Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.055560

ARTICLE

Efficient User Identity Linkage Based on Aligned Multimodal Features and Temporal Correlation

Jiaqi Gao[1], Kangfeng Zheng[1,*], Xiujuan Wang[2], Chunhua Wu[1] and Bin Wu[2]

[1]School of Cyberspace Security, Beijing University of Posts and Telecommunications, Beijing, 100876, China

[2]Faculty of Information Technology, Beijing University of Technology, Beijing, 100124, China

*Corresponding Author: Kangfeng Zheng. Email: kfzheng@bupt.edu.cn

Received: 01 July 2024 Accepted: 13 August 2024 Published: 15 October 2024

ABSTRACT

User identity linkage (UIL) refers to identifying user accounts belonging to the same identity across different social media platforms. Most of the current research is based on text analysis, which fails to fully explore the rich image resources generated by users, and the existing attempts touch on the multimodal domain, but still face the challenge of semantic differences between text and images. Given this, we investigate the UIL task across different social media platforms based on multimodal user-generated contents (UGCs). We innovatively introduce the efficient user identity linkage via aligned multi-modal features and temporal correlation (EUIL) approach. The method first generates captions for user-posted images with the BLIP model, alleviating the problem of missing textual information. Subsequently, we extract aligned text and image features with the CLIP model, which closely aligns the two modalities and significantly reduces the semantic gap. Accordingly, we construct a set of adapter modules to integrate the multimodal features. Furthermore, we design a temporal weight assignment mechanism to incorporate the temporal dimension of user behavior. We evaluate the proposed scheme on the real-world social dataset TWIN, and the results show that our method reaches 86.39% accuracy, which demonstrates the excellence in handling multimodal data, and provides strong algorithmic support for UIL.

KEYWORDS

User identity linkage; multimodal models; attention mechanism; temporal correlation

1 Introduction

In recent years, with the development of mobile Internet technology and the increase in user demand, there are more and more virtual accounts in cyberspace, and the same user owns multiple accounts in different applications or even on the same platform. UIL refers to identifying user accounts belonging to the same identity across different platforms. UIL is essential for downstream tasks such as information diffusion prediction [1] and cross-platform recommendation [2].

Users frequently generate content with concurrent intrinsic relevance across diverse social media platforms, marked by temporal stamps and predominantly manifested as textual or visual content, as shown in Fig. 1. UIL research is bifurcated into approaches leveraging: (i) unimodal data, focusing on text-based elements such as user IDs and posts [3–7]; and (ii) multimodal data, integrating textual and

visual information to holistically profile user identities and behaviors through data complementarity [8–12]. Acknowledging the temporal clustering of contextually related user posts, recent studies [6,12] have incorporated temporal dimensions, enhancing the precision of behavioral pattern recognition, rendering user profiles more nuanced, and augmenting cross-platform user identification accuracy and efficacy.

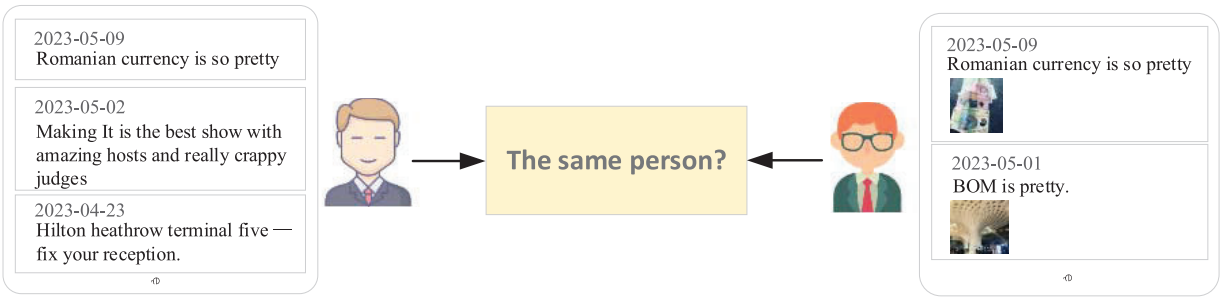

Figure 1: UIL based on multimodal UGCs

In response to the current state of research, this study focuses on two key points: (1) Exploring an innovative method that utilizes multimodal UGCs. (2) Analyzing the effect of the time factor on user behavior, incorporating the timestamps of UGCs into the model design. For the first point, current UIL methods based on multimodal UGCs generally adopt textual or visual models trained on unimodal data to extract features from heterogeneous data such as text and images, and then directly train feature extraction or modal fusion models based on anchor user pairs to build up an interaction and fusion mechanism between heterogeneous modalities. However, these approaches exist the following problems: (1) Due to the natural semantic gap between text and image modalities, the desired synergistic effect is hard to achieve by integrating multimodal features based on a limited number of anchor pairs across social platforms. (2) The current method needs to train both textual and visual models, and the number of parameters is large, which makes the training inefficient. (3) Different service providers focus on different features, e.g., the content posted on Instagram is mainly images with short text, while the posts posted on Twitter are mainly text supplemented by images, and the text of the posts on Instagram is shorter than Twitter. Coping with the variability of data modality and the imbalance of data volume is another problem to be faced by UIL. For the second point, existing methods usually use a simplified function to summarize the effect of time on user behavior, ignoring the changes that may exist within the period.

To solve the above problems, we propose EUIL. As shown in Fig. 2, the method includes the BLIP-based text enhancement module, the aligned feature extraction module, the multimodal adapter module, and the temporal correlation-based cross-modal similarity calculation module. Specifically, the BLIP-based text enhancement module first utilizes the powerful multimodal model BLIP [13] to generate captions for images in graphic posts, which not only compensates for the problem of missing textual data but also further explores the deep semantic information of images; then, the alignment feature extraction module unifies the features of texts and images with the pre-trained cross-modal comparative learning model CLIP [14]. Then, the alignment feature extraction module unifies the texts and images with the pre-trained cross-modal contrast learning model CLIP, which can maintain cross-modal semantic consistency while capturing the unique features of the texts and images, effectively bridging the semantic gap between the texts and the images. Furthermore, we design a multimodal adapter module optimized for UIL based on the multi-head self-attention mechanism to generate highly customized user feature representations by fine-tuning limited parameters without changing the

underlying model. Finally, the cross-modal similarity computation module based on temporal linkage constructs the inter-user similarity matrix based on the generated multimodal features and dynamically adjusts the importance of the information within each time window in the UIL process by introducing a time factor assignment mechanism. We conduct a series of exhaustive experiments on real datasets to confirm the efficient UIL capability of the proposed method in coping with heterogeneous modal data imbalance.

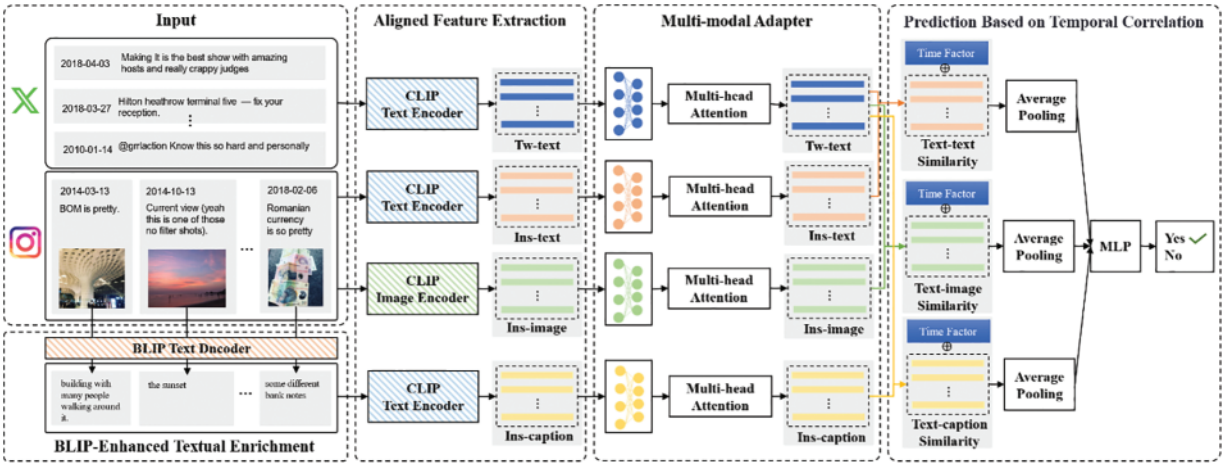

Figure 2: Illustration of the proposed scheme for UIL

The main contributions of this paper are as follows:

- Presents an efficient UIL method for multimodal data. Based on the pre-trained multimodal large model CLIP to extract text and image features on social media platforms, the unique features of each text and image are captured while maintaining cross-modal semantic consistency, effectively bridging the semantic gap between text and image, and providing excellent feature initialization for subsequent fine-tuning.
- Innovatively employs the BLIP text generation model to enhance short text on social media platforms. Generating image captions compensates for the problem of insufficient text information and deeply integrates the semantic information of images.
- Presents a multimodal adapter module for UIL tasks. The module effectively fuses text and image features from CLIP based on a multi-head self-attention mechanism. The training of the lightweight multimodal adapter module makes it possible to generate comprehensive feature representations suitable for user identity association without changing the parameters of the pre-trained multimodal model, making the training efficient.
- Innovatively designs a segmented time decay function that distinguishes and meticulously models the differential effects over different periods. This approach not only better fits the complex characteristics of real-world user behavior dynamically changing over time, but also strengthens the model's sensitivity and adaptability to UIL matching under different time windows, thus improving the accuracy of UIL and the practicality of the model.
- Evaluates the proposed method on real social datasets, and the experimental results demonstrate the superiority of the proposed model over state-of-the-art models.

2 Related Work

2.1 User Identity Linkage

Existing UIL studies can be categorized into unimodal data-based methods and multimodal data-based methods based on the data modality utilized.

2.1.1 Unimodal Data-Based Methods

Current unimodal UIL methodologies predominantly harness users' textual data, encompassing attributes and published content. Based on the fact that people prefer to choose similar usernames across social platforms, EEUPL [3] utilizes usernames and profiles for linkage, assigns similar data to the same bucket via the minHashLSH algorithm, and generates similarity graphs by calculating only the similarity of user pairs within the buckets, which reduces the number of user pairs to be compared. MAUIL [7] categorizes the social media texts into character-, word-, and topic-level attributes, extracting features via unsupervised methodologies tailored to each category. Gao et al. [6] link entities in the texts to a knowledge graph through entity recognition and entity linking techniques, subsequently constructing word similarity matrices and entity similarity matrices using GloVE [15] and TransE [16], respectively. AsyLink [4] first introduces external text location pairs by associating words with locations using the topic modeling technique, reducing the bias caused by sparse link labels. Then it constructs the user-user interaction tensor as the basis for linking and captures the matching patterns in the user interaction tensor with a 3D convolutional neural network. Huang et al. [5] incorporate user attributes, generated content, and check-ins, mitigating local semantic noise via semantic feature extraction and enhancing noise resilience through multi-view graph data augmentation. In addition, graph neural networks are also used for UIL, where the social relationship graph is constructed by considering users as nodes and the relationships between users as edges. To model the influence of the neighbors, MEgo2Vec [17] designs three mechanisms for representing different attributes, distinguishing different neighbors and capturing structural information of the social network, which addresses the problem of error propagation and the presence of noise in social networks. Long et al. [18] propose a degree-aware graph neural network model DegUIL, which effectively solves the challenges brought by tail nodes and super head nodes in the UIL task. By supplementing and correcting the neighborhood information of nodes, DegUIL improves the quality of node representation and thus improves the accuracy of user identity links across social networks.

Unimodal data-based UIL methods rely only on textual modal data for UIL, ignoring the rich semantic information embedded in images, and failing to comprehensively capture and understand users' multifaceted behavioral characteristics and personalized preferences.

2.1.2 Multimodal Data-Based Methods

Multimodal data-based methods utilize both textual modal data and visual modal data of users. LinkSocial [8] extracts username features based on bi-grams, crops images to extract faces using OpenFace, an open-source image similarity framework, and uses deep learning to represent the face features on a 128-dimensional unit hypersphere. AHGNet [11] combines several deep learning techniques for extracting user features from multimodal social data, where textual information is captured using (Bidirectional Encoder Representations from Transformers, BERT) [19] and TextCNN [20] fusion models deep semantic features, image information is encoded using ResNet [21] to extract high-dimensional image features, check-in data is represented by constructing spatio-temporal co-occurrence matrix and applying GRU [22] model to form spatio-temporal sequences and social relationships are transformed into feature vectors revealing the structure of the social network using

DeepWalk [23] algorithm. UserNet [12] is used to analyze the user's textual information using BiLSTM [24] to encode users' text posts to capture contextual information as well as potential emotional and semantic features in the posts, and uses a pre-trained ResNet model to extract the depth features of the images uploaded by users, and then constructs the similarity matrices based on textual features and the similarity matrices based on image features, respectively, and predicts whether the users from the two platforms are the same person by fusing the similarity matrices of the two modalities whether they are the same person or not. AMSA [9] uses the BLIP model to generate textual descriptions for the images and extracts the subject of the text using the Latent Dirichlet Allocation (LDA) [25] model. In addition, AMSA extracts the textual features from the textual collection using BERT and extracts the image features of the user using ConvNeXT [26]. GRU-based model extracts the user's check-in features and fuses them into a user representation vector. MFlink [10] utilizes usernames and user-generated multimodal data to represent the user. Specifically, the method employs the bag-of-words model to extract the username features and extracts the user's posting text and image content information based on the BERT and ResNet models, respectively, and finally integrates the three modalities with the help of graph neural network and attention mechanism to integrate the three modalities. The current research extracts the features of each modal data based on the corresponding unimodal model and then trains the model based on the anchor user data for inter-modal interaction and fusion, however, there is a semantic gap between text and image, and due to the lack of multimodal social data and the difficulty in acquiring the anchor user pairs, it is more difficult to train the feature extraction model and the multimodal fusion model based on text and image features extracted from unimodal models, making the UIL unsatisfactory.

2.2 Multimodal Models

Given that visual and linguistic modalities often convey complementary insights, the synergy achieved through joint multimodal representation learning has demonstrated remarkable efficacy across various tasks, including visual question answering, image captioning, and quotation interpretation.

The CLIP (Contrastive Language-Image Pre-training) model is a multimodal pre-training model proposed by OpenAI, designed to bridge the semantic gap between text and images through contrastive learning. The innovation of CLIP lies in its use of unsupervised learning methods, training on large-scale data pairs of image and text scraped from the Internet, thereby learning cross-modal representations. The CLIP model comprises two main components: an image encoder and a text encoder. The image encoder can be any pre-trained convolutional neural network (CNN), such as ResNet, while the text encoder is typically based on the Transformer architecture. These two encoders map input images and text into the same vector space, allowing for a direct comparison of their similarities.

During training, CLIP utilizes a contrastive loss function, which aims to maximize the similarity of positive pairs while minimizing the similarity of negative pairs. Suppose we have a set of images I and corresponding textual descriptions T, where each image i has a corresponding text description t_i. The loss function can be represented as:

$$L_{CLIP} = -log\frac{exp(S(i,t_i)/\tau)}{\sum_j exp(S(i,t_j)/\tau)} - log\frac{exp(S(t_i,i)/\tau)}{\sum_k exp(S(t_i,k)/\tau)} \quad (1)$$

Here, $S(i,t_i)$ and $S(t_i,i)$ denote the similarity score between the image i and its corresponding text description t_i, and τ is the temperature parameter used to scale the similarity scores before applying the

softmax function. For each image i, we compute its similarity with all text descriptions, and for each text description t_i, we calculate its similarity with all images. Then, we normalize these similarities using the softmax function to obtain probability distribution. The goal of this loss function is to make the similarity of paired image-text descriptions significantly higher compared to non-paired ones. Minimizing this loss function allows the CLIP model to learn meaningful representations in the multimodal space, effectively bridging the semantic gap between text and images.

Building upon CLIP, the BLIP Series Models [13,27] represent an advanced multimodal pre-training architecture, incorporating refined training paradigms and architectural enhancements. Beyond joint representation learning, BLIP emphasizes improvements in text generation rooted in visual understanding, thus achieving a harmonious integration of visual analysis and linguistic expression, marking a significant advance toward comprehensive multimodal intelligence. BLIP2 is an upgraded version of BLIP, further enhancing the model's performance and efficiency while retaining its original strengths. BLIP2 achieves higher cross-modal matching accuracy and faster inference speed through improvements in model architecture and optimized training strategies.

3 Preliminaries

To clearly articulate our research problem and its underlying components, we first introduce the necessary notation and then define the research problem itself.

3.1 Notation

We establish a set of symbols that will be used throughout our discussion to ensure clarity and consistency. Table 1 summarizes the main notations used in this paper.

Table 1: Summary of the main notations

Notation	Explanation
$\mathcal{O}_1$	The first social media platform.
$\mathcal{O}_2$	The second social media platform.
u_1^i	The i-th user on $\mathcal{O}_1$.
u_2^i	The i-th user on $\mathcal{O}_2$.
t_k^i	The k-th textual post of u_1^i.
v_g^i	Image of the g-th post of u_2^i.
c_g^i	Text of the g-th post of u_2^i.
d_g^i	Image caption of the g-th post of u_2^i.
$\mathcal{T}_i$	K textual posts by u_1^i.
$\mathcal{S}_i$	G multimodal posts by u_2^i.
$\widehat{\mathcal{S}}_i$	Expanded posts by u_2^i.

3.2 Problem Formulation

In this work, we aim to address the problem of UIL based on multimodal UGCs across different social media platforms. Without loss of generality, we focus on UIL between $\mathcal{O}_1$ and $\mathcal{O}_2$ platforms, and we choose the popular social media Twitter and Instagram as $\mathcal{O}_1$ and $\mathcal{O}_2$ platforms for illustration

of our scheme. Due to the different service focuses, on Twitter people tend to broadcast news or join topic discussions through text tweets, while on Instagram they tend to post images with short text descriptions to share their daily lives or ideas. To make the model more representative, we specifically assume that users prefer to generate unimodal text posts on $\mathcal{O}_1$ and multimodal paired image and text posts on $\mathcal{O}_2$.

Suppose we have a set of training user account pairs $\mathcal{U} = \{(u_1^1, u_2^1), (u_1^2, u_2^2), \cdots, (u_1^N, u_2^N)\}$ where each pair consists of two user accounts from different social media platforms ($\mathcal{O}_1$ and $\mathcal{O}_2$) are composed of. Also, suppose that for each user account u_1^i on $\mathcal{O}_1$, we have his/her K unimodal text posts $\mathcal{T}_i = \{t_1^i, t_2^i, \cdots, t_K^i\}$. Similarly, for each user account u_2^i on $\mathcal{O}_2$, we collect his/her G multimodal posts $\mathcal{S}_i = \{(v_1^i, c_1^i), (v_2^i, c_2^i), \cdots, (v_G^i, c_G^i)\}$, where v_g^i and c_g^i denote the image and text of the $i-th$ post, respectively. The training user account pairs are labeled by $\boldsymbol{y} = (y_1, y_2, \cdots, y_N)^{\boldsymbol{T}} \in \mathbb{R}^N$, where y_i stands for the real situation of the first i pair of users (u_1^i, u_2^i). Specifically, $y_i = 1$ if accounts u_1^i and u_2^i refer to the same identity in the physical world, and $y_i = 0$ otherwise. In a sense, our goal is to use labeled pairs of training user accounts to learn the projection $f: u_1^i \times u_2^i \to \{0, 1\}$.

4 Methodology

Facing the task of UIL in multimodal UGCs scenarios, we propose an efficient UIL method based on aligned multimodal features and temporal correlation, including the BLIP-based text enhancement module, the aligned feature extraction module, the multimodal adapter module, and the temporal correlation-based cross-modal similarity calculation module. To address the problem of data modality variability and data volume imbalance in the content posted by users on different social media, the BLIP-based text enhancement module first utilizes the BLIP model to generate a detailed description for the image content as a supplement to the text modality data. Then, to address the semantic gap between image and text, the aligned feature extraction module extracts the representation of text and image under a unified feature space with the CLIP model. Then, we design and train the multimodal adapter module to generate feature representations suitable for UIL. Finally, the temporal correlation-based cross-modal similarity computation module constructs the multimodal user similarity matrix, and we introduce a time factor to reweight the similarity matrix. We input the multimodal similarity matrix into the classification network to classify the input user pairs. The specific implementation of each module is described in detail below.

4.1 BLIP-Based Text Enhancement

Diverse social media platforms are characterized by distinct functionalities; for instance, Instagram prioritizes visual content, with user-generated multimodal posts predominantly featuring images accompanied by brief or even absent textual accompaniments. To bolster textual content on such platforms and leverage the copious semantic information embedded in images, we introduce a text augmentation strategy grounded in the BLIP model. This approach taps into the decoder component of a pre-trained BLIP model to generate descriptive captions for individual images within posts. BLIP, a robust bidirectional text-image pre-training model, boasts a decoder capable of translating image substance into coherent natural language narratives. By doing so, it effectively transforms non-verbal image data into textual representations, thereby enabling the interpretation of image content through the lens of linguistic semantics and enriching the overall semantic context.

Specifically, for the G multimodal posts $\mathcal{S}_i = \{(v_1^i, c_1^i), (v_2^i, c_2^i), \cdots, (v_G^i, c_G^i)\}$ for each user u_2^i on the $\mathcal{O}_2$ platform, a pre-trained BLIP model based on the pre-training BLIP model first used ResNet to process the input image v_g^i into a high dimensional feature vector. This process captures the spatial and

appearance information of the image. Next, the decoder Transformer receives the image feature vector and generates a sequence of word embeddings to form the final caption d_g^i. The caption d_g^i not only captures the specific visual elements in the image such as objects, scenes, and actions, but also expresses deeper contextual information and implicit relationships. The generated caption d_g^i is paired with the original data (v_g^i, c_g^i) to get the extended data pair (v_g^i, c_g^i, d_g^i), which in turn constructs an extended set of posts notated as $\widehat{\mathcal{S}}_i = \{(v_1^i, c_1^i, d_1^i), (v_2^i, c_2^i, d_2^i), \cdots, (v_G^i, c_G^i, d_G^i)\}$. $\widehat{\mathcal{S}}_i$ contains all the modal information for each post: the image, the original text, and the caption of the image generated by BLIP.

By transforming image information into text, the semantic gap between two different modalities, image and text, can be narrowed, which facilitates the comparison and fusion of data from the two modalities in a unified semantic space, and is conducive to improving the accuracy of cross-modal UIL. Especially on social media platforms, image posts posted by users are often accompanied by short or even no textual descriptions. BLIP-generated captions can effectively supplement this lack of information and provide more valuable information clues for UIL. In addition, image-based caption generation can be viewed as a kind of data augmentation of text, which increases the diversity of data and helps to improve the generalization ability of the subsequent model and the ability to cope with unseen complex situations.

4.2 Aligned Feature Extraction

Traditional work tends to use targeted image models and language models to extract image and text features respectively, e.g., extracting 2048-D image feature vectors using the ResNet model and 768-D or 1024-D text features using the BERT model, and then training modal fusion networks or classification networks based on these two features. However, due to the semantic gap between text and images, high quality as well as a high quantity of multimodal data from anchor users is required to achieve the desired training results. Since social data tends to have a lot of noise and anchor users are expensive to acquire, the current results of training fused multimodal features based on unimodal models are not satisfactory.

To address the above problem, we design the alignment feature extraction module to provide a unified feature representation of text and images on social media with the help of a pre-trained cross-modal comparative learning model, CLIP. CLIP employs a comparative learning strategy, the basic idea of which is to maximize the similarity between positive sample pairs (an image and the text that describes it correctly), while minimizing the similarity between negative sample pairs (an image and an irrelevant text description). This approach motivates the model to learn representations that efficiently map images and text into a shared semantic space. The process of extracting features based on the CLLP model can be formalized as Eqs. (2) and (3):

$$Z_1 = \{Z^v, Z^c, Z^d\} = \{f_{img}(v), f_{txt}(c), f_{txt}(d)\} \tag{2}$$

$$Z_2 = Z^t = f_{txt}(t) \tag{3}$$

where f_{img} and f_{txt} denote the image encoder and text encoder of CLIP, respectively. Specifically, for the image v of $\widehat{\mathcal{S}}_i$ in the extended set, the image features are obtained using the image encoder f_{image} of CLIP, denoted as $Z^v \in \mathbb{R}^{G \times d_{img}}$, with d_{img} being the f_{image} output image feature's dimension. For the original text c and image caption d in $\widehat{\mathcal{S}}_i$, the text features are obtained using CLIP's text encoder f_{txt}, respectively, denoted as $Z^c \in \mathbb{R}^{G \times d_{txt}}$ and $Z^d \in \mathbb{R}^{G \times d_{txt}}$, with d_{txt} being the dimension of the f_{txt} output text feature. Similarly, for text t in the set of $\mathcal{T}_i$, the text encoder of CLIP is also utilized to obtain its text features $Z^t \in \mathbb{R}^{K \times d_{txt}}$, which are more comparable between multimodal features Z_1 and Z_2 than those extracted by the unimodal model. CLIP can capture the unique features of each of the text and image

while maintaining cross-modal semantic consistency, so that content expressing the same concept or topic will have similar representations even in different modalities, effectively bridging the semantic gap between text and image and providing high-quality input features for subsequent UIL.

4.3 Multimodal Adapter

The contents posted on different social media platforms by users are often different, and for the UIL task, some of the posts are extremely similar in content, which possesses a positive impact on the UIL effect, while some of the posts are equivalent to noise, which harms the UIL effect. Considering the different degrees of contribution of each post to end-UIL, we design a lightweight Adapter module to personalize and optimize the features extracted from the CLIP model to further enhance the performance of these features in the UIL task. The Adapter module is typically a pre-trained model that is supplemented by one or more layers of a small network structure, allowing for the targeting of the UIL without changing the parameters of the original model. Adapter modules typically add one or more layers of small network structures to a pre-trained model, allowing fine-tuning of the model's performance for a specific task without changing the original model parameters. In the multimodal UIL scenario, we design the multimodal Adapter module as a four-terminal, multi-headed self-attention module, with each end as shown in Eq. (4):

$$\widehat{Z} = Adapter\,(Z) = MultiHead\,(MLP\,(Z)) \tag{4}$$

where *MultiHead* stands for Multi-Head Self-Attention Operation, *MLP* stands for Multilayer Vector Machine Operation, $Z \in \{Z^v, Z^c, Z^d, Z^t\}$, Z^c is the text feature of user u_1^i, Z^v, Z^c, Z^d are the image feature, text feature, and image caption feature of user u_2^i, $\widehat{Z} \in \{\widehat{Z}^v, \widehat{Z}^c, \widehat{Z}^d, \widehat{Z}^t\}$ which correspond to the outputs of the features in Z after passing through the Adapter module, respectively.

Specifically, we set up the Adapter module for the text end of the $\mathcal{O}_1$ platform, the text end of the $\mathcal{O}_2$ platform, the image end, and the caption end, respectively. First, the features of each end are passed into an MLP network, as shown in Eq. (5):

$$Z' = MLP\,(Z) = \sigma\,(W_2 \cdot \sigma\,(W_1 \cdot Z + b_1) + b_2) \tag{5}$$

The MLP consists of two fully connected layers, where W_1 and W_2 are the weight matrices of the first and second layers, respectively, and b_1 and b_2 are the bias terms of the corresponding layers. The $\sigma(\cdot)$ denotes the activation function. The MLP layer performs more complex nonlinear transformations on the multimodal features extracted by CLIP to fully explore and combine the high-level abstraction relationships among the features.

The feature Z' output from the MLP network is then passed to the Multihead Self-Attention module, and the Multihead Self-Attention is calculated as shown in Eq. (6):

$$\hat{Z} = MultiHead\,(Z') = Concat\,(head_1, \cdots, head_h)\,W^O \tag{6}$$

where h denotes the number of heads, $head_i$ denotes the output of the *i-th* head, and W^O is the output transformation matrix. The output of each head $head_i$ can be calculated by Eq. (7):

$$head_i = Attention\left(Z'W_i^Q, Z'W_i^K, Z'W_i^V\right) = softmax\left(\frac{Z'W_i^Q(Z'W_i^K)^T}{\sqrt{d_k}}\right)Z'W_i^V \tag{7}$$

Attention is the attention computation function, W_i^Q, W_i^K, W_i^V are the query, key, and value transformation matrices of the *i-th* head, d_k is the dimension of the key vector $Z'W_i^K$, *softmax*

normalizes the similarity, the weight of each key vector is computed, and then multiply the weights by the value vectors, and finally perform a weighted summation to get the Attention output.

The Adapter module utilizes a multi-head self-attention mechanism to assign dynamic weights to the features of each modality of each post so that those post features that are more representative and distinguishable in determining the identity of a user will receive higher attention and weighting. The multi-head self-attention mechanism processes the input features in parallel from multiple different perspectives, and each head outputs a weighted subset of feature vectors, and then stitches together the results from the individual heads to ultimately generate a new representation of the features that reflect the importance of each post.

The multimodal Adapter module makes full use of the powerful cross-modal features provided by CLIP and fine-tunes the differentiation of the importance of image features based on the actual application scenarios, which improves the accuracy of UIL, and does not require fine-tuning of the large multimodal model during training, but only needs to train the lightweight Adapter module, which improves the performance of UIL.

4.4 Temporal Correlation-Based Cross-Modal Similarity

In response to the behavioral patterns of users posting content on different social media platforms, it is found that users tend to post content with some kind of intrinsic relevance within the same period. Based on this observation, we construct a multimodal similarity computation framework incorporating a time factor for measuring users' content consistency across platforms.

First, we construct a user similarity matrix based on text posts on the $\mathcal{O}_1$ platform and multimodal posts containing text, images, and captions on the $\mathcal{O}_2$ platform as shown in Eq. (8):

$$sim = \left[CosSim\left(\widehat{Z}^t, \widehat{Z}^v\right), CosSim\left(\widehat{Z}^t, \widehat{Z}^c\right), CosSim\left(\widehat{Z}^t, \widehat{Z}^d\right)\right] \tag{8}$$

where $CosSim\left(\widehat{Z}^t, \widehat{Z}^v\right)$ is the cosine similarity between the image content of u_2^j and the text content of u_1^i, $CosSim\left(\widehat{Z}^t, \widehat{Z}^c\right)$ is the cosine similarity between the text content of u_2^j and u_1^i, and $CosSim\left(\widehat{Z}^t, \widehat{Z}^d\right)$ is the cosine similarity between u_2^j's caption content and u_1^i's text content. Eq. (6) measures the content consistency of users between the two platforms.

However, considerations based solely on content similarity are not sufficient to fully capture the temporal nature of user behavior. Given that content posted by users within the same period is more relevant, we introduce a time factor r to adjust the similarity weights. This time factor r is defined as shown in Eq. (9):

$$r = \begin{cases} 2/(1 + e^{\alpha \cdot x - \alpha \cdot t_1}), 0 < \mathrm{x} < t_1 \\ 1, t_1 < \mathrm{x} < t_2 \\ e^{-\beta \cdot x + \beta \cdot t_2}, x > t_2 \end{cases} \tag{9}$$

where x represents the publishing time difference between the two posts, t_1 and t_2 are the set time difference threshold parameters, and α and β control the magnitude of the change in the decay function. The image of the time decay function is shown in Fig. 3, when the two posts are closer to each other in terms of publishing time, the larger the time factor value is, and the higher the corresponding similarity confidence is. Specifically, when the publishing time difference x between two posts is small enough, i.e., $0 < x < t_1$, we assign a reward value greater than 1 to the similarity between the posts, and as the time difference increases, the time factor gradually decays, but because the time difference is still small at this time, the time factor decays slowly. When the time difference x between two posts is large, i.e., $x > t_2$, we assign a penalty value less than 1 to the similarity between the posts, and as

the time difference increases, the time factor gradually decays, and it decays rapidly, converging to 0. When the time difference is within the fuzzy period, i.e., $t_1 < x < t_2$, we retain the original content similarity, i.e., we set the time factor to 1.

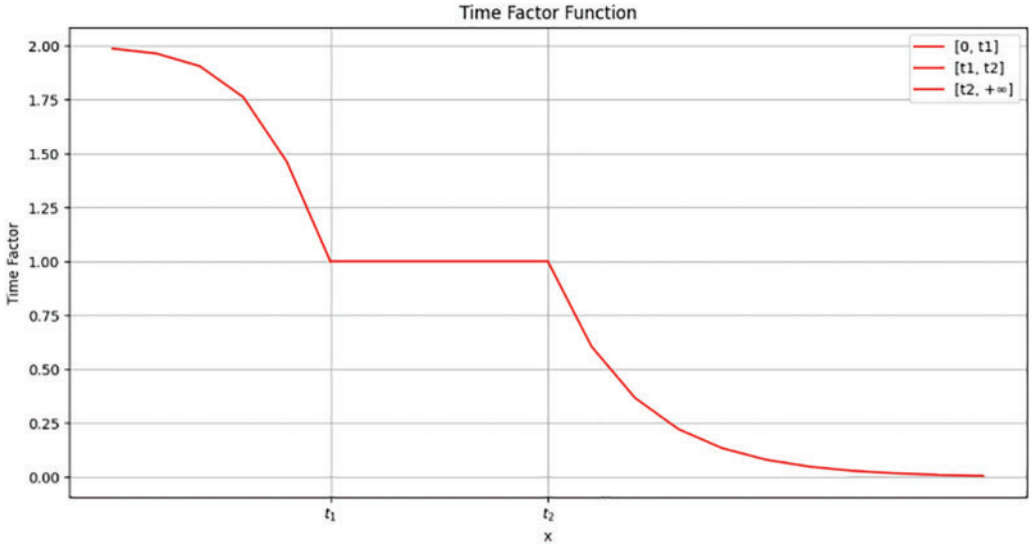

Figure 3: Image of the time decay function. The horizontal coordinate is the time interval and the vertical coordinate is the time factor value

According to Eq. (7), we calculate the time factor matrix R between K textual posts and G multimodal posts, and finally, we combine the time factor-adjusted cross-modal similarity defined as a list of content similarities weighted by the time factor as shown in Eq. (10) as follows:

$$sim^{final} = \left[R \odot CosSim\left(\widehat{Z}^t, \widehat{Z}^v\right), R \odot CosSim\left(\widehat{Z}^t, \widehat{Z}^c\right), R \odot CosSim\left(\widehat{Z}^t, \widehat{Z}^d\right)\right] \tag{10}$$

where $\odot$ denotes multiplication by elements. The final prediction is shown in Eq. (11):

$$o = softmax\left(W_f \cdot AvgPooling\left(sim^{final}\right) + b_f\right) \tag{11}$$

$AvgPooling(\cdot)$ performs mean pooling on the final similarity matrix to reduce the dimensionality, thus forming a vector that reflects the overall similarity between user u_1 and user u_2. Finally, this dimensionally reduced user similarity vector is input to the fully connected layer network for classification, with W_f denoting the weight of the classification layer, b_f denoting the bias of the classification layer, and the output of the fully connected layer transformed into a probability distribution by the *softmax* function. After optimized training, the proposed network can effectively predict whether the user pairs from two platforms are the same person.

The cross-modal similarity calculation method based on the time factor not only effectively addresses the limitations of single-modal similarity calculations but also successfully leverages temporal information to enhance the accuracy and effectiveness of the UIL task. By incorporating the time factor into the similarity calculation, this method provides a more nuanced understanding of the relationship between different modalities, particularly when dealing with unaligned image-text pairs. Algorithm 1 outlines the complete process of the EUIL algorithm.

Algorithm 1: Overall Procedure of EUIL

Input: K textual posts $\mathcal{T}_i$ by u_1^i, G multimodal posts $\mathcal{S}_i$ by u_2^i.
Output: The prediction result o

$\widehat{\mathcal{S}}_i \leftarrow \phi$
for $(\boldsymbol{v}^i, \boldsymbol{c}^i)$ in $\mathcal{S}_i$ **do**
  $\boldsymbol{d}^i \leftarrow f_{BLIP}(\boldsymbol{v}^i)$
  $\widehat{\mathcal{S}}_i \leftarrow \widehat{\mathcal{S}}_i \cup (\boldsymbol{v}^i, \boldsymbol{c}^i, \boldsymbol{d}^i)$
end
Compute Z_1 and Z_2 according to Eqs. (1) and (2)
Compute $\widehat{Z}^v$, $\widehat{Z}^c$, $\widehat{Z}^d$, $\widehat{Z}^t$ according to Eqs. (5) and (6)
Compute user similarity matrix according to Eqs. (7)∼(9)
Predict the probability o that the u_1^i and u_2^i are consistent according to Eq. (10).
return o

5 Experiments

To demonstrate the effectiveness of the proposed method, we conducted experiments on real-world datasets to prove the effectiveness of the proposed method. All experiments were conducted on servers with Intel (R) Xeon (R) Gold 6330 CPUs and three Nvidia A800 GPUs.

5.1 Setup

Dataset. We validate our approach on the TWIN dataset [12] collected by Chen et al. The TWIN dataset collects information about users on two popular heterogeneous social media platforms, Twitter and Instagram. After filtering out some low-quality data, 5765 user pairs were obtained on Twitter and Instagram, along with 1,729,500 UGCs and corresponding timestamps.

Training settings. We divided the user account pairs into three parts: 80% for training, 10% for validation, and 10% for testing. These are considered as positive samples, while negative samples are randomly generated with the same number of positive and negative samples. We converted the UIL into a binary classification task, using accuracy as the evaluation metric. For optimization, we used the Adam optimizer with a learning rate of 0.0001.

5.2 Model Comparison

Due to the limited research done on the issue of linking user identities in UGCs, we compare our scheme with the following baseline:

- WSF-GBDT: This baseline is derived from the method in [28], which introduced four writing-style features, including lexical, syntactic, structural, and content-specific features, to characterize users and employed the support vector machine [29] for the UIL.
- WHOLE: This baseline also characterizes each user account's textual/visual modality by considering all UGCs as a whole. WHOLE employs BiLSTM to derive the textual representation, conducts the text-text similarity and the text-image similarity separately by MLP, and obtains the final user similarity by fusing the text-text similarity and the text-image similarity.
- UserNet: This baseline utilizes two bidirectional long and short-term memory networks to encode text posts from users of the $\mathcal{O}_1$ and $\mathcal{O}_2$ platforms, extracts the depth features of the images uploaded by users of the $\mathcal{O}_2$ platforms using a pre-trained ResNet model, and then constructs the similarity matrices based on the text features and the similarity matrix based on

image features, and makes introduce time factor to weight the similarity. The similarity matrices of the two modalities are fused through the attention mechanism to predict whether the users from the two platforms have the same identity.

Table 2 shows the performance comparison of the different methods. Based on this table, we obtain the following observations: (1) Our proposed method and UserNet are significantly better than WSF-GBDT. this may be attributed to the fact that the former two methods integrate multimodal user representations, not just writing style features. (2) Our proposed method outperforms WHOLE, demonstrating the advantages of fine-grained similarity modeling in the context of user identity correlation, and can further incorporate temporal correlation. (3) Our proposed method significantly outperforms the optimal baseline UserNet. This may be attributed to the fact that UserNet extracts text features using the text-specific unimodal model BiLSTM and image features using the visual unimodal model ResNet, respectively, and computes the similarity matrices directly on top of the two unimodal features, ignoring the semantic gaps between the text and image. semantic gap between text and image. In contrast, our proposed method, which uses the text encoder and image encoder of the multimodal model CLIP to extract features, allows the features of different modalities to be in the same embedding space that can be compared and is more capable of capturing the similarity relationship between image and text. To visually demonstrate the effectiveness of the multimodal model in extracting aligned features on social media, we use a similarity matrix to show the similarity of the image-text features extracted by CLIP, where the lighter the color of the block of elements in the matrix, the higher the image similarity. Where image-text pairs come from heterogeneous multimodal posts made by users on social media, there is often a potential correlation between images and text. As shown in the Fig. 4 below, the features extracted by the CLIP model better capture this correlation.

Table 2: Performance comparison among different models in terms of accuracy

Model	Accuracy
WHOLE	0.6836
WSF-GBDT	0.7552
UserNet	0.8369
EUIL (ours)	**0.8639**

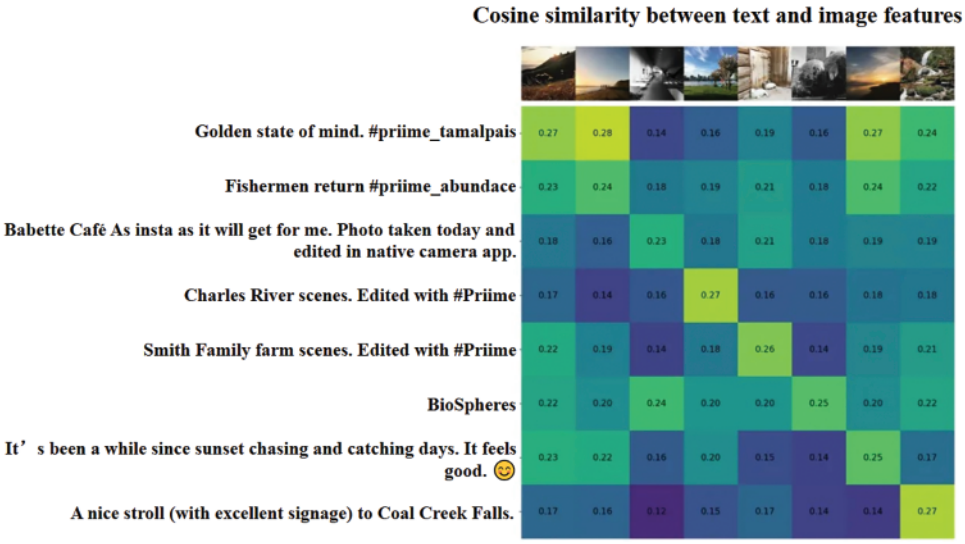

Figure 4: Similarity matrix of multimodal posts based on features extracted from CLIP model

5.3 Ablation Experiments

To gain a deeper understanding of our proposed model, we compare our proposed method EUIL with several derived methods. Specifically, we explore the effects of modal diversity, text enhancement, multimodal Adapter, and time factor on UIL performance.

- EUIL$_{text}$: Only the text modal of the $\mathcal{O}_2$ platform is considered and only the original text posted by the user is taken into account, ignoring the image modal data of the $\mathcal{O}_2$ platform with the text of the generated image description.
- EUIL$_{img}$: Only the image modal of the $\mathcal{O}_2$ platform is considered, ignoring the text modal data of the $\mathcal{O}_2$ platform, which includes the original text posted by the user with the generated caption.
- EUIL$_{text\&img}$: Only the raw data posted by the users of the $\mathcal{O}_2$ platform is considered, including the raw text data posted by the users along with the image data, ignoring the captions generated for the images.
- EUIL$_{img\&cap}$: Only the image modality of the $\mathcal{O}_2$ platform and the captions generated for the images were considered, ignoring the raw text posted by the users of the $\mathcal{O}_2$ platform.
- EUIL$_{w/o-adapter}$: We disable the multimodal Adapter module of EUIL to directly construct a similarity matrix based on the features extracted from the pre-trained model CLIP to predict whether the users have the same identity.
- EUIL$_{w/o-time}$: We disable the time factor of EUIL and unweight the time factor on the similarity when calculating the similarity between posts.

5.3.1 On Modal

To explore the contribution of each modality to UIL, we compared the results between EUIL, EUIL$_{text}$, and EUIL$_{img\&cap}$, which are shown in Table 3.

The primary observation that the accuracy of the EUIL model is higher than that of EUIL$_{text}$ may be rooted in the fact that the EUIL$_{text}$ model is limited to analyzing only the textual information of the $\mathcal{O}_2$ platform, whereas the EUIL model not only integrates the textual information but also

incorporates the rich resources of image data. Image data, as a complement, can fill in the potential contextual gaps and detail omissions in textual descriptions, and this fine-grained information is often difficult to fully convey by pure textual representations. Second, the observation that the accuracy of EUIL is higher than that of the $\text{EUIL}_{img\&cap}$ model emphasizes the central role of textual data in decoding the user's intention and contextual framework. User-generated text is rich in direct emotional expressions and specific points of view, elements that are essential for fine-grained characterization of users' personalities and pinpointing their needs. Taken together, these analyses point to the conclusion that the fusion of multiple data modalities (e.g., text and images) is decisive for building a more comprehensive user profile, which significantly enhances the performance and accuracy of user-matching models. This conclusion emphasizes the significance of cross-modal data integration in enhancing the effectiveness of UIL systems.

Table 3: Results on the effect of different modal combinations on the accuracy

Model	K = G = 150	K = G = 100	K = G = 50
EUIL	0.8815	0.8671	0.8191
EUIL_{text}	0.7654	0.7446	0.7134
$\text{EUIL}_{img\&cap}$	0.6549	0.5717	0.5709

5.3.2 On Image Caption

To validate the contribution of the BLIP-based text enhancement module to UIL, we compare the results of EUIL, $\text{EUIL}_{text\&img}$, EUIL_{img}, $\text{EUIL}_{img\&cap}$ and the results are shown in Table 4.

Table 4: Results on the effect of captions on the accuracy

Model	K = G = 150	K = G = 100	K = G = 50
EUIL	**0.8815**	**0.8671**	**0.8191**
$\text{EUIL}_{text\&img}$	0.8647	0.8471	0.7862
EUIL_{img}	0.5124	0.5116	0.4604
$\text{EUIL}_{img\&cap}$	0.6549	0.5717	0.5709

Preliminary observations show that the accuracy of EUIL model outperforms $\text{EUIL}_{text\&img}$, while the performance of $\text{EUIL}_{img\&cap}$ is significantly higher than that of EUIL_{img}. This finding strongly proves the validity of the textual descriptions of the images and indicates that the textual descriptions of the images deeply mine and enrich the semantic content behind the image, which in turn enhances the expressiveness of user features. In addition, the superior performance of $\text{EUIL}_{img\&cap}$ further highlights the strategy of expanding textual data by integrating textual descriptions of images, which is especially crucial to compensate for the data limitations of social media platforms that rely on a single textual modality. In summary, our proposed BLIP-Enhanced Textual Enrichment module significantly enhances the accuracy and efficacy of user matching by generating textual descriptive information of images, validating its effectiveness in enhancing UIL.

5.3.3 On Multimodal Adapter

To validate the contribution of the multimodal Adapter module to UIL, we compared the results of EUIL and $\text{EUIL}_{w/o-adapter}$, which are shown in Table 5.

Table 5: Results on the effect of multimodal adapter on accuracy

Model	K = G = 150	K = G = 100	K = G = 50
EUIL	**0.8815**	**0.8671**	**0.8191**
$\text{EUIL}_{w/o-adapter}$	0.5532	0.5468	0.5460

Observations show that the EUIL model exhibits a significant advantage in accuracy over its variant $\text{EUIL}_{w/o-adapter}$. This performance difference may be rooted in the simplifying assumption that $\text{EUIL}_{w/o-adapter}$ is adopted in processing the characteristics of K/G user posts, i.e., it defaults to the same importance of the information contained in all the posts, which fails to adequately take into account the intrinsic differences and noise specific to user-generated content (UGC) in different social media platforms. On the contrary, the EUIL model explicitly identifies the information noise embedded in the K/G samples through deep insights and is highly sensitive to the uneven contribution of each post to the user matching task, which plays a central role in the innovative multimodal adapter mechanism introduced by EUIL, which dynamically and adaptively assigns differentiated weights to each user's K/G posts. This strategy not only optimizes the adjustment based on the relevance and quality of each post but also effectively utilizes multimodal information to enhance feature representation. Therefore, EUIL not only refines the learning of each post's features but also significantly improves the accuracy and robustness of UIL through this series of fine-tuning, verifying its effectiveness and superiority.

5.3.4 On Time Factor

To validate the contribution of introducing the time factor to UIL, we compared the results of EUIL with $\text{EUIL}_{w/o-time}$, which are shown in Table 6.

Table 6: Results on the effect of time factor on the accuracy

Model	K = G = 150	K = G = 100	K = G = 50
EUIL	**0.8815**	**0.8671**	**0.8191**
$\text{EUIL}_{w/o-time}$	0.8639	0.8311	0.7958

Observations show that the accuracy of the EUIL model outperforms that of the $\text{EUIL}_{w/o-time}$ model, a difference that may stem from the fundamental difference in their approaches to user posts. The $\text{EUIL}_{w/o-time}$ model evaluates only the direct similarity in the content of users' posts, ignoring the effect of the time dimension. In contrast, the EUIL model incorporates a temporal dimension into the consideration of content relevance, i.e., it assumes that there is an intrinsic relevance of the content posted by users in a similar period. By integrating the time information of posts to construct a temporal relevance model, EUIL effectively enhances the granularity of post-similarity modeling, which in turn optimizes the effectiveness of user matching. This finding highlights the importance of integrating time-series features with content features to enhance UIL accuracy in user behavior analysis.

6 Discussion

6.1 Potential Improvements

Our approach leverages aligned multimodal features and temporal correlation to enhance the performance of the UIL task. By utilizing aligned multimodal features and incorporating temporal information, we have achieved significant improvements. However, despite these advancements, there remain some limitations and areas for potential improvement:

(1) Integration of Multi-Dimensional User Features: EUIL primarily focuses on studying the contribution of multi-modal user-generated content to user alignment. However, different social media platforms emphasize various types of content, which poses significant challenges to the proposed methods, potentially affecting their generalization and robustness. Beyond user-generated content, other user features on social media platforms, such as user profiles and social network structures, may provide crucial clues for enhancing the accuracy of UIL. To build more universally applicable models, future work should consider incorporating a wider range of user features into the model.

(2) Construction of Datasets with Multi-Dimensional User Features: Training and evaluating models using more diverse datasets ensures their generality and robustness. However, due to user privacy protection, publicly available datasets that meet the requirements are limited. Therefore, constructing datasets that include comprehensive features is necessary for future work, including collecting data from multiple social media platforms, and covering information from diverse user groups, thereby enhancing the representativeness and breadth of research. In this way, we can more comprehensively evaluate the model's performance in real-world scenarios, ensuring its adaptability to real-world environments and improving its recognition capabilities in complex situations.

(3) The dynamic nature of social media user behavior and preferences poses significant challenges for UIL. Capturing long-term trends and short-term changes is crucial for maintaining the accuracy and relevance of these systems. In future work, we can pursue research along the following three directions: (1) Longitudinal Tracking Mechanism: We plan to develop a mechanism capable of tracking user behavior over extended periods. This will involve collecting and analyzing historical data to identify patterns and trends in user activities. By doing so, the system can adapt to long-term changes in user behavior. (2) Dynamic Learning Algorithms: We aim to design algorithms that can continuously update user models based on incoming data. This dynamic learning capability will enable the system to quickly adapt to short-term changes in user behavior, such as shifts in interests or sudden changes in activity patterns. (3) Feedback Loop for Model Refinement: We propose implementing a feedback loop to refine user models based on real-time feedback. This could involve engaging users with the system, such as through corrections or verifications of linked identities, to improve the accuracy of the models. We believe that these enhancements will significantly improve the robustness and adaptability of our UIL system. They will enable the system to maintain its performance even as user behavior evolves.

6.2 Privacy Security

The performance of EUIL on the TWIN dataset indicates that UIL relying solely on UGC across Twitter and Instagram platforms is feasible. Unlike user profiles, which can be disclosed or contain false information, UGC often represents genuine user behavior and is publicly available on social platforms, which makes UGC-based UIL techniques susceptible to malicious exploitation for consolidating sensitive personal information across social media platforms. Therefore, it is necessary to take measures to ensure privacy security. Research on data privacy protection in other fields is relatively advanced [30], but research on user social data privacy protection is relatively insufficient.

There are specific UIL adversarial attack methods targeting user attributes, such as the DeLink [31] framework, which draws from adversarial text generation ideas to help users modify their social media usernames, thus defending against UIL. Additionally, there are adversarial attack methods targeting the structure of social networks, such as the TOAK [32] strategy, which leverages the topological structure information of networks by carefully perturbing network structures to reduce the matching accuracy of UIL models. However, there are currently adversarial methods specifically targeting UGC. To prevent UGC from being maliciously utilized, our next steps will explore how to generate adversarial UGC by introducing imperceptible perturbations without disrupting the intended expression of the user, thereby reducing the effectiveness of UGC-based UIL strategies and further protecting user data privacy.

7 Conclusion

In this paper, we propose an efficient UIL scheme based on aligned multimodal features and temporal correlation to address the challenges of multimodal UIL in social media platforms. The method utilizes the BLIP model for image-generated captions, which compensates for the lack of textual information in social media platforms and significantly improves the expressive power of image modality in UIL tasks. Moreover, the method makes full use of the pre-trained multimodal large model CLIP to realize the representation of data in both text and image modalities under a unified feature space, achieving an effective fusion of cross-modal features. The subsequently designed multimodal adapter module, by its lightweight and flexible features, can extract a comprehensive feature representation that is highly correlated and suitable for user identity association from the multimodal features extracted by CLIP by training with a small number of additional parameters. In addition, we introduce a time factor assignment mechanism to construct a user similarity prediction model that can dynamically reflect the temporal dynamics of user interest evolution and social behavior, thus enhancing the timeliness and accuracy of the UIL method. Experiments demonstrate that this method performs well on large-scale real-world social media datasets, not only overcoming the problem of data modal differences and imbalance but also surpassing the existing state-of-the-art methods in terms of accuracy and efficiency.

Acknowledgement: Not applicable.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization and methodology, Jiaqi Gao; software, Jiaqi Gao and Bin Wu; validation, Xiujuan Wang, Kangfeng Zheng and Chunhua Wu; writing—original draft preparation, Jiaqi Gao; writing—review and editing, Xiujuan Wang and Chunhua Wu; supervision, Kangfeng Zheng. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are available from the corresponding author, Kangfeng Zheng, upon reasonable request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] X. Chen, F. Zhou, K. Zhang, G. Trajcevski, T. Zhong and F. Zhang, "Information diffusion prediction via recurrent cascades convolution," presented at the 2019 IEEE 35th Int. Conf. Data Eng. (ICDE), Macao, China, Apr. 8–11, 2019, pp. 770–781.

[2] T. H. Lin, C. Gao, and Y. Li, "CROSS: Cross-platform recommendation for social E-commerce," presented at the Proc. 42nd Int. ACM SIGIR (Special Interest Group Inf. Retr.) Conf. Res. Develop. Inf. Retr., Paris, France, Jul. 21–25, 2019, pp. 515–524.

[3] M. Wang, W. Wang, W. Chen, and L. Zhao, "EEUPL: Towards effective and efficient user profile linkage across multiple social platforms," *World Wide Web*, vol. 24, no. 5, pp. 1731–1748, Jun. 2021. doi: 10.1007/s11280-021-00882-7.

[4] J. Shao, Y. Wang, H. Gao, B. Shi, H. Shen and X. Cheng, "AsyLink: User identity linkage from text to geo-location via sparse labeled data," *Neurocomputing*, vol. 515, no. 6, pp. 174–184, Jan. 2023. doi: 10.1016/j.neucom.2022.10.027.

[5] Y. Huang, P. Zhao, Q. Zhang, L. Xing, H. Wu and H. Ma, "A Semantic-enhancement-based social network user-alignment algorithm," *Entropy*, vol. 25, no. 1, Jan. 2023, Art. no. 172. doi: 10.3390/e25010172.

[6] H. Gao, Y. Wang, J. Shao, H. Shen, and X. Cheng, "User identity linkage across social networks with the enhancement of knowledge graph and time decay function," *Entropy*, vol. 24, no. 11, Nov. 2022, Art. no. 1603. doi: 10.3390/e24111603.

[7] B. Chen and X. Chen, "MAUIL: Multilevel attribute embedding for semisupervised user identity linkage," *Inf. Sci.*, vol. 593, pp. 527–545, May 2022. doi: 10.1016/j.ins.2022.02.023.

[8] V. Sharma and C. Dyreson, "LINKSOCIAL: Linking user profiles across multiple social media platforms," presented at the 2018 IEEE Int. Conf. Big Knowl. (ICBK), Singapore, Nov. 17–18, 2018, pp. 260–267.

[9] Y. Li, G. Gou, G. Xiong, Z. Li, and M. Cui, "The potential utility of image descriptions: User identity linkage across social networks based on MultiModal self-attention fusion," presented at the 2023 IEEE Int. Perform., Comput., Commun. Conf. (IPCCC), Anaheim, CA, USA, Nov. 17–19, 2023, pp. 265–273.

[10] S. Li, D. Lu, Q. Li, X. Wu, S. Li and Z. Wang, "MFLink: User identity linkage across online social networks via multimodal fusion and adversarial learning," *IEEE Trans. Emerg. Top. Comput. Intell.*, pp. 1–10, Mar. 2024. doi: 10.1109/TETCI.2024.3440057.

[11] X. Chen, X. Song, G. Peng, S. Feng, and L. Nie, "Adversarial-enhanced hybrid graph network for user identity linkage," presented at the Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr., Jul. 11–15, 2021, pp. 1084–1093.

[12] X. Chen, X. Song, S. Cui, T. Gan, Z. Cheng and L. Nie, "User identity linkage across social media via attentive time-aware user modeling," *IEEE Trans. Multimedia*, vol. 23, pp. 3957–3967, Nov. 2020. doi: 10.1109/TMM.2020.3034540.

[13] J. Li, D. Li, C. Xiong, and S. Hoi, " BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," presented at the Proc. 39th Int. Conf. Mach. Learn. (PMLR), Baltimore, MD, USA, Jul. 17–23, 2022, pp. 12888–12900.

[14] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," presented at the Proc. 38th Int. Conf. Mach. Learn. (ICML), Jul. 18–24, 2021, pp. 8748–8763.

[15] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," presented at the Proc. 2014 Conf. Empir. Methods Nat. Lang. Process. (EMNLP), Doha, Qatar, Oct. 25–29, 2014, pp. 1532–1543.

[16] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," presented at the Proc. 26th Int. Conf. Neural Inf. Proc. Syst., Lake Tahoe, NV, USA, Dec. 05–10, 2013, pp. 2787–2795.

[17] J. Zhang *et al.*, "MEgo2Vec: Embedding matched ego networks for UIL across social networks," presented at the Proc. 27th ACM Int. Conf. Inf. Knowl. Manag., Torino, Italy, Oct. 22–26, 2018, pp. 327–336.

[18] M. Long *et al.*, " DegUIL: Degree-aware graph neural networks for long-tailed user identity linkage," presented at the Proc. Eur. Conf. Mach. Learn. Princ. Pract. Knowl. Discov. Databases, Turin, Italy, Sep. 18–22, 2023, pp. 122–138.

[19] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.
[20] Y. Kim, "Convolutional neural network for sentence classification," 2015, *arXiv:1408.5882*.
[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," presented at the Proc. 2016 IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, Jun. 27–30, 2016, pp. 770–778.
[22] J. Chung *et al.*, "Empirical evaluation of gated recurrent neural networks on sequence modeling," 2014, *arXiv:1412.3555*.
[23] B. Perozzi, R. Al-Rfou, and S. Skiena, "DeepWalk: Online learning of social representations," presented at the Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., NY, USA, Aug. 24–27, 2014, pp. 701–710.
[24] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997. doi: 10.1162/neco.1997.9.8.1735.
[25] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.
[26] Z. Liu, H. Mao, C. Y. Wu, C. Feichtenhofer, T. Darrell and S. Xie, "A convnet for the 2020s," presented at the Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), New Orleans, LA, USA, Jun. 18–24, 2022, pp. 11976–11986.
[27] J. Li *et al.*, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," presented at the Proc. Int. Conf. Mach. Learn. (ICML), Honolulu, HI, USA, Jul. 23–29, 2023, pp. 19730–19742.
[28] R. Zheng, J. Li, H. Chen, and Z. Huang, "A framework for authorship identification of online messages: Writing style features and classification techniques," *J. Am. Soc. Inf. Sci. Technol.*, vol. 57, no. 3, pp. 378–393, Dec. 2005. doi: 10.1002/asi.20316.
[29] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995. doi: 10.1007/BF00994018.
[30] F. Buccafurri, V. De Angelis, and S. Lazzaro, "MQTT-A: A broker-bridging P2P architecture to achieve anonymity in MQTT," *IEEE Internet Things J.*, vol. 10, no. 17, pp. 15443–15463, Sep. 2023. doi: 10.1109/JIOT.2023.3264019.
[31] P. Zhang *et al.*, "DeLink: An adversarial framework for defending against cross-site user identity linkage," *ACM Trans. Web*, vol. 18, no. 2, pp. 1–34, Mar. 2024. doi: 10.1145/3643828.
[32] J. Shao *et al.*, "TOAK: A topology-oriented attack strategy for degrading user identity linkage in cross-network learning," presented at the Proc. 32nd ACM Int. Conf. Inf. Knowl. Manag. (CIKM), Birmingham, UK, Oct. 21–25, 2023, pp. 2208–2218.