# Distilling Meta Knowledge on Heterogeneous Graph for Illicit Drug Trafficker Detection on Social Media

**Anonymous Author(s)**
Affiliation
Address
email

## Abstract

Driven by the considerable profits, the crime of drug trafficking (a.k.a. illicit drug trading) has co-evolved with modern technologies, e.g., social media such as Instagram has become an important platform for marketing and selling illicit drugs. The activities of online drug trafficking are nimble and resilient, which call for novel techniques to effectively detect, disrupt, and dismantle illicit drug trades. In this paper, we propose a holistic framework named MetaHG to automatically detect illicit drug traffickers on social media (i.e., Instagram), by tackling the following two new challenges: (1) different from existing works which merely focus on analyzing post content, MetaHG is capable of jointly modeling post content and relational structured information on social media for illicit drug trafficker detection; (2) in addition, through the proposed meta-learning technique, MetaHG addresses the issue of requiring sufficient data for model training. More specifically, in our proposed MetaHG, we first build a heterogeneous graph (HG) to comprehensively characterize the complex ecosystem of drug trafficking on social media. Then, we employ a relation-based graph convolutional neural network to learn node (i.e., user) representations over the built HG, in which we introduce graph structure learning to compensate sparse relational information among entities in the HG for more robust node representation learning. Afterwards, we propose a meta-learning algorithm to optimize model parameters. A self-supervised module and a knowledge distillation module are further designed to exploit unlabeled data for improving the model. Extensive experiments based on the real-world data collected from Instagram demonstrate that the proposed MetaHG outperforms state-of-the-art methods.

## 1 Introduction

As the market of illicit drugs (e.g., heroin, synthetic opioids such as Fentanyl) is considerably lucrative, the crime of drug trafficking (a.k.a. illicit drug trading) has never stopped but co-evolved with the advance of modern technologies, e.g., the practice of illicit drug trade has transformed from the physical world to online platforms. It has been shown that major social media platforms, including Instagram, Twitter, Snapchat, and Facebook, have become a direct-to-consumer marketing medium for illicit drug traffickers [27]. As illustrated in Figure 1, due to the convenience and popularity, drug traffickers can easily create accounts on these platforms to advertise and sell drugs by posting codewords (e.g., street name) and images of drugs that are in stock (shown in Figure 1.(a)); they can also post their drug sale websites and utilize social media platforms for promotion (shown in Figure 1.(b)). Buyers can access illicit drugs and trade with vendors easily through these social media platforms. Illegal trading of drugs has turned into a serious global and national concern because of its catastrophic consequences on society, from unimaginable violent crimes to public health (e.g., more than 87,000 people died from drug overdose in 2020 in the U.S. and it broke the record with a 20% increase in a year [41]).

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

Despite the persistent effort by law enforcement agencies, due to the considerable profits, activities of online drug trafficking are nimble and resilient: as shown in Figure 1.(c), to avoid being banned, skilled drug traffickers (e.g., with 1,002 followers, 118 following, and 630,000 like) rarely post drug description or images on social media, but instead, they advertise drugs implicitly using the slang (e.g., using "KET" to refer ketamine) and leave encrypted chat tool contact information (e.g., Kik Messenger) through their comments to other users' posts. To combat online drug trafficking, there's an imminent need to develop novel techniques for effective detection of drug traffickers on social media and thus enable law enforcement for proactive interventions to disrupt and dismantle illicit drug trades.

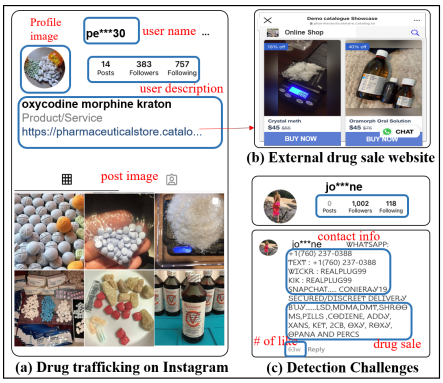

Figure 1: Showcase of drug trafficking on social media platform.

To tackle the above problem, there have been many research efforts on the investigation of online drug trafficking. Some existing works focus on analyzing the drug trafficking activities on darknet markets [45, 46, 10], while some methods have been proposed for detecting illicit drug trading activities on social media [22, 23, 24]. Most of these works usually rely on a single type of content feature (e.g., text or image) while ignoring the structured relationship information among entities on social media. For instance, Instagram monitors drug traffickers simply by filtering drug related keyword hashtags (e.g., weed4sale) [15]. Unfortunately, as illustrated in Figure 1.(c), drug traffickers may always invent new tactics to evade detection. Those methods that fail to consider multi-modal features (e.g., both text and image) and the structural relationships among entities (e.g., users and posts) could limit the effectiveness of detecting drug trafficking activities. Additionally, most recent works [24, 25, 22] require sufficient labeled samples to train the model for the detection of drug trafficking. For example, Mackey *et al.* [25] collected almost 612,000 posts on Twitter and only found 1,778 drug trafficking-related posts. Obtaining labeled samples on social media is always expensive, and existing models may suffer from the constraint of few labeled data to detect drug traffickers.

To address the above challenges, in this paper, we propose a holistic framework named MetaHG (as shown in Figure 2) to automatically detect illicit drug traffickers on social media (i.e., using Instagram as a showcase). First, to overcome the shortcomings of merely analyzing a single type of content information (e.g., text or image), we propose to consider the multi-modal content features (i.e., both text and image) and introduce a heterogeneous graph (HG) to model the relationships among three types of entities (i.e., user, post, and keyword). The constructed HG is able to comprehensively characterize the complex ecosystem of drug trafficking on social media. And then we exploit relation-based graph convolutional neural network (R-GCN) [33] to fuse both relational information among entities and content features to obtain initial node embeddings over the built HG. Due to spare relational connections among some entities, we further introduce graph structure learning (GSL) [26, 18] to learn more robust node representations on the HG. Furthermore, to address the few labeled data issue, we propose a meta-learning framework to transfer knowledge from training tasks and effectively adapt them to testing tasks (e.g., new types of drug traffickers with few labeled samples). More specifically, to exploit unlabeled information for better performance, our proposed model consists of two additional modules: an attribute similarity-based self-supervised learning module to augment R-GCN for node representation refinement, and a knowledge distillation module to facilitate better model optimization. In summary, the major contributions of our work include:

- We study the illicit drug trafficker detection on social media and create a new dataset on Instagram (including text, image and relation information) for the problem, which is novel and urgent.
- To solve the problem, we develop a holistic yet novel framework (i.e., MetaHG) to jointly model both structured relations and unstructured content information for illicit drug trafficker detection, accounting for the constraints of graph sparsity and limited labeled data for model training.
- Comprehensive experiments on the real-world dataset demonstrate the outstanding performance of MetaHG by comparison with the state-of-arts methods. To the best of our knowledge, MetaHG is the first work that automatically detects drug traffickers on social media using few labeled data.
- As drug overdose deaths have continued to increase over the past decade across the country, our proposed technique will have significant societal impact to help address this critical issue.

## 2 Related Work

In this section, we briefly introduce the relevant studies in three aspects: drug trafficker detection, graph neural networks, and meta-learning.

**Drug Trafficker Detection.** Most existing works analyze drug traffickers on darknet markets [45, 46, 5, 47, 4]. For instance, Zhang et al. [45] developed a system by leveraging user's styles to detect drug traffickers on darknet markets. As drug trafficking appears on surface web dramatically recently, some recent works [24, 22, 25] detect illicit drug traffickers on social media platforms by analyzing text or image of posts. For example, Li et al. [22] used a RNN model to learn the text pattern of posts to detect drug dealers on Instagram. However, these methods seldom consider multi-modal features and the relational information among entities. Different from existing works, we study the characters of drug traffickers by incorporating structural relation and unstructured content information.

**Graph Neural Networks.** Most existing graph neural networks (GNNs) learn the node embedding by aggregating the features of neighborhood nodes [19, 13, 43, 39]. GCN [19] implements layer-wise propagation rule to learn the node embedding, and GAT[39] proposes to learn different attention scores for neighbors when aggregating neighborhood information. While most of these existing models work with homogeneous graphs, some recent models [33, 34, 42, 44] are proposed to deal with heterogeneous graphs which are more practical in reality. Inspired by GCN, Weighted-GCN [34] and R-GCN [33] propose to learn node embedding based on multiple relational neighborhoods. Additionally, inspired by GAT, HAN [42] leverages attention mechanism to learn the importance between nodes and meta-path simultaneously in heterogeneous graphs. Motivated by these studies, we employ R-GCN as our base model to learn the node embedding.

**Meta-Learning.** Recent meta-learning models generally can be divided into two groups: metric-based meta-learning [20, 40, 36, 30, 6] and gradient-based meta-learning [11, 21, 12, 3, 29, 9]. For metric-based models, they implement a generalized metric and matching functions from training tasks to train the model. For instance, Matching Networks [40] learns a network that maps a small labeled support set and an unlabelled example to its label. For gradient-based models, they employ existing tasks data to learn well initialized model parameters that can be fast updated to new tasks with few data [12] or directly implement a meta-optimizer to learn the optimization process [38]. For example, Finn et al. [11] proposed MAML, in which the parameters of the model are explicitly trained such that several gradient steps with few training data from a new task will produce good generalization performance on that task. In this paper, motivated by MAML, we introduce a meta-learning framework to address small labeled data challenge of illicit drug traffickers detection.

## 3 Preliminary

### 3.1 Problem Definition

Given a social media dataset, we build a heterogeneous graph (HG) to comprehensively depict the rich information among the dataset for learning the user embedding. Let $G = (\mathcal{V}, \mathcal{E}, \mathcal{X})$ denotes a HG where $\mathcal{V}$ is the set of different types of nodes, $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the set of edges, and $\mathcal{X}$ is the node attribute set. Specifically, there are three types of nodes (user, post, and keyword) and eight types of relations (e.g., user-reply-post) in HG (see Figure 2.(a)). A node can be regarded as any type of entities and an edge can be regarded as any type of relations between two nodes. The node attribute can be considered as the feature applied to the node. The goal is to learn the user embedding which can be fed to a classifier for drug traffickers detection. Given the features of all user nodes $X_r = (x_1, \ldots, x_N)$ ($N$: number of users) and their binary labels $Y = (y_1, \ldots, y_N)$ ($y_i = 1$ denotes drug trafficker and $y_i = 0$ represents regular user), we target learning a mapping function (detection function) $U_\phi : X_r \rightarrow Y$ (with parameter $\phi$). Unlike existing works that use sufficient samples for model training, we consider a more practical scenario that only few labeled data are available because the number of drug trafficker samples collected from social media is limited. Following the definition of few-shot learning [11], given the labeled data of existing drug traffickers of different types (e.g., stimulants trafficker (refer to the Supplementary Material 1.1)), we aim to build a classifier which can effectively detect drug traffickers of new types with few labeled data (e.g., opioid trafficker). Formally, the problem is defined as follows.

**Problem 1.** ***Drug Trafficker Detection on Social Media.*** *Given a social media data denoted as HG $G$ and a set of different types of drug traffickers with their features $X_r = (x_1, \ldots, x_N)$ and*

*corresponding labels $Y = (y_1, \ldots, y_N)$ (training data), the problem is to build a machine learning model to detect illicit drug traffickers of new types that only have few labeled samples (testing data).*

### 3.2 Relational Graph Convolutional Networks

Relational graph convolutional network (R-GCN) [33] has been proved to be powerful for learning the node representation in heterogeneous graphs. Different from GCN, R-GCN considers the different types of relationships between two nodes when aggregating features from neighborhood. Thus, we employ R-GCN as the base model to learn user representations in HG. According to the definition of R-GCN, The layer-wise propagation rule can be formulated as follows:

$$h_i^{l+1} = \sigma(\sum_{r \in \mathcal{R}} \sum_{j \in \mathcal{N}_i^r} \frac{1}{|N_i^r|} W_r^{(l)} h_j^{(l)} + W_0^{(l)} h_i^{(l)}) \tag{1}$$

where $h_i^{l+1}$ denotes the representation of node $i$ at $l+1$ layer, $N_i^r$ denotes the set of neighbors of node $i$ under relation $r \in \mathcal{R}$, and $\mathcal{R}$ is the relation type set of HG. For simplicity, we use $Z = \text{R-GCN}(X, A, \mathcal{R})$ to denote a R-GCN model where $Z$ is the node embedding, $X$ is the attribute feature matrix, and $A$ is the adjacency matrix based on $\mathcal{R}$.

## 4 Methodology

In this section, we present the details of our proposed model MetaHG (Figure 2). At first, we construct a HG to depict both relation and content information on social media, and introduce graph structure learning (GSL) to generate a refined HG for compensating the spare connection among some entities. Then we utilize R-GCN over the learned graph to learn node embedding and augment the model by an attribute similarity based self-supervised learning module (AS-SSL). Lastly, we design a meta-learning algorithm for model optimization, which is further enhanced by a knowledge distillation module (KD).

### 4.1 HG Construction on Social Media

To comprehensively describe illicit drug traffickers on social media, besides text and image content, we also consider the structural relation on social media. As shown in Figure 2.(a), we build a HG with three types of entities (user, post, and keyword) and eight types of relations as well as features of each entity, such that both content and relation information can be exploited simultaneously. Now we introduce the content feature and relation information in detail. **Text Feature.** For each type of node in HG, we extract different text features (as listed in Table 1) and implement pre-train language model *BERT* [7] to obtain the text feature vector. Afterwards, we concatenate them to denote feature vector of node. **Image Feature.** Most users on social media post images and text simultaneously. Thus we employ the pre-trained image model *VGG19* [35] and *PCA* [32] to acquire the image feature for each post and user. **Relation.** Besides text and image features, relations among different entities are also indispensable to distinguish drug traffickers on social media. Hence, we leverage eight types of edges (R1-R8 in Table 1) to describe the informative relations among three types of entities. Detailed description of features and relations in HG are discussed in the Supplementary Material 1.2.

Table 1: Content and relation information in HG.

| Content Feature | |
|---|---|
| User | username, profile information,is-business<br># of followers/followings/posts |
| Post | text content, image information<br>comment content, # of likes/comments |
| Keyword | keywords from all posts<br>(oxycodone, fentanyl, codine, morphine, LSD) |

| Relation | |
|---|---|
| R1: user-follow/followed-user | R2: user-tagger-user |
| R3: user-reply-post | R4: user-mention-post |
| R5: user-have-post | R6: user-profile-keyword |
| R7: post-include-keyword | R8: post-tag-keyword |

### 4.2 HG Structure Refinement via Metric Learning

The HG constructed in the previous step is sparse in connections among some nodes since some users appear to be inactive on social media. That is, the original HG is not optimal for detecting drug traffickers on social media. To address this problem, we propose to generate better HG by graph structure learning (GSL). Specifically, we first calculate the similarities between node features to

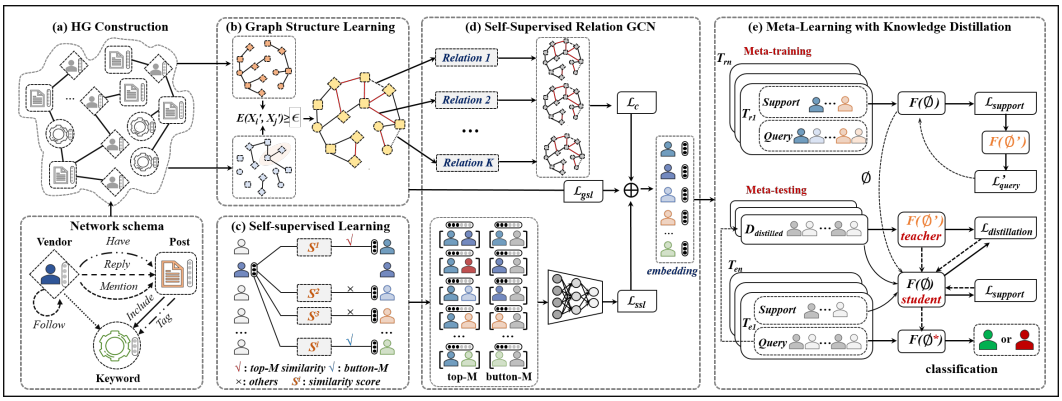


Figure 2: The overall framework of MetaHG: (a) HG construction; (b) Graph structure learning for HG refinement; (c) and (d) Self-supervised R-GCN for generating node embedding; (e) Meta-learning with knowledge distillation for model optimization.

infer the potential connections between nodes by metric learning and obtain the feature similarity graph (FSG). Then, the original HG and FSG are combined to obtain a new and refined graph (see Figure 2.(b)). The generation process of FSG is described as below.

**Feature Similarity Graph.** Since the dimension of different types of node features is different, we first adopt a type-specific mapping layer to project all of the node features $X$ to a $d_c$-dimensional common space:

$$X_i^{'} = \sigma(X_i \cdot W_k + b_k), \tag{2}$$

where $\sigma$ is the *sigmoid* activation function, $X_i$ is the feature vector of node $v_i$, $W_k \in \mathbb{R}^{d_k \times d_c}$ and $b_k \in \mathbb{R}^{d_c}$ denote the projection matrix and the bias vector for node type $k$ respectively. Afterwards, we implement metric learning on node features in the common space and obtain a FSG $\mathcal{G}'$ where the edge between node pair $v_i$ and $v_j$ is obtained by:

$$\mathcal{E}'_{i,j} = \begin{cases} 1 & E(X_i^{'}, X_j^{'}) \geq \epsilon \\ 0 & \text{otherwise} \end{cases}, \tag{3}$$

where $\epsilon \in [0, 1]$ is the threshold value controlling the density of the generated FSG. $E(X_i^{'}, X_j^{'})$ is a parameterized function to calculate the similarity between two nodes:

$$E(X_i^{'}, X_j^{'}) = \mathbf{\Gamma}(W_{s1} \odot X_i^{'}, W_{s2} \odot X_j^{'}) \tag{4}$$

where $\mathbf{\Gamma}$ denotes a similarity function ($cosine$ similarity in this work), $W_{s1}$ and $W_{s2}$ are the parameters which measure the importance of different feature dimensions, and $\odot$ denotes the Hadamard product. The learned graph $\mathcal{G}^l$ is obtained by combining the original HG $\mathcal{G}$ and the generated FSG $\mathcal{G}'$. In order to make $\mathcal{G}^l$ be relatively sparse and avoid overfitting for the downstream task, we introduce a structure sparsity regularizer via $L_1$ norms:

$$\mathcal{L}_{gsl} = \left\| A^l \right\|, \tag{5}$$

where $A^l$ is the adjacency matrix of $\mathcal{G}^l$.

### 4.3 HG Representation Learning with Self-Supervised Augmentation

With the refined graph $\mathcal{G}^l$ obtained in the previous step, we implement R-GCN to learn node embedding for drug trafficker detection. In particular, we stack R-GCN layers to generate node embedding (Equation 1). For each user node $v_i \in \mathcal{V}$, we feed $h_i$ into a fully-connected layer and *softmax* function to obtain predicted illegal value of drug trafficker, i.e., $\hat{Y}_i = h_i W_c$. The objective is to jointly optimize the drug trafficker detection loss (i.e., the cross-entropy loss between the given drug trafficker labels $Y$ and the predicted results $\hat{Y}$) and the GSL loss $\mathcal{L}_{gsl}$:

$$\mathcal{L} = \mathcal{L}_c + \lambda_{gsl}\mathcal{L}_{gsl} = -\sum_{i \in \mathcal{V}^c} Y_i \log(\hat{Y}_i) + \lambda_{gsl} \left\| A^l \right\| \tag{6}$$

where $\mathcal{V}^c$ is the set of labeled user nodes in HG and $Y$ is the label set, $\lambda_{gsl}$ is a trade-off weight for balancing the two objectives.

**Self-Supervised Augmentation.** Existing graph neural networks [19, 13, 28] utilize the local neighborhood information and attribute information to generate the node embedding. However, the representation learning is dominated mainly by local neighborhood information. We wish to achieve a balance among the importance of local neighborhood and attribute of nodes [17]. Hence, we introduce a node attributed based self-supervised module (AS-SSL) using unlabeled data to enhance the HG representation learning (see Figure 2.(c)). Specifically, given a node $v_i$, we define two sets of nodes $\mathcal{V}_i^h$ and $\mathcal{V}_i^l$ having the $M$ highest similarities and $M$ lowest similarities with $v_i$ respectively. The similarity score is calculated as $cosine$ similarity between two node features. Then we introduce a classification task based on computed similarity. Specifically, we concatenate the attribute feature for each node pair $(v_i,\ v_q)$ $(i \neq q)$ as $X_{iq}$ and assign each pair with label $Y_{iq}$. $Y_{iq} = 1$ when $v_q \in \mathcal{V}_i^h$ otherwise $Y_{iq} = 0$ when $v_q \in \mathcal{V}_i^l$. Afterwards, we feed the node pair features $X_{iq}$ into a fully-connected neural network, i.e., $\hat{Y}_{iq} = X_{iq}W_{ssl}$, to obtain the predicted label. Finally, the final objective of node representation learning is to minimize the joint loss including drug trafficker detection loss, the AS-SSL loss, and the GSL loss:

$$\mathcal{L}_{joint} = \mathcal{L}_c + \lambda_{ssl}\mathcal{L}_{ssl} + \lambda_{gsl}\mathcal{L}_{gsl} = -\sum_{i\in\mathcal{V}^c} Y_i\ \log(\hat{Y}_i) - \lambda_{ssl}\sum_{i\in\mathcal{V}}\sum_{q\in\mathcal{V}_i^s} Y_{iq}\ \log(\hat{Y}_{iq}) + \lambda_{gsl}\left\|A^l\right\| \tag{7}$$

where $\mathcal{V}_i^s = \mathcal{V}_i^h \cup \mathcal{V}_i^l$, and $\lambda_{gsl}$ is a hyper-parameters of AS-SSL. By minimizing $\mathcal{L}_{joint}$, parameters of R-GCN, AS-SSL, and GSL are optimized jointly for downstream drug trafficker classification task.

### 4.4 Model Optimization with Meta-Learning

We propose to leverage meta-learning technique [11] for model optimization since the labeled data of drug traffickers is limited. In particular, it applies gradient-based algorithm that learns well initialized model parameters (using training tasks data) which can be quickly adapted to unseen new tasks. In this paper, given a set of tasks $\mathcal{T}$ defined as drug trafficker detection problem (binary classification) of different trafficker types, we choose three drug trafficker types with relatively large number of samples (e.g., stimulants trafficker) as training tasks and the other two types with few samples (e.g., opioid trafficker) as testing tasks. Then, we divide the data of each task $\tau \in \mathcal{T}$ into support set $\mathcal{S}_\tau$ and query set $\mathcal{Q}_\tau$. Afterwards, the classifier is first updated to task-specific model (three training tasks) with support set $\mathcal{S}_\tau$, and further optimized to task-agnostic model using $\mathcal{Q}_\tau$ of all tasks in training data, which is called meta-training. After sufficient training, the learned model can further adapt to new tasks (i.e., two testing tasks) with few data samples in support set, which is called meta-testing. Let $\phi$ be the set of model parameters. For a certain classification task $\tau$, we begin with feeding $\mathcal{S}_\tau$ to the model and calculate the loss $\mathcal{L}_\tau$ on $\mathcal{Q}_\tau$ to update $\phi$ to $\phi_\tau^{'}$ through gradient descent:

$$\phi_\tau' = \phi - \alpha \bigtriangledown_\phi \mathcal{L}_\tau(\phi) \tag{8}$$

where $\alpha$ is the step size for inner-level meta-training. Later, the model (with parameter $\phi$) is further updated in task-agnostic manner:

$$\phi \leftarrow \phi - \beta \bigtriangledown_\phi \sum_{\tau\in\mathcal{T}} \mathcal{L}_\tau(\phi_\tau') \tag{9}$$

where $\beta$ is the learning rate. After sufficient training over training tasks, the well initialized parameters is further adapted for prediction over testing tasks.

**Meta Knowledge Distillation for Model Adaptation.** The above optimization strategy directly adapts the classifier learnt from meta-training to new testing tasks, while ignoring the unlabeled data. Inspired by the knowledge distillation technique [14], we develop a meta knowledge distillation module (KD) that utilizes the unlabeled data in meta-learning to transfer soft knowledge from the meta-training classifier (teacher model) to the meta-testing model (student model) with the same structure (Figure 2.(e)). Mimicking teacher's prediction results enables student model to learn the secondary information that cannot be expressed by the labeled samples in meta-testing data alone. Soft knowledge from teacher model is formulated as the predicted probability of each drug trafficker in test data:

$$P_l^T(Z_i, t) = \text{Softmax}(f(Z_i), t)) = \frac{\exp\left[f_l(Z_i/t)\right]}{\sum_{c=0}^{1}\exp\left[f_c(Z_i/t)\right]} \tag{10}$$

where $Z_i$ is the embedding of node $v_i$ in $\mathcal{Q}_\tau$ for test task $\tau$, $f(Z_i)$ is the score logit that $Z_i$ achieves on label $l$, and $t$ is the temperature index to soften the peaky softmax distribution [14]. Thus, the

knowledge distillation loss of teacher model and student model is defined as follows:

$$\mathcal{L}_{kd} = -t^2 \sum_{i \in \mathcal{Q}_\tau} \sum_{c=0}^{1} P_c^T(Z_i, t) \log(P_c^S(Z_i, t)) \tag{11}$$

where $P_c^T$ and $P_c^S$ are the predicted distributions of teacher model and student model respectively. In addition, the labeled examples is still significant. Thus, we combine the KD loss $\mathcal{L}_{kd}$ over query data (unlabeled samples) and the cross-entropy loss $\mathcal{L}_{ce}$ over support data (labeled samples):

$$\mathcal{L}_{total} = \mathcal{L}_{ce} + \lambda_{kd} \mathcal{L}_{kd} \tag{12}$$

where $\lambda_{kd}$ is the trade-off weight for balancing the two losses. Taking the model (with parameter $\phi$) obtained by Equation 9 as the initial model, the student model for testing tasks are further updated (fine-tuned) as follows:

$$\phi^* = \phi - \alpha \bigtriangledown_\phi \mathcal{L}_{total}(\phi) \tag{13}$$

where $\alpha$ is the learning rate. Finally, the fine-tuned student model is used to detect drug trafficker in testing data. The detailed model optimization process is provided in the Supplementary Material 2.

# 5 Experiments

In this section, we first create a novel dataset for the problem. Then we conduct extensive experiments to evaluate the performance of our model. Further analysis are provided to show effectiveness of each model component, hyper-parameter sensitivity, and model stability.

## 5.1 Novel Dataset for Drug Trafficker Detection

Existing works on drug trafficker detection on social media merely rely on text or image data and most of the datasets are not publicly accessible. Thus, we create a new dataset crawled from Instagram (the most popular social media platform worldwide [16]) for the problem. Specifically, at first, we utilize a set of illegal drug related keywords (e.g., LSD, cannabis, and fentanyl) to crawl the public user profiles and the corresponding posts on Instagram using the public official Instagram API [1] from Jan 2020 to Jan 2021. Note that all the collected data are public and have been fully anonymized using hash values of user names. To obtain the groundtruth labels, we (including 3 groups, 2 annotators in each group) spent two months to collect and manually label these users into six groups (regular user, stimulants trafficker, depressants trafficker, hallucinogens trafficker, opioids trafficker and mixture trafficker) according to the criterion in [2]. In total, we obtain 8,651 users (including 3,242 drug traffickers and 5409 regular users) and 79,705 posts. Based on this data, we construct a HG which has 129,894 nodes (including 8,651 user nodes, 79,705 post nodes, and 41,538 keyword nodes) and 218,039 edges of eight relations (Table 1). The details of this dataset is provided in the Supplementary Material 1.1.

## 5.2 Baseline Methods

We compare our model with twenty baseline methods spanning several categories:

- **Traditional Classification Methods**. We take text features (tFeature), image features (iFeature), and the combination of them (cFeature) as the feature vector for each user respectively and feed them to a 3-layer deep neural network [37] (**DNN**) (B1). We also consider two existing works [22, 24] using tFeature for drug traffickers detection (B2).
- **Meta-Learning Models**. For meta-learning baseline methods, besides **MAML** [11], we also employ two popular few-shot learning models, i.e., Matching Network (**MatchingNet**) [40] and Prototypical Network (**ProtoNet**) [36] to train a meta-learner with cFeature (B3).
- **Graph Representation Learning Models**. We take cFeature as node attribute information and apply six popular graph representation learning models to learn user embeddings in HG, i.e., **Deepwalk** [31], **metapath2vec** [8], **GCN** [19], **GAT** [39], **HAN** [42], and **R-GCN** [33]. The learned user embeddings are further fed to a 3-layer DNN (B4) for classification. In addition, we also apply MAML to train these graph representation learning models (B5).

The details of all models are discussed in the Supplementary Material 3.1.

Table 2: Performance comparisons of all methods with different support data sizes (shot numbers).

| Setting | | 1-shot | | 5-shot | | 10-shot | | 20-shot | |
|---|---|---|---|---|---|---|---|---|---|
| Group | Model | ACC | F1 | ACC | F1 | ACC | F1 | ACC | F1 |
| B1 | tFeature+DNN [37] | 0.4399 | 0.1507 | 0.4641 | 0.2157 | 0.4768 | 0.2646 | 0.4835 | 0.2865 |
| | iFeature+DNN | 0.4458 | 0.1672 | 0.4701 | 0.2269 | 0.4849 | 0.2781 | 0.4957 | 0.2963 |
| | cFeature+DNN | 0.4646 | 0.2078 | 0.4991 | 0.2670 | 0.5017 | 0.3127 | 0.5209 | 0.3315 |
| B2 | Li *et al.* [22] | 0.4579 | 01947 | 0.4818 | 0.2493 | 0.473 | 0.2919 | 0.5079 | 0.3195 |
| | Rokon *et al.* [24] | 0.4835 | 0.2253 | 0.5171 | 0.2814 | 0.5279 | 0.3405 | 0.5478 | 0.3497 |
| B3 | cFeature+ProtoNet [36] | 0.5832 | 0.5534 | 0.5918 | 0.5686 | 0.6128 | 0.5732 | 0.6278 | 0.5924 |
| | cFeature+MatchingNet [40] | 0.6015 | 0.5738 | 0.6153 | 0.5789 | 0.6242 | 0.5805 | 0.6308 | 0.6105 |
| | cFeature+MAML [11] | 0.6343 | 0.6031 | 0.6584 | 0.6045 | 0.6636 | 0.6277 | 0.6716 | 0.6386 |
| B4 | [31] Deepwalk+DNN | 0.5112 | 0.2952 | 0.5631 | 0.3025 | 0.5783 | 0.3314 | 0.5809 | 0.3661 |
| | [8] metapath2vec+DNN | 0.5425 | 0.3275 | 0.5978 | 0.3315 | 0.6017 | 0.3634 | 0.6118 | 0.3989 |
| | [19] GCN+DNN | 0.6045 | 0.3835 | 0.6538 | 0.3915 | 0.6718 | 0.4208 | 0.6753 | 0.4429 |
| | [39] GAT+DNN | 0.6118 | 0.3808 | 0.6609 | 0.4042 | 0.6785 | 0.4314 | 0.6874 | 0.4561 |
| | [42] HAN+DNN | 0.6345 | 0.4058 | 0.6874 | 0.4231 | 0.6971 | 0.4551 | 0.7025 | 0.4717 |
| | [33] R-GCN+DNN | 0.6638 | 0.4368 | 0.7157 | 0.4581 | 0.7204 | 0.4853 | 0.7348 | 0.5051 |
| B5 | Deepwalk+MAML | 0.6523 | 0.6201 | 0.6705 | 0.6253 | 0.6878 | 0.6485 | 0.7028 | 0.6658 |
| | metapath2vec+MAML | 0.6819 | 0.6502 | 0.7078 | 0.6538 | 0.7152 | 0.6725 | 0.7358 | 0.6914 |
| | GCN+MAML | 0.7235 | 0.7157 | 0.7658 | 0.7487 | 0.7762 | 0.7605 | 0.7902 | 0.7768 |
| | GAT+MAML | 0.7348 | 0.7221 | 0.7705 | 0.7548 | 0.7817 | 0.7721 | 0.8019 | 0.7986 |
| | HAN+MAML | 0.7552 | 0.7481 | 0.7914 | 0.7745 | 0.8048 | 0.7979 | 0.8371 | 0.8246 |
| | R-GCN+MAML | 0.7831 | 0.7742 | 0.8216 | 0.8029 | 0.8335 | 0.8203 | 0.8658 | 0.8552 |
| Ours | **MetaHG** | **0.8489** | **0.8480** | **0.8855** | **0.8438** | **0.9196** | **0.9122** | **0.9354** | **0.9311** |

### 5.3 Performance Comparison

Table 2 reports the performances of all models. The best performances are highlighted in bold. All baseline models are divided into five groups (B1-B5) and the shot number denotes the support data size in model training and testing. According to the table, we can conclude that (i) The combination of text and image features (cFeature) contributes larger than single text or image features for drug traffickers detection (B1). Because both text and image are very informative to describe users on social media. (ii) By comparing results in B1 and B4, we find that considering the relationships among entities can help improve the model performance. Also, R-GCN has the best performance to learn the node embeddings for the problem. The reason may be that R-GCN considers multi-typed relationships when aggregating neighborhood information. (iii) The performance of meta-learning models with combined features (B3) or node embeddings (B5) are much better than traditional classification models (B1 and B2), showing the effectiveness of meta-learning for solving this problem. (iv) In all cases, MetaHG significantly outperforms all baseline methods, demonstrating the strongest capability of our model for drug traffickers detection on social media.

### 5.4 Ablation Studies

Our model MetaHG integrates three crucial components, i.e., graph structure learning (GSL), self-supervised learning in R-GCN (AS-SSL), meta-learning (MAML), and knowledge distillation (KD). To verify the effectiveness of each component, we conduct ablation studies by removing each of them independently. The results of different model variants are reported in see Table 3. At first, we replace MAML by a 3-layer DNN (A1) and remove KD from MAML (A2) respectively. As both performances drop dramatically, we can conclude that both MAML and KD have great contribution to MetaHG. Then, we remove GSL (A3) and AS-SSL (A4) from MetaHG respectively and we find that the performances of A3 and A4 drop almost 4% and 3% respectively by comparing with MetaHG, which demonstrates that both GSL and SSL are effective to enhance our model for solving the problem.

| Model | 1-shot | | 5-shot | | 20-shot | |
|---|---|---|---|---|---|---|
| | ACC | F1 | ACC | F1 | ACC | F1 |
| **MetaHG** | **0.8489** | **0.8480** | **0.8855** | **0.8438** | **0.9354** | **0.9311** |
| – MAML (A1) | 0.6948 | 0.4778 | 0.7428 | 0.4872 | 0.7523 | 0.5374 |
| – KD (A2) | 0.7914 | 0.7922 | 0.8370 | 0.7968 | 0.8804 | 0.8689 |
| – GSL (A3) | 0.8078 | 0.8018 | 0.8423 | 0.8069 | 0.8925 | 0.8771 |
| – AS-SSL (A4) | 0.8147 | 0.8052 | 0.8554 | 0.8125 | 0.9056 | 0.9005 |

Table 3: Results of different model variants.

### 5.5 Embedding Visualization

To examine the effectiveness of our model intuitively, we visualize embeddings of drug traffickers generated by cFeature + DNN, R-GCN + DNN, R-GCN + MAML, and MetaHG respectively in Figure 3. The blue points represent the embeddings for opioids traffickers while gray points show the embeddings for regular users. We can find that MetaHG generates the most distinct boundaries and the smallest overlapping area between opioids traffickers and regular users, demonstrating the superiority of our model.

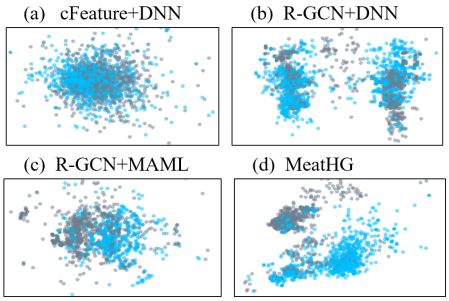


Figure 3: Embedding visualizations.

### 5.6 Hyper-parameter Sensitivity and Model Stability

To explore the hyper-parameter sensitivity, we conduct three analysis experiments w.r.t. $\epsilon$ on GSL, $\lambda$ on SSL, and $\lambda_{kd}$ on meta-learning. Specifically, in Figure 4(a), we vary $\epsilon$ in Eq. 3 to control the density of the generated graph in GSL. We find that the model performance increases with the increment of $\epsilon$ and the optimal value is 0.95, while the performance decreases when $\epsilon$ goes beyond the optimal value because the graph will be too sparse with large similarity threshold. Additionally, in Figure 4(b), we vary the value of $\lambda_{ssl}$ in Eq. 7 of SSL. We can find that SSL enhances the model by comparing results of $\lambda_{ssl} = 0$ (without SSL) and $\lambda_{ssl} = 5$, while the performance drops with larger $\lambda_{ssl}$ due to the overfitting on the SSL task. Moreover, in Figure 4(c), we vary the value value of $\lambda_{kd}$ in Eq. 12 of KD and the optimal value $\lambda_{kd}$ is 0.01 for our model, indicating that it is necessary to incorporate appropriate distilled knowledge. We also analyze the stability of our model by by comparing with three baseline methods on different testing tasks (T1: opioids trafficker detection, T2: hallucinogens trafficker detection), as shown in Fig. 4 (d). MetaHG is obviously better than baseline methods, showing the stability of MetaHG.

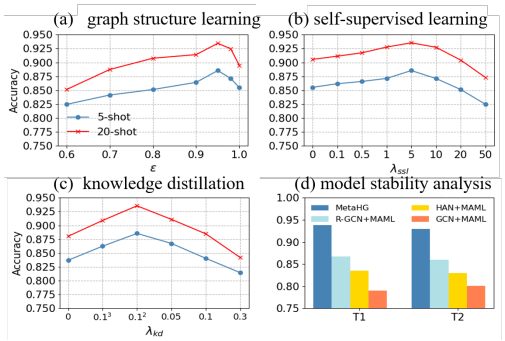


Figure 4: Hyper-parameter sensitivity (a)-(c) and model performance over different test tasks (d).

### 5.7 Case Study

We further analyze the drug traffickers detected by MetaHG. Figure 5 illustrates the superiority of our model by comparing with baseline models for real cases. MetaHG is able to detect hidden drug traffickers on social media although they pretend to be regular user. In this figure, drug trafficker "se***11" pretends to be regular user who does not have any illicit information on his homepage. However, he advertise at least four drugs (e.g., LSD) and leaves encrypted chat tool contact information (e.g., WICKR) to others' posts. These hidden drug traffickers cannot be detected by existing models, while we detect 421 hidden drug traffickers with MetaHG.

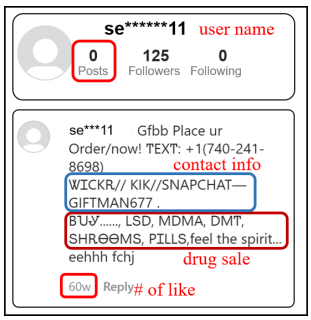


Figure 5: A real case of hidden drug trafficker.

## 6 Conclusion

In this paper, to solve the problem of illicit drug trafficker detection with limited labeled data constraint, we create a new dataset and develop a novel system called MetaHG. Specifically, we first construct a HG to describe the relation and content information in the dataset, and employ graph structure learning to learn a refined HG. Then we employ R-GCN with self-supervised module to learn the node embedding. Afterwards, we design a meta-learning algorithm to optimize the model. A knowledge distillation module is further introduced to improve model optimization. The extensive results demonstrate the effectiveness of our model by comparison with numerous baseline models.

## References


[1] Official instagram graph api, 2020. https://developers.facebook.com/docs/instagram-api.

[2] Brian Wu American Addiction Centers Editorial Staff, Kindra Sclar. Drug classifications. https://luxury.rehabs.com/drug-abuse/classifications/, 2019.

[3] Antreas Antoniou, Harrison Edwards, and Amos Storkey. How to train your maml. *arXiv preprint arXiv:1810.09502*, 2018.

[4] Roderic Broadhurst, Matthew Ball, and Harshit Trivedi. Fentanyl availability on darknet markets. *Broadhurst R, Ball M & Trivedi H*, 2020.

[5] Julian Broséus, Damien Rhumorbarbe, Caroline Mireault, Vincent Ouellette, Frank Crispino, and David Décary-Hétu. Studying illicit drug trafficking on darknet markets: structure and organisation from a canadian perspective. *Forensic Science International*, 2016.

[6] Jiaxin Chen, Li-Ming Zhan, Xiao-Ming Wu, and Fu-lai Chung. Variational metric scaling for metric-based meta-learning. In *AAAI*, 2020.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[8] Yuxiao Dong, Nitesh V Chawla, and Ananthram Swami. metapath2vec: Scalable representation learning for heterogeneous networks. In *KDD*, 2017.

[9] Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. On the convergence theory of gradient-based model-agnostic meta-learning algorithms. In *AISTATS*, 2020.

[10] Yujie Fan, Yanfang Ye, Qian Peng, Jianfei Zhang, Yiming Zhang, Xusheng Xiao, Chuan Shi, Qi Xiong, Fudong Shao, and Liang Zhao. Metagraph aggregated heterogeneous graph neural network for illicit traded product identification in underground market. In *ICDM*, 2020.

[11] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*, 2017.

[12] Chelsea Finn, Kelvin Xu, and Sergey Levine. Probabilistic model-agnostic meta-learning. In *NeurIPS*, 2018.

[13] William L Hamilton, Rex Ying, and Jure Leskovec. Inductive representation learning on large graphs. *arXiv preprint arXiv:1706.02216*, 2017.

[14] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[15] Instagram. How we enforce against illicit drug sales. https://about.fb.com/news/2018/09/enforcing-against-drug-sales/, 2020.

[16] Instagram. Instagram: distribution of global audiences 2021, by age group. https://www.statista.com/statistics/325587/instagram-global-age-group/, 2021.

[17] Wei Jin, Tyler Derr, Haochen Liu, Yiqi Wang, Suhang Wang, Zitao Liu, and Jiliang Tang. Self-supervised learning on graphs: Deep insights and new direction. *arXiv preprint arXiv:2006.10141*, 2020.

[18] Wei Jin, Yao Ma, Xiaorui Liu, Xianfeng Tang, Suhang Wang, and Jiliang Tang. Graph structure learning for robust graph neural networks. In *KDD*, 2020.

[19] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *ICLR*, 2017.

[20] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. Siamese neural networks for one-shot image recognition. In *ICML*, 2015.

[21] Yoonho Lee and Seungjin Choi. Gradient-based meta-learning with learned layerwise metric and subspace. In *ICML*, 2018.

[22] Jiawei Li, Qing Xu, Neal Shah, and Tim K Mackey. A machine learning approach for the detection and characterization of illicit drug dealers on instagram: model evaluation study. *Journal of medical Internet research*, 21(6):e13803, 2019.

[23] Tim Mackey, Janani Kalyanam, Josh Klugman, Ella Kuzmenko, and Rashmi Gupta. Solution to detect, classify, and report illicit online marketing and sales of controlled substances via twitter: using machine learning and web forensics to combat digital opioid access. *Journal of Medical Internet Research*, 2018.

[24] Tim K Mackey and Janani Kalyanam. Detection of illicit online sales of fentanyls via twitter. *F1000Research*, 2017.

[25] Tim K Mackey, Janani Kalyanam, Takeo Katsuki, and Gert Lanckriet. Twitter-based detection of illegal online sale of prescription opioid. *American Journal of Public Health*, 2017.

[26] Qi Mao, Li Wang, Steve Goodison, and Yijun Sun. Dimensionality reduction via graph structure learning. In *KDD*, 2015.

[27] ABC NEWS. Social media has provided a new marketplace for drugs and police are struggling to keep up, 2020. https://www.abc.net.au/news/2020-08-24/rise-and-challenge-of-social-media-drug-dealing/12545320.

[28] Tu Dinh Nguyen, Dinh Phung, et al. Unsupervised universal self-attention network for graph classification. 2019.

[29] Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018.

[30] Boris N Oreshkin, Pau Rodriguez, and Alexandre Lacoste. Tadam: task dependent adaptive metric for improved few-shot learning. In *NeurIPS*, 2018.

[31] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *KDD*, 2014.

[32] Sam Roweis. Em algorithms for pca and spca. *NeurIPS*, 1998.

[33] Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks. In *ESWC*, 2018.

[34] Chao Shang, Yun Tang, Jing Huang, Jinbo Bi, Xiaodong He, and Bowen Zhou. End-to-end structure-aware convolutional networks for knowledge base completion. In *AAAI*, 2019.

[35] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[36] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *NeurIPS*, 2017.

[37] Donald F Specht et al. A general regression neural network. *IEEE Transactions on Neural Networks*, 1991.

[38] Pinzhuo Tian, Zhangkai Wu, Lei Qi, Lei Wang, Yinghuan Shi, and Yang Gao. Differentiable meta-learning model for few-shot semantic segmentation. In *AAAI*, volume 34, pages 12087–12094, 2020.

[39] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

[40] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *NeurIPS*, 2016.

[41] Nick Vinzant. A growing crisis: Drug overdose deaths in america, 2021. https://www.aarp.org/money/scams-fraud/info-2020/coronavirus/.

[42] Xiao Wang, Houye Ji, Chuan Shi, Bai Wang, Yanfang Ye, Peng Cui, and Philip S Yu. Heterogeneous graph attention network. In *WWW*, 2019.

[43] Chengfeng Xu, Pengpeng Zhao, Yanchi Liu, Victor S Sheng, Jiajie Xu, Fuzhen Zhuang, Junhua Fang, and Xiaofang Zhou. Graph contextualized self-attention network for session-based recommendation. In *IJCAI*, 2019.

[44] Chuxu Zhang, Dongjin Song, Chao Huang, Ananthram Swami, and Nitesh V Chawla. Heterogeneous graph neural network. In *KDD*, 2019.

[45] Yiming Zhang, Yujie Fan, Wei Song, Shifu Hou, Yanfang Ye, Xin Li, Liang Zhao, Chuan Shi, Jiabin Wang, and Qi Xiong. Your style your identity: Leveraging writing and photography styles for drug trafficker identification in darknet markets over attributed heterogeneous information network. In *WWW*, 2019.

[46] Yiming Zhang, Yiyue Qian, Yujie Fan, Yanfang Ye, Xin Li, Qi Xiong, and Fudong Shao. dstyle-gan: Generative adversarial network based on writing and photography styles for drug identification in darknet markets. In *ACSAC*, 2020.

[47] Panpan Zheng, Shuhan Yuan, Xintao Wu, and Yubao Wu. Identifying hidden buyers in darknet markets via dirichlet hawkes process. *arXiv preprint arXiv:1911.04620*, 2019.

# Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See the Supplementary Material 4.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See the Supplementary Material 4.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] See the Supplementary Material 4.
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] The code and instructions will be available upon publication. Please see the Supplementary Material 4 about code and data availability.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 5.6 and the Supplementary Material 3.2.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We run 50 times for each experiment by changing the values of random seeds and average the results. Due to space limit, we report results with error bars in Supplementary Material 3.3.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See the Supplementary Material 3.2.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes] See the Supplementary Materials 4.

(c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We create a new dataset and introduce the dataset in Section 5.1. See the Supplementary Materials 1.1 and 4 for more details.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] See the Supplementary Material 4.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] See Section 5.1 and the Supplementary Material 4.

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]