Exploring and Detecting Self-disclosure in Multi-modal posts on Chinese Social Media

Jingbao Luo[1], Ming Liu[2], Aoli Huo[3], Fujing Hu[3], Gang Li[2], Peng Wu[4]*,
[1] School of Cyber Science and Engineering, Nanjing University of Science and Technology
[2]School of Information Technology, Deakin University
[3]School of Economics & Management, Nanjing University of Science and Technology
[4]School of Intelligent Manufacturing, Nanjing University of Science and Technology
{luojingbao, hufujing, houaoli, wupeng}@njust.edu.cn
{m.liu, gang.li}@deakin.edu.au

Abstract

Self-disclosure can provide psychological comfort and social support, but it also carries the risk of unintentionally revealing sensitive information, leading to serious privacy concerns. Research on self-disclosure in Chinese multimodal contexts remains limited, lacking high-quality corpora, analysis, and methods for detection. This work focuses on self-disclosure behaviors on Chinese multimodal social media platforms and constructs a high-quality text-image corpus to address this critical data gap. We systematically analyze the distribution of self-disclosure types, modality preferences, and their relationship with user intent, uncovering expressive patterns unique to the Chinese multimodal context. We also fine-tune five multimodal large language models to enhance self-disclosure detection in multimodal scenarios. Among these models, the Qwen2.5-omni-7B achieved a strong performance, with a partial span F1 score of 88.2%. This study provides a novel research perspective on multimodal self-disclosure in the Chinese context.

1 Introduction

An increasing number of users engage in self-disclosure within online social platforms, openly sharing information about themselves (Cozby, 1973; Wang et al., 2016). The types of disclosed information are diverse, ranging from sensitive attributes such as health status (De Choudhury and De, 2014), gender (Mejova and Hommadova Lu, 2023), and sexual orientation to less sensitive content like emotional expressions, life experiences, and personal preferences. Online self-disclosure is often driven by multiple motivations, including the desire for belonging, emotional support, social connection, and adherence to community norms and interaction conventions (Luo and Hancock, 2020; Lee et al., 2023). While self-disclosure offers certain psychological and social benefits, it also carries significant privacy risks (Wood et al., 2014; Tay et al., 2018). In particular, under unintentional circumstances, users may inadvertently reveal sensitive personal information, which, once identified or misused by others, can lead to adverse real-world consequences such as harassment, discrimination, or limited employment opportunities. Moreover, as multimodal contents within online communities continue to proliferate, the dimensions of potential privacy violations become increasingly complex, making identifying and mitigating these emerging risks crucial.

Existing research on self-disclosure has primarily concentrated on single-modal textual content within English-language contexts (Valizadeh et al., 2021; Cho et al., 2022; Staab et al., 2024). There is a notable lack of studies regarding self-disclosure in other language environments, particularly in the Chinese context. As multimodal content becomes increasingly prevalent on social media platforms, users often rely on a combination of images, text, emojis, and other media to express themselves, which raises higher demands for self-disclosure detection models. Traditional approaches primarily rely on single-text modality models such as BERT and RoBERTa, which exhibit limited capability in modeling cross-modal semantic correlations (Dou et al., 2024; Haq et al., 2025), leading to suboptimal performance in multimodal data processing tasks. To more accurately characterize self-disclosure expressions in Chinese multimodal social media and identify potential privacy risks associated with them, there is an urgent need for more in-depth and systematic research.

In our work, we design a crowdsourcing task to systematically collect user-generated multimodal (text and image) posts related to personal information from the rednote platform, thereby constructing a multimodal corpus. Based on this corpus, we leverage the existing self-disclosure category

*Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 21510–21527

framework (Dou et al., 2024) and conduct an in-depth analysis of the distribution patterns and expressive forms of various types of self-disclosure. Finally, we fine-tune Multimodal large language models (MLLMs) to develop an automated self-disclosure detection model to alert users to potential privacy risks.

Specifically, we create a high-quality dataset with human annotations on 4,870 rednote multimodal posts[1]. We systematically analyse their overall distribution patterns, modality-specific posting preferences, inter-category correlations, and distributional variations under user intent scenarios. With this corpus, we fine-tune five MLLMs to identify the self-disclosures in the given post, achieving over 80% partial span F1 and select the best-performing model as an automatic tool for identifying self-disclosures, aiming to assist users in detecting potential privacy risks. Our key contributions include:

- We construct a brand-new Chinese multimodal self-disclosure corpus, which includes over 4,870 user-generated texts and image contents from the rednote social platform, filling the research gap where no Chinese multimodal self-disclosure corpus existed.
- We construct a social media posting motivation framework based on multiple psychological theories and analyze self-disclosure patterns, including overall distribution, modality preferences, inter-category correlations, and motivation-driven variations to address the lack of analysis in multimodal scenarios.
- We fine-tune five pre-trained MLLMs, benchmark the task of self-disclosure detection on Chinese social media, preliminarily address the challenge in existing research of effectively capturing multimodal information. It offers a new view on privacy identification in Chinese multimodal social media settings.

2 Related Work

Online Self-Disclosure Types Previous studies on self-disclosure have predominantly focused on a single entity category within specific communities, such as healthcare or mental health(Balani and De Choudhury, 2015; Klein et al., 2017; Benton et al., 2017; Yates et al., 2017; Reuel et al., 2022; Valizadeh et al., 2023). Although some research has begun to explore multi-category self-disclosure (Blose et al., 2021; Maddela et al., 2023), much of it remains tied to specific event-driven contexts, such as online discussions during the Covid-19 pandemic. Only a limited number of studies have attempted to systematically expand the scope to cover a broader range of personal information (Dou et al., 2024; Haq et al., 2025), encompassing up to 19 distinct categories. These studies primarily focus on text-only modalities in English-language contexts, with limited systematic exploration of self-disclosure in Chinese multimodal settings. We construct a Chinese multimodal corpus aimed at advancing the identification and analysis of self-disclosure behaviors within Chinese multimodal contexts.

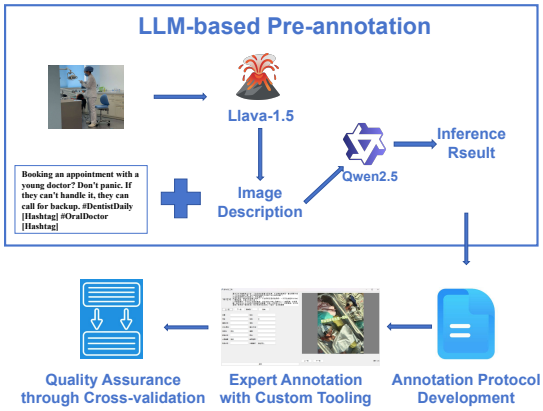

Figure 1: Annotation process includes four steps: LLM-based pre-annotation, annotation protocol development, expert annotation with custom tooling, and quality assurance through cross-validation.

3 Chinese Multimodal Self-Disclosure Dataset

We have constructed a small-scale Chinese self-disclosure multimodal dataset, which includes 4,870 user-generated posts combining text and images, shared on the Rednote social media platform.

Crowdsourcing data collection We selected rednote[2] as the platform for social media data collection because its content is lifestyle-oriented and emphasizes sharing personal experiences. Unlike platforms that primarily feature commercial promotions or entertainment content, rednote encourages users to actively document and discuss various aspects of daily life, such as interpersonal relationships, health conditions, educational experiences,

[1]The dataset contains personal and sensitive information. Access is restricted; please contact the corresponding author via email if needed.

[2]https://www.xiaohongshu.com/explore

Category	Attribution	Social Post	Identified Value
Personal Identity	Name	The picture is a medical record, including information Name: Wang Yuanyuan Gender: Female Age: 29 years old Clinic number: 2023019616	Wang Yuanyuan
	Age	The picture is a medical record, including information Name: Wang Yuanyuan Gender: Female Age: 29 years old Clinic number: 2023019616	29
	Gender	I started dating with my male roommate. I met this good man #Daily love of two boys	Male
	Location	I love Suzhou and Xiangcheng. I just moved here for half a month and I successfully made an appointment.	Suzhou
	Sexual Orientation	I started dating with my male roommate. I met this good man #Daily love of two boys	Homosexual
Family and Social Relationship	Marital Status	Don't worry, my future boyfriend, I'm still reading literature in the library during National Day.	Unmarried
	Husband/Boyfriend	Don't worry, my future boyfriend, I'm still reading literature in the library during National Day.	None
	Wife/Girlfriend	Family, we had a two-hour video call tonight. Family, family, I am almost hooked by my wife. She has asked me to call her sister many times before, but I was so shy that I couldn't open my mouth. Today, in order to video chat with me, she actually called me "sister". The point is, she hopes that we will end our long-distance relationship sooner!	Girlfriend
	Family Status	Since the beginning of my pregnancy, I have been thinking about how I can be a good mother after my two babies are born.	Two children
	Pet	The first time I went to the flower, bird and fish market, I brought back a golden sun. I didn't know anything, so I started by feeding it, watching videos while making milk powder. Let's give the parrot a name first.	Parrot
Health	Physical Health	I was diagnosed with advanced cancer at the age of 35 and I am now 41. It is fate that I met you all in this vast crowd of people. Thank you all for your support and encouragement. Come on to yourself, don't give up!	Cancer
	Mental Health	I was diagnosed with moderate depression and anxiety at the end of September 2021.	Moderate depression and anxiety
Occupation and Education	Occupation	I have finished sewing the grapefruit peel, but I feel like my hands are still not satisfied. What should I sew next time? Give me a new challenge. #Surgeon#	Surgeon
	Education Information	After I finish my master's degree, I will work hard to get a doctorate.	Master student
Economic and Financial Status	Financial Status	State-owned and private enterprises. Monthly income of 40,000 yuan, taking advantage of the opportunity is saving money∼	Monthly income: 40,000 yuan

Table 1: The table presents a multidimensional self-disclosure attribute classification framework, consisting of five categories: personal identity information, family and social relationship information, health information, occupation and education information, and economic and financial status information.

and consumption habits. Therefore, the platform is well-suited for collecting authentic, life-related social media data.

We designed a crowdsourcing task and recruited a team of 20 university students with rich experience using rednote from college communities. The task's goal was to manually collect social media posts with life-related attributes from rednote, excluding content containing advertising links and brand sponsorship statements. Participants were advised to prioritize posts from users sharing their personal lives. They were required to collect information, including post titles, main content, and related images. A payment of $2 was paid for each successful post. We successfully collected and curated 4,870 high-quality multimodal (text-image) posts. Further details about the crowdsourcing task can be found in the Appendix A.

Self-Disclosure Classification We drew on the self-disclosure attribute classification proposed in previous research (Dou et al., 2024), removed specific inapplicable attributes, and reorganized the remaining ones to construct a new classification framework consisting of five core categories and fifteen key personal information attributes. Table 1 lists the definitions of these attributes, categorized into five distinct groups: personal identity information, family and social relationship information, health information, occupation and education information, and economic and financial status information, which comprehensively cover various types of personal information in real life. The detailed definitions are shown in the Appendix B.

Data annotation We designed an annotation process to ensure quality and privacy standards, as shown in the figure 1.

Step1: LLM-based Pre-annotation We implemented a multimodal annotation pipeline beginning with visual analysis. Images were processed through the LLaVA-1.5 [3](Liu et al., 2023) model using structured prompts to generate comprehensive visual descriptions. These image captions were subsequently aligned with their corresponding textual content to form integrated data inputs. The combined inputs were then fed into the Qwen2.5-7B-Instruct [4](Yang et al., 2024) model for structured annotation generation, adhering to predefined output templates. Detailed prompt engineering specifications are available in Appendix C.1.

Step2: Annotation Protocol Development Leveraging preliminary annotations from Step 1, we established a rigorous annotation framework document. This protocol precisely defines: (1) personal information taxonomy with operational definitions, (2) attribute value domains, and (3) multimodal evidence integration rules. The document ensures consistent interpretation of explicit data elements and implicit inferences across modalities. Complete annotation guidelines are provided in Appendix C.2.

Step3: Expert Annotation with Custom Tooling We independently designed and developed a practical annotation tool, which integrates image and text viewing with annotation functions, displaying the images, text, and annotation options for self-discourse attributes in a single window. By optimizing the interface design, annotators can quickly browse pictures and text in the same window and annotate each entry directly. We hired six annotators from our internal team. We divided them into two groups, which received training tutorials and annotation example exercises to carry out the formal data annotation. We required annotators to maintain confidentiality of the annotated content and analyze the post content based on the annotation guidelines document, including intuitive presentations and inferable personal attributes.

Category	Human annotations (%)	LLM pre-annotations (%)
Location	74.23	57.70
Name	55.34	44.46
Age	84.52	62.47
Gender	100	87.78
Marital Status	91.02	67.35
Pet	82.30	81.73
Husband/Boyfriend	87.91	73.27
Wife/Girlfriend	89.25	87.89
Sexual Orientation	92.58	77.62
Physical Health	74.85	69.32
Family Status	89.26	78.78
Occupation	73.61	63.04
Mental Health	67.37	65.79
Education Information	73.23	62.26
Financial Status	75.08	54.30
Overall	**80.64**	**68.92**

Table 2: The table shows both the inter-annotator agreement scores for human annotations and LLM pre-annotations, calculated using the Two Agree method.

Step 4: Quality Control We performed consistency verification on the two sets of human-annotated results, as well as between the large language model's pre-annotations and the final annotated results, using the **Two Agree** method (Dou et al., 2024) to ensure the reliability and accuracy of the annotations. The calculation formula is as

[3]https://github.com/haotian-liu/LLaVA

[4]https://huggingface.co/Qwen/Qwen2.5-7B-Instruct

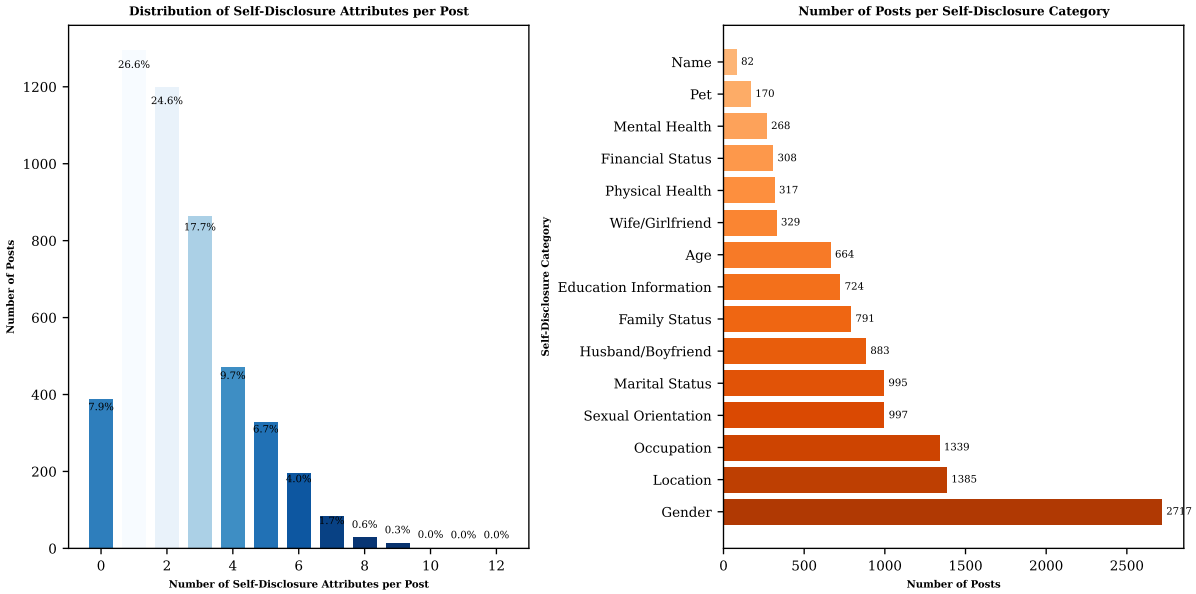

Figure 2: Distribution differences in self-disclosure categories on social media platforms

follows:

$$\text{Two Agree} = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|}$$

where S_1 and S_2 denote the sets of words annotated by the two groups, $|S_1 \cap S_2|$ represents the number of words both groups labeled as disclosure (the intersection), and $|S_1 \cup S_2|$ represents the total number of unique words labeled as disclosure by either annotator (the union). The results are shown in Table 2, the overall performance of the pre-annotations is satisfactory, largely due to the strong reasoning capabilities of the large language model. However, because of the complexity of Chinese semantics and the implicit nature of information conveyed between images and text, the pre-annotations still show certain omissions and shortcomings.

4 Self-disclosure exploration

Distribution of Self-Disclosure Types Figure 2 presents a statistical analysis of the number of self-disclosed attributes within individual social media posts. The analysis reveals that the vast majority of posts (approximately 59%) contain only one to two self-disclosed attributes, indicating that the scope of personal information users reveal in a single post is typically limited. As the number of attributes contained in a post increases, its frequency of occurrence declines significantly, suggesting that posts containing multidimensional and rich personal information are relatively rare. This finding reflects a prevalent behavioral pattern on social media: users tend to engage in moderate information sharing, while in-depth, multi-faceted self-disclosure is less common.

Further analysis, also depicted in the figure, finds that the distribution of different self-disclosure categories shows significant variation. Among these, the disclosure of gender information is the most prominent, potentially because users unconsciously employ gender-specific expressions when describing daily activities, professional experiences, or personal interests. In contrast, the disclosure rate for sensitive information, such as financial status and mental health, is markedly lower. This phenomenon can be primarily attributed to two factors: first, users actively avoid publicizing such private information; second, these topics are more likely to be discussed within private social circles or on anonymous platforms. Furthermore, information related to names is the most rarely disclosed. This can be explained from two perspectives: first, when describing social relationships, users are more inclined to use relational pronouns (e.g., "my friend," "my colleague") rather than specific names; second, mainstream social media platforms generally permit users to register with nicknames or pseudonyms, which effectively reduces the risk of exposing their real names.

Distribution Differences Across Different Modalities. In addition to text-based content, personal information can be disclosed through

Figure 3: The post's main content is "Happy times with my roommate. Five years of breaking up and getting back together—thank you for your love."

images or a combination of text and images. According to the annotation results, gender is the most commonly disclosed personal information through images. When users share their life updates, they often include photos of themselves. In cases where gender is not explicitly mentioned, the visual information in the image frequently provides sufficient clues for inferring gender. Furthermore, information related to sexual orientation often necessitates the integration of both textual and visual modalities for accurate inference. For instance, a post, as shown in figure 3 depicting two pairs of men's hands forming a heart shape, paired with the caption "Happy times with my roommate. Five years of breaking up and getting back together—thank you for your love," conveys sexual orientation only through the combined interpretation of image and text.

Inter-Category Correlations of Self-Disclosure The correlation analysis presented in Figure 4 reveals significant associations among different categories of self-disclosure, particularly between marital status, sexual orientation, and partner relationships. For example, when users disclose information involving terms like "husband/boyfriend," there is a 57% likelihood of inferring their marital status and a 39% likelihood of inferring their sexual orientation. This correlation stems from the strong semantic coupling among these attributes—mentioning "husband/boyfriend" inherently implies both marital status and sexual orientation. It is important to note that this study adopts a binary existence determination standard when annotating relationship-related terms such as "husband/boyfriend" and "wife/girlfriend." Under this standard, any mention of such relationships is treated as a disclosure, regardless of whether a specific identity is identified. Given the strong interconnections between these attributes, the privacy risk is significantly heightened: a single piece of

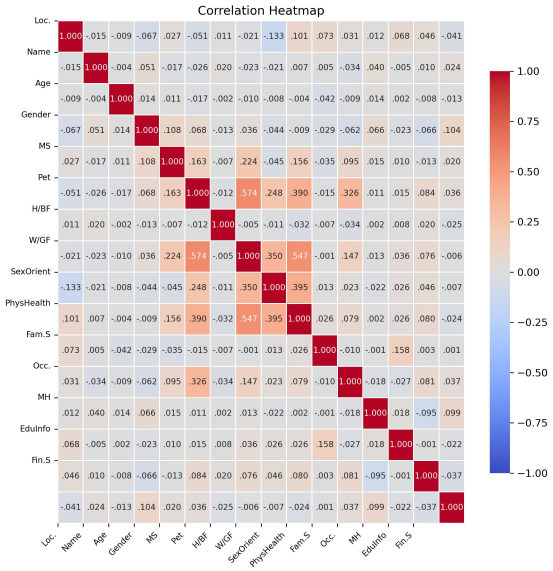

Figure 4: Heatmap depicting the correlations between different self-disclosure categories. The colors represent the strength of the relationship, with darker shades indicating stronger correlations. This analysis illustrates how attributes such as marital status, sexual orientation, and partner relationships are semantically linked and influence the inferability of other personal information.

disclosed information may trigger a chain of inferences, leading to the indirect exposure of multiple sensitive attributes. For instance, the statement "celebrating an anniversary with my husband" not only indicates marital status and heterosexual orientation but may also suggest details about fertility, family planning, or even age range.

Differentiated Distribution of Self-Disclosure Under Different Motivations The degree of self-disclosure in user-generated content varies significantly depending on users' underlying motivations for posting on social media platforms. To gain a deeper understanding of this phenomenon, this study draws upon several classic psychological theories, including Maslow's hierarchy of needs (Maslow, 1943), the theory of emotional social sharing (Harber and Cohen, 2005), self-determination theory (Deci and Ryan, 2012), and social identity theory (Islam, 2014). Based on these theoretical frameworks, we identify and construct six core categories of user posting motivations: **Social Media Posting Motivation Framework**, as summarized in Table 3.

As shown in Figure 5, user-generated posts driven by different motivations exhibit significant variability in self-disclosure. Specifically, posts categorized under "Emotional Release and Reso-

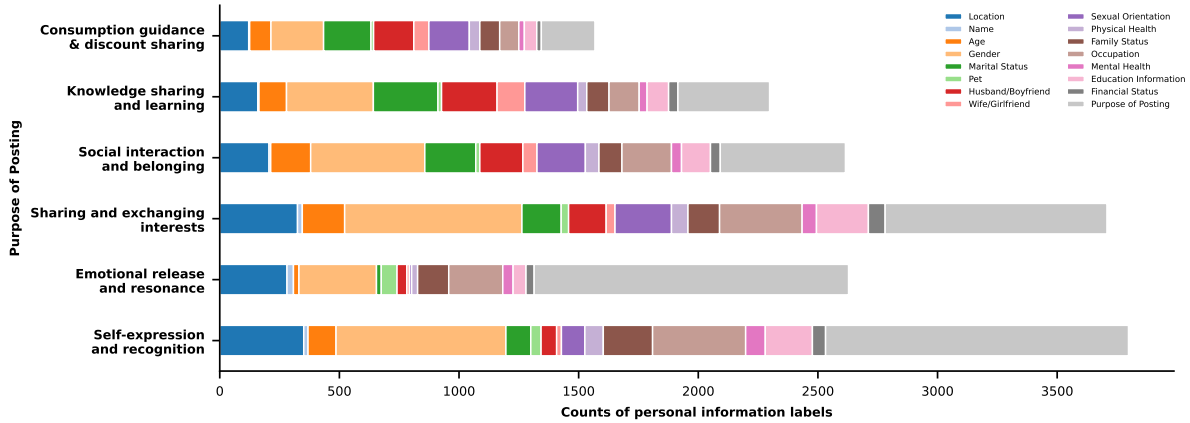

Figure 5: Variations in self-disclosure attributes across different posting motivations. Posts related to emotional expression and social interaction reveal more sensitive information (e.g., gender, marital status), while those focused on knowledge sharing and consumption show less personal disclosure.

Motivation	Description
Self-Expression and Recognition	Includes achievements, honors, workplace, school, and family information.
Emotional Release and Resonance	Involves emotional experiences, family conflicts, and relationship issues.
Interest Sharing and Exchange	Share hobbies, frequent places, and participation in activities.
Knowledge Sharing and Learning	Posts professional knowledge, work background, and research fields.
Consumption Guidance and Discount Sharing	Covers discount information and consumption habits.
Social Interaction and Sense of Belonging	Shares photos, contact info, and social activities.

Table 3: This table summarizes six core categories of user motivations for content posting on social media platforms, along with brief descriptions of each category.

nance" and "Social Interaction and Sense of Belonging" show higher levels of personal identifiers being disclosed, such as gender, marital status, intimate relationships (husband/boyfriend, wife/girlfriend), and sexual orientation. This phenomenon can be attributed to users' tendency to disclose personal narratives about intimate relationships and marital experiences when seeking emotional resonance or social recognition.

In contrast, posts associated with "Knowledge Sharing and Learning" and "Consumption Guidance and Discount Sharing" exhibit a markedly different pattern of disclosure. These categories predominantly involve occupational, educational, and financial information, indicating that users engaged in disseminating professional knowledge or commercial decision-making are less likely to expose personally identifiable details. This distinction highlights the significant differences in self-disclosure distribution under varying posting motivations, providing valuable insights for understanding and protecting online self-disclosure behaviors.

5 LLM-Based Self-Disclosure Detection Model

We frame the self-disclosure detection task as a question-answering task (specific examples are detailed in Appendix D.1) and fine-tune five multimodal large language models using an annotated corpus of constructed self-disclosure. Considering the limitations of resources, we selected models with a parameter size below 11 B. We chose those that performed relatively well on other tasks for fine-tuning: Gemm3-4b-it [5](Team, 2025), InternVL3-8B [6](Zhu et al., 2025), Llama-3.2-11B-Vision [7], Qwen2.5-VL-7B-Instruct and Qwen2.5-Omni-7B [8](Xu et al., 2025). Additionally, we further fine-tuned the RoBERTa model(Cui et al., 2020)[9] using the fused results of image-to-text conversion and original text generated during the annotation process. In this setup, the self-disclosure identification task is formulated as a sequence labeling problem to locate self-disclosure information in the text precisely.

[5]https://huggingface.co/google/gemma-3-4b-it
[6]https://huggingface.co/OpenGVLab/InternVL3-8B
[7]https://huggingface.co/meta-llama/Llama-3.2-11B-Vision
[8]https://huggingface.co/Qwen/Qwen2.5-Omni-7B
[9]https://huggingface.co/hfl/chinese-RoBERTa-wwm-ext

Model	Random Split			Stratified Split		
	Token F1	Span F1	Partial F1	Token F1	Span F1	Partial F1
Finetune						
Gemm3-4b-it	77.97	86.19	87.09	54.08	67.21	67.43
InternVL3-8B	78.00	86.24	87.21	65.86	**77.30**	**77.66**
Llama-3.2-11B-Vision	77.27	85.50	86.02	**70.30**	75.16	75.33
Qwen2.5-VL-7B-Instruct	75.61	85.08	85.95	68.14	76.88	76.99
Qwen2.5-Omni-7B	**79.37**	**87.35**	**88.20**	63.36	76.25	76.61
RoBERTa	75.66	83.13	83.13	55.06	62.86	62.86
No Finetune						
Gemm3-4b-it	63.59	75.54	76.32	53.43	66.67	66.89
InternVL3-8B	72.29	81.19	81.85	64.72	74.54	74.69
Llama-3.2-11B-Vision	65.99	74.00	74.46	65.92	74.17	74.17
Qwen2.5-VL-7B-Instruct	73.94	81.78	82.12	67.81	76.78	76.92
Qwen2.5-Omni-7B	73.31	81.60	82.15	65.83	75.24	75.39

Table 4: Performance comparison of various models on privacy inference tasks under random split and multi-label stratified split. The results exhibit an overall trend: fine-tuned models generally outperform their non-fine-tuned counterparts, and Chinese large models demonstrate relatively stronger performance on Chinese datasets.

Dataset split We use a random split to divide the data into train and test sets(train:4383, test:487). Additionally, we introduce a multi-label stratified split to construct another pair of training, test, and validation sets. The label distributions for both splitting strategies are shown in appendix D.2. Multi-label stratified sampling better preserves the consistency of label proportions between the training and test sets and the original dataset. For example, key labels such as "Location," "Gender," and "Occupation" maintain distributions in the split datasets that are very close to the original data, indicating the method's effectiveness in controlling label distribution. In contrast, random splitting leads to larger fluctuations in some labels—for instance, the proportions of "Education Information" and "Sexual Orientation" in the test set deviate more significantly from the original distribution, which may negatively impact the model's generalization ability on these labels.

Experiment settings We used four A100 80G GPUs and fine-tuned several multimodal large language models using LlamaFactory [10](Zheng et al., 2024). Additionally, we fine-tuned the RoBERTa model using two A100 80G GPUs.

We fine-tuned all multimodal large language models using LoRA. We enabled Flash Attention 2 (flash_attn2) and DeepSpeed ZeRO Stage 3 to improve memory efficiency. These two optimizations significantly enhanced both the training speed and scalability.

[10]https://github.com/hiyouga/LLaMA-Factory

For the RoBERTa model, we adopt a sentence segmentation combined with a sliding window approach to handle long-text inputs and overcome token length limitations.

More parameter settings are in the appendix D.3.

Models Performances We adopted Partial F1 as the primary evaluation metric. Compared to the traditional Token-level F1(Token F1), this metric emphasizes the model's ability to identify the boundaries of disclosed information. It allows for slight boundary deviations, preventing the model from being completely penalized due to minor boundary mismatches. Specifically, a prediction is considered valid if the predicted disclosure span overlaps with the reference span and the overlapping portion exceeds half of the longer span.

Table 4 summarizes the performance of various models on the self-disclosure test set. The evaluation results indicate that the fine-tuned multimodal large language models perform decently across all three metrics (Token F1, Span F1, and Partial F1). Although there are differences between the stratified sampling and random sampling test sets, resulting in some variation in results, the overall trend remains consistent: finetuned models generally outperform their non-finetuned counterparts, and Chinese large models demonstrate relatively stronger performance on Chinese datasets. It is worth noting that despite the different test sets, all models show limited performance in recognizing self-disclosure categories, indicating substantial room for improvement.

Model	Partial F1 (Sexual Orientation)
Finetune	
Gemm3-4b-it	71.43
InternVL3-8B	66.67
Llama-3.2-11B-Vision	41.67
Qwen2.5-VL-7B-Instruct	50.00
Qwen2.5-Omni-7B	75.00
RoBERTa	0.00
No Finetune	
Gemm3-4b-it	33.33
InternVL3-8B	0.00
Llama-3.2-11B-Vision	25.00
Qwen2.5-VL-7B-Instruct	0.00
Qwen2.5-Omni-7B	25.00

Table 5: Partial F1 scores for Sexual Orientation on implicit self-disclosure posts.

Analysis of More "Implicit" Types of Self-Disclosure We define "more implicit" self-disclosure as cases where neither text nor image alone provides sufficient semantic information to determine the disclosure category, and inference must be made by combining both modalities. According to this criterion, we find that such implicit disclosures most frequently appear in the Sexual Orientation attribute, with 148 related posts identified, accounting for approximately 3.04% of the total sample of 4,870 posts. During the model fine-tuning phase, we created a test set consisting of 487 posts, among which 16 contain this type of implicit disclosure. We also report the performance of various models on these posts.

Experimental results in table 5 demonstrate that fine-tuning significantly enhances the ability of multimodal models to understand implicit information, particularly in tasks requiring cross-modal reasoning, such as identifying disclosures related to sexual orientation. After fine-tuning, Qwen2.5-Omni-7B achieved an F1 score of 75.00, showcasing strong capabilities in modality fusion and semantic inference. Similarly, Gemm3-4b-it and InternVL3-8B reached F1 scores of 71.43 and 66.67, respectively, far surpassing their non-fine-tuned counterparts (33.33 and 0.00). In contrast, the text-only model RoBERTa completely failed on this task , highlighting the limitations of unimodal approaches in handling implicit expressions. Moreover, models exhibit varying degrees of sensitivity to such implicit disclosures, with Qwen2.5-Omni-7B again standing out due to its superior performance. These implicit self-disclosures impose higher demands on model capabilities, requiring not only the ability to "see" and "understand" but also to apply appropriate protective measures to safeguard the disclosed information.

6 Conclusion

We systematically constructed a novel Chinese multimodal self-disclosure corpus based on social media data. We conducted an in-depth analysis of the distribution patterns and expressive forms of various types of self-disclosure. It provides a solid data foundation and theoretical reference for subsequent self-disclosure recognition and protection efforts. Building on this corpus, we fine-tuned several multimodal large language models and evaluated their performance using partial span-level F1 scores. Experimental results show that all fine-tuned LLMs achieved a partial span F1 score exceeding 80%. We ultimately best fine-tuned LLM as the automatic recognition model to assist users in detecting potential privacy leakage risks.

Limitation

We recognize that the dataset exhibits imbalanced sample distributions, which may affect the model's generalization capability. In the future, we will start from data collection to expand the dataset, with a focus on balancing attribute distributions. Despite certain limitations, our work addresses the scarcity of Chinese multimodal self-disclosure data, significantly enhancing the value of the dataset and enriching the linguistic and cultural diversity of resources available to the research community.

Ethics Statement

Data collection approval was received from an ethics review board. All codes and data used in this paper comply with the license for use.

Acknowledgments

This paper was supported by the National Natural Science Foundation of China (Project No. 62202075, 72274096, 71774084, 72301136, and 72174087), the Foreign Cultural and Educational Expert Program of the Ministry of Science and Technology of China (G2022182009L), the Natural Science Foundation of Chongqing, China (No. CSTB2022NSCQ-MSX1404), Fundamental Research Funds for the Central Universities (No. SWU-KR24008), Key Laboratory of Data Science and Smart Education, Hainan Normal University, Ministry of Education (No. DSIE202206).

References

Sairam Balani and Munmun De Choudhury. 2015. Detecting and characterizing mental health related self-disclosure in social media. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '15, page 1373–1378, New York, NY, USA. Association for Computing Machinery.

Adrian Benton, Margaret Mitchell, and Dirk Hovy. 2017. Multitask learning for mental health conditions with limited social media data. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 152–162, Valencia, Spain. Association for Computational Linguistics.

Taylor Blose, Prasanna Umar, Anna Squicciarini, and Sarah Rajtmajer. 2021. A study of self-disclosure during the coronavirus pandemic. *First Monday*, 26(7).

Won Ik Cho, Soomin Kim, Eujeong Choi, and Younghoon Jeong. 2022. Assessing how users display self-disclosure and authenticity in conversation with human-like agents: A case study of luda lee. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 145–152, Online only. Association for Computational Linguistics.

Paul C. Cozby. 1973. Self-disclosure: a literature review. *Psychological bulletin*, 79 2:73–91.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 657–668, Online. Association for Computational Linguistics.

Munmun De Choudhury and Sushovan De. 2014. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1):71–80.

Edward L Deci and Richard M Ryan. 2012. Self-determination theory. *Handbook of theories of social psychology*, 1(20):416–436.

Yao Dou, Isadora Krsek, Tarek Naous, Anubha Kabra, Sauvik Das, Alan Ritter, and Wei Xu. 2024. Reducing privacy risks in online self-disclosures with language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13732–13754, Bangkok, Thailand. Association for Computational Linguistics.

Ehsan-Ul Haq, Shalini Jangra, Suparna De, Nishanth Sastry, and Gareth Tyson. 2025. Unpacking the layers: Exploring self-disclosure norms, engagement dynamics, and privacy implications. *Preprint*, arXiv:2502.10701.

Kent D Harber and Dov J Cohen. 2005. The emotional broadcaster theory of social sharing. *Journal of Language and Social Psychology*, 24(4):382–400.

Gazi Islam. 2014. *Social Identity Theory*, pages 1781–1783. Springer New York, New York, NY.

Ari Klein, Abeed Sarker, Masoud Rouhizadeh, Karen O'Connor, and Graciela Gonzalez. 2017. Detecting personal medication intake in Twitter: An annotated corpus and baseline classification system. In *BioNLP 2017*, pages 136–142, Vancouver, Canada,. Association for Computational Linguistics.

Jooyoung Lee, Sarah Rajtmajer, Eesha Srivatsavaya, and Shomir Wilson. 2023. Online self-disclosure, social support, and user engagement during the covid-19 pandemic. 6(3–4).

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved baselines with visual instruction tuning.

Mufan Luo and Jeffrey T. Hancock. 2020. Self-disclosure and social media: motivations, mechanisms and psychological well-being. *Current Opinion in Psychology*, 31:110–115. Privacy and Disclosure, Online and in Social Interactions.

Mounica Maddela, Yao Dou, David Heineman, and Wei Xu. 2023. LENS: A learnable evaluation metric for text simplification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16383–16408, Toronto, Canada. Association for Computational Linguistics.

Abraham Harold Maslow. 1943. A theory of human motivation. *Psychological review*, 50(4):370.

Yelena Mejova and Anya Hommadova Lu. 2023. Gender in the disclosure of loneliness on twitter during covid-19 lockdowns. *Frontiers in Digital Health*, Volume 5 - 2023.

Ann-Katrin Reuel, Sebastian Peralta, João Sedoc, Garrick Sherman, and Lyle Ungar. 2022. Measuring the language of self-disclosure across corpora. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1035–1047, Dublin, Ireland. Association for Computational Linguistics.

Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. 2024. Beyond memorization: Violating privacy via inference with large language models. In *The Twelfth International Conference on Learning Representations*.

Stacie Tay, Kat Alcock, and Katrina Scior. 2018. Mental health problems among clinical psychologists: Stigma and its impact on disclosure and help-seeking. *Journal of Clinical Psychology*, 74(9):1545–1555.

Gemma Team. 2025. Gemma 3.

Mina Valizadeh, Xing Qian, Pardis Ranjbar-Noiey, Cornelia Caragea, and Natalie Parde. 2023. What clued the AI doctor in? on the influence of data source and quality for transformer-based medical self-disclosure detection. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1201–1216, Dubrovnik, Croatia. Association for Computational Linguistics.

Mina Valizadeh, Pardis Ranjbar-Noiey, Cornelia Caragea, and Natalie Parde. 2021. Identifying medical self-disclosure in online communities. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4398–4408, Online. Association for Computational Linguistics.

Yi-Chia Wang, Moira Burke, and Robert Kraut. 2016. Modeling self-disclosure in social networking sites. CSCW '16, page 74–85, New York, NY, USA. Association for Computing Machinery.

Benjamin T Wood, Olivia Bolner, and Phillip Gauthier. 2014. Student mental health self-disclosures in classrooms: Perceptions and implications. *Psychology Learning & Teaching*, 13(2):83–94.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Andrew Yates, Arman Cohan, and Nazli Goharian. 2017. Depression and self-harm risk assessment in online forums. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2968–2978, Copenhagen, Denmark. Association for Computational Linguistics.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, and 32 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *Preprint*, arXiv:2504.10479.

A Crowdsourcing data collection task

A.1 Pre-screening Procedure

We implemented a pre-screening procedure before the formal launch of the crowdsourcing task to ensure that participants possessed sufficient experience with rednote and the ability to identify personal life-related content accurately. The procedure consisted of the following two stages:

Questionnaire Screening. We designed a structured online questionnaire to identify participants with sufficient experience using the Rednote platform and the ability to recognize life-related content. The questionnaire consists of two parts: the first assesses platform familiarity, and the second evaluates content recognition ability.

Part I: Platform Usage Experience

- **Q1. How frequently do you use rednote?**
  A. Daily B. 3–6 times per week C. Occasionally D. Rarely or never
- **Q2. Approximately which year did you start using rednote?** (Open-ended)
- **Q3. Have you ever posted original content on rednote?**
  Yes / No
- **Q4. What types of content do you usually follow on Rednote?** (Multiple choice)
  A. Emotional expression B. Study/life logs C. Product reviews and recommendations D. Campus gossip E. Advertisements or prize draws

Part II: Content Identification Ability

- **Q5. Which of the following are more likely to be categorized as "life-sharing" posts on Rednote?** (Multiple choice)
  A. "Looking for recommendations on affordable quality earphones"
  B. "Fought with my boyfriend today, feeling down."
  C. "Click the link to join a giveaway livestream!"
  D. "100 days left before the graduate entrance exam—time to go all in!"
- **Q6. Briefly describe what you think are the key characteristics of a "life-sharing" post.** (Within 50 characters)

Category	Example Content
Emotional expression	"I had another argument with my roommate today. So frustrating."
Study/Work status	"30 days left until the entrance exam—library grind every day."
Consumption sharing	"This affordable perfume smells amazing! Had to share with the girls."

Table 6: Examples of life-related post content

Training and Agreement Signing. After a comprehensive evaluation, we distributed the questionnaire through online communities and successfully recruited 20 university students to participate in the task. All selected participants were required to attend an online operational training session covering the task objectives, data collection boundaries, privacy protection protocols, and the content review workflow.

Upon completion of the training, all participants signed a confidentiality and data usage agreement, pledging the following:

- Not to use automated tools such as web crawlers or any non-human means for data collection;
- Not to delegate the task to any third party;
- To collect only publicly available content from the Rednote platform, excluding private or paywalled information.

A.2 Data collection types and format

The crowdsourcing task was to collect social media posts with identifiable life-related attributes. We defined "life-related attributes" as naturalistic narratives by users about their personal life experiences, emotional expressions, social interactions, learning activities, or consumption behaviors to ensure consistency in understanding and execution. Representative examples are provided in the Table 6. To ensure the authenticity of the data and eliminate commercial interference, we explicitly excluded posts containing the following:

- Content featuring brand names, promotional codes, or shopping links;
- Posts explicitly or implicitly indicating brand sponsorship, giveaways, or product seeding;
- Advertisements or reposts using templated or promotional language.

We also specified the required submission fields, as outlined in Table 7.

Field Name	Description
Post Title	The title or headline of the post, typically a user-generated summary.
Post Content	The main textual body of the post, describing personal experiences, thoughts, or daily activities.
Image Files	The source files of any associated images.

Table 7: Required Submission Fields for Collected Posts

A.3 Quality Control and Incentive Mechanism

We manually reviewed the content submitted by each participant, rigorously filtering out low-quality samples that did not meet the required standards to ensure overall data quality. We implemented an incentive mechanism for the validated submissions, rewarding contributors at a rate of $2 per approved post. As a result, we successfully collected and curated 4,870 high-quality multimodal (text-image) posts.

B Self-disclosure attribute classification

We systematically categorized the types of self-disclosure and provided clear definitions for each category, as shown in Table 8.

C Corpus annotation

C.1 LLM-based Pre-annotation prompt

Image_description_prompt We employ the LLaVA-1.5 model to perform image captioning for multimodal posts and design the following prompts to guide the model in generating more accurate descriptions.

> You are now an image describer. Please describe the content of the image in Chinese.
>
> Please avoid using any language other than Chinese, unless the image explicitly contains text in another language.

Category	Attribution	Definition
Personal Identity	Name	A person's real name, including the poster's name, which can directly point to a specific individual.
	Age	A person's life age status, including exact age, age range, traditional age terms, age group, period, or birth date descriptions.
	Gender	A person's physiological gender or gender identity status, including clear gender or unknown gender descriptions.
	Location	Descriptions of geographic locations related to a person, including place names, addresses, geographic coordinates, administrative divisions, landmarks, transportation routes, and other specific details.
	Sexual Orientation	A person's emotional and sexual attraction tendencies to genders, including heterosexual, homosexual, bisexual, asexual, and other sexual orientations.
Family and Social Relationship	Marital Status	A person's marital relationship status, including unmarried, married, widowed, divorced, and unknown marital status descriptions.
	Husband/Boyfriend	The status of a person's relationship with a male partner, including having a husband, having a boyfriend, no partner, or unknown situation.
	Wife/Girlfriend	The status of a person's relationship with a female partner, including having a wife, having a girlfriend, no partner, or unknown situation.
	Family Status	Descriptions of a person's family-related status, including family composition, family structure, family relationships, economic status, and family employment status.
	Pet	Whether a person owns pets and the specific types of pets (e.g., cat, dog, etc.), or explicit statements of no pets or unknown pet status.
Health	Physical Health	A person's physical health status, including descriptions of being healthy or unhealthy (involving specific physical issues and manifestations).
	Mental Health	A person's mental health status, including healthy and unhealthy descriptions (involving specific mental issues and manifestations).
Occupation and Education	Occupation	A person's occupation title, including job type or unemployed status.
	Education Information	A person's educational background, including academic level or whether they have overseas study experience.
Economic and Financial Status	Financial Status	Descriptions of a person's economic situation, including income, assets, liabilities, consumption level, savings, etc., classified by economic class.

Table 8: The table presents a multidimensional self-disclosure attribute classification framework, consisting of five categories: personal identity information, family and social relationship information, health information, occupation and education information, and economic and financial status information. It includes 15 personal information attributes, such as location, age, sexual orientation, etc.

When describing the image, faithfully record the information presented. Respect the original content and refrain from speculation, summarization, or subjective interpretation.

Do not include introductory phrases such as "This image shows..." at the beginning of the output.

Special reminder: If the image displays a screenshot of a conversation or chat from a social media application, please return "This is a screenshot of a chat conversation" and provide the screenshot's content.

In addition, for any textual content in the image, please transcribe it accurately without paraphrasing or simplifying to ensure completeness and objectivity.

Pre-annotaion prompt We employ the Qwen2.5-7B-Instruct model to perform pre-annotation of self-disclosure categories. To enhance the quality and consistency of the annotations, we design the following **prompt**:

You are an information annotator responsible for analyzing and labeling the poster's private information based on the input text content.

If any item of private information is missing or cannot be determined, return "Unknown" as the value. Do not infer or fabricate any information.

Only return a structured JSON file as the output.

Output Format:

The result must strictly follow the JSON key-value format below:

```
{
  "Location": "",
  "Name": "",
  "Age": "",
  "Gender": "",
  "Marital Status": "",
  "Pet": "",
  "Husband/Boyfriend": "",
  "Wife/Girlfriend": "",
  "Sexual Orientation": "",
  "Health": "",
  "Family Status": "",
  "Occupation": "",
  "Mental Health": "",
  "Education": "",
  "Financial Status": ""
}
```

Instructions:

- Only return the JSON in the specified format. Do not include any other output.
- Input format:

```
{
  "text": "(main text content)
     ",
  "image_description_1": "
     description",
  "image_description_2": "
     description",
  "image_description_3": "
     description"
}
```

- If the output format does not match the JSON structure, return a JSON where all values are "Unknown".
- Each value in the JSON must not exceed 20 Chinese characters. Avoid lengthy explanations.
- Image descriptions may be inaccurate and should weigh less than 20%. Prioritize the main text.
- If there is a conflict between text and image description, follow the text.

Value Constraints for Specific Fields:

- Gender: <Unknown / Male / Female>
- Marital Status: <Unknown / Married / Single>
- Pet: <Unknown / None / Cat / Dog / ...>
- Husband/Boyfriend: <Unknown / None / Has Boyfriend / Has Husband>
- Wife/Girlfriend: <Unknown / None / Has Girlfriend / Has Wife>
- Sexual Orientation: <Unknown / Homosexual / Heterosexual>

C.2 Annotation Protocol documents

Annotators were provided with comprehensive annotation documents during the manual annotation process, as shown in Table 9. It offers precise definitions of various self-disclosure categories, delineates criteria for boundary determination, and presents representative examples.

Self-disclosure category	Value
Location	(A) Place Name Description (B) Address Description (C) Geographic Coordinates (D) Administrative Division Description (E) Landmark Building Description (F) Traffic Route Description *Example: UK, Maldives, Beijing, Wuhan, Lijiang, New Zealand, Ningbo Women and Children's Hospital, etc.*
Name	Real personal name (including names of non-posters)
Age	(A) Exact age (e.g., 20 years old) (B) Age range (e.g., 20–30 years old) (C) Traditional term (e.g., primary school age) (D) Age group (e.g., middle-aged) (E) Period (e.g., "lived for thirty years") (F) Birth year/month (e.g., born in 1988)
Gender	(1) Male (2) Female (3) Unknown
Marital status	(1) Unmarried (2) Married (3) Widowed (4) Divorced (5) Unknown
Pet	(1) Has pets (e.g., cat, dog) (2) No pets (3) Unknown
Husband/Boyfriend	(1) Has husband (2) Has boyfriend (3) None (4) Unknown
Wife/Girlfriend	(1) Has wife (2) Has girlfriend (3) None (4) Unknown
Sexual Orientation	(1) Heterosexual (2) Homosexual (3) Bisexual (4) Asexual (5) Other (6) Unknown
Health	(1) Healthy (2) Non-healthy (e.g., leg disability, chronic disease)
Family Situation*	(A) Family members (e.g., parents, children) (B) Family structure (e.g., single-parent) (C) Family relations (e.g., conflict, harmony) (D) Economic situation (e.g., income level, savings) (E) Occupational status of family
Occupation	Specific occupation name (including student, unemployed)
Mental Health	(1) Healthy (2) Non-healthy (e.g., depression, bipolar)
Educational Information	(A) Educational level: primary school to doctorate (B) Study abroad: international student / not
Financial situation*	Monthly income, assets, liabilities, expenditure By type: wealthy, rich, middle class, well-off, ordinary, poor

Table 9: This table outlines the defined self-disclosure categories and their corresponding annotation boundaries used during the labeling process.

D Further experiment details

D.1 LLMs Finetune detail

We fine-tune the large language model (LLM) using a supervised fine-tuning (SFT) approach. The original raw text, images, and their corresponding annotation results are transformed into a question–answer (QA) format to better align with the model's training paradigm.

QA pair example Here is an example of the constructed QA pair. We follow the LlamaFactory training data format.


```
{
"messages":
[
{
"content": "The following content
    is a post published by a
   blogger on social media:\n<
   image><image><image><image>\nI
    love Suzhou, I love
   Xiangcheng. I just moved here
   half a month ago and
   successfully made an
   appointment! Previously, it
   was necessary to make an
   appointment offline, but I
   couldn't catch up with others.
    After two failed attempts,
   the third time I finally
   grabbed the appointment on the
    mini-program at the scheduled
    time! It was super easy, and
   both my friend and I got it.
   After securing the appointment
   , I immediately took the
   landlord to get the mobile
   population proof, and the next
    day, it was done! No
   connections, no extra money
   spent, all by myself~ I saw
   that other sisters had boxes,
   but I was too shy to ask for
   one. The method is shown in
   the last image~ First, follow
   the Yuanhe Community Health
   public account, then the Miao
   public account will notify you
    in advance, just grab the
   appointment on the designated
   platform~ #hpv #NineValent #
   SuzhouNineValentHPV",
"role": "user"
},
{
    "content": "Okay, I have read
        this post.",
    "role": "assistant"
    },
    {
    "content": "Please identify
       the following personal
       information categories of
       the blogger based on the
       content of the post: - \"
       Location\" - \"Name\" - \"
       Age\" - \"Gender\" - \"
       Marital Status\" - \"Pet\"
        - \"Husband/Boyfriend\" -
        \"Wife/Girlfriend\" - \"
       Sexual Orientation\" - \"
       Physical Health\" - \"
       Family Status\" - \"
       Occupation\" - \"Mental
       Health\" - \"Education
       Information\" - \"
       Financial Status\" The
       requirements are as
       follows: 1. Try to extract
        the above 15 pieces of
       personal information from
       the post content. 2. If a
       category cannot be
       identified, fill it in as
       \"Unknown\", and leave no
       blanks. 3. Regardless of
       whether information can be
        identified or not, all 15
        fields must be output. 4.
        The result must be output
        only in JSON format. 5.
       The keys in the JSON
       should be in English, and
       the values should be in
       Chinese. 6. Absolutely no
       extra explanations, text,
       or non-JSON format content
       . Example of output format
       : {\"Location\": \"Beijing
       \", \"Name\": \"Unknown\",
        \"Age\": \"25 years\", \"
       Gender\": \"Female\", \"
       Marital Status\": \"Single
```

```
    \", \"Pet\": \"Cat\", \"
    Husband/Boyfriend\": \"
    Unknown\", \"Wife/
    Girlfriend\": \"Unknown\",
     \"Sexual Orientation\":
    \"Heterosexual\", \"
    Physical Health\": \"
    Healthy\", \"Family Status
    \": \"Only child\", \"
    Occupation\": \"Software
    Engineer\", \"Mental
    Health\": \"Unknown\", \"
    Education Information\":
    \"Undergraduate\", \"
    Financial Status\": \"
    Financially Stable\"} Note
    : - Even if there is no
    identifiable information
    in the post, the complete
    15 items must still be
    output, with each filled
    as \"Unknown\". - Do not
    add or remove fields, and
    do not adjust the field
    order.",
"role": "user"
},
{
"content": "{\n    \"
    personal_information\": {\
    n        \"Location\": \"
    Suzhou, Xiangcheng\",\n
           \"Name\": \"Unknown
    \",\n        \"Age\": \"
    Unknown\",\n        \"
    Gender\": \"Unknown\",\n
           \"Marital Status\":
     \"Unknown\",\n        \"
    Pet\": \"Unknown\",\n
           \"Husband/Boyfriend
    \": \"Unknown\",\n
    \"Wife/Girlfriend\": \"
    Unknown\",\n        \"
    Sexual Orientation\": \"
    Unknown\",\n        \"
    Physical Health\": \"
    Unknown\",\n        \"
    Family Status\": \"Unknown
    \",\n        \"Occupation
    \": \"Unknown\",\n
    \"Mental Health\": \"
    Unknown\",\n        \"
    Education Information\":
    \"Unknown\",\n        \"
    Financial Status\": \"
    Unknown\"\n    }\n}",
    "role": "assistant"
    }
    ],
"images": [
    "personal_info/329_1.jpg",
    "personal_info/329_2.jpg",
    "personal_info/329_3.jpg",
    "personal_info/329_4.jpg"
]
}
```


D.2 Label distributions for both splitting strategies

We use a random split to divide the data into training and validation sets. Additionally, we introduce a multi-label stratified split to construct another pair of training and validation sets. The label distributions for both splitting strategies are shown in Table D.2.

D.3 Implementation Details

LLMs-finetune We fine-tuned all multimodal large language models using LoRA. We enabled Flash Attention 2 (flash_attn2) and DeepSpeed ZeRO Stage 3 to improve memory efficiency. These two optimizations significantly enhanced both the training speed and scalability.

This training setup has eight epochs and an initial learning rate of 1e-4. We use bfloat16 (bf16) precision to reduce memory consumption further and accelerate training.

While the maximum input length is typically set to 10,000, for Qwen2.5-VL-7B-Instruct, which features joint embeddings of text and images, the cutoff_len is specially configured to 16,000 to ensure adequate context handling. Each GPU processes a batch size of 2 for batch configuration, and with gradient_accumulation_steps=8, the adequate batch size per update becomes 16. It allows us to maintain a larger adequate batch size under memory constraints. All other parameters follow LLaMAFactory's default settings, including the optimizer configuration, regularization strategy, caching policy, and logging options.

RoBERTa finetune For the RoBERTa model, we adopt a sentence segmentation combined with a sliding window approach to handle long-text inputs

Label	Original	Multi-label Stratified Train	Multi-label Stratified Test	Random Train	Random Test
Location	11.57	11.63	11.09	11.60	11.28
Gender	22.70	22.82	21.71	22.73	22.48
Occupation	11.19	11.24	10.69	11.22	10.88
Education Information	6.05	6.08	5.75	6.17	5.03
Financial Status	2.57	2.58	2.47	2.59	2.44
Marital Status	8.31	8.18	9.42	8.30	8.44
Husband/Boyfriend	7.38	7.36	7.50	7.32	7.87
Sexual Orientation	8.33	8.28	8.78	8.16	9.82
Age	5.55	5.46	6.30	5.55	5.52
Physical Health	2.65	2.66	2.55	2.70	2.19
Family Status	6.61	6.64	6.30	6.56	7.06
Name	0.69	0.69	0.64	0.70	0.57
Pet	1.42	1.43	1.36	1.43	1.30
Wife/Girlfriend	2.75	2.72	3.03	2.74	2.84
Mental Health	2.24	2.22	2.39	2.24	2.27

Table 10: Comparison of label distributions between random split and multi-label stratified split.

Model	Metric	Location	Name	Age	Gender	Marital Status	Pet	Husband /Boyfriend	Wife /Girlfriend	Sexual Orientation	Physical Health	Family Status	Occupation	Mental Health	Education Information	Financial Status
Gemm3-4b-it	Partial F1	75.36	90.14	76.59	27.31	73.1	89.12	76.8	92.2	77.82	80.29	80.49	70.23	68.99	81.31	85.01
Gemm3-4b-it-sft		77.82	**98.97**	87.89	69.82	79.67	97.33	83.78	94.25	79.67	93.63	88.71	78.44	93.63	90.35	92.4
InternVL3-8B		76.8	95.48	87.68	41.27	79.67	94.46	78.03	92.4	75.15	86.24	81.72	76.59	87.89	85.63	88.71
InternVL3-8B-sft		80.08	98.36	**88.5**	65.3	79.67	97.54	85.22	**94.25**	79.47	93.84	89.94	78.23	94.87	89.94	93.02
Llama-3.2-11B-Vision		62.42	97.95	71.25	30.18	65.5	88.71	79.88	92.2	74.13	72.9	76.18	62.83	92.61	71.66	78.44
Llama-3.2-11B-Vision-sft		72.9	98.77	86.65	63.45	79.47	97.33	85.22	93.02	79.47	94.25	86.24	76.18	94.05	89.94	93.43
Qwen2.5-VL-7B-Instruct		77	94.87	87.27	39.01	**93.43**	93.43	77	90.55	75.15	93.02	80.49	76.18	91.79	87.27	92.2
Qwen2.5-VL-7B-Instruct-sft		77.21	97.54	88.09	66.12	76.8	97.13	79.88	90.55	79.06	94.05	86.86	79.88	**94.87**	88.5	92.81
Qwen2.5-Omni-7B		**80.29**	95.28	87.47	40.45	72.69	93.63	77.41	92.2	76.39	91.17	80.7	78.03	92.81	86.24	87.47
Qwen2.5-Omni-7B-sft		79.67	98.56	88.3	**71.46**	81.31	**97.54**	**86.04**	93.84	**82.14**	**94.46**	**90.14**	**80.29**	94.46	**90.97**	**93.84**
RoBERTa-sft		71.46	98.56	86.04	43.12	78.64	96.71	80.08	92.81	75.15	**94.46**	82.14	72.48	94.25	87.27	**93.84**

Table 11: This table presents the Partial F1 scores for various self-disclosure categories evaluated on different models.

and overcome token length limitations. The dataset is divided into training, validation, and test sets in an 8:1:1 ratio. Fine-tuning is performed with a learning rate of 1e-5, a per-device training batch size of 16, a per-device evaluation batch size of 64, and the model is trained for 15 epochs.

D.4 More result analysis

In contrast, while the traditional RoBERTa model integrates text information derived from images, its core structure is still based on single-modal text sequence labeling. Therefore, it performs relatively stably when processing self-disclosure information with a clear structure and explicit expression, reflecting a certain level of language understanding and entity recognition capability. However, RoBERTa still falls short in overall recognition performance, especially when faced with implicit disclosure information dependent on contextual reasoning or multimodal cues, where it tends to miss labels or make inaccurate boundary predictions.

Further evidence supporting the above observations can be found in Table 11, which reports the F1 scores for different self-disclosure attributes. Among these, the performance variation in the education information category is the most notable, which is primarily because such information is often expressed implicitly (e.g., "just finished my thesis" or "preparing for graduate exams") rather than directly stating the degree or school name. Recognizing this type of information requires a deep understanding of the language and demands the model's contextual reasoning and semantic abstraction abilities. Due to the lack of multimodal cues and task-specific fine-tuning, RoBERTa struggles to make accurate judgments in such contexts, which is one of the main reasons we chose multimodal large language models as the core recognition tool.