# Multimodal Coreference Resolution for Chinese Social Media Dialogues: Dataset and Benchmark Approach

**Xingyu Li, Chen Gong*, Guohong Fu**
Institute of Artificial Intelligence, School of Computer Science and Technology
Soochow University, Suzhou, China
{xingyuli7007}@gmail.com;
{gongchen18, ghfu}@suda.edu.cn

## Abstract

Multimodal coreference resolution (MCR) aims to identify mentions referring to the same entity across different modalities, such as text and visuals, and is essential for understanding multimodal content. In the era of rapidly growing multimodal content and social media, MCR is particularly crucial for interpreting user interactions and bridging text-visual references to improve communication and personalization. However, MCR research for real-world dialogues remains unexplored due to the lack of sufficient data resources. To address this gap, we introduce TikTalkCoref, the first Chinese multimodal coreference dataset for social media in real-world scenarios, derived from the popular Douyin short-video platform. This dataset pairs short videos with corresponding textual dialogues from user comments and includes manually annotated coreference clusters for both person mentions in the text and the coreferential person head regions in the corresponding video frames. We also present an effective benchmark approach for MCR, focusing on the celebrity domain, and conduct extensive experiments on our dataset, providing reliable benchmark results for this newly constructed dataset. We release the TikTalkCoref dataset to facilitate future research on MCR for real-world social media dialogues at https://github.com/lxystaruni/TikTalkCoref.


Figure 1: An example of TikTalkCoref.

## 1 Introduction

Coreference resolution (CR) aims to identify mentions and cluster those referring to the same entity. For example, in Figure 1, the highlighted phrases represent mentions, with those sharing the same color indicating that they refer to the same person. Coreference resolution is essential for enhancing natural language understanding and is widely applied in downstream tasks such as summarization (Huang and Kurohashi, 2021), sentiment analysis (Cai et al., 2024) and entity linking (Chen et al., 2017). Currently, most existing coreference resolution methods focus on the text modality (Lee et al., 2017; Bohnet et al., 2023; Martinelli et al., 2024). However, with the prevalence of multimodal content in both offline real-world and online social media platforms, traditional text-based coreference resolution can no longer meet the demands of understanding and interacting with growing multimedia content.

Therefore, multimodal coreference resolution (MCR) has recently gained widespread attention and made significant advancements (Goel et al., 2023; Willemsen et al., 2023; Ueda et al., 2024). However, previous MCR work has primarily focused either on human-machine dialogue (Ueda et al., 2024), movie narrations (Rohrbach et al., 2017), or images descriptions (Goel et al., 2023), which may make it difficult to fully capture the complexity and diversity of naturally occurring multimodal interactions in real-world scenarios. Unlike task-oriented dialogues with clear goals, chat-style dialogues, characterized by their open-ended and spontaneous nature, are more prevalent and natural in real-world social interactions, yet remain underexplored in current MCR research. Moreover, most existing studies have concentrated on English, with relatively little attention paid to

*Corresponding author



*Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 31513–31525
July 27 - August 1, 2025 

Chinese. This imbalance has resulted in a scarcity of MCR research for Chinese social media dialogues in real-world scenarios.

To address this gap, we focus on multimodal coreference resolution for person entities in real-world social media dialogues, a task that holds significant potential in misinformation detection (Wu et al., 2022), stance identification (Weinzierl and Harabagiu, 2023), and emotion-driven response generation (Zhang et al., 2024). We propose TikTalkCoref, to the best of our knowledge, the first Chinese multimodal coreference resolution dataset for real-world social media dialogues, derived from the popular Douyin[1] short-video platform. TikTalkCoref includes annotations of textual person clusters in dialogues and their corresponding visual regions in videos. As illustrated in Figure 1, we manually annotate the textual mentions of two persons "Charlene Choi" and "Gillian Chung", clustering mentions that refer to the same person (in Figure 1 we mark them with the same color), and link these clusters to the head regions of the corresponding persons in the video frames, establishing the alignment of textual references and head regions, which we denote as cross-modal coreference relationships.

Based on our dataset, we propose an effective benchmark approach and conduct extensive experiments to provide reliable benchmark results on the TikTalkCoref dataset under both zero-shot and fine-tuning settings. To ensure the accuracy and objectivity of the evaluation, our benchmark focuses on multimodal coreference resolution for celebrities. In-depth analysis is also conducted to gain more insights.

## 2 Related Work

### 2.1 Coreference Resolution Datasets

Textual coreference resolution datasets such as OntoNotes 5.0 (Hovy et al., 2006), LitBank (Bamman et al., 2020), GAP (Webster et al., 2018), GUM (Zeldes, 2017), WikiCoref (Ghaddar and Langlais, 2016), OntoGUM (Zhu et al., 2021), WinoBias (Zhao et al., 2018), and PreCo (Chen et al., 2018a) have achieved notable success in their respective domains and provided high-quality annotations, but they are limited to the text modality.

With the increasing richness of multimedia content, multimodal coreference resolution has become a new research hotspot. Current multimodal coreference resolution datasets mainly focus on image caption coreference resolution (Ramanathan et al., 2014; Rohrbach et al., 2017; Goel et al., 2023; Ueda et al., 2024) and visual dialogue coreference resolution (Yu et al., 2019; Kottur et al., 2021). Among these, Yu et al. (2019), Kottur et al. (2021) and Goel et al. (2023) focus on visual coreference resolution for general objects, while Ramanathan et al. (2014), Rohrbach et al. (2017), and Ueda et al. (2024) are dedicated to visual coreference resolution for persons. All of these datasets provide high-quality annotations of textual clusters and their corresponding visual regions. However, these datasets share a common issue: they either originate from human-computer dialogues, movie narrations, or image descriptions, which may not adequately represent the complexity and diversity of naturally occurring multimodal interactions in real-world scenarios.

To address this problem, we propose TikTalkCoref, the first Chinese multimodal coreference resolution dataset based on real-world social media dialogues.

### 2.2 Coreference Resolution Methods

Early work in textual coreference resolution relied on rule-based systems and handcrafted features (Hobbs, 1978). The rise of deep learning advanced the field, with Lee et al. (2017) introducing an end-to-end model that jointly learns mention detection and coreference resolution, eliminating manual feature engineering. Recent studies have explored fine-tuning pre-trained models like BERT for coreference tasks (Joshi et al., 2020). As large language models emerged, studies such as Bohnet et al. (2023) and Zhang et al. (2023) showed performance improvements with larger models. Unlike the trend of using LLMs for coreference resolution, Martinelli et al. (2024) proposed an efficient pipeline that identifies mention boundaries and applies mention pruning strategies to reduce computational overhead.

Different from text coreference resolution, multimodal coreference resolution combines textual context with visual information to identify cross-modal coreferential relationships. For cross-modal coreference relationships between text and video, Ramanathan et al. (2014) and Rohrbach et al. (2017) proposed using trajectory prediction methods to align character references in narrations with corresponding video regions. Recently, Guo et al. (2022) and Goel et al. (2023) focus on cross-modal coref-

[1]https://www.douyin.com/

| Dataset | #Dialog | Dur.(min) | #Mention | #Cluster | #Bbox |
|---|---|---|---|---|---|
| TikTalkCoref | 1,012 | 519.65 | 2,179 | 1,435 | 958 |
| TikTalkCoref-celeb | 338 | 158.33 | 731 | 488 | 426 |

Table 1: Statistics of TikTalkCoref dataset and the sub-dataset TikTalkCoref-celeb.

erence between text and images by incorporating additional information such as object metadata and mouse trajectories. These methods excel in their specific tasks; however, they struggle with real-world dialogues on social media. This is because, in real-world dialogues, speakers often omit descriptions of visible objects' appearance or position, making it difficult for models to obtain visual cues from the text to locate the mentioned objects.

To address this, we propose a novel benchmark for multimodal coreference resolution based on our newly constructed TikTalkCoref dataset, aimed at exploring cross-modal coreference resolution with implicit visual cues in real-world dialogues from social media.

# 3 Construction of TikTalkCoref Dataset

In this section, we provide a detailed description of the annotation methodology and process to construct the TikTalkCoref dataset.

## 3.1 Data Selection

Our TikTalkCoref dataset is built upon the TikTalk dataset (Lin et al., 2023), a multimodal dialogue dataset derived from Douyin, a Chinese short video platform. The dialogues in TikTalk are user comments responding to video content, reflecting a real-world social chat environment. Most dialogues are single-turn interactions between two speakers. The TikTalk dataset contains a total of 367k dialogues commenting on 38k videos.

From this dataset, we randomly select 4,000 samples, with each dialogue paired with its associated video. High-quality dialogues are manually filtered for subsequent coreference resolution annotation based on the following criteria: (1) Excludes content with personal identifying information or sensitive details. (2) Excludes videos with significant blurriness or noise, ensuring faces are identifiable. (3) Excludes dialogues that do not clearly mention person entities for coreference resolution.

Finally, we select a total of 1,012 high-quality dialogues for annotation.

## 3.2 Annotation Guidelines

In our annotation task, we focus on annotating mentions and clusters related to persons in both textual dialogues and their corresponding videos. For example, in Figure 1, we have the following mentions: "阿娇(AJiao)", "她(she)", "阿sa (Asa), and "钟欣潼(Gillian Chung)". Here, "阿娇(AJiao)", "她(she)", and "钟欣潼(Gillian Chung)" both refer to the same person, Gillian Chung, so they form a coreference cluster {"阿娇(AJiao)", "她(she)", "钟欣潼(Gillian Chung)"}. On the other hand, "阿sa (Asa)" refers to Charlene Choi, but since there are no other mentions in the dialogue that corefer with it, "阿sa (Asa)" forms a singleton cluster {"阿sa (Asa)"}. In Figure 1, the mentions within the same cluster are highlighted in the same color, and their visual regions in the video are marked with bounding boxes (bbox) of the same color.

We conduct an in-depth investigation into existing mainstream coreference annotation guidelines, such as OntoNotes (Hovy et al., 2006) and ACE (Walker et al., 2006), which have been widely adopted in the field of coreference resolution. Considering the characteristics of the TikTalk dataset, where most data consists of short dialogues and many persons are mentioned only once within a dialogue, we develop annotation guidelines tailored to our task. Our annotation guidelines focus primarily on coreferential relationships involving persons in the videos, and are outlined as follows:

(1) For mention annotation, noun phrases and pronoun phrases referring to a person are treated as potential mentions. This includes proper names (such as "蔡卓妍(Charlene Choi)" and "钟欣潼(Gillian Chung)"), common noun phrases referring to a person (such as "歌手(singer)" and "那个人(that person)"), and pronouns (such as "他(he)", "她(she)" and "他们(they)"). For nested mentions, we employ an efficient simplified annotation strategy that focuses solely on annotating inner submentions, then reconstructs the complete nested mentions through mention indexing. (2) For cluster annotation, mentions that appear only once in the dialogue are treated as singleton cluster.

In addition, we ask annotators to annotate whether a person belongs to the "celebrity" category. The "celebrity" should have the following characteristics: (1) The person is regularly featured in public media and has a widely recognized public image. (2) Based on the common knowledge or publicly available information, annotators all agree

| Dataset | Domain | Language | Modalities | Text Type | Object Type | Mention Type | #Dialog |
|---|---|---|---|---|---|---|---|
| MPII-MD (Rohrbach et al., 2017) | Movies | English | Text, Video | Caption | People | PNs, PRs | - |
| VisPro (Yu et al., 2019) | Open-world | English | Text, Image | Dialogue | General objects | CNs, PRs | 5,000 |
| Simmc2 (Kottur et al., 2021) | Shopping | English | Text, Image | Dialogue | Clothing | CNs | 11,244 |
| CIN (Goel et al., 2023) | Open-world | English | Text, Image | Caption | General objects | CNs, PRs | - |
| J-CRe3 (Ueda et al., 2024) | Household | Japanese | Text, Video | Dialogue | General objects | CNs, PRs | 96 |
| TikTalkCoref (Ours) | Social media | Chinese | Text, Video | Dialogue | People | PNs, CNs and PRs | 1,012 |

Table 2: Comparison with main multimodal coreference resolution datasets. Mention types include Proper names (PNs), Common nouns (CNs), and Pronouns (PRs).

that the person is a celebrity.

### 3.3 Annotation Process

We employ three undergraduate and graduate students as annotators for annotating the TikTalkCoref dataset and select one expert annotator to resolve any inconsistencies. All annotators are paid based on the quality and quantity of their annotations. Our annotation process follows a rigorous independent double annotation workflow, where two annotators independently annotate each dialogue, and an experienced annotator, acting as an expert, resolves any inconsistencies between their annotations. Specifically, our annotation process is divided into two steps:

(1) Watch the video to understand its content, then annotate the mentions and clusters of persons in the dialogue.

(2) Select key frames that clearly show the faces of the person mentioned in the video and draw their head regions. Note that the selected key frame should have a clear frontal view of the person.

To ensure the quality and efficiency of the annotations, we have developed an annotation system based on Labelstudio[2] to support the entire process. We report more details of our annotation tool in Appendix A.3.

### 3.4 Data Statistics and Analysis

**Overall Statistics.** Table 1 presents a statistical overview of our newly constructed TikTalkCoref dataset. TikTalkCoref consists of 1,012 dialogues with a total video duration of 519.65 minutes. It contains 1,435 clusters, 2,179 mentions, and 958 bounding boxes. In addition, we divide the dialogues mentioning celebrities into a sub-dataset named TikTalkCoref-celeb, which contains 338 dialogues with a total video duration of 158.33 minutes, as well as 731 clusters, 488 mentions, and 426 bounding boxes.

[2]https://labelstud.io/

This dataset integrates coreferential relationships of persons from both dialogues and video frames, making it ideal for multimodal learning and coreference resolution in dialogue systems. More statistical information, such as the distribution of gender, age, and occupation is in Appendix A.2.

**Inter-Annotator Agreement.** As previously mentioned, we employ an independent double annotation approach, with a third expert annotator resolving inconsistencies to ensure high annotation quality. Following previous work (Pradhan et al., 2012), we measure inter-annotator agreement using the average MUC score between two annotators, achieving a score of 78.19. This result demonstrates the effectiveness of our double annotation workflow in ensuring annotation quality.

**Mention Type Distribution.** In our dataset, proper names make up 33.51%, pronouns 44.41%, and common nouns 22.08%. The high frequency of proper names and pronouns can be attributed to the fact that speakers have typically already watched the video and are familiar with the persons in the video, and thus tend to use these references to identify and refer to the persons. This reflects a more natural form of dialogue, where speakers rely on previously established context to efficiently refer to people and objects in the video.

**Comparison with Existing MCR Datasets.** We compare our TikTalkCoref dataset with other multimodal coreference resolution datasets in Table 2. Unlike other existing datasets, TikTalkCoref offers annotations for a rich variety of mention types, including proper names, common nouns, and pronouns. Moreover, according to our survey, it is the first Chinese multimodal coreference resolution dataset on social media.

## 4 Method

To provide reliable benchmark results on our newly constructed TikTalkCoref dataset and facilitate future research, we present a simple yet effective

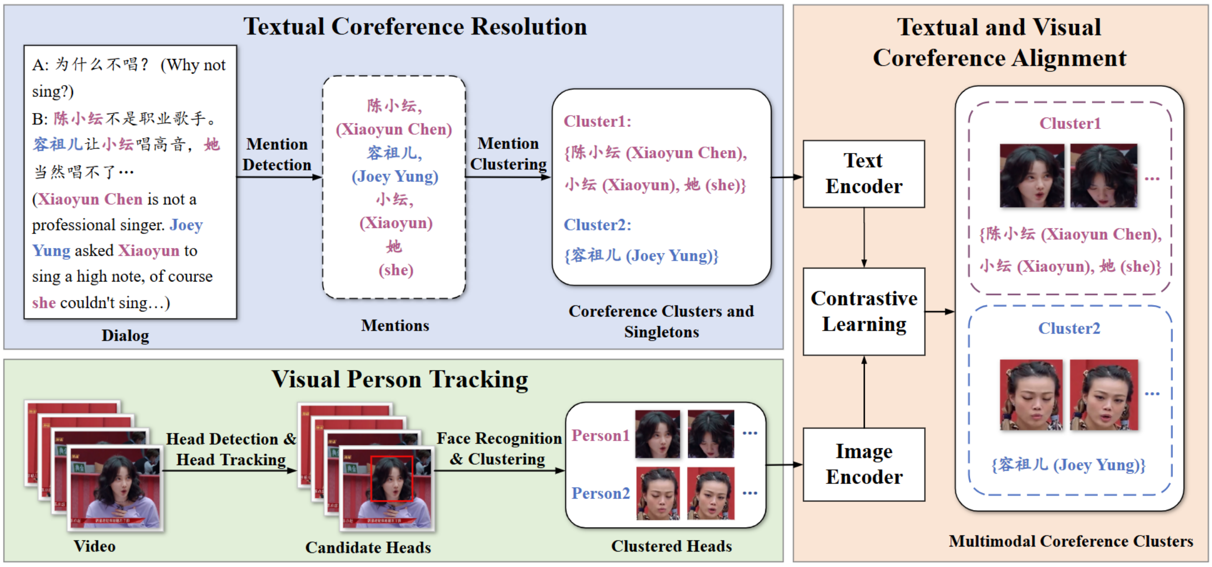


Figure 2: The overview of our model architecture.

multimodal coreference resolution pipeline method (illustrated in Figure 2), consisting of three modules: (1) a textual coreference resolution module to extract mentions and cluster those referring to the same person in the dialogue; (2) a visual person tracking module to detect person regions in the video and clustering the regions representing the same person; and (3) a textual and visual coreference alignment module to link the textual clusters to the video clusters to establish cross-modal coreferential relationships.

### 4.1 Textual Coreference Resolution

To identify and cluster mentions referring to the same person entity in the textual dialogue, we adopt the state-of-the-art Maverick model (Martinelli et al., 2024) as our textual coreference resolution module, which is a pipeline method that first detects the mentions and then groups those referring to the same person.

**Mention Detection.** In the mention detection phase, we use a DeBERTa encoder (He et al., 2021) to obtain the hidden representation $\mathbf{x}_i$ for each token $t_i$ in the dialogue. We then apply two fully connected layers to $\mathbf{x}_i$ to compute the probability of token $t_i$ being a start of a mention, and subsequently extract its possible ends. Tokens with a start probability $p_{start}(t_i) > 0.5$ are considered as candidate mention starts $t_s$, while tokens with an end probability $p_{end}(t_j \mid t_s) > 0.5$ are considered as candidate mention ends $t_e$. In this way, we obtain all the candidate mentions.

**Mention Clustering.** In the mention clustering phase, after obtaining candidate mentions from the mention detection phase, we cluster them using the coarse-to-fine mention-antecedent method of Lee et al. (2018). This method formulates mention clustering as a binary classification task, where the model predicts whether two mentions are coreferential. For each mention $m_i$, we first coarsely identify its top $K$ antecedents (i.e., the coreferential mentions $m_j$ that appear earlier than $m_i$ in the text) using a bilinear scoring function. Then, the resulting mention pairs $(m_i, m_j)$ are evaluated more finely with a mention-pair scorer based on fully connected layers. If the score exceeds a threshold of 0.5, the two mentions are considered coreferential, otherwise, they are not. All mentions sharing coreferential relationships are grouped into a cluster. If a cluster contains only a single mention, meaning no other mention is coreferential with it, we refer to such clusters as singletons.

### 4.2 Visual Person Tracking

The visual person tracking module identifies and tracks persons in videos by detecting their head regions across frames and clustering precisely the head regions that correspond to the same person based on facial features.

**Head Detection and Tracking.** We employ a YOLOv5-based head detector[3] to identify candidate person head regions in each video frame and employ the DeepSORT algorithm (Wojke et al., 2017) to track and link the detected heads belonging to the same person across frames. However,

[3] https://gitee.com/wallebus/yolo

since our dataset is sourced from social media and typically consists of non-continuous video segments stitched together, the above head detection and tracking method is limited to identifying and tracking persons within each segment. As a result, the same person appearing across different segments is often treated as distinct persons due to significant positional shifts and pose variations between segments. To address this limitation, we perform an additional face recognition and clustering step to accurately group heads belonging to the same person across different video segments, as detailed in the following paragraph.

**Face Recognition and Clustering.** To improve clustering accuracy, we use MTCNN (Zhang et al., 2016) and MobileFaceNet (Chen et al., 2018b) to detect and extract facial features from candidate head regions. For each trajectory, the extracted facial features are averaged, and cosine similarity is calculated between trajectories. Trajectories with a similarity score exceeding 0.6 are grouped as belonging to the same person. In this way, for each video, head regions representing the same person are grouped into the same cluster.

### 4.3 Textual and Visual Coreference Alignment

The textual and visual coreference alignment module links textual clusters from the coreference resolution module with visual head region clusters from the visual person tracking module using contrastive learning, thereby generating multimodal coreference clusters.

To achieve this, we use Chinese CLIP (Yang et al., 2022), a model based on CLIP (Radford et al., 2021) and pre-trained on a large-scale Chinese image-text dataset. Specifically, for each dialogue and its corresponding video, we obtain the cluster set $C = \{C_1, C_2, \ldots, C_K\}$ of the dialogue through the textual coreference resolution module and select one image of each person $P_j$ from all frames of the video as the representative image $I_j$ for that person through the visual person tracking module, forming the candidate image set $I = \{I_1, I_2, \ldots, I_J\}$. To optimize computational efficiency, we employ random sampling (rather than pooling) to select representative person images, as the clustering phase already preserves highly similar facial samples.

Both $C$ and $I$ are encoded into a shared multimodal embedding space using the textual and visual encoders of Chinese CLIP. Contrastive learning is applied to maximize the similarity for matching pairs of textual cluster $C_k$ and their corresponding person head regions $I_j$, while minimizing similarity for non-matching pairs. This process aligns the textual clusters with their corresponding person head regions, thereby establishing cross-modal coreference relationships to form the multimodal coreference clusters $\{(C_{k_1}, I_{j_1}), (C_{k_2}, I_{j_2}), \ldots, (C_{k_N}, I_{j_N})\}$.

|  | #Dialog | #Mention | #T-Cluster | #M-Cluster |
|---|---|---|---|---|
| Train-all | 910 | 1,952 | 1,289 | - |
| Train-celeb | 236 | 504 | 342 | 270 |
| Dev | 35 | 76 | 49 | 40 |
| Test | 67 | 151 | 97 | 80 |
| Total | 1,012 | 2,179 | 1,435 | 390 |

Table 3: Data statistics: the number of dialogues (#Dialog), mentions (#Mention), textual clusters (#T-cluster) and multimodal clusters (#M-cluster) in each split of TikTalkCoref.

### 4.4 Training Loss

The total loss of our multimodal coreference resolution model is defined by the sum of textual coreference resolution loss $\mathcal{L}_{coref}$ and multimodal alignment loss $\mathcal{L}_{align}$:

$$\mathcal{L} = \mathcal{L}_{coref} + \mathcal{L}_{align} \tag{1}$$

For the textual coreference resolution loss $\mathcal{L}_{coref}$, it is calculated as described in Maverick (Martinelli et al., 2024):

$$\mathcal{L}_{coref} = \mathcal{L}_{start} + \mathcal{L}_{end} + \mathcal{L}_{clust} \tag{2}$$

where $\mathcal{L}_{start}$ and $\mathcal{L}_{end}$ are the mention start and end loss, $\mathcal{L}_{clust}$ represents the cluster loss. All of them are computed using binary cross-entropy.

For multimodal alignment loss $\mathcal{L}_{align}$, we adopt the normalized temperature-scaled cross-entropy loss (Radford et al., 2021). For each textual cluster $T_m$, the loss aims to maximize its similarity with the matching image $I_m^*$ of its candidate images $\{I_{m,i}\}_{i=1}^{N_m}$ from the same video, while minimizing its similarity with other candidates.

## 5 Experiments

### 5.1 Experimental Settings

**Data.** From our constructed TikTalkCoref data, we select 338 dialogues mentioning celebrities (TikTalkCoref-celeb) and randomly split them into Train-celebrity set (Train-celeb), Dev set and Test

| Model | MUC | | | $B^3$ | | | CEAF$_{\phi4}$ | | | Avg.F1 |
|---|---|---|---|---|---|---|---|---|---|---|
| | P | R | F1 | P | R | F1 | P | R | F1 | |
| e2e-coref | 69.39 | 62.96 | 66.02 | 18.75 | 83.36 | 30.61 | 11.77 | 82.37 | 20.59 | 39.07 |
| - Coref clusters | 60.00 | 50.00 | 54.54 | 65.26 | 46.15 | 54.07 | 44.40 | 49.34 | 46.74 | 51.78 |
| - Singletons | - | - | - | 9.78 | 81.97 | 17.47 | 9.88 | 69.95 | 17.31 | 11.59 |
| Maverick | 54.00 | 50.00 | 51.92 | 68.78 | 67.51 | 68.14 | 73.76 | 79.06 | 76.30 | 65.46 |
| - Coref clusters | 58.49 | 50.00 | 53.91 | 58.78 | 46.59 | 51.98 | 68.86 | 45.89 | 55.06 | 53.65 |
| - Singletons | - | - | - | 68.49 | 83.61 | 75.30 | 75.12 | 82.51 | 78.65 | 51.31 |

Table 4: Textual coreference resolution results on our TikTalkCoref dataset.

| | R@1 | R@2 | R@3 | Mean |
|---|---|---|---|---|
| **Zero-shot** | | | | |
| R2D2 | **52.50** | **71.50** | **76.25** | **66.67** |
| CN-Clip | 45.00 | 65.00 | 68.75 | 59.58 |
| **Fine-tuning** | | | | |
| R2D2 | 56.25 | 73.75 | **80.00** | 70.00 |
| CN-Clip | **60.83** | **75.83** | 78.75 | **71.81** |

Table 5: Textual and visual coreference alignment results under zero-shot and fine-tuning settings on our TikTalkCoref dataset.

set in a 7:1:2 ratio. The remaining 674 dialogues which do not mention celebrities, are combined with Train-celeb to form the Train-all. During training, we use either Train-all or Train-celeb for the text coreference resolution module to evaluate the impact of data augmentation, and use Train-celeb to train the text and visual coreference alignment module. Data statistics are provided in Table 3.

**Settings.** For textual coreference resolution, we fine-tune the Maverick model with DeBERTa-Chinese-Large using the Adafactor optimizer (lr=3e-4, weight decay=0.01) for 50 epochs. For visual person tracking, we employ a YOLOv5-based head detector with DeepSORT algorithm to extract and track person head regions, and MobileFaceNet for accurate person clustering. For textual and visual coreference alignment, we use the pre-trained CN-CLIP (ViT-B/16) with a ViT-B/16 image encoder and RoBERTa-wwm-Base text encoder, trained with a cosine learning rate schedule (initial lr=5e-6, 100-step warmup) for 3 epochs. The pipeline is trained on two NVIDIA V100 GPUs in approximately 6 hours. We conduct experiments using three different random seeds and report the average performance.

**Evaluation metrics.** Following previous works, we use MUC, $B^3$, and CEAF$_{\phi4}$, along with their average F1 score, as evaluation metrics for textual coreference resolution, and employ R@K (Recall@K) as the evaluation metric for textual and visual coreference alignment. For significance test, we use Dan Bikel's randomized parsing evaluation comparator (Noreen, 1989). Detailed explanation of metrics is in Appendix A.1.

**Comparison Models.** For comparison, we use the End-to-End Coreference Resolution model (e2e-coref) (Lee et al., 2018) for textual coreference resolution due to its effectiveness and wide adoption, and the R2D2 model (Xie et al., 2023) for textual and visual coreference alignment due to its strong text-to-image alignment performance.

## 5.2 Main Results

**Results of Textual Coreference Resolution.** Table 4 presents the results of textual coreference resolution using the e2e-coref and Maverick models on our constructed TikTalkCoref dataset. We also report the clustering performance of these models on two types of clusters: coreference clusters (coref clusters), which contain more than one mention, and singletons, which contain only one mention. From Table 4, we observe that the Maverick model significantly outperforms the e2e-coref model ($p < 0.001$) in textual coreference resolution, especially in clustering singletons. This may be attributed to the large number of singletons in our TikTalkCoref dataset and Maverick's clustering approach, which independently handles singletons during clustering. This allows Maverick to effectively distinguish singletons from coreference clusters. As a result, Maverick performs better in $B^3$ and CEAF$_{\phi4}$, which evaluate mention matching accuracy and cluster alignment quality. However, the MUC metric of e2e-coref is higher than that of the Maverick model. This may be because e2e-coref is limited in handling singletons precisely, leading it to generate more coreference clusters and resulting in better performance on MUC, a metric based on

| | MUC.F1 | $B^3$.F1 | CEAF$_{\phi4}$.F1 | Avg.F1 |
|---|---|---|---|---|
| e2e-coref | **66.02** | **30.61** | **20.59** | **39.07** |
| - w/o DA | 20.56 | 24.63 | 17.98 | 21.06 |
| Maverick | **51.92** | **68.14** | **76.30** | **65.46** |
| - w/o DA | 32.32 | 58.63 | 67.06 | 52.67 |

Table 6: Comparison of performance with and without data augmentation (DA). Metrics include MUC.F1, $B^3$.F1, CEAF$_{\phi4}$.F1, and the average F1 score of the three metrics (Avg.F1).

coreference link.

**Results of Textual and Visual Coreference Alignment.** Table 5 presents the results of textual and visual coreference alignment. Since the alignment is essentially achieved by retrieving the person head regions in the video that are coreferential with the text clusters, we report retrieval results using R@1, R@2, and R@3. We compare the performance of the CN-Clip and R2D2 models under both zero-shot and fine-tuning settings. R2D2 significantly outperforms CN-Clip in the zero-shot setting ($p < 0.001$), particularly excelling in metrics like R@1 and R@2. This may be due to R2D2's fine-grained ranking strategy, which captures more detailed feature similarities between images and texts, allowing it to perform well in cross-modal retrieval tasks without additional training. However, despite R2D2's outstanding performance in the zero-shot setting, CN-Clip performs better in R@1 and overall Mean scores after fine-tuning. This is likely due to our use of non-matching images from the corresponding video as negative samples during training. CN-Clip seems better able to adapt to this limited negative sample setting during fine-tuning, leading to better performance in the R@1 and Mean metrics.

### 5.3 Impact of Data Augmentation on Textual Coreference Resolution

To investigate the impact of adding non-celebrity data on textual coreference resolution, we compare the performance of both e2e-coref and Maverick models training with Train-all and Train-celeb, as shown in Table 6. The results suggest that non-celebrity data shares certain characteristics with celebrity data, which helps enhance the textual coreference resolution performance. In contrast to celebrity data, which mainly involves well-known persons and familiar contexts, non-celebrity data introduces a broader range of linguistic contexts and mention types. It includes more common nouns, pronouns, and generalized references. This diversity enables the model to better understand and resolve coreference relations across various contexts, making it more robust and accurate when handling celebrity-related dialogues.

| | Name(*) | Noun(*) | Pronoun(*) |
|---|---|---|---|
| **Zero-shot** | | | |
| R2D2 | **68.03** | **51.51** | **80.00** |
| CN-Clip | 61.90 | 36.36 | 66.67 |
| **Fine-tuning** | | | |
| R2D2 | 67.35 | **60.61** | 81.67 |
| CN-Clip | **70.52** | 54.54 | **82.22** |

Table 7: Retrieval accuracy of clusters with different mention types on our TikTalkCoref dataset. Name(*), Noun(*), and Pronoun(*) represent the name-central clusters, the noun-central clusters, and the pronoun-central clusters.

### 5.4 Image Retrieval Performance Across Clusters with Different Mention Types

To analyze image retrieval performance across clusters containing different types of mentions, we categorize the clusters into three types base on their central mention: name-central (clusters with names), noun-central (clusters with common nouns and pronouns, but no names), and pronoun-central (clusters with only pronouns). The image retrieval accuracy of CN-CLIP and R2D2 for these three cluster types is shown in Table 7.

In the zero-shot setting, R2D2 outperforms CN-CLIP across all cluster types ($p < 0.001$), demonstrating its strong zero-shot retrieval capability, consistent with the findings in Table 5. In the fine-tuning setting, CN-CLIP improves its performance across different cluster types by effectively leveraging negative examples from our dataset during contrastive learning, surpassing R2D2 in name-centered and pronoun-centered clusters. However, R2D2 retains its advantage in noun-centered clusters. This may be attributed to R2D2's pretraining dataset, which exhibits a closer alignment with the characteristics of our TikTalkCoref dataset, particularly in containing a substantial number pairs of indefinite common nouns and their corresponding images, allowing R2D2 to more effectively align common nouns referring to persons with their corresponding visual regions in our task.

## 6 Conclusion

In this work, we introduce TikTalkCoref, the first Chinese multimodal coreference dataset designed for social media. TikTalkCoref addresses the challenges of multimodal coreference resolution in real-world scenarios by providing detailed annotations of textual clusters of persons and their corresponding visual regions. We propose an effective benchmark approach to tackle the multimodal coreference resolution problem and conduct extensive experiments to provide reliable benchmark results on the TikTalkCoref dataset under both zero-shot and fine-tuning settings. We also conduct in-depth analysis of the experimental results, providing valuable insights for future research. We hope that TikTalkCoref will facilitate future research in multimodal understanding of real-world dialogues.

## 7 Limitations

There are two main limitations in our work. First, despite significant annotation efforts, our TikTalkCoref dataset is relatively small in scale and relies solely on data from the Douyin platform, which may limit its diversity. As another major Chinese social media platform, Weibo, unlike Douyin's video-centric format, features more logically structured discussions due to its blog-based design. In the future, we plan to collect and annotate multimodal conversational data from Weibo to improve the scale and diversity of the dataset.

Second, the supervised training approach used in this work may not fully exploit the potential of the model in low-resource scenarios. To address this, we will explore semi-supervised or unsupervised techniques to improve the model performance in low-resource scenarios.

## Acknowledgements

We would like to thank the anonymous reviewers for the helpful comments. This work was supported by National Natural Science Foundation of China (Grant No. 62476187 and 62306202), and a Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

## References

Amit Bagga and Breck Baldwin. 1998. Entity-based cross-document coreferencing using the vector space model. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 79–85, Montreal, Quebec, Canada. Association for Computational Linguistics.

David Bamman, Olivia Lewke, and Anya Mansoor. 2020. An annotated dataset of coreference in English literature. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 44–54, Marseille, France. European Language Resources Association.

Bernd Bohnet, Chris Alberti, and Michael Collins. 2023. Coreference resolution through a seq2seq transition-based system. *Transactions of the Association for Computational Linguistics*, 11:212–226.

Hongjie Cai, Heqing Ma, Jianfei Yu, and Rui Xia. 2024. A joint coreference-aware approach to document-level target sentiment analysis. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12149–12160. Association for Computational Linguistics.

Henry Y. Chen, Ethan Zhou, and Jinho D. Choi. 2017. Robust coreference resolution and entity linking on dialogues: Character identification on TV show transcripts. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 216–225. Association for Computational Linguistics.

Hong Chen, Zhenhua Fan, Hao Lu, Alan Yuille, and Shu Rong. 2018a. PreCo: A large-scale dataset in preschool vocabulary for coreference resolution. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 172–181, Brussels, Belgium. Association for Computational Linguistics.

S. Chen, Y. Liu, X. Gao, and Z. Han. 2018b. Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. In *Biometric Recognition. CCBR 2018. Lecture Notes in Computer Science*, volume 10996, pages 329–338. Springer, Cham.

Abbas Ghaddar and Phillippe Langlais. 2016. WikiCoref: An English coreference-annotated corpus of Wikipedia articles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 136–142, Portorož, Slovenia. European Language Resources Association (ELRA).

Arushi Goel, Basura Fernando, Frank Keller, and Hakan Bilen. 2023. Who are you referring to? coreference resolution in image narrations. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15201–15212.

Danfeng Guo, Arpit Gupta, Sanchit Agarwal, Jiun-Yu Kao, Shuyang Gao, Arijit Biswas, Chien-Wei Lin, Tagyoung Chung, and Mohit Bansal. 2022. GRAVL-BERT: Graphical visual-linguistic representations for multimodal coreference resolution. In *Proceedings* *of the 29th International Conference on Computational Linguistics*, pages 285–297, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: decoding-enhanced bert with disentangled attention. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Jerry R. Hobbs. 1978. Resolving pronoun references. *Lingua*, 44(4):311–338.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. OntoNotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60, New York City, USA. Association for Computational Linguistics.

Yin Jou Huang and Sadao Kurohashi. 2021. Extractive summarization considering discourse and coreference relations based on heterogeneous graph. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3046–3052. Association for Computational Linguistics.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Satwik Kottur, Seungwhan Moon, Alborz Geramifard, and Babak Damavandi. 2021. SIMMC 2.0: A task-oriented dialog dataset for immersive multimodal conversations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4903–4912, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark. Association for Computational Linguistics.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 687–692, New Orleans, Louisiana. Association for Computational Linguistics.

Hongpeng Lin, Ludan Ruan, Wenke Xia, Peiyu Liu, Jingyuan Wen, Yixin Xu, Di Hu, Ruihua Song, Wayne Xin Zhao, Qin Jin, and Zhiwu Lu. 2023. Tiktalk: A video-based dialogue dataset for multi-modal chitchat in real world. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, page 13031313, New York, NY, USA. Association for Computing Machinery.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 25–32, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Giuliano Martinelli, Edoardo Barba, and Roberto Navigli. 2024. Maverick: Efficient and accurate coreference resolution defying recent trends. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13380–13394, Bangkok, Thailand. Association for Computational Linguistics.

Eric W. Noreen. 1989. *Computer-intensive methods for testing hypotheses: An introduction*. John Wiley & Sons, Inc., New York.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Joint Conference on EMNLP and CoNLL - Shared Task*, Jeju Island, Korea. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

V. Ramanathan, A. Joulin, P. Liang, and L. Fei-Fei. 2014. Linking people in videos with "their" names using coreference resolution. In *Computer Vision - ECCV 2014*, volume 8689 of *Lecture Notes in Computer Science*, pages 96–110, Cham. Springer.

Anna Rohrbach, Marcus Rohrbach, Siyu Tang, Seong Joon Oh, and Bernt Schiele. 2017. Generating descriptions with grounded and co-referenced people. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4196–4206.

Label Studio. 2024. Label studio. Accessed: 2024-11-14.

Nobuhiro Ueda, Hideko Habe, Akishige Yuguchi, Seiya Kawano, Yasutomo Kawanishi, Sadao Kurohashi, and Koichiro Yoshino. 2024. J-CRe3: A Japanese conversation dataset for real-world reference resolution. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9489–9502, Torino, Italia.

Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. In *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*, Columbia, Maryland.

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. Ace 2005 multilingual training corpus. *Linguistic Data Consortium*, 57(45).

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the GAP: A Balanced Corpus of Gendered Ambiguous Pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617.

Maxwell Weinzierl and Sanda Harabagiu. 2023. Identification of multimodal stance towards frames of communication. pages 12597–12609. Association for Computational Linguistics.

Bram Willemsen, Livia Qian, and Gabriel Skantze. 2023. Resolving references in visually-grounded dialogue via text generation. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 457–469, Prague, Czechia. Association for Computational Linguistics.

Nicolai Wojke, Alex Bewley, and Dietrich Paulus. 2017. Simple online and realtime tracking with a deep association metric. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3645–3649. IEEE.

Xueqing Wu, Kung-Hsiang Huang, Yi Fung, and Heng Ji. 2022. Cross-document misinformation detection based on event graph reasoning. pages 543–558. Association for Computational Linguistics.

Chunyu Xie, Heng Cai, Jincheng Li, Fanjing Kong, Xiaoyu Wu, Jianfei Song, Henrique Morimitsu, Lin Yao, Dexin Wang, Xiangzheng Zhang, et al. 2023. Ccmb: A large-scale chinese cross-modal benchmark. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4219–4227.

An Yang, Junshu Pan, Junyang Lin, Rui Men, Yichang Zhang, Jingren Zhou, and Chang Zhou. 2022. Chinese clip: Contrastive vision-language pretraining in chinese. *CoRR*, abs/2211.01335.

Xintong Yu, Hongming Zhang, Yangqiu Song, Yan Song, and Changshui Zhang. 2019. What you see is what you get: Visual pronoun coreference resolution in dialogues. In *Proceedings of EMNLP-IJCNLP 2019*, pages 5122–5131. Association for Computational Linguistics.

Amir Zeldes. 2017. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503.

Tenggan Zhang, Xinjie Zhang, Jinming Zhao, Li Zhou, and Qin Jin. 2024. ESCoT: Towards interpretable emotional support dialogue systems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13395–13412. Association for Computational Linguistics.

Wenzheng Zhang, Sam Wiseman, and Karl Stratos. 2023. Seq2seq is all you need for coreference resolution. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11493–11504, Singapore. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

Yilun Zhu, Sameer Pradhan, and Amir Zeldes. 2021. OntoGUM: Evaluating contextualized SOTA coreference resolution on 12 more genres. In *Proceedings of ACL-IJCNLP 2021*, pages 461–467, Bangkok, Thailand.

# A Appendix

## A.1 Evaluation metrics

For textual coreference resolution evaluation, we use MUC (Vilain et al., 1995), $B^3$ (Bagga and Baldwin, 1998), and $\text{CEAF}_{\phi 4}$ (Luo, 2005), along with their average F1 score.

MUC focuses on coreferential relations (i.e., coreferential links) between mentions, evaluating coreference resolution by comparing the overlap of links between predicted and true coreference chains.

$B^3$ is a mention-based metric, which evaluates the overall precision and recall by calculating the precision and recall of each individual mention.

$\text{CEAF}_{\phi 4}$ includes mention-based $\text{CEAF}_\text{m}$ and entity-based $\text{CEAF}_\text{e}$ metrics, and we use the latter, which focuses on the overall overlap between predicted and true clusters, assessing the model's performance in recognizing the same entity.

For the evaluation of textual and visual coreference alignment, we use Recall@K. Recall@K (R@K) is a widely used evaluation metric in information retrieval, designed to measure a model's ability to correctly retrieve target entities within its top-K predicted results. In our task, it evaluates the model's capability to accurately match textual clusters with corresponding images.

## A.2 Detailed Statistics of TikTalkCoref

We report the gender, age, and occupation distribution of the persons in our dataset in Figure 3. Due to the difficulty in collecting age and occupation statistics for non-celebrities, we only present the age and occupation distribution for the celebrities.

In our dataset, males make up 59.9% and females make up 40.1%. This gender disparity may be largely due to the dataset containing a significant number of videos related to the NBA, where male athletes are more prominently featured. The higher representation of male NBA players compared to female athletes could contribute to the observed gender imbalance.

Our dataset contains a total of 245 celebrities. Among the 245 celebrities, 51.02% are aged 18-35, 32.65% are 35-50, and 16.33% are over 50. This result aligns with the trends of current social media: young and middle-aged celebrities have higher activity levels and influence, making their videos more likely to be promoted on social media. Most of these celebrities are in the entertainment industry (85.71%), followed by sports (12.24%), business (1.63%), and agriculture (0.41%). This reflects the high public exposure of celebrities in entertainment and sports.

| Dataset | Avg_P | Avg_R | Avg_F |
|---|---|---|---|
| celeb_100 | 43.35 | 43.23 | 43.28 |
| celeb_200 | 46.82 | 49.69 | 48.02 |
| celeb & no_celeb_200 | 46.70 | 57.56 | 51.50 |

Table 8: Comparative Results with Celebrity vs. Non-Celebrity Data Augmentation

## A.3 Annotation Tool

Our annotation system is based on Labelstudio(Studio, 2024) which is an HTML-based tool that allows us to annotate coreference chains in text and bounding boxes in videos simultaneously. The annotation system interface is shown in Figure 4.

The annotation system is mainly divided into two modules: dialogue annotation and video annotation. In the dialogue annotation module, annotators first select a cluster number and then highlight mentions in the dialogue text. Mentions with the same cluster number belong to the same cluster. In the video annotation module, annotators first review the video, select the frames where the persons mentioned in the dialogue appear, and then use the same cluster number as in the dialogue to draw head bounding boxes. If the mentioned persons do not appear in the video or if the faces in the video are not clear enough, annotators mark "Person not found" or "Person found but face unidentifiable".

## A.4 Analysis of Performance Improvement After Adding Non-Celebrity Data

To verify the unique properties of celebrity vs. non-celebrity, we conduct a comparative experiment as with the following setup:

1) Base data: 100 celebrity samples;
2) Base data + 100 celebrity samples;
3) Base data + 100 non-celebrity samples.

As shown in Table 8, both celebrity and non-celebrity data improve model performance, with non-celebrity data slightly outperforming. This discrepancy may be due to the fact that the 100 non-celebrity samples used for data augmentation contain more coreference chains compared to the 100 celebrity samples, thereby providing richer supervisory signals. The results partially validate our hypothesis in Section 5.3 that non-celebrity data shares similar characteristics with celebrity data.

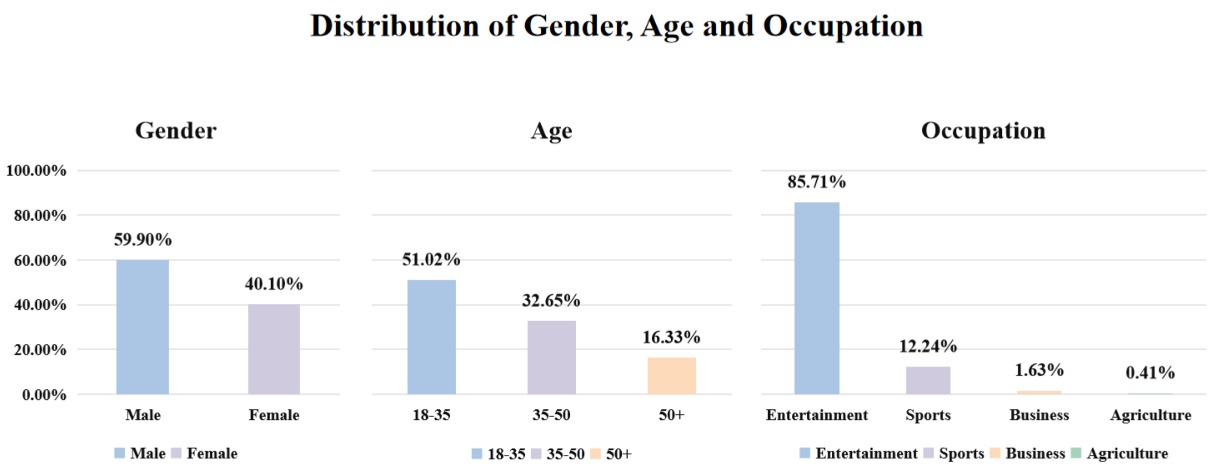


Figure 3: The distribution of gender, age and occupation of persons in the TikTalkCoref. Note that the distribution of gender is counted on all persons, and the distribution of age and occupation is counted on celebrities.

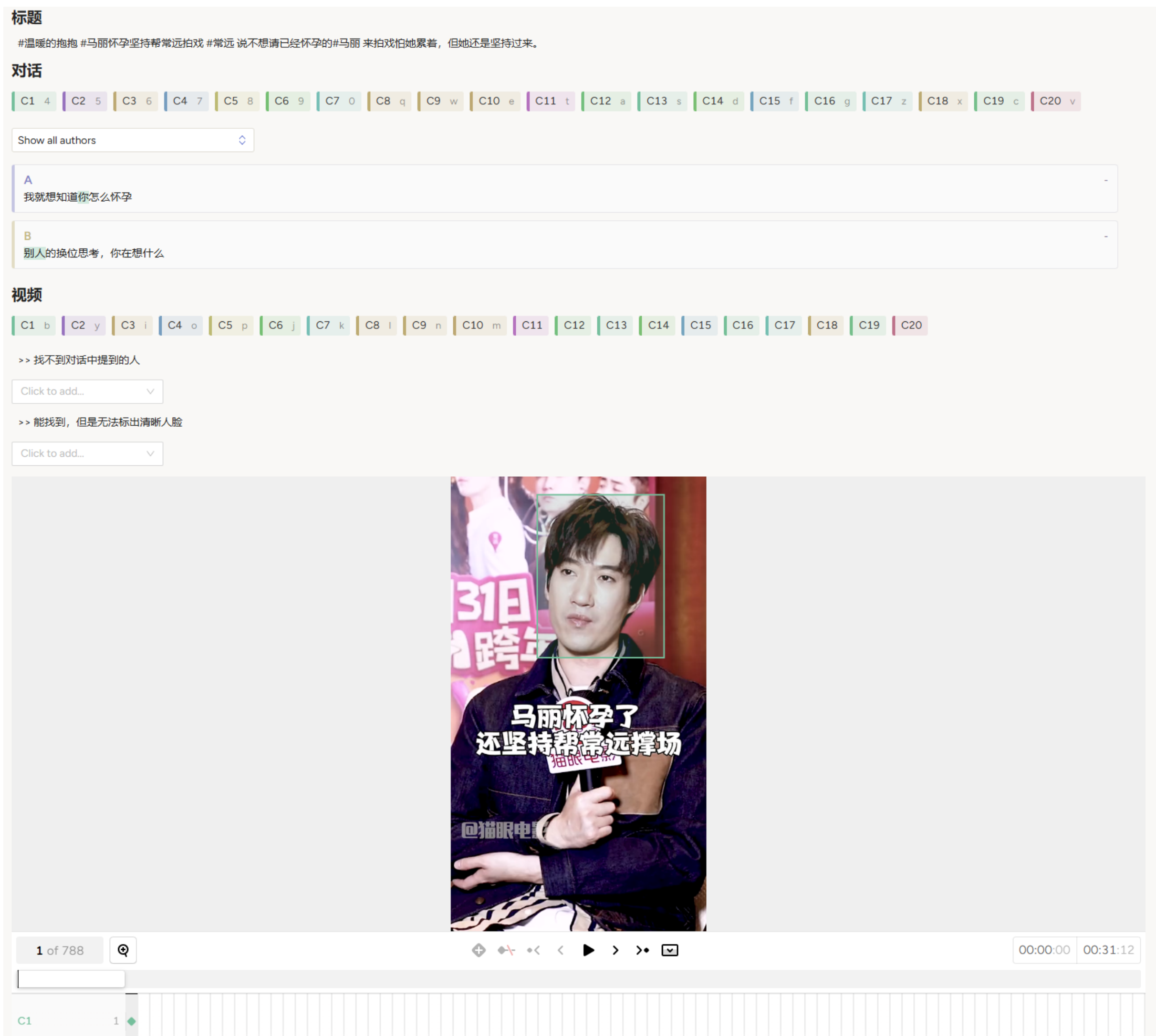


Figure 4: Annotation interface of our annotation tool.