Interesting Culture: Social Relation Recognition from Videos via Culture De-confounding

Yuxuan Zhang[1], Yangfu Zhu[2], Haorui Wang[1], Bin Wu[1]*
[1]Beijing University of Posts and Telecommunications,
[2]Capital Normal University
{yxzhang, wang_harry_cn, wubin}@bupt.edu.cn, zhuyangfu@cnu.edu.cn

Abstract

Social relationship recognition, as one of the fundamental tasks in video understanding, contributes to the construction and application of multi-modal knowledge graph. Previous works have mainly focused on two aspects: generating character graphs and multi-modal fusion. However, they often overlook the impact of cultural differences on relationship recognition. Specifically, relationship recognition models are susceptible to being misled by training data from a specific cultural context. This can result in the learning of culture-specific spurious correlations, ultimately restricting the ability to recognize social relationships in different cultures. Therefore, we employ a customized causal graph to analyze the confounding effects of culture in the relationship recognition task. We propose a Cultural Causal Intervention (CCI) model that mitigates the influence of culture as a confounding factor in the visual and textual modalities. Importantly, we also construct a novel video social relation recognition (CVSR) dataset to facilitate discussion and research on cultural factors in video tasks. Extensive experiments conducted on several datasets demonstrate that the proposed model surpasses state-of-the-art methods.

1 Introduction

Social relationship recognition task initially focused on extracting relationship categories between characters in text and still images (Tang et al., 2025; Yu et al., 2024). However, due to the widespread use of social media, there has been a significant growth trend in video data. The temporal sequence information and multi-modal semantic information present in videos make relation recognition more challenging, but provide more reliable evidence.

In previous research works, researchers focused on multi-modal fusion and character graphs construction (Hu et al., 2023; Qin et al., 2024). But they have overlooked the impact of cultural differences (Zheng et al., 2023) on model performance.

Figure 1: Cultural differences can affect model predictions. In Chinese culture (green area), intimate behavior between the opposite sex often signals a couple relationship, while in American culture (blue area), it might indicate friendship or another type of relationship.

Culture is a set of attitudes, values, beliefs, and behaviors shared by a group of people and communicated from one generation to the next (Matsumoto and Juang, 1996). Cultural bias is universal, which means that subjective opinions from a particular culture may offend or misinterpret other cultures. In the task of relationship recognition, due to the influence of different cultures, individuals with the same social relationships may engage in significantly different interactions and language communication, which can even lead to confusion in relationship discrimination.

We extracted the cultural background of the ViSR dataset (Liu et al., 2019), operationalizing it into country or region reflected in the video data (Sawaya et al., 2017). Nationality or region is not always coextensive with cultural background, but it was typically the only clue of cultural belonging in the video data (Peters and Carman, 2024). Simple statistical analysis revealed significant imbalance in the distribution of cultural categories within the ViSR dataset. Figure 1 illustrates how cultural bias affect predictions.

More intrigued, we conducted an experiment us-

*Corresponding Author.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 5174–5184
November 4-9, 2025 ©2025 Association for Computational Linguistics

Group	Train	Test	mAP
Balance	A, C, I, J	A, C, I, J	**0.464**
Imbalance	A, C, I	J	0.296
Imbalance	A, I, J	C	0.342

Table 1: Baseline model results controlling for cultural attributes. American Culture (A), Chinese Culture (C), Indian Culture (I), Japanese Culture (J).

ing baseline model (Liu et al., 2019) to provide more objective illustration of how cultural bias impact the predictive performance of the model. To highlight the impact of cultural bias further, we constructed a new dataset with four distinct cultures and designed three different control groups for experimental research. As shown in Table 1, the balanced group achieved the highest mAP value. This indicates that even though the video data does not explicitly include any cultural information, the trained model indeed learned false correlations between culturally influenced features and labels.

Based on the observations mentioned above, we attempted to improve the relation recognition model by applying causal intervention (Pearl, 2009a) rather than aiming to beat them. However, the implementation of causal intervention in video relation recognition task is challenging, such as the definition and extraction of culturally biased features, and lack of cultural annotations in existing datasets. Therefore, we propose a Cultural Causal Intervention model (CCI) to mitigate the impact of cultural differences by incorporating visual and textual modalities. To facilitate the study of cultural factors in video tasks, we construct a new dataset for video social relation recognition. And conduct cultural annotations on multiple datasets based on the consensus. We evaluated the effectiveness of the CCI framework on the several datasets. Extensive experiments demonstrate that CCI can effectively mitigate the impact of cultural bias on model performance and achieve SOTA performance. We summarize our contributions as follows:

- We are the first to investigate the issue of cultural bias in video social relationship recognition task using causal graphs.
- To validate our framework and facilitate research, we build a high-quality Video Social Relation dataset, named CVSR.
- Based on the causal theory of backdoor adjustment, we propose the CCI framework to mitigate impact of cultural bias on model performance. Extensive experiments on datasets show the effectiveness of CCI framework.

2 Related Work

2.1 Social Relation Recognition for Videos

With social relation recognition achieving remarkable achievements on text and still images, researchers turned to social relation recognition for videos. Datasets play a vital role as the foundation for social relationship recognition task. SRIV (Lv et al., 2018) contributed the first dataset for video-based social relationship recognition. ViSR (Liu et al., 2019) dataset restricts the relationship labels to eight categories. Previous works proposed solutions for video social relationship recognition from various perspectives. Liu et al.(2019) proposed a multi-scale spatio-temporal reasoning model based on triple graphs to extract relationships. Wu et al.(2021) generated character graphs from multimodal perspective to recognize relationships. Wang et al.(2023) proposed novel relational graphs and focused on continuous reasoning in long videos. Qin et al.(2024) proposed Dynamic-Evolutionary Graph Attention Network to capture the evolutionary trajectory of relations. However, they overlooked the impact of cultural differences on the performance of relation recognition models.

2.2 Debiasing

The existing debiasing methods are mainly divided into two categories (General debiasing (Sun et al., 2023) and Specific debiasing (Yang et al., 2024a)). This paper will focus on the confounding effects of specific bias types (i.e., cultural bias) in social relationship recognition task. Exploring causal relationships between variables is an effective way to identify and explain biases. Existing methods are mainly divided into causal intervention and counterfactual reasoning (Da et al., 2024). Counterfactuals depict imagined outcomes produced by factual variables under different treatments. Intervention aims to change the original distribution of the independent variable to eliminate the detrimental effects of specific bias. Long et al. (2023) employed a dual sampling method to alleviate the confounding effects of identity bias in facial anti-spoofing task. Yang et al. (2024b) proposed a causal intervention module to alleviate the subject bias in multimodal intention recognition. Inspired by existing researches, we will make the first attempt to apply

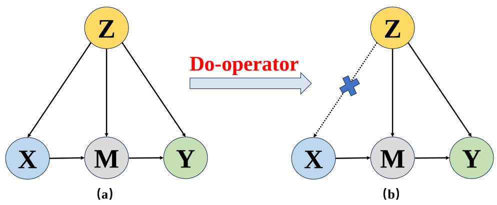

Figure 2: The proposed causal graph explains the causal effects in relationship recognition task. (a) The conventional likelihood $P(Y|X)$. (b) The causal intervention $P(Y|do(X))$.

causal intervention to explain and address the influence of cultural bias on relationship recognition. However, we face challenges such as culturally biased feature extraction, lack of cultural labels in existing datasets.

3 Method

3.1 Task Description

We pre-define the basic concepts and properties required for the target task to facilitate the construction and description of our framework. F, the set of frame in a video. $F = \{f_1, f_2, \cdots, f_n\}$, where f_i represents the i^{th} frame in the video. C, the set of characters in a video. $C = \{c_1, c_2, \cdots, c_k\}$, where c_i represents the i^{th} character in the video. R, the set of relationships between characters. $R = \{r_{1,2}, \cdots, r_{k-1,k}\}$, where $r_{i,j}$ represents the social relationship between c_i and c_j.

Task Definition: Given a video V, our goal is to predict relationship $\mathbf{r}_{i,j}$ between the pairs of characters $(\mathbf{c}_i, \mathbf{c}_j)$ that appear in the video, where the relationships $\mathbf{r}_{i,j}$ belong to our predefined social relationships set M.

3.2 Causal View at Relationship Recognition

To clearly depict the confounding effects in relationship recognition task, we utilize a customized causal graph to summarize the causal relationships between variables. In particular, we follow the same graphical notation in structured causal model (Pearl, 2009b). Specifically, the causal graph $G = \{N, E\}$ is a directed acyclic graph, where nodes N represent variables, and edges E represent direct causal effects. As shown in Figure 2, the causal graph involves four variables, namely the input video X, video modality features M, confounder Z, and prediction Y.

$\mathbf{Z} \rightarrow \mathbf{X}$: Due to the influence of cultural differences, which leads to prediction biases in relationship categories, culture is identified as a harmful confounding factor Z. For the input video X, Z represents the recorded culture-related bias, denoting $Z \rightarrow X$ in the causal graph.

$\mathbf{Z} \rightarrow \mathbf{M} \leftarrow \mathbf{X}$: The directed edge $X \rightarrow M$ indicates that the modality feature M is a part of the multi-modal input X. The directed edge $Z \rightarrow M$ signifies the detrimental Z confounding model, capturing the culturally related features embedded in M, thereby generating false semantic correlations.

$\mathbf{M} \rightarrow \mathbf{Y} \leftarrow \mathbf{Z}$: The directed edge $M \rightarrow Y$ represents the further influence of M, which is affected by Z, on the final prediction Y. The directed edge $Z \rightarrow Y$ indicates that the detrimental confounding factor Z implicitly interferes with the prediction Y in the training data.

According to the causal theory, the confounders Z are the common cause of X and corresponding predictions Y. The positive effect of the multi-modal semantic features for relationship recognition provided by M follows the expected causal pathway, which is what we aim for. Unfortunately, the confounding factor Z misguides the trained model to learn culture-related misleading semantic information instead of pure causal effects. This further leads to biased predictions towards unseen cultures. The harmful influence follows the backdoor paths $X \leftarrow Z \rightarrow Y$ and $X \leftarrow Z \rightarrow M \rightarrow Y$.

3.3 Causal Intervention

Ideally, a solution would involve collecting a large number of samples to ensure that data influenced by various cultural factors is included in both the training and testing sets. Due to practical constraints, this seems like an impractical task to accomplish. To address this, we employ causal interventions using $P(Y|do(X))$. By employing backdoor adjustment techniques, we can intervene on X to break the backdoor paths between X and Y, thereby alleviating the adverse influence of Z on the prediction Y. The $do(\cdot)$ operator is an efficient approximation to implement the empirical intervention. In Figure 2(a), existing relationship recognition methods rely on the likelihood $P(Y|X)$. This process is formulated by Bayes rule:

$$P(Y|X) = \sum_z P(Y|X, M = F_m(X, z))P(z|X) \quad (1)$$

where $F_m(\cdot)$ denotes general model to learn the multi-modal representations M. z is a stratum of confounders (i.e., a culture), which introduces the observational bias via $P(z|X)$.

In the task of relationship recognition, backdoor adjustment involves calculating the causal effects at each layer of the cultural confounding factors. Then, based on the prior proportions of samples from different cultures in the training data, a weighted integration is performed to estimate the average causal effect. From Figure 2(b), the impact from Z to X is cut off since the model would enable the cultural prototype as the confounder in each stratum to contribute equally to the predictions Y by $P(Y|do(X))$. By applying the Bayes rule on the new graph, Equation(1) with the intervention is formulated as:

$$P(Y|do(X)) = \sum_z P(Y|X, M = F_m(X, z))P(z) \quad (2)$$

Since the confounding factor Z no longer influences X, the model is no longer poisoned by false correlations specific to certain cultures along the backdoor paths. $P(z)$ is the prior probability that depicts the proportion of each z in the whole.

3.4 Framework Overview

As shown in Figure 3, our framework mainly consists of four parts: construction of video-level features, cultural feature disentanglement, causal intervention, and construction of video-level relationship graphs. In our framework, we first build frame-level character graphs, and extract global visual features and video-level character graphs. Text features are extracted through the pre-trained model. To facilitate the extraction of bias features, we use cultural labels as supervised signals to encourage model to classify bias attributes with bias features. At the same time, we maintain two cache matrices to represent culturally relevant information. Through the backdoor adjustment technology, the visual and textual features are intervened to construct pure multimodal features. Finally, we use relations as nodes and introduce text features to construct video-level relation graph to realize relation classification.

3.5 Video-level Feature Construction

Construct global visual and text features: We construct a frame set $\mathbf{F}$ by sampling multiple frames uniformly from a video. The character features $\mathbf{T_{c_i}}$ in the frame are extracted by pre-training the model ViT. As shown in the blue area in Figure 3, since the relationship between characters can naturally be expressed by the graph structure, we construct the frame-level character graph $\mathbf{G_{f_i}}$ based on the pre-processed features. With characters as nodes and relations as edges, the deep feature interaction between characters is realized through GCN model. We formalize the definition as follows:

$$F = \{f_1,\ f_2, \cdots,\ f_n\} \quad (3)$$

$$G_{f_i} = \{T_{c_1},\ T_{c_2}, \cdots,\ T_{c_k}\} \quad (4)$$

$$GCN(G_{f_i}) = G'_{f_i} = \left\{T'_{c_1},\ \cdots,\ T'_{c_k}\right\} \quad (5)$$

where n represents the number of frames sampled and k represents the number of characters appearing. A frame-level character graph is a complete graph. In order to obtain the global visual features, we construct the frame feature $\mathbf{T_{f_i}}$ based on frame-level character graph and each frame feature is used as a token, which can be formulated as follows:

$$T_{f_i} = \Gamma(G'_{f_i}) \quad (6)$$

where $\mathbf{\Gamma}$ denotes global average pooling operation.

Inspired by the existing pre-trained models, we add a learnable special token (CLS token) before the token sequence to construct the complete token sequence. Input the token sequence into the transformer encoder to learn the global visual representation. We formalize the definition as follows:

$$T_F = \{T_{cls}, T_{f_1},\ \cdots,\ T_{f_n}\},\ V_g = \Phi(T_F) \quad (7)$$

where T_{cls} represents the learnable special token and T_F represents the complete token sequence. Φ represents the transformer encoder and V_g represents the global visual feature, which is the representation of the CLS token output by the encoder.

We use the pre-trained model BERT to process the caption information in the video as the global text feature TE_g of the video. We formalize the definition as follows:

$$TE_g = BERT(text) \quad (8)$$

where $\mathbf{text}$ represents the caption information.

Video-level character graph: Our goal is to predict relationships between characters in a video. In order to aggregate global information, we construct a video-level character graph based on the frame-level character graph. Specifically, we treat the feature information of the character $\mathbf{c}_i$ in multiple frames as a sequence. The character sequence is processed through the temporal model to obtain the global features of the characters, thereby constructing the video-level character graph $\mathbf{G}_v$.

$$T^g_{c_i} = \left\{T^{f1}_{c_i},\ T^{f2}_{c_i}, \cdots,\ T^{fn}_{c_i}\right\} \quad (9)$$

$$T^{g'}_{c_i} = \Psi(T^g_{c_i}), G_v = \left\{T^{g'}_{c1},\ \cdots,\ T^{g'}_{c_k}\right\} \quad (10)$$

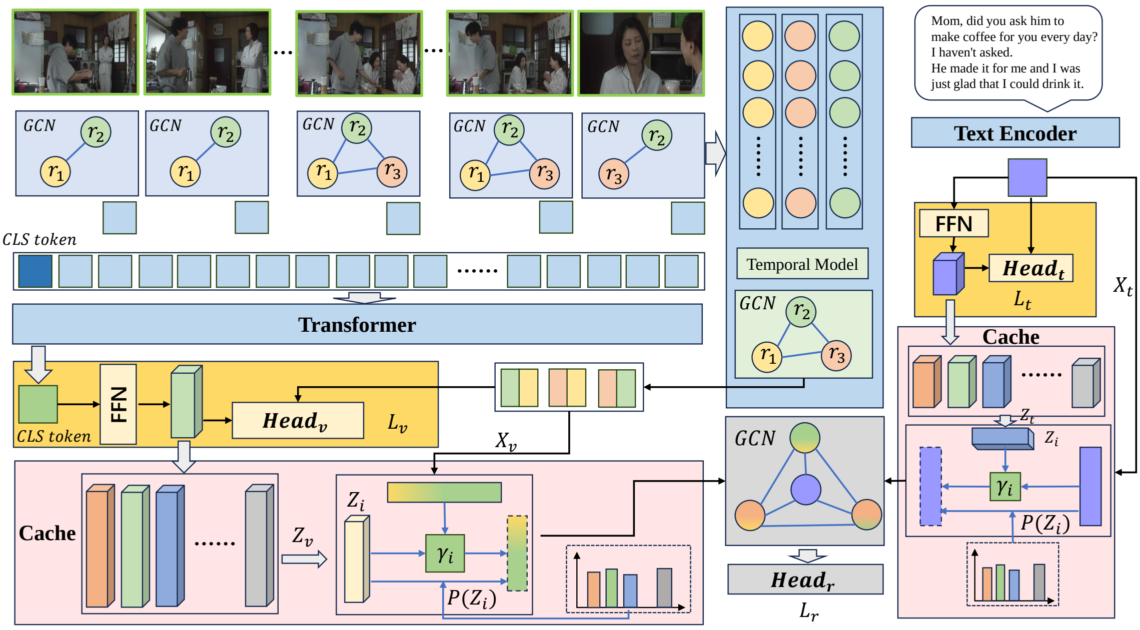

Figure 3: An overview of our proposed CCI framework. The blue region represents video-level feature construction component. The yellow region represents the feature disentanglement component. The red region represents the causal intervention component. The gray region represents relationship graph component.

where Ψ represents the temporal model. We process video-level character graphs through GCN to achieve high-level character feature interaction.

$$GCN(G_v) = G_v^{'} = \left\{ T_{c_1}^{g''}, \cdots, T_{c_k}^{g''} \right\} \quad (11)$$

3.6 Disentangling Cultural Feature

To make causal interventions on the visual and textual features of characters, we need to disentangle the cultural bias features from original visual and textual features. We manually annotate the cultural labels of video data as a supervision signal to encourage the model to learn culturally biased features. Specifically, we feed video-level visual and textual features into the feedforward layer to extract culturally biased features $\mathbf{b}_v$ and $\mathbf{b}_t$, which facilitate the construction of confusion dictionary. Both biased features and original features are input into the classification head to achieve cultural classification. We formalize the definition as follows:

$$b_v = FFN(V_g), b_t = FFN(TE_g) \quad (12)$$

where b_v and b_t represent culturally biased features extracted from visual and textual features, FFN represents the feedforward layer.

$$\begin{aligned} \mathcal{L}_v = \mathcal{CE}(Head_v(b_v, T_{c_i}^{g''}, T_{c_j}^{g''}) \\ - Head_v(b_0, T_{c_i}^{g''}, T_{c_j}^{g''}), y_c) \end{aligned} \quad (13)$$

$$\begin{aligned} \mathcal{L}_t = \mathcal{CE}(Head_t(b_t, TE_g) \\ - Head_t(b_0, TE_g), y_c) \end{aligned} \quad (14)$$

where $\mathcal{CE}(\cdot)$ represents the cross-entropy loss function, $Head_v$ and $Head_t$ represent the classification heads for visual and text, respectively. b_0 denotes the zero vector and y_c denotes the cultural category.

3.7 Causal Intervention

Confounding Construction: The confounding construction aims to measure the causal effects of the cultural confounding factors between different strata during the training process, in order to avoid prediction biases related to culture. Intuitively, cultural features within the same stratum are similar, while cultural features differ across different strata. We consider all features associated with the same culture as feature prototypes, representing the general attributes of a specific culture. Specifically, during the training process, we will construct a confounding dictionary represented as follows:

$$Z = [z_1, z_2, \cdots, z_l], \quad (15)$$

where l represents the number of cultural categories, and $z_i = \frac{1}{N_i}\sum_{k=1}^{N_i} b_k^i$ represents a prototype of specific culture. N_i is the number of training samples for the i-th culture, and b_k^i denotes the k-th feature of the i-th culture. z_i is updated at the end of each epoch. So we will obtain the corresponding confusion matrices Z_v and Z_t based on b_v and b_t.

De-confounding Training: Since computing X involves costly forward computations for each pair of X and Z, we introduce the Normalized Weighted Geometric (2015) Mean as an intervention approximation at the feature level. This helps to mitigate the computational overhead:

$$P(Y|do(X)) \approx P(Y|X, M = \sum_{z} F_m(X, z)P(z)) \quad (16)$$

Causal intervention aims to enable fair predictions of Y by utilizing each z. We approximate Equation(16) through a parameterized neural network model as follows:

$$P(Y|do(X)) = W_m x + W_h E[h(z)] \quad (17)$$

where W_m and W_h are the learnable parameters. x represents a modality feature. This approximation attributes the impact on Y to the combination of M and Z in the causal graph and then adaptsively aggregates all confounding factors based on the backdoor adjustment theory:

$$E[h(z)] = \textstyle\sum_{i=1}^{l} \gamma_i z_i p(z)_i \quad (18)$$

$$\gamma_i = \Phi(\frac{(W_q x)^T (W_k z_i)}{\sqrt{d(\cdot)}}) \quad (19)$$

where W_q and W_k are mapping matrices. $P(z_i) = \frac{Ni}{N}$, where N is the number of training samples. $d(\cdot)$ represents the embedding dimension. Actually, x from one sample queries each z_i in the confounder dictionary Z to obtain the sample-specific attention set $\{\gamma_i\}_{i=1}^{l}$. In other words, samples from a particular culture will be impacted by confounding factors from other cultures to varying degrees.

Based on the above analysis, we will conduct causal intervention on the visual and text features in Equation(8) and (11). Since our goal is to predict the relationship between character pairs (c_i, c_j) and is inspired by Wang et al.(2023), we fuse the features of character pairs as relationship features for de-confounding training. We formalize the definition as follows:

$$\begin{aligned} T_{(c_i,c_j)} &= P(Y|do(T_{c_i}^{g''}, T_{c_j}^{g''})) \\ T_{text} &= P(Y|do(TE_g)) \end{aligned} \quad (20)$$

where $T_{(c_i,c_j)}$ represents the relation features and T_{text} represents the text features after intervention.

3.8 Video-level Relationship Graph

In this part, we construct a video-level relationship graph based on the features after the previous intervention. Compared with the traditional character graph, the relationship graph can better adapt to our tasks and promote the reasoning effect between relationships. Specifically, we take the relationship features of (c_i, c_j) as nodes, and text features are regarded as a special node, thereby constructing relationship graph G_r:

$$G_r = \left\{T_{text},\ T_{(c_1,c_2)},\ \cdots,\ T_{(c_{k-1},c_k)}\right\} \quad (21)$$

To facilitate inter-relation reasoning, we use GCN to process the relation graph G_r, thus enabling the interaction of features between relations.

$$GCN(G_r) = G_r^{'} = \left\{T_{(c_1,c_2)}^{'}, \cdots, T_{(c_{k-1},c_k)}^{'}\right\} \quad (22)$$

And the relation classification is realized by the classification head.

$$\mathcal{L}_r = \mathcal{CE}(Head_r(T_{(c_1,c_2)}^{'}), y_r) \quad (23)$$

where $\mathcal{CE}(\cdot)$ represents the cross-entropy loss function, $Head_r$ represent the relation classification head. y_r denotes the relation category.

3.9 Training Objective

We employ the cross-entropy loss for training the framework. The loss function can be formulated as follows:

$$\mathcal{L} = \mathcal{L}_r + \alpha(\mathcal{L}_v + \mathcal{L}_t) \quad (24)$$

where α is the hyper-parameter used to adjust the weight of the culture-biased classification loss.

4 Experiment

4.1 CVSR Dataset

The existing video datasets for social relation recognition mainly include ViSR(2018) and MovieGraphs(2019) , however, they lack the annotation of cultural labels. Meanwhile, they show a long-tail distribution in terms of cultural categories, which is not conducive to the discussion and research of cultural factors in the task of video social relationship recognition. Therefore, we have constructed a high-quality video-based social relationship dataset called CVSR.

The CVSR dataset defines eight types of social relationships based on domain-specific theories. The construction process involves three main steps: 1) Firstly, collecting various types of movies, including genres such as family, action, romance, comedy, etc., excluding surreal genres like science fiction. 2) Video clips are extracted from the movies by five trained annotators. Each clips is required to have a duration between 10-30 seconds and involve at least two interacting individuals. Ultimately, over 5,000 candidate video clips are obtained for annotation. 3) Each candidate video clip is annotated by five annotators individually. If a video clip receives more than two different relationship labels, it is discarded. Otherwise, the video

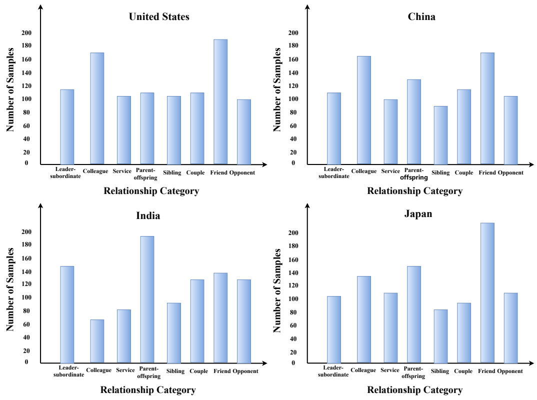

Figure 4: Statistics of the sample number of relationship categories under each culture.

clip is assigned the relationship label that has the majority of votes among the annotators. We annotate the cultural categories on the MovieGraphs and ViSR datasets, and the annotation process is basically consistent with the relationship category annotation process on the CVSR dataset.

As shown in Figure 4, we count the number of relation categories under different cultural categories in the CVSR dataset. To better understand each class of relation labels, we show examples of partial relation categories in Figure 5. The CVSR dataset has several advantages: 1) All video clips are kept within 10-30 seconds to maintain stable video scenes. 2) The dataset consists of more than 4,000 video clips, including four cultural backgrounds (American Culture, Chinese Culture, Indian Culture and Japanese Culture). Each cultural background contains more than 1,000 video clips. 3) The dataset shows strong generalization, being useful not only for relationship recognition task but also for tasks like video question answering.

The CVSR dataset is only available for academic research and not for commercial use. Researchers can use it according to the CC BY-NC protocol. At the same time, our dataset is derived from publicly accessible movie resources, which does not involve harmful content such as privacy information. The use of multiple languages is involved in the video data, including but not limited to English, Chinese, Japanese, etc. Our data is at the full risk of the authors and annotators will not be involved in any type of risk. Our annotators are drawn from undergraduate students in computer science and technology who will be compensated for course credit. At the same time we will protect the essential information of our annotators. We maintained

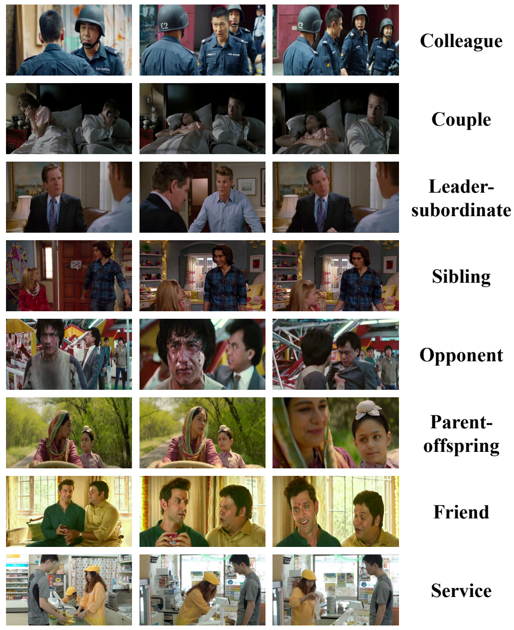

Figure 5: Some examples of videos in the CVSR dataset.

ownership of the data and explained the purpose of the data to the annotators. Our data comes from ethically censored movies, so our data does not involve any ethical risk.

4.2 Experiment Setting

We compare the CCI framework with the following baseline methods. GCN (2016) is a standard Graph Convolutional Network. PGCN (2019) applies multi-scale graph convolutions to Triple Graphs. MSTR (2019) combines PGCN and TSN to aggregate temporal and spatial information. TSN-ST (2018) follows the same framework as TSN-Spatial but introduces optical flow to fuse spatial and temporal features. HC-GCN (2021) constructs a hierarchical character graph. SGCAT-CT (2023) transforms character graph into relation graph and incorporates hierarchical accumulative memory to model temporal information. LiReC (2020) learns contextual features by jointly predicting interactions and relationships. MRR (2021) constructs multi-entity relation graphs to recognize and predict relationships. PMFL (2022) introduces self-supervised learning to learn multimodal features. DE-GAT (2024) proposes a Dynamic-Evolutionary Graph Attention Network to capture changes in relationships between characters.

To ensure a fair comparison, all baseline models are kept with their original settings. The text feature extractor utilizes the Bert model. The vi-

Dataset	Method	Top-1 Accuracy								
		Leader-sub	Colleague	Service	Parent-offs	Sibling	Couple	Friend	Opponent	mAP
MovieGraphs	GCN	0.295	0.365	0.132	0.325	0.280	0.167	0.391	0.158	0.264
	PGCN	0.313	0.374	0.290	0.137	0.320	0.250	0.407	0.375	0.308
	MSTR	0.409	0.392	0.342	0.407	0.434	0.326	0.373	0.357	0.380
	LiReC	0.352	0.329	0.244	0.435	0.301	0.423	0.317	0.269	0.334
	MRR	0.454	0.423	0.428	0.385	0.392	0.446	0.439	0.365	0.417
	PMFL	0.363	0.484	0.485	0.333	0.161	0.368	0.440	0.386	0.401
	SGCAT	0.425	0.226	0.317	0.523	0.333	0.429	0.318	0.284	0.357
	SGCAT-CT	0.387	0.573	0.415	0.382	**0.459**	0.406	0.509	0.570	0.463
	DE-GAT	0.503	0.500	0.682	**0.562**	0.261	0.521	**0.615**	0.606	0.510
	CCI	**0.672**	**0.631**	**0.768**	0.508	0.420	**0.684**	0.476	**0.884**	**0.579**
ViSR	GCN	0.562	0.495	0.271	0.368	0.416	0.344	0.398	0.500	0.435
	PGCN	0.541	0.549	0.257	0.408	0.348	0.333	0.453	0.483	0.447
	MSTR	**0.575**	0.511	0.300	0.456	0.393	0.387	0.532	0.474	0.478
	TSN-ST	0.411	0.333	0.300	0.328	0.458	0.292	0.638	0.329	0.432
	HC-GCN	0.493	0.542	0.356	0.496	0.405	0.365	0.623	0.408	0.487
	SGCAT-CT	0.420	**0.625**	**0.473**	**0.529**	0.514	0.489	0.542	0.417	0.501
	CCI	0.425	0.525	0.442	0.449	**0.574**	**0.503**	**0.633**	**0.600**	**0.560**
CVSR (Imbalance)	GCN	0.380	0.229	0.262	0.492	0.196	0.298	0.396	0.433	0.294
	PGCN	**0.483**	0.257	0.388	0.449	0.260	0.244	0.394	**0.488**	0.305
	MSTR	0.421	0.282	0.494	0.439	0.272	0.237	0.321	0.453	0.346
	SGCAT-CT	0.426	0.383	**0.529**	0.321	0.257	0.211	0.341	0.353	0.383
	CCI	0.433	**0.647**	0.377	**0.476**	**0.607**	**0.324**	**0.534**	0.449	**0.516**
CVSR (Balance)	GCN	0.434	0.554	0.424	0.381	0.429	0.357	0.480	0.532	0.464
	PGCN	0.393	0.471	0.424	0.345	0.420	0.286	0.465	0.524	0.453
	MSTR	0.483	0.306	0.305	0.470	0.446	0.294	0.520	0.411	0.413
	SGCAT-CT	**0.593**	0.408	**0.559**	**0.583**	0.339	0.317	0.510	**0.645**	0.486
	CCI	0.467	**0.577**	0.453	0.339	**0.555**	**0.457**	**0.640**	0.479	**0.526**

Table 2: Comparisons of top-1 accuracy on the MovieGraphs, ViSR and CVSR datasets.

Method	mAP			
	CVSR_J	CVSR_C	CVSR_I	CVSR_A
GCN	0.296	0.342	0.294	0.341
PGCN	0.303	0.354	0.305	0.363
MSTR	0.260	0.285	0.346	0.376
SGCAT-CT	0.370	0.394	0.383	0.404
CCI	**0.482**	**0.431**	**0.516**	**0.493**

Table 3: Comparisons of mAP on the CVSR dataset.

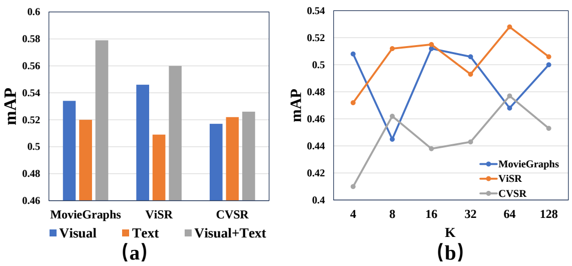

Figure 6: Ablation study of CCI

sual feature extractor employs pre-trained Faster R-CNN and ViT for person detection. The text feature dimension is 768, while the visual feature dimension is 2048. The optimizer used is Adam, with a batch size of 128. We implemented the selected methods and the CCI framework using the PyTorch toolkit on an NVIDIA A100 Tensor Core GPU. In the evaluation phase, given a test set sample, we rely on global visual features G_v' and text features TE_g to build a video-level relation graph for relation classification.

4.3 Experiment Results

Results on MovieGraphs and ViSR Dataset: As shown in Table 2, our method surpasses the SOTA performance by 6.9% and 5.9% in terms of average precision, respectively. Meanwhile, our method achieves the best performance in most categories. Notably, the CCI framework shows excellent performance on the couple and opponent relationships, which indicates that there are significant differences in intimacy behavior and conflict handling between characters in different cultures. At the same time, the construction process of the video-level relationship graph also makes the CCI framework have excellent representation ability.

Results on CVSR Dataset: The results are shown in Table 2. In the imbalanced partitioning, the data of three cultures are used for training, while the data from the remaining culture is used for testing. The test set includes only Indian culture. Our method surpasses the SOTA performance by 13.3% in terms of average precision and achieves the best performance in most categories. Indian culture has a strong religious color and pays attention to social status and traditional customs, leads to distinct features in various social relationships compared to other cultures. It is significant in colleague and sibling relationships. As a result, the baseline models show performance decrease, and our method can mitigate cultural biases to a certain extent. In the balanced partitioning, both the train and test sets include data from all four cultures. Our method surpasses the SOTA performance by 4% in terms of average precision.

To avoid randomness under imbalanced partitioning, we conduct more experiments (the test set includes only a specific culture). The results are shown in Table 3. Our method surpasses the SOTA performance by 11.2%, 3.7%, 13.3% and 8.9% in terms of average precision, respectively. Due to cultural differences, there is a decline in model performance. Specifically, American culture advocates freedom and independence; Chinese culture emphasizes tradition and collective consciousness; Japanese culture emphasizes teamwork and etiquette; Indian culture is deeply religious and hierarchical. Our method attempts to mitigate cultural bias and achieves SOTA performance.

4.4 Ablation Study

Importance of Modality De-confounding: As shown in Figure 6(a), when text de-confounding and visual de-confounding are individually removed, and the initial features are used, the model performance exhibits varying degrees of decline. This indicates that false dependencies between potential cultural features and labels exist simultaneously in the visual and text modalities.

Importance of Disentangling: We attempted an alternative approach by not using feature disentanglement component for supervision. Instead, we directly applied K-means++ clustering to the global visual and text features. To avoid the influence of the number of categories K, we conducted multiple sets of experiments. As shown in Figure 6(b), without the feature disentanglement component, the average precision of model fluctuates irregularly. This indicates that directly using global features cannot highlight cultural attributes, thus affecting the quality of confusion dictionary construction.

4.5 Case Study

As shown in Figure 7, we selected two typical examples to demonstrate the performance of the model before and after debiasing. 1) Example of the distinct dressing styles between superiors and subordinates in the Indian workplace, which can mislead the model to develop incorrect dependencies. Such dressing distinctions may not be emphasized under American culture. 2) In the United States, sibling terms are typically used within sibling relationships. In China, however, such terms can have a broader range of possibilities and may be used in different contexts.

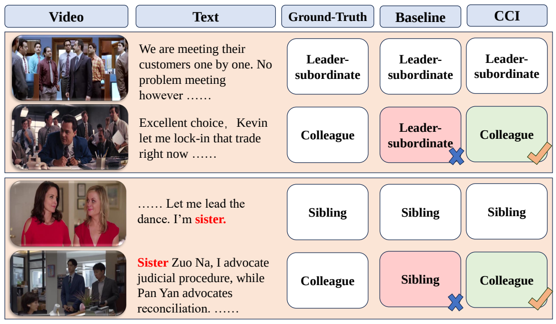

Figure 7: Difference between the model approximate $P(Y|X)$ and $P(Y|do(X))$

5 Conclusion

This paper is the first to present a analysis and identification of cultural bias in the video social relationship recognition task from a causal perspective. It proposes a Cultural Causal Intervention model (CCI) to mitigate the adverse effects of cultural bias as a confounding factor by utilizing causal graphs and backdoor adjustment techniques. Extensive experiments conducted on several datasets demonstrate that CCI effectively alleviates the impact of bias on model performance and enhances the accuracy of relationship classification. In the future, we will focus on the separation of causal and biased features under self-supervision and design more general frameworks for multimodal bias.

Acknowledgments

This work is supported by the National Natural Science Foundations of China under Grant (62372060, 61972047).

Limitations

Culture is an abstract and complex concept, and it is challenging to define the cultural properties of a video. In the paper, we refer to the existing literature to provide cultural annotations for the video data, but there is a lack of more precise and fine-grained cultural definitions. In the future, we will further expand the scale and cultural diversity of the CVSR dataset while annotating fine-grained cultural properties.

Research on a single bias will limit the development of the video social relationship recognition task. In the future, we aim to design more general frameworks to deal with multi-modal biases, while introducing Multi-modal Large Language Model for efficient and accurate relation recognition.

Ethical Considerations

In this paper, we discuss and study the influence of cultural factors on the task of video-based social relation recognition. We aim to analyze the expression differences of relations in different cultures, so as to provide a more accurate basis for relation recognition. However, we fully respect and understand the differences between cultures and are committed to being neutral and objective in the research process. We hope that this study can provide useful reference and inspiration for the attention and discussion of cultural factors in various tasks.

References

Yifei Da, Matías Nicolás Bossa, Abel Díaz Berenguer, and Hichem Sahli. 2024. Reducing bias in sentiment analysis models through causal mediation analysis and targeted counterfactual training. *IEEE Access*.

Yibo Hu, Chenyu Cao, Fangtao Li, Chenghao Yan, Jinsheng Qi, and Bin Wu. 2023. Overall-distinctive gcn for social relation recognition on videos. In *International Conference on Multimedia Modeling*, pages 57–68. Springer.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Anna Kukleva, Makarand Tapaswi, and Ivan Laptev. 2020. Learning interactions and relationships between movie characters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9849–9858.

Xinchen Liu, Wu Liu, Meng Zhang, Jingwen Chen, Lianli Gao, Chenggang Yan, and Tao Mei. 2019. Social relation recognition from videos via multi-scale spatial-temporal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3566–3574.

Xingming Long, Jie Zhang, Shuzhe Wu, Xin Jin, and Shiguang Shan. 2023. Dual sampling based causal intervention for face anti-spoofing with identity debiasing. *IEEE Transactions on Information Forensics and Security*.

Jinna Lv, Wu Liu, Lili Zhou, Bin Wu, and Huadong Ma. 2018. Multi-stream fusion model for social relation recognition from videos. In *MultiMedia Modeling: 24th International Conference, MMM 2018, Bangkok, Thailand, February 5-7, 2018, Proceedings, Part I 24*, pages 355–368. Springer.

David Matsumoto and Linda Juang. 1996. Culture and psychology. *Pacific Grove*, pages 266–270.

Judea Pearl. 2009a. Causal inference in statistics: An overview. *Statistics Surveys*, 3:96–146.

Judea Pearl. 2009b. *Causality: Models, Reasoning and Inference*, 2nd edition. Cambridge University Press, USA.

Uwe Peters and Mary Carman. 2024. Cultural bias in explainable ai research: A systematic analysis. *Journal of Artificial Intelligence Research*, 79:971–1000.

Penggang Qin, Shiwei Wu, Tong Xu, Yanbin Hao, Fuli Feng, Chen Zhu, and Enhong Chen. 2024. When i fall in love: Capturing video-oriented social relationship evolution via attentive gnn. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(6):5160–5175.

Yukiko Sawaya, Mahmood Sharif, Nicolas Christin, Ayumu Kubota, Akihiro Nakarai, and Akira Yamada. 2017. Self-confidence trumps knowledge: A cross-cultural study of security behavior. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 2202–2214.

Teng Sun, Juntong Ni, Wenjie Wang, Liqiang Jing, Yinwei Wei, and Liqiang Nie. 2023. General debiasing for multimodal sentiment analysis. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 5861–5869.

Wang Tang, Linbo Qing, Pingyu Wang, Lindong Li, and Yonghong Peng. 2025. Graph-based interactive knowledge distillation for social relation continual learning. *Neurocomputing*, 634:129860.

Yiyang Teng, Chenguang Song, and Bin Wu. 2022. Learning social relationship from videos via pre-trained multimodal transformer. *IEEE Signal Processing Letters*, 29:1377–1381.

Paul Vicol, Makarand Tapaswi, Lluis Castrejon, and Sanja Fidler. 2018. Moviegraphs: Towards understanding human-centric situations from videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8581–8590.

Haorui Wang, Yibo Hu, Yangfu Zhu, Jinsheng Qi, and Bin Wu. 2023. Shifted gcn-gat and cumulative-transformer based social relation recognition for long videos. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 67–76.

Shiwei Wu, Joya Chen, Tong Xu, Liyi Chen, Lingfei Wu, Yao Hu, and Enhong Chen. 2021. Linking the characters: Video-oriented social graph generation via hierarchical-cumulative gcn. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4716–4724.

Ke Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*.

Chenghao Yan, Zihe Liu, Fangtao Li, Chenyu Cao, Zheng Wang, and Bin Wu. 2021. Social relation analysis from videos via multi-entity reasoning. In *Proceedings of the 2021 International Conference on Multimedia Retrieval*, pages 358–366.

Dingkang Yang, Mingcheng Li, Dongling Xiao, Yang Liu, Kun Yang, Zhaoyu Chen, Yuzheng Wang, Peng Zhai, Ke Li, and Lihua Zhang. 2024a. Towards multimodal sentiment analysis debiasing via bias purification. In *European Conference on Computer Vision*, pages 464–481. Springer.

Dingkang Yang, Dongling Xiao, Ke Li, Yuzheng Wang, Zhaoyu Chen, Jinjie Wei, and Lihua Zhang. 2024b. Towards multimodal human intention understanding debiasing via subject-deconfounding. *arXiv e-prints*, pages arXiv–2403.

Xiaotian Yu, Hanling Yi, Qie Tang, Kun Huang, Wenze Hu, Shiliang Zhang, and Xiaoyu Wang. 2024. Graph-based social relation inference with multi-level conditional attention. *Neural Networks*, 173:106216.

Yueyuan Zheng, Sarah de la Harpe, Angeline Y Yang, William G Hayward, Romina Palermo, and Janet Hsiao. 2023. Cultural differences in the effect of mask use on face and facial expression recognition. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 45.