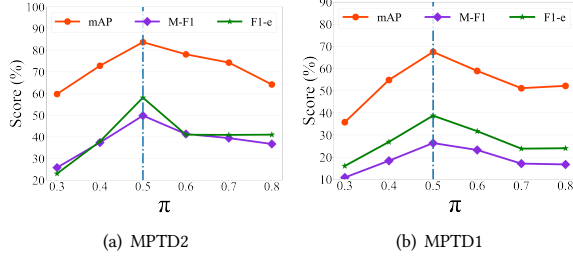
**5.4.4 Hyperparameter Tuning.** For the important hyperparameters in M3PT, we have studied their impacts on the performance of POI tagging. Due to space limitation, we only display the results of  $\tau_1$  and  $\pi$ .

We first focus on the impact of the temperature in  $s(p, t)$  in Equation 15, i.e.,  $\tau_1$ . As we mentioned before, the temperature in the scoring function of contrastive learning is used to maintain the balance between the alignment and consistency of contrasted samples. The results depicted in Fig. 5 show that M3PT performs better when



**Table 6: Tag ranking performance comparisons between M3PT and the best baseline in each group.**

Model	MPTD2			MPTD1		
	mAP	1-Error	RL	mAP	1-Error	RL
ERNIE	60.05	0.3315	0.1620	78.21	0.2403	0.1461
M3PT(text)	<b>64.15</b>	<b>0.2251</b>	<b>0.1390</b>	<b>80.41</b>	<b>0.1528</b>	<b>0.1230</b>
ViT(Q2L)	13.90	0.6701	0.3902	18.90	<b>0.2271</b>	0.3610
M3PT(image)	<b>16.73</b>	<b>0.6278</b>	<b>0.3894</b>	<b>20.05</b>	0.2315	<b>0.3521</b>
TAILOR	63.28	0.3025	0.1242	81.50	0.2362	0.0936
M3PT	<b>69.14</b>	<b>0.2143</b>	<b>0.0827</b>	<b>84.32</b>	<b>0.1264</b>	<b>0.0541</b>

**Figure 5: The tuning results of temperature  $\tau_1$  in  $\mathcal{L}_{PTC}$  on the two datasets.****Figure 6: The tuning results of prediction threshold  $\pi$  on the two datasets.**

$\tau_1 = 0.08 \sim 0.12$ . It indicates that a suitable temperature should not be too big. Otherwise, it would make the model less focused on discriminating the hard negative samples and thus fail to obtain satisfactory performance.

We also display the tuning results of the prediction threshold ( $\pi$ ) in Fig. 6. It shows that  $\pi = 0.5$  is the best setting, which just equals the default threshold in generic binary classifications.

### 5.5 Case Study

We further display the tagging results of M3PT and the best baseline in each group for a specific POI ‘龙泉山城市森林公园/Longquan Mountain Urban Forest Park’ from MPTD2. The top-5 tags predicted by the compared models are listed in Fig. 7, where the incorrect tags are marked red, and the green tags were assessed as correct but not in the POI’s gold tag set. Moreover, the rest black tags are gold

POI's texts		Top-5 predicted tags			
POI name: 龙泉山城市森林公园 Longquan Mountain Urban Forest Park Comment 1: 行近观景台，一对巨大的热气球映入眼帘。Saw a pair of huge fire balloons when close to observation deck. Comment 2: 景色好，是个徒步的好地方。Good scenery, a good place for hiking. Comment 3: 在龙泉山游玩了五凤古镇，风景不错。Traveled Wufeng ancient town when in Longquan Mountain.		ERDIE	VIT	TAILOR	M3PT
POI's images		徒步 hiking	森林 forest	山景 mountain scenery	观景台 observation deck
		森林 dolphin	动物园 zoo	观景台 observation deck	山景 mountain scenery
		热气球 fire balloon	河流 river	热气球 fire balloon	登山 climbing
		古镇 ancient town	水景 water view	茶园 tea garden	森林 forest
		湖泊 lake	玩水 dabble	森林 forest	徒步 hiking

**Figure 7: The top-5 tags predicted by the models for a POI from MPTD2. The incorrect tags are marked red, the green tags were assessed as correct but not in the POI’s gold tag set. And the rest tags are its gold tags (better viewed in color). It shows that M3PT predicts more gold tags.**

tags. At the same time, we display some texts and images of this POI on the left, which were fed into the models for tag prediction. Obviously, the incorrect tag ‘热气球/fire balloon’ and ‘古镇/ancient town’ were predicted by ERNIE and TAILOR, due to the misleading of Comments 1 and 3. While the incorrect tag ‘玩水/dabble’ and ‘茶园/tea garden’ were predicted by ViT and TAILOR, due to the misleading of some images. In fact, these incorrect tags only have the similar semantics to uni-modal (either textual or visual) data. They can be filtered out through the alignment (matching) of cross-modality. Compared with the baselines, our M3PT achieves the full fusion and precise matching between the textual and visual features, and are less disturbed by the noise of each modal data. Consequently, all of the 5 tags predicted by M3PT are correct.

## 6 CONCLUSION

Towards the POI tagging in the real-world tour scenario of Ali Fliggy, in this paper we propose a novel multi-modal model, namely M3PT, which incorporates the textual and visual features of POIs simultaneously to achieve the tagging task. In M3PT, we specially devise a domain-adaptive image encoder (DIE) to generate the image embeddings to better adapt to the requirements of the real-world scenario. In addition, we build the text-image fusion module (TIF) in our model, to achieve the full fusion and precise matching between textual and visual features. We further adopt a contrastive learning strategy to bridge the gap among the cross-modal representations in the model. Our extensive experiments with two datasets that were constructed from the Fliggy platform, not only demonstrate M3PT’s advantage, but also justify the rationalities and effectiveness of its important components.

## ACKNOWLEDGMENTS

This work was supported by Alibaba Group through Alibaba Innovative Research Program, Chinese NSF Major Research Plan (No.92270121), Shanghai Science and Technology Innovation Action Plan (No.21511100401), and Shanghai Sailing Program (No. 23YF1409400).

## REFERENCES

- [1] Kobus Barnard, Pinar Duygulu, David A. Forsyth, Nando de Freitas, David M. Blei, and Michael I. Jordan. 2003. Matching Words and Pictures. *J. Mach. Learn. Res.* 3 (2003), 1107–1135. <http://jmlr.org/papers/v3/barnard03a.html>
- [2] Emanuel Ben-Baruch, Tal Ridnik, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. 2020. Asymmetric loss for multi-label classification. *arXiv preprint arXiv:2009.14119* (2020).
- [3] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: UNiversal Image-Text Representation Learning. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX (Lecture Notes in Computer Science, Vol. 12375)*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Springer, 104–120. [https://doi.org/10.1007/978-3-030-58577-8\\_7](https://doi.org/10.1007/978-3-030-58577-8_7)
- [4] Qimin Cheng, Qian Zhang, Peng Fu, Conghuan Tu, and Sen Li. 2018. A survey and analysis on automatic image annotation. *Pattern Recognit.* 79 (2018), 242–259. <https://doi.org/10.1016/j.patcog.2018.02.017>
- [5] Karan Desai and Justin Johnson. 2020. ViTEx: Learning Visual Representations from Textual Annotations. *CoRR abs/2006.06666* (2020). [arXiv:2006.06666](https://arxiv.org/abs/2006.06666) <https://arxiv.org/abs/2006.06666>
- [6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
- [7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xi-aohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3–7, 2021*. OpenReview.net. <https://openreview.net/forum?id=YicbFdNTTy>
- [8] Zheyun Feng, Songhe Feng, Rong Jin, and Anil K. Jain. 2014. Image Tag Completion by Noisy Matrix Recovery. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part VII (Lecture Notes in Computer Science, Vol. 8695)*, David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.). Springer, 424–438. [https://doi.org/10.1007/978-3-319-10584-0\\_28](https://doi.org/10.1007/978-3-319-10584-0_28)
- [9] Liming Gao, Dongliang Liao, Gongfu Li, Jin Xu, and Hankz Hankui Zhuo. 2022. Semantic IR fused Heterogeneous Graph Model in Tag-based Video Search. In *Companion of The Web Conference 2022, Virtual Event / Lyon, France, April 25–29, 2022*, Frédérique Laforest, Raphaël Troncy, Elena Simperl, Deepak Agarwal, Aristides Gionis, Ivan Herman, and Lionel Médini (Eds.). ACM, 94–98. <https://doi.org/10.1145/3487553.3524214>
- [10] Shantanu Godbole and Sunita Sarawagi. 2004. Discriminative Methods for Multi-labeled Classification. In *Advances in Knowledge Discovery and Data Mining, 8th Pacific-Asia Conference, PAKDD 2004, Sydney, Australia, May 26–28, 2004, Proceedings (Lecture Notes in Computer Science, Vol. 3056)*, Honghua Dai, Ramakrishnan Srikant, and Chengqi Zhang (Eds.). Springer, 22–30. [https://doi.org/10.1007/978-3-540-24775-3\\_5](https://doi.org/10.1007/978-3-540-24775-3_5)
- [11] Kingshy Goh, Edward Y. Chang, and Beita Li. 2005. Using One-Class and Two-Class SVMs for Multiclass Image Annotation. *IEEE Trans. Knowl. Data Eng.* 17, 10 (2005), 1333–1346. <https://doi.org/10.1109/TKDE.2005.170>
- [12] 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*. IEEE Computer Society, 770–778. <https://doi.org/10.1109/CVPR.2016.90>
- [13] Tiek He, Hongzhi Yin, Zhenyu Chen, Xiaofang Zhou, Shazia W. Sadiq, and Bin Luo. 2016. A Spatial-Temporal Topic Model for the Semantic Annotation of POIs in LBSNs. *ACM Trans. Intell. Syst. Technol.* 8, 1 (2016), 12:1–12:24. <https://doi.org/10.1145/2905373>
- [14] Vinod Hegde, Josiane Xavier Parreira, and Manfred Hauswirth. 2013. Semantic Tagging of Places Based on User Interest Profiles from Online Social Networks. In *Advances in Information Retrieval - 35th European Conference on IR Research, ECIR 2013, Moscow, Russia, March 24–27, 2013, Proceedings (Lecture Notes in Computer Science, Vol. 7814)*, Pavel Serdyukov, Pavel Braslavski, Sergei O. Kuznetsov, Jaap Kamps, Stefan M. Rüger, Eugene Agichtein, Ilya Segalovich, and Emine Yilmaz (Eds.). Springer, 218–229. [https://doi.org/10.1007/978-3-642-36973-5\\_19](https://doi.org/10.1007/978-3-642-36973-5_19)
- [15] Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. 2021. Seeing Out of the Box: End-to-End Pre-Training for Vision-Language Representation Learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19–25, 2021*. Computer Vision Foundation / IEEE, 12976–12985. <https://doi.org/10.1109/CVPR46437.2021.01278>
- [16] Yuqi Huo, Manli Zhang, Guangzhen Liu, Haoyu Lu, Yizhao Gao, Guoxing Yang, Jingyuan Wen, Heng Zhang, Baogui Xu, Weihao Zheng, Zongzheng Xi, Yueqian Yang, Anwen Hu, Jinming Zhao, Ruichen Li, Yida Zhao, Liang Zhang, Yueling Song, Xin Hong, Wanqing Cai, Dan Yang Hou, Yingyan Li, Junyi Li, Peiyu Liu, Zheng Gong, Chuhao Jin, Yuchong Sun, Shizhe Chen, Zhiwu Lu, Zhicheng Dou, Qin Jin, Yanyan Lan, Wayne Xin Zhao, Ruihua Song, and Ji-Rong Wen. 2021. WenLan: Bridging Vision and Language by Large-Scale Multi-Modal Pre-Training. *CoRR abs/2103.06561* (2021). [arXiv:2103.06561](https://arxiv.org/abs/2103.06561) <https://arxiv.org/abs/2103.06561>
- [17] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18–24 July 2021, Virtual Event (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 4904–4916. <http://proceedings.mlr.press/v139/jia21b.html>
- [18] Yvonne Kammerer, Rowan Nairn, Peter Piroli, and Ed Huai-hsin Chi. 2009. Signpost from the masses: learning effects in an exploratory social tag search browser. In *Proceedings of the 27th International Conference on Human Factors in Computing Systems, CHI 2009, Boston, MA, USA, April 4–9, 2009*, Dan R. Olsen Jr., Richard B. Arthur, Ken Hinckley, Meredith Ringel Morris, Scott E. Hudson, and Saul Greenberg (Eds.). ACM, 625–634. <https://doi.org/10.1145/1518701.1518797>
- [19] John Krumm and Dany Rouhana. 2013. Placer: semantic place labels from diary data. In *The 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp '13, Zurich, Switzerland, September 8–12, 2013*, Friedemann Mattern, Silvia Santini, John F. Canny, Marc Langheinrich, and Jun Rekimoto (Eds.). ACM, 163–172. <https://doi.org/10.1145/2493432.2493504>
- [20] John Krumm, Dany Rouhana, and Ming-Wei Chang. 2015. Placer++: Semantic place labels beyond the visit. In *2015 IEEE International Conference on Pervasive Computing and Communications, PerCom 2015, St. Louis, MO, USA, 23–27 March, 2015*. IEEE Computer Society, 11–19. <https://doi.org/10.1109/PERCOM.2015.7146504>
- [21] Nikolaos Lagos, Salah Ait-Mokhtar, and Ioan Calapodescu. 2020. Point-Of-Interest Semantic Tag Completion in a Global Crowdsourced Search-and-Discovery Database. In *ECAI 2020 - 24th European Conference on Artificial Intelligence, 29 August–8 September 2020, Santiago de Compostela, Spain, August 29 - September 8, 2020 - Including 10th Conference on Prestigious Applications of Artificial Intelligence (PAIS 2020) (Frontiers in Artificial Intelligence and Applications, Vol. 325)*, Giuseppe De Giacomo, Alejandro Catalá, Bistra Dilkina, Michela Milano, Senén Barro, Alberto Bugarín, and Jérôme Lang (Eds.). IOS Press, 2993–3000. <https://doi.org/10.3233/FAIA200474>
- [22] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942* (2019).
- [23] Junnan Li, Dongxu Li, Caimeing Xiong, and Steven C. H. Hoi. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *International Conference on Machine Learning, ICML 2022, 17–23 July 2022, Baltimore, Maryland, USA (Proceedings of Machine Learning Research, Vol. 162)*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (Eds.). PMLR, 12888–12900. <https://proceedings.mlr.press/v162/li22n.html>
- [24] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. VisualBERT: A Simple and Performant Baseline for Vision and Language. *CoRR abs/1908.03557* (2019). [arXiv:1908.03557](https://arxiv.org/abs/1908.03557) <https://arxiv.org/abs/1908.03557>
- [25] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. 2021. UNIMO: Towards Unified-Modal Understanding and Generation via Cross-Modal Contrastive Learning. 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*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, 2592–2607. <https://doi.org/10.18653/v1/2021.acl-long.202>
- [26] Xue Li, Bin Shen, Bao-Di Liu, and Yu-Jin Zhang. 2016. A Locality Sensitive Low-Rank Model for Image Tag Completion. *IEEE Trans. Multim.* 18, 3 (2016), 474–483. <https://doi.org/10.1109/TMM.2016.2518478>
- [27] Junyang Lin, Rui Men, An Yang, Chang Zhou, Ming Ding, Yichang Zhang, Peng Wang, Ang Wang, Le Jiang, Xianyan Jia, Jie Zhang, Jianwei Zhang, Xu Zou, Zhikang Li, Xiaodong Deng, Jie Liu, Jinbao Xue, Huiling Zhou, Jianxin Ma, Jin Yu, Yong Li, Wei Lin, Jingren Zhou, Jie Tang, and Hongxia Yang. 2021. M6: A Chinese Multimodal Pretrainer. *CoRR abs/2103.00823* (2021). [arXiv:2103.00823](https://arxiv.org/abs/2103.00823) <https://arxiv.org/abs/2103.00823>
- [28] Rongcheng Lin, Jing Xiao, and Jianping Fan. 2018. Nextvlad: An efficient neural network to aggregate frame-level features for large-scale video classification. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*. 0–0.
- [29] Zijia Lin, Guiguang Ding, Mingqing Hu, Jianmin Wang, and Xiaojun Ye. 2013. Image Tag Completion via Image-Specific and Tag-Specific Linear Sparse Reconstructions. In *2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23–28, 2013*. IEEE Computer Society, 1618–1625. <https://doi.org/10.1109/CVPR.2013.212>
- [30] Shilong Liu, Lei Zhang, Xiao Yang, Hang Su, and Jun Zhu. 2021. Query2Label: A Simple Transformer Way to Multi-Label Classification. *CoRR abs/2107.10834* (2021). [arXiv:2107.10834](https://arxiv.org/abs/2107.10834) <https://arxiv.org/abs/2107.10834>
- [31] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6–9, 2019*. OpenReview.net. <https://openreview.net/forum?id=Bkg6RiCqY7>

- [32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 8748–8763. <http://proceedings.mlr.press/v139/radford21a.html>
- [33] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.* 21 (2020), 140:1–140:67. <http://jmlr.org/papers/v21/20-074.html>
- [34] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett (Eds.). 91–99. <https://proceedings.neurips.cc/paper/2015/hash/14bfa6bb14875e45bba028a21ed38046-Abstract.html>
- [35] Tal Ridnik, Hussam Lawen, Asaf Noy, Emanuel Ben Baruch, Gilad Sharir, and Itamar Friedman. 2021. TResNet: High Performance GPU-Dedicated Architecture. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021*. IEEE, 1399–1408. <https://doi.org/10.1109/WACV48630.2021.00144>
- [36] Mert Bülent Sariyildiz, Julien Perez, and Diane Larlus. 2020. Learning Visual Representations with Caption Annotations. *CoRR* abs/2008.01392 (2020). [arXiv:2008.01392](https://arxiv.org/abs/2008.01392) <https://arxiv.org/abs/2008.01392>
- [37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
- [38] Feng Wang and Huaping Liu. 2021. Understanding the Behaviour of Contrastive Loss. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 2495–2504. <https://doi.org/10.1109/CVPR46437.2021.00252>
- [39] Yan Wang, Zongxu Qin, Jun Pang, Yang Zhang, and Jin Xin. 2017. Semantic Annotation for Places in LBSN through Graph Embedding. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017, Singapore, November 06 - 10, 2017*, Ee-Peng Lim, Marianne Winslett, Mark Sanderson, Ada Wai-Chee Fu, Jimeng Sun, J. Shane Culpepper, Eric Lo, Joyce C. Ho, Debora Donato, Rakesh Agrawal, Yu Zheng, Carlos Castillo, Aixin Sun, Vincent S. Tseng, and Chenliang Li (Eds.). ACM, 2343–2346. <https://doi.org/10.1145/3132847.3133075>
- [40] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2022. SimVLM: Simple Visual Language Model Pretraining with Weak Supervision. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net. [https://openreview.net/forum?id=GUrhfTuf\\_3](https://openreview.net/forum?id=GUrhfTuf_3)
- [41] Na Yan, Honglong Chen, Kai Lin, Zhichen Ni, Zhe Li, and Huansheng Xue. 2023. BFSearch: Bloom filter based tag searching for large-scale RFID systems. *Ad Hoc Networks* 139 (2023), 103022. <https://doi.org/10.1016/j.adhoc.2022.103022>
- [42] Dingqi Yang, Bin Li, and Philippe Cudré-Mauroux. 2016. POISketch: Semantic Place Labeling over User Activity Streams. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, Subbarao Kambhampati (Ed.). IJCAI/AAAI Press, 2697–2703. <http://www.ijcai.org/Abstract/16/383>
- [43] Yiming Yang. 1999. An Evaluation of Statistical Approaches to Text Categorization. *Inf. Retr.* 1, 1-2 (1999), 69–90. <https://doi.org/10.1023/A:1009982220290>
- [44] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. 2022. FILIP: Fine-grained Interactive Language-Image Pre-Training. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net. <https://openreview.net/forum?id=cpDhcsEDC2>
- [45] Mao Ye, Dong Shou, Wang-Chien Lee, Peifeng Yin, and Krzysztof Janowicz. 2011. On the semantic annotation of places in location-based social networks. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, August 21-24, 2011*, Chid Apté, Joydeep Ghosh, and Padhraic Smyth (Eds.). ACM, 520–528. <https://doi.org/10.1145/2020408.2020491>
- [46] Jiaming Zhang, Shuhui Wang, and Qingming Huang. 2017. Location-Based Parallel Tag Completion for Geo-Tagged Social Image Retrieval. *ACM Trans. Intell. Syst. Technol.* 8, 3 (2017), 38:1–38:21. <https://doi.org/10.1145/3001593>
- [47] Yi Zhang, Mingyuan Chen, Jundong Shen, and Chongjun Wang. 2022. Tailor versatile multi-modal learning for multi-label emotion recognition. *arXiv preprint arXiv:2201.05834* (2022).
- [48] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D. Manning, and Curtis P. Langlotz. 2020. Contrastive Learning of Medical Visual Representations from Paired Images and Text. *CoRR* abs/2010.00747 (2020). [arXiv:2010.00747](https://arxiv.org/abs/2010.00747) <https://arxiv.org/abs/2010.00747>
- [49] Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced Language Representation with Informative Entities. 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*, Anna Korhonen, David R. Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, 1441–1451. <https://doi.org/10.18653/v1/p19-1139>
- [50] Jiawei Zhao, Yifan Zhao, and Jia Li. 2021. M3tr: Multi-modal multi-label recognition with transformer. In *Proceedings of the 29th ACM International Conference on Multimedia*. 469–477.
- [51] Jingbo Zhou, Shan Gou, Renjun Hu, Dongxiang Zhang, Jin Xu, Airong Jiang, Ying Li, and Hui Xiong. 2019. A Collaborative Learning Framework to Tag Refinement for Points of Interest. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*, Ankur Teredesai, Vipin Kumar, Ying Li, Rómer Rosales, Evimaria Terzi, and George Karypis (Eds.). ACM, 1752–1761. <https://doi.org/10.1145/3292500.3330698>
- [52] Yin Zhu, Erheng Zhong, Zhongqi Lu, and Qiang Yang. 2013. Feature engineering for semantic place prediction. *Pervasive Mob. Comput.* 9, 6 (2013), 772–783. <https://doi.org/10.1016/j.pmcj.2013.07.004>