Through Their Eyes: User Perceptions on Sensitive Attribute Inference of Social Media Videos by Visual Language Models

Shuning Zhang
Tsinghua University
Beijing, China
zsn23@mails.tsinghua.edu.cn

Gengrui Zhang
Tsinghua University
Beijing, China
zgr23@mails.tsinghua.edu.cn

Yibo Meng
Tsinghua University
Beijing, China
mengyb22@mails.tsinghua.edu.cn

Ziyi Zhang
Southeast University
Nanjing, China
ziyiz13@illinois.edu

Hantao Zhao
Southeast University
Nanjing, China
htzhao@seu.edu.cn

Xin Yi*
Tsinghua University
Beijing, China
yixin@tsinghua.edu.cn

Hewu Li
Tsinghua University
Beijing, China
lihewu@cernet.edu.cn

Abstract

The rapid advancement of Visual Language Models (VLMs) has enabled sophisticated analysis of visual content, leading to concerns about the inference of sensitive user attributes and subsequent privacy risks. While technical capabilities of VLMs are increasingly studied, users' understanding, perceptions, and reactions to these inferences remain less explored, especially concerning videos uploaded on the social media. This paper addresses this gap through a semi-structured interview (N=17), investigating user perspectives on VLM-driven sensitive attribute inference from their visual data. Findings reveal that users perceive VLMs as capable of inferring a range of attributes, including location, demographics, and socioeconomic indicators, often with unsettling accuracy. Key concerns include unauthorized identification, misuse of personal information, pervasive surveillance, and harm from inaccurate inferences. Participants reported employing various mitigation strategies, though with skepticism about their ultimate effectiveness against advanced AI. Users also articulate clear expectations for platforms and regulators, emphasizing the need for enhanced transparency, user control, and proactive privacy safeguards. These insights are crucial for guiding the development of responsible AI systems, effective privacy-enhancing technologies, and informed policymaking that aligns with user expectations and societal values.

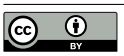
CCS Concepts

• **Security and privacy → Usability in security and privacy**.

Keywords

Privacy Inference, Visual Language Models, Privacy Attributes

*Corresponding author.



HAIPS '25, Taipei, Taiwan

ACM ISBN 979-8-4007-1905-9/25/10
https://doi.org/10.1145/3733816.3760751

ACM Reference Format:
Shuning Zhang, Gengrui Zhang, Yibo Meng, Ziyi Zhang, Hantao Zhao, Xin Yi, and Hewu Li. 2025. Through Their Eyes: User Perceptions on Sensitive Attribute Inference of Social Media Videos by Visual Language Models. In *Proceedings of the 2025 Workshop on Human-Centered AI Privacy and Security (HAIPS '25), October 13–17, 2025, Taipei, Taiwan.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3733816.3760751

1 Introduction

The proliferation of digital cameras and ubiquitous sharing of visual content on online platforms have created an unprecedented volume of user-generated images and videos [22, 56]. Concurrently, Visual Language Models (VLMs) have emerged, demonstrating remarkable capabilities in interpreting and generating human-like inferences from visual data by integrating computer vision and natural language processing [36, 37, 52]. These models, often pre-trained on vast datasets, can discern subtle visual cues indicative of sensitive attributes ranging from demographic information (age, gender) to more nuanced details like health status [16, 32], religious beliefs, or even personality traits [34]. Such inferences can be leveraged for constructing detailed personal profiles [41], potentially leading to (re)identification or cross-app tracking [19, 39].

While the technical process of VLMs in attribute inference presents a significant research area [41, 43], a critical yet often underexplored dimension is the user's perspective on these capabilities [61]. Understanding how users perceive the possibility of their sensitive attributes being inferred by AI, their comfort levels with such practices, their privacy concerns, and their expectations for control and transparency is paramount for fostering user trust and developing responsible AI systems. If users are unaware of or uncomfortable with the inferences being made, it can lead to distrust not only in specific applications but in AI technologies more broadly. Moreover, insights into user perspectives can directly inform the design of more effective privacy-enhancing technologies that resonate with user needs and mental models.

This paper aims to bridge the existing gap by systematically investigating users' perspectives, concerns, and expectations regarding the inference of their sensitive attributes by VLMs from their visual data, through exploring the following research questions (RQs):

RQ1: How do users perceive the capabilities of VLMs to infer their sensitive attributes from visual data and what are the associated privacy risks they identify?

RQ2: What strategies do users employ to manage these perceived risks?

RQ3: What are users' challenges to manage their privacy risks of VLM inferences on their videos, and what's their trade-offs and expectations?

To address these research questions, as a preliminary step, we conducted a semi-structured interview on Chinese users (N=17) to gather insights into their perspectives and experiences regarding AI-driven attribute inference from videos updated to social media. Towards RQ1, our findings indicate that users perceive VLMs as capable of inferring a wide array of sensitive attributes with often unsettling accuracy, including geographical locations and contextual activities, demographic profiles like age and gender, and even socioeconomic status. These perceptions are coupled with significant privacy concerns, such as fears of unauthorized identification and doxxing, misuse of personal information, pervasive surveillance, and potential harm from inaccurate algorithmic judgments.

Towards RQ2, users reported adopting diverse mitigation strategies such as proactive content modification (e.g., blurring, mosaics), selective online exposure, and strategic interaction with AI systems. However, they often expressed skepticism regarding the ultimate effectiveness of these personal efforts against advanced AI. Concurrently, users articulate clear expectations for platforms and developers to implement robust, automated privacy safeguards, enhance transparency in data handling, and provide granular user controls.

Towards RQ3, our study reveals that users engage in a complex and individualized trade-off when balancing the utility of VLMs against their privacy concerns. This often involves a hierarchical assessment of information sensitivity, where disclosure for functional benefits is more acceptable if tied to explicit utility and user control, especially over highly sensitive or directly identifiable information. To sum up, the primary contributions of this paper are:

- We qualitatively detail user experiences and perceptions of VLM inference capabilities.
- We document user-identified privacy risks, their mitigation strategies, and perceived effectiveness.
- We articulate user expectations for the design of trustworthy and privacy-respecting VLM systems.

2 Related Work

Research on visual privacy issues can generally be categorized into two main approaches: one that focuses on the technology itself such as system-side policy and visual data collection control, and the other that examines users' understanding of AI-driven visual privacy inferences, which is generally limited and inaccurate.

2.1 Privacy Risks in Visual Data and AI Systems

There are various ways to capture visual data, including recording social media videos and wearing smart glasses. These visual data, provided to AI systems or collected by smart devices themselves, often pose privacy risks to the individuals featured in them [8, 15, 20, 25, 40, 53]. For wearers or video shooters, privacy concerns arise largely from the mismatch between their mental models and the actual stealthy tracking practices of the devices. Harborth et al. [20] highlighted the gap between users' limited understanding and the reality of covert data collection. For bystanders, they also face privacy risks, particularly in public spaces where images or videos captured may contain sensitive information about them. In many if not all instances, bystanders are unaware of being recorded, raising significant privacy issues [24].

The advent of LLMs and VLMs has further compounded privacy concerns. LLMs have enabled privacy-infringing inferences from previously unseen texts [9], while VLMs have shown significant improvements on various PAR datasets [11, 12, 46], outpacing earlier, non-VLM-based approaches. For instance, Tomekcce et al. [43] explored the potential for privacy attribute inference from visual datasets, extending the privacy concerns associated with textual data to the visual modality.

Therefore, privacy concerns are no longer limited to direct data capture but extend into latent inferential risks that users cannot easily perceive or defend against in real-time interactions. Regarding these threats, researchers have proposed various solutions to mitigate the privacy risks. For instance, Jung and Philipose [23] developed a system that adjusts the privacy settings of the camera based on social context, while PrivacEye [42] employed deep learning to disable video recording based on visual cues. Other approaches, such as the "Life-Tags" system by Vatavu et al. [1], abstracted visual data into concepts and tags rather than storing raw images.

Although many of these solutions aim to mitigate privacy exposure, they often focus on system-side policy controls rather than enhancing user-side awareness. For example, Privacy-Eye [42] uses visual scene classification to automatically disable the camera without providing explicit user feedback, limiting user transparency over system actions. In addition, there is still a lack of AI interaction system designs focusing on user operations and feedback, indicating a critical gap in privacy interaction design.

2.2 User Perspectives on AI-driven Inferences and Visual Privacy

A significant body of research indicates that users often possess a limited or inaccurate understanding of how their data is used to generate AI-driven inferences, which is quite opaque for normal users without clear reminder. This gap can result in an "uninformed" conception of a service's data practices and the mechanisms by which inferences are made [49, 54]. Recent work by Zhang et al. [59] also finds that users' mental models regarding the inference process, particularly concerning private memories, are notably opaque. This limited awareness is problematic, as it can lead to misinformed online behaviors that not only fail to preserve privacy but may also degrade the quality of personalized services [26].

In such a situation of limited cognition, users may adopt various protective strategies in response to privacy concerns, ranging from actively withholding or strategically sharing information, such as limiting posts on social media [33, 35], to completely avoiding a service, for instance, by declining to purchase a product [5]. In other cases, users may adopt a stance of resignation, feeling powerless and hoping that online services will act in their best interests [27]. These strategies are often radical and negative for both users and service providers.

Conversely, interventions that increase transparency can significantly alter user perceptions and behaviors. When users can interact with and make sense of their inferred data, for example, through realistic advertising profiles, it can foster a more informed awareness of privacy implications and heighten interest in privacy-protective actions [5, 50]. Due to improved cognition, users can adopt more positive approaches to inducing risks. Studies presenting users with their own inferred data via retrospective surveys and dashboards revealed a tendency to reject the user of sensitive inferences, such as location or demographics, for third-party applications [51]. Highlighting this potential, Ma et al. [30] demonstrated that an LLM-powered intervention effectively increased participants' awareness of location privacy risks posed by LLMs and encouraged them to value greater control.

Despite these insights, a limitation persists in much of the existing research. While numerous studies employing interviews, dairy studies and surveys have successfully elucidated how users make sense of inferences in contexts like online behavioral advertising [17, 38, 54], voice assistants [27, 31, 45], and chatbots [18, 29], they often stop short of revealing the concrete privacy decisions users make regarding their data in the face of AI inferences.

3 Methodology

We detailed the participant recruitment, interview design and procedure, and the analysis methods sequentially.

3.1 Participant Recruitment and Demographics

This IRB-approved study recruited Chinese participants through distributing posters on online platforms. We recruited 17 participants (9 males, 8 females). The demographics of participants were shown in Table 1. Each participant was compensated 100 RMB for their time.

3.2 Interview Design and Procedure

We focused on identifying users' mental model towards the inference, their perceived privacy risks and harms of the inference, their own inference capabilities and the comparison with LLMs, their mitigation strategies and the effectiveness. The interview was carried out in a semi-structured manner, with all interviews audio-recorded and transcribed. Each interview lasted about 30 minutes on average.

3.3 Data Analysis

We adopted thematic analysis [14] on the transcribed data and the original recordings. Two researchers first read through the data for several times to get familiar with the data. They then open coded a subset (3 participants), discussed and solved the disagreements, forming the initial codebook. Using this codebook, these two researchers separately coded the rest of the data, reaching an inter-rater reliability of 0.9 using Cohen's Kappa.

Table 1: Demographic information.

User ID	Use Case	Demographic	Highest Education
P1	AI editor, Doubao	Student	Bachelor
P2	VLMs	Student	Bachelor
P3	ChatGPT, Doubao	Student	Master
P4	Doubao, kimi, qwen	NA	Bachelor
P5	Doubao, deepseek, Wenxin Yiyan, qwen	Teacher	Bachelor
P6	deepseek, Doubao, Wenxin Yiyan, kimi	Student	Bachelor
P7	Sora, deepseek, GPT	Student	Master
P8	GPT4, Claude 2.0	Student	Bachelor
P9	gpt	Product Manager	Master
P10	Doubao	Student	Bachelor
P11	NA	Worker	High School or Lower
P12	Doubao	Driver	High School or Lower
P13	Deepseek, Doubao	NA	High School or Lower
P14	NA	Student	Bachelor
P15	Doubao	Student	Bachelor
P16	NA	Student	Bachelor
P17	Deepseek	Student	Bachelor

4 Results

We presented the results according to the research questions, first analyzing the perceptions of inference capabilities and privacy risks, and then users' mitigation strategies and their effectiveness, and finally the challenges, trade-offs and expectations.

4.1 RQ1: User Perceptions of VLM Inference Capabilities and Associated Privacy Risks

4.1.1 Users' Perception of VLM Inference. Users' perception of VLM inference are centered around specific inference types and cases, reflecting their mental model of the ubiquitous inference in the real world.

Ubiquitous location and contextual activity inference. A predominant theme was the VLM's strong capability to infer geographical location and contextual activities, a finding reported by a significant majority of participants (14/17, 82%). Models achieved this by leveraging distinct visual cues, such as identifying a specific city street from billboards (P1), deducing a tour group's location from local snacks like *"Roujiamo"* (P8), or recognizing a hometown's specific *"blacksmithing flower"* festival and its cultural status from the scenery (P17). This capability often reached a startling level of precision, with a substantial subset of participants (9/17, 52.9%) reporting instances of such direct location inference. P15 expressed shock when a VLM identified their specific street from a photo of

a pigeon, stating, *"I was dumbfounded, felt like a complete privacy leak."* While some users found this specificity useful for identifying places (P10), the accuracy of the capability was inconsistent. As P12 noted, a VLM could correctly identify a province but fail in the specific city, demonstrating variable reliability: *"the first time it was right, but later it wasn't."*

Demographic profile inference: age and gender. A majority of participants (13/17, 76%) reported that VLMs can infer basic demographic information like age and gender from visual cues. The inference of gender was perceived to be straightforward. P2 suggested, *"if a person appears, it should be very easy to tell,"* while P9 stated that the probability of correctly identifying male or female *"is quite high."* P8 provided a detailed account of visual features used by models for analysis, like *"wrinkles, crow's feet"* for age and facial structure and presentation: *"males have that jawline, a bit squarer ... female faces are finer ... then there's hairstyle and makeup."* for gender.

Age inference is also frequent, as discussed by 11/17 participants (64.7%), though often guessed as a range. P8 noted that VLMs could infer age based on visual cues (*"facial features ... clothing"*) and even speech patterns: *"Young people like us, born in the 00s or 90s, speak faster. And the words used are very trendy... This model will classify us as younger."* P16 observed, *"From the background, and looking at the person's general facial features, age can probably be judged."* P10 humorously noted a behavioral inference of age: *"Riding a bike with one hand while filming, only someone under 25 would do that."*

Assessing socioeconomic status and lifestyle indicators. VLMs are perceived to infer socioeconomic status (income, education, occupation) and lifestyle traits (marital status, hobbies), though with variability in accuracy and reliance on indirect cues. 7/17 (41%) mentioned income inference, 5/17 (29%) for occupation, and 5/17 (29%) for education. For socioeconomic attributes such as occupation, education level, income, and personal interests, these are generally considered challenging to infer accurately without explicit cues. Occupation might be guessed from attire or context, as P2 mused: *"Occupation might consider logic ... the clothes worn, the person's actions, or the logical connection of their actions."* However, P6 found an inferred occupation ("engineer") inaccurate. P17 suggested that past interactions could inform such inferences: *"If you are a student ... and you've asked it questions about your homework ... the large model might already have an impression of you ... and estimate your education level and identity."* P2 also theorized that education level could be inferred if *"someone attending a class in a university or displaying class content."* was showed. For income, which P2 noted is *"quite abstract"*, direct proof like a *"payslip in the video"* would be definitive. P8 suggested income could be inferred from home decor: *"For example, uploading a video about home decoration and renting a house ... it can infer a Nordic furniture style, and then infer the family's income based on the furniture."* P17 similarly proposed that VLMs could *"analyze my income situation through the items filmed,"* distinguished between *"high-end places"* and *"crowded activities"* to gauge financial status. Lifestyle and interests are inferred from the activities or objects depicted. P5 noted that a video with a cat might lead to inferences about gender (*"young female"*), age and preferences (*"like small animals"*). P8 also mentioned that preferences for *"minimalist or modern style"* versus more traditional items could be indicative of age and lifestyle. Overall, while these attributes are subjects of VLM inference, their accuracy is more questionable and context-dependent, as expressed or implied by 11/17 participants (64.7%). P7 for example, stated that inferring education requires specific evidence like a *"student ID"*, and location inference needs *"landmarks or road names"* to be accurate, implying difficulty otherwise.

Perceived Sophistication of Inference and Emergent Privacy Concerns. Users also found the inference mechanisms advanced and unsettling, leading to negative emotions such as fear, shock or violated feelings (5/17, 29%). Several participants suspected that VLMs might access more data than explicitly provided, or cross-reference information from various sources. P5 described a particularly *"terrifying"* experience: *"Once I only sent my own photo, it directly stated the photo's location ... it started showing my personal information not given in its "deep thinking" process ... It displayed my name and shooting location."* This user also speculated about photo album access: *"maybe it can read my photo album information, where the photos were taken, where my permanent residence is, it might all be read."* P2 was aware of metadata like GPS in images but believed their VLM (Doubao) did not typically use it for image content analysis. P17 theorized about data aggregation from other platforms: *"it might collect data from Xiaohongshu or Douyin where users are synchronously posting similar videos, and then analyze my video based on that."* P11 also believed AI uses *"big data comparison to verify a user's age and other basic information."* The unexpected revelation of personal data, as in P5's case, caused considerable alarm.

4.1.2 *Perceived Privacy Risks and Harms.*

Based on the thematic analysis of participants' discussions about VLMs, several distinct categories of privacy risks and underlying concerns emerge. They were shown in Table 2. Users' mental models of VLMs often depict them as powerful data aggregation and inference engines that learn from vast datasets, store uploaded information, and possess capabilities surpassing human understanding or control, leading to these apprehensions. The primary categories of perceived risks identified are: (1) Unauthorized identification, doxxing and associated harassment or safety threats; (2) Misuse and exploitation of personal information for commercial or malicious ends; (3) Pervasive surveillance, algorithmic profiling, and loss of anonymity; (4) Harm from inaccurate inferences and algorithmic misjudgment.

Unauthorized identification, doxxing, and harassment. 8/17 participants expressed significant concern that VLMs could enable precise user identification and *"doxxing,"* leading to severe real-world consequences. They feared VLMs can meticulously analyze content to *"locate you accurately, and even excavate all your information"* (P3). This capability is perceived as dangerous because it could democratize doxxing. P4 worried if *"AI can infer this information easily, ordinary people can also infer user information, which is very scary."* The underlying mental model is that VLMs compile and cross-reference data points to make non-obvious connections. P1 provided an example, from *"a family photo, or a graduation photo ... [it] can deduce which university I attended, and by inferring the time, when I graduated."*

Table 2: User perceptions of VLM inference privacy risks.

VLM privacy risks	Description
Unauthorized identification, doxxing and associated harassment or safety threats	VLMs enable user identification, leading to doxxing, harassment, and real-world safety threats
Misuse and exploitation of personal information for commercial or malicious ends	Exploitation of inferred data for commercial manipulation and malicious scams
Pervasive surveillance, algorithmic profiling, and loss of anonymity	Pervasive data collection and profiling erodes anonymity by linking digital and real-world identities
Harm from inaccurate inferences and algorithmic misjudgment	Inaccurate VLM inferences create false profiles, leading to misrepresentation and flawed social categorization

Participants primarily feared the downstream effects of such identification, including online harassment and *"cyberbullying"* (P4). P12 worried about *"social death"* from viral content and the general risk of exposure: *"I am very worried that current online technology is too powerful. I would be worried about my personal information being posted."* P17 noted this risk can be gendered, fearing that if men *"determine I'm in the same city, they might harass me via private messages."* Beyond online threats, 3 participants explicitly linked location inference to their physical safety. P6 stated, *"If it can infer a specific Chinese location correctly, for someone with malicious intent, my personal safety is not guaranteed."* Similarly, after a VLM identified their garden, P15 feared not only theft but also direct *"harassment or stalking."*

Misuse and exploitation of personal information for commercial or malicious ends. 7 participants feared that entities could misuse their VLM-inferred personal data for exploitative purposes, ranging from manipulative advertising to outright scams. Participants viewed their data as a valuable commodity, with P2 identifying an *"information asymmetry"* where companies can *"sell this information as a resource."* This fear of commercialization was specific: P9 believed targeted job content serves to *"commercialize later,"* while P16 strongly objected to their information being sold to *"advertisers or telemarketers,"* leading to unwanted *"promotional calls."*

These concerns escalated to criminal exploitation. P8 expressed a concrete fear that leaked personal details could allow scammers to *"impersonate bank staff ... and take all my money."* This reflects a deep mistrust of third-party data handling, which P16 found *"quite uncomfortable"* and P9 worried would lead to *"information leakage or being sold."* P9 contextualized this ultimate risk by referencing *"Snowden's case"* and the potential for *"espionage."*

Pervasive surveillance, algorithmic profiling, and loss of anonymity. 6 participants (P1,P2,P5,P7,P9,P11) expressed a fundamental anxiety that VLMs erode privacy through perceived surveillance, deep profiling, and indefinite data storage. This anxiety stems from the mental model that *"AI learns through big data, and things sent are stored in a database"* (P1), fueling fears of *"privacy leakage."* P5 directly questioned the necessity and risk of this data collection: *"Why does it need to collect my information? If someone searches for something related to me, will it directly tell the searcher my information? Isn't that a huge risk for me?"* This concern highlights a core desire to anonymity online, as another participant stated, *"If other people know everything when I go online, then there's no privacy."*

Worries also target the volume and granularity of the data processed. P7 feared the leakage of *"professional information, or other physical privacy,"* while P9 expressed concern over corporate surveillance via methods like taking *"periodically take screenshots of my computer."* The VLM's ability to infer non-explicit information, especially location (P11,P13,P15), further amplifies this sense of lost anonymity. P15 explained that an AI's superior database allows it to *"know which city you are in"* from a picture when a human cannot, enabling services that could also facilitate unwanted tracking.

Harms from inaccurate inferences and algorithmic misjudgment. Participants described a distinct risk: harm caused not by accurate surveillance, but by inaccurate VLM inferences that lead to misrepresentation. P8 provided a specific example, worrying that a video filmed near a landmark could generate a false profile: *"If in my video ... there's Batu Caves ... which has some religions like Buddhism, and some Indians, it might identify me as a local Indian who likes Buddhism. This is wrong information."*

P8 further explained the consequences of such misjudgment could range from receiving irrelevant, targeted content (*"on social platforms and my recommendations, I might get more local Indians and Buddhists"*) to having this *"erroneous information"* negatively affect social perceptions or be *"disclosed to others."* This highlights a nuanced concern where the harm stems not from an exposed truth, but from a propagated algorithmic falsehood that misrepresents a user's digital identity.

4.1.3 Places Where Participants Perceive Privacy Risks. Participants perceive VLM-related privacy risks across various digital activities where they create or share multimedia content, as detailed in Table 3. These contexts fall into three primary categories: (1) the public sharing of everyday life and social activities, (2) the dissemination of content rich in personal and environmental identifiers, and (3) direct engagement with AI systems for personalized assistance.

Table 3: Scenarios where participants perceive privacy risks.

Scenarios of privacy risks	Description
Public sharing of everyday life and social activities	Inference risks from publicly shared daily life content.
Content rich in explicit personal and environmental identifiers	VLM extraction of explicit identifiers from shared content.
Direct engagement with AI for personalized tasks or information disclosure	Direct interaction with AI enables unintended user profiling.

Public sharing of everyday life and social activities. 5 participants (P1,P2,P12,P13,P17) identified significant privacy risks in publicly sharing content about their everyday lives and social activities. They

worried that casual documentation could lead to unintended data breaches. P1 noted that by *"causally filming life vlogs,"* a VLM could easily infer their *"place of residence and lifestyle"* from landmarks and signs. Similarly, P17 described how a video of a social outing might *"leak our city, our ages, and regional information,"* or even *"directly pinpoint the exact location of the bar I was in."*

The underlying fear is that VLMs extract and synthesize these seemingly innocuous details to build comprehensive user profiles. This can lead to severe social consequences, as P12 feared *"social death"* if a casual video posted on Douyin *"suddenly becomes popular"* because the AI might *"push it to everyone."* This illustrates a core concern that VLMs dangerously amplify both the audience and the analytical depth applied to content shared in public digital spaces.

Content rich in explicit personal and environmental identifiers. 9 participants expressed heightened privacy risk when their content contains explicit personal or environmental identifiers that VLMs adeptly process. They identified risks in sharing recognizable faces, specific locations like homes or schools, and textual information on tickets or signs. P4 emphasized the VLM's superior capability, stating that *"if the image contains specific text, it's easy to expose information,"* and that even obscured *"surrounding buildings ... can often be inferred."*

P5 shared a direct experience where a VLM processed their life photos to *"directly infer my age ... and by inferring the appearance of trees or the campus outside, it could deduce where my neighborhood is, exposing my address."* This reflects a core fear that VLMs can easily extract and link these explicit identifiers to build detailed profiles. Participants listed many examples of such risky content, including photos of a *"first flight, train journey, friends' faces"* (P10) or videos featuring a *"university platform"* (P16).

Direct engagement with AI for personalized tasks or information disclosure. 4 participants (P3,P5,P6,P7) identified significant privacy risks when directly engaging VLMs with sensitive personal data for personalized tasks. They worried that providing such information could lead to unexpected and deep inferences. P5 recounted how uploading *"personal photos"* could lead to the inference of an *"ID card number,"* while P6 described *"telling it my identity, student ID, or even photos of myself"* as a high-risk scenario. Similarly, P7 feared that asking *"questions related to my work or study"* would allow a VLM to infer their location.

This anxiety is rooted in the mental model that directly provided data is not just used for the immediate task but is permanently stored and learned from. As P1 stated, *"AI learns through big data, and things sent are stored in a database."* This makes the VLM's inference capabilities feel more direct and intrusive. P3 added a further nuance, noting that risk is heightened even when merely inquiring about data, as *"when I ask it some specific questions about this video,"* the act of inquiry itself can guide the VLM toward sensitive disclosures.

4.2 RQ2: User Mitigation Strategies and Their Effectiveness

To mitigate the perceived privacy risks of VLMs processing their multimedia content, participants employed various strategies, as detailed in Table 4. These strategies fall into five primary categories: (1) proactive content modification and obfuscation, (2) selective disclosure and controlled online exposure, (3) strategic interaction with and data minimization for AI systems, (4) a combination of reliance on external controls and personal vigilance, and (5) an acknowledgement of the inherent limitations in fully mitigating these risks.

Table 4: Mitigation strategies and their effectiveness.

Mitigation strategies	Effectiveness
Proactive content modification and obfuscation	Directly sanitizes data, disrupting VLM data extraction
Selective disclosure and controlled online exposure	Fundamentally limits data available for VLM analysis
Strategic interaction and data minimization with AI systems	Offers partial control during necessary AI interactions
Reliance on external controls and vigilant personal review	Strong protection through meticulous self-monitoring and review

4.2.1 Proactive content modification and obfuscation. 7/17 participants (P1,P2,P3,P4,P10,P13,P15) described directly modifying or obfuscating their multimedia content to prevent VLMs from extracting sensitive information. The most common technique was applying mosaics or blurring. P1 would *"apply mosaics to places or road signs that I clearly don't want to reveal,"* while P13 blurred an image for a medical consultation to *"only leave the part to be checked."* P2 considered this the *"simplest form of desensitization,"* believing it causes *"permanent destruction to the image layer."* Other modifications included removing sensitive audio segments (P3) or stripping image metadata, which P2 suggested as a technical form of desensitization. These actions demonstrate a hands-on approach by users to sanitize their data before it undergoes potentially intrusive VLM analysis.

4.2.2 Selective disclosure and controlled online exposure. 10/17 participants practice selective disclosure, consciously deciding what content is too sensitive to record or upload to limit the data available to VLMs. This can be an absolute refusal, as P2 noted the *"risk of privacy infringement is an important reason preventing me from doing self-media."* More commonly, participants practice selective creation. P5 tries to *"avoid filming my life scenes, or things that might expose my address,"* and P16 attempts to *"avoid personal facial information and landmark objects appearing [in videos]."*

This strategy extends to controlling the audience. 4 participants use platform privacy settings to share content with *"close friends only"* (P8), *"only in smaller circles"* (P9), or use niche tags to *"reduce traffic to the minimum"* (P12). These practices demonstrate an understanding that limiting exposure reduces the surface area for VLM analysis and its subsequent privacy risks.

4.2.3 Strategic interaction and data minimization with AI systems. 5 participants (P3,P5,P6,P7,P9) described strategies for minimizing

data disclosure during necessary interactions with VLMs. To control their input, they limit the specificity of their queries. P3 would *"choose to ask more general questions,"* and P7 would try to *"minimize giving it useful information."* P6 advised to *"not give the large model too much specific information"* and even suggested instructing the VLM to *"anonymize all privacy-implicating information with symbols"* in its output.

Participants also use technical controls. P5 detailed a multi-pronged approach including managing *"camera permissions,"* clearing *"big data history,"* and reducing *"interaction with it."* While acknowledging the difficulty of complete avoidance on a single account, P9 suggested partial measures like *"information hiding."* These strategies demonstrate a cautious engagement model, where users treat VLMs as entities requiring careful information sharing.

4.2.4 Reliance on external controls and vigilant personal review. A notable group of participants indicated reliance on external controls, such as platform policies or emphasized personal vigilance through careful content review. Among them, 2 participants (P6,P7) considered the privacy policies or statements of VLM providers. P7 stated, *"I tend to choose models that [their company] first declares they will not steal my privacy. If they don't have such a declaration ... I will not transmit information to such models."* This shows a desire for VLM developers to take responsibility for user privacy. Complementing this, 3 participants (P8,P14,P17) highlighted the importance of personal review and post-upload control. P8 would *"first check the uploaded content, then apply mosaics to things I feel cannot be shown, and then upload it."* P14, if accidental exposure was found, would *"immediately delete the work."* P17 described a meticulous approach: *"I might review my videos multiple times to see if any points that could leak information... were accidentally filmed... and I might be less active in uploading to media."* This suggests that some users combine cautious selection of tools with diligent self-monitoring.

4.3 Acknowledged limitations, risk acceptance, and perceived high mitigation costs

Participants acknowledged the inherent difficulty and high cost of fully mitigating VLM-related privacy risks, often leading to a pragmatic acceptance of exposure. P9 explained this trade-off, noting that the *"cost of hiding information is very high,"* especially for video, so *"many people cede a part of their privacy rights."*

This difficulty can result in a passive stance, a sentiment shared by two participants (P2,P9). For instance, despite knowing about desensitization techniques, P2 admitted to *"currently not taking any measures to avoid risks"* for personal content, citing a *"not very strong privacy awareness."* This suggests that for some users, the high effort required for protection, combined with the fact that *"short-term risks from privacy leakage are not visible,"* leads to a conscious acceptance of a certain level of privacy risk.

4.4 Effectiveness of Mitigation

Participants held varied and nuanced views on the effectiveness of their strategies to mitigate VLM-related privacy risks. While some believed their measures offered tangible protection against casual observation or simple inference, a significant skepticism remained about their robustness against advanced AI or determined actors. Overall, participants viewed any effectiveness as conditional and partial, sometimes valuing their strategies more for psychological comfort than for providing foolproof security.

4.4.1 Conditional and partial efficacy of common mitigation techniques. Many participants believe common strategies like content modification and selective sharing offer a conditional or partial degree of effectiveness. 4 participants (P1,P4,P10,P16) viewed content obfuscation, primarily mosaics, as a key defense. They opined that *"AI is not yet strong enough to directly see through mosaics"* (P1) and that these methods have *"some effect"* (P4), with P10 estimating them to be *"effective about 80% of the time."* However, they acknowledged clear limitations. Mosaics become *"merely symbolic"* for highly recognizable landmarks (P10), and P2 warned that from many desensitized images, *"some information can be inferred because omissions are very common."*

Selective sharing was also seen as a valuable control; P3 felt that with certain measures, *"at least strangers won't know,"* and P6 trusted their methods were effective with *"reputable companies."* P8 highlighted a more advanced tactic of using obfuscation to induce misidentification, noting a VLM might misidentify a landmark, and this *"wrong information, in turn, protected my privacy."* Ultimately, some participants recognized absolute data minimization as the most reliable strategy, as P15 stated simply, *"not filming is very effective."*

4.4.2 Skepticism regarding efficacy against advanced AI and determined adversaries. Despite the perceived benefits of countermeasures, a strong skepticism emerged, with 4 participants (P9,P12,P16,P17) doubting the ultimate effectiveness of their mitigation strategies against sophisticated VLMs. This skepticism is rooted in a belief that the VLM's analytical power, derived from *"big data,"* is overwhelming. P12 was particularly pessimistic, arguing that because *"current technology is very powerful,"* any measure short of total abstinence from posting is *"useless."* P16 echoed this, feeling that *"effectiveness is not that great, big data is so powerful, it feels like it can analyze all my info at once."* Similarly, P17 acknowledged that while their efforts might have *"some effect... perhaps not that much"* because the model's analytical ability *"is still very strong."*

4.4.3 The role of avoidance measures in providing psychological comfort. For some participants (2/17, P11, P13), the adoption of privacy-preserving measures serves an important function in providing psychological comfort, even if the actual security benefits are uncertain or perceived as limited. P11 admitted that complete avoidance is *"impossible... but relatively speaking, it gives oneself psychological comfort."* P13 echoed this sentiment: *"I don't know if it's reliable, but psychologically it feels effective."* This suggests that engaging in these practices, such as applying mosaics or being cautious with uploads, can create a sense of agency and control in an environment where users often feel their data is vulnerable. The act of taking precautions, regardless of its ultimate impenetrability, can alleviate anxiety and allow users to continue participating in digital spaces, albeit with a cautious mindset. This psychological dimension is an important aspect of user behavior, indicating that the perceived, rather than solely the actual, effectiveness of privacy measures plays a role in their adoption.

4.5 RQ3: Challenges, Trade-offs and Expectations

4.5.1 Perceived Trade-offs and the High Cost of Complete Privacy. Users also highlighted the trade-offs between privacy, usability and the effort required for robust protection. While not posting content at all is recognized as highly effective (P15: *"not filming is very effective"*), it fundamentally curtails participation in digital content creation, a trade-off not all users are willing to make. P9 articulated the high cost of comprehensive mitigation: *"The cost of hiding information is very high. If you upload a picture, you can apply a mosaic. If you upload a video of several minutes, it's very difficult to simply and specifically hide various kinds of information."* This implies that users often engage in a practical cost-benefit analysis, where the effort of meticulous desensitization for every piece of content might outweigh the perceived immediate risk, especially as P2 noted, when *"omissions are very common"* even with desensitization efforts. This can lead to a degree of risk acceptance, as noted by P9, *"many people cede a part of their privacy rights because it's very hard to 100% eliminate this concern,"* and they may *"default to [accepting] there's a risk, and it doesn't matter if there's a risk."* This underscores the challenge users face in balancing their desire for privacy with the content sharing aim with pervasive VLMs.

4.5.2 Expectations of Platforms, Stakeholders and Supervisors. Participants envision a multi-stakeholder approach to fortifying privacy in the age of advanced VLMs, expressing expectations for proactive measures from platforms and developers, enhanced user control and transparency, strong regulatory frameworks, and a continued evolution of user vigilance. The expectations were shown in Table 5. Their future outlook emphasizes a shared responsibility model where technological solutions, ethical design principles, clear governance, and informed user practices converge to mitigate the risks associated with sophisticated AI-driven content analysis.

Table 5: An overview of user expectations for privacy protection

Expectations	Description
Proactive platform-integrated privacy safeguards and automated moderation	Allow AI identify and inform users which information constitutes privacy violations
AI-assisted content modification tools	Provide tools to allow users edit parts of the image
Enhanced transparency, user control, and granular consent mechanisms	Let users explicitly know the data handling practices and offer more granular control
Strengthened developer responsibility and regulatory oversight	Privacy-by-design principles to be embedded by developers; Government oversight and legal monitoring
Empowered and cautious user practices	Users take a heightened sense of awareness and more deliberate decision-making regarding content creation and sharing

Proactive platform-integrated privacy safeguards and automated moderation. A dominant expectation, voiced by a significant majority of participants (10/17) is for VLM-integrated platforms and editing software to incorporate proactive and automated privacy-enhancing features. Users anticipate that AI itself should be leveraged to protect privacy. P1 suggested platforms should *"edit AI's own content, allowing AI to identify which information constitutes user privacy or violations,"* and then offer users options to *"replace, hide, or delete"* problematic sections. This includes automated detection and flagging of sensitive information. P8 envisioned a system where the VLM *"detects and highlights potential leaks (e.g., logos, location, faces), making it obvious to me where information might be exposed."* P7 suggested VLMs could *"automatically help me mosaic or otherwise prevent access to"* inadvertently captured sensitive details like ID cards.

Furthermore, there's a desire for AI-assisted content modification tools. P4 hoped editing software like Jianying, which already incorporates AI functions, could *"develop reminder functions"* that *"automatically detect aspects that can help privacy, reminding users to apply mosaics or edit parts of the footage."* P5 suggested platforms could automatically *"apply mosaics to areas imposing risks ... or tell users which parts that might expose privacy should be cropped."* P8 also proposed features like *"background blur, just showing my face,"* allowing users to *"decide whether to show a face or an outdoor landmark."* The underlying mental model is that since VLMs are powerful at analysis, they should also be powerful at automated safeguarding, shifting some of the burden from the user to the technology provider.

Enhanced transparency, user control, and granular consent mechanisms. Complementing automated safeguards, participants strongly desire increased transparency regarding data handling practices and granular control over their personal information, a sentiment expressed by 5 participants. P5 articulated a need for applications to *"clearly tell me they will not read my information, and what penalties they might face if they do,"* advocating for *"clear terms ... letting users explicitly know."* This extends to permission management, where users are informed *"what permissions are needed, what privacy they involve, whether it's an acceptable range, whether they can be turned off, and if it will affect usage"* (P5). P2, while skeptical around the efficacy of lengthy user agreements that *"99% of users won't read,"* suggested that "reminder boxes" coupled with options like *"one-click desensitization before uploading"* could be effective.

A key aspect of control is data deletion and purpose limitation. P3 hoped for models to *"delete their memory, for example, delete user data"* after processing. P6 wished for an option where *"after generation is complete, a prompt appears asking if the large model should automatically delete this data ... so the data I generate can only be used by myself and is deleted after use."* P12 envisioned an "incognito mode" for AI apps like Doubao, and settings to ensure *"what I send is not authorized for others or for big data, making it completely private."* These expectations reflect a user desire to not only be informed but also to actively manage the lifecycle and accessibility of their data within VLM systems.

Strengthened developer responsibility and regulatory oversight. Participants also look towards developers (the "business-end") and

regulatory bodies to establish a safer VLM ecosystem, a view supported by 4 participants (P2,P9,P14,P17). There's a strong sentiment that the onus of privacy protection should not solely rest on users. P2 asserted that risk mitigation should depend *"more not on customer-end users figuring it out themselves, but on how the business-end ... can avoid company ethics, safety, and privacy risks."* This implies a call for privacy-by-design principles to be embedded by developers. P9 detailed technical expectations such as anonymization for sensitive information, end-to-end encryption for data transmission, and clarity on data storage practices (local vs. cloud, and potential sharing with other entities or across borders, citing *"political risks"* and data sovereignty concerns like *"data placed in Ireland or Singapore"*).

The role of government and regulatory oversight is also highlighted. P2 stated, *"the government should do more risk screening and control."* P14, suggesting AI algorithms for risk detection on social platforms, implied a need for standards or enforcement: *"if platforms can detect that you have some risks when you click publish, and a warning box pops up, I think this might be a better way."* P17 acknowledged that VLM data use *"will also be subject to national supervision and legal monitoring,"* expressing a hope that this oversight ensures platforms *"rigorously review uploads for accidental PII ... and provide feedback to the user."* This indicates an expectation that external authorities will enforce accountability and baseline privacy standards for VLM operations.

Empowered and cautious user practices. While emphasizing the responsibilities of platforms and regulators, 5 participants also foresee a future where users themselves adopt empowered and cautious practices. This involves a heightened sense of awareness and more deliberate decision-making regarding content creation and sharing. P10 suggested temporal or geographical displacement of posts (*"If I went to Shanghai today, I'll post about it a few days later ... from a different region"*) or even altering self-presentation (*"dress to look poorer"*). P11 indicated a willingness to *"modify personal information before publishing again"* or *"no longer use AI to make videos in that area"* if risks are identified.

P17 described a future approach involving increased personal diligence: *"I might pay more attention to my uploading behavior... review my videos multiple times... to see if any points that could leak information ... and I might be less active in uploading to media... [I will] control the desire to share and not upload all my daily life to social media."* This suggests a move towards more cautious participation, where users actively curate their online persona and manage their "sharing desire" in response to the capabilities of VLMs.

4.5.3 Privacy-utility balancing perspectives. Participants engage in a nuanced and often highly individualized calculus when balancing the perceived utility of VLM functionalities against their privacy concerns. This balancing act is not uniform. It varies significantly based on the type of information in question, the specific context of VLM use, and individual user thresholds for privacy.

Hierarchical sensitivity of information and utility-driven disclosure. A primary factor governing the privacy-utility balance is the perceived sensitivity of the information, with users demonstrating a tiered approach to what they are willing to share. P16 and P17 clearly distinguish between general categorizations or functional data and core personally identifiable or highly sensitive information. P16, for instance, is amenable to VLMs using *"broad category"* information, such as identifying them as a *"student or a running enthusiast,"* to receive "helpful" personalized recommendations for *"student supplies"* or *"products they like."* However, this acceptance sharply contrasts with a strong objection to the disclosure of data that enables direct, unsolicited contact: *"I don't like it when they use my phone number or email for promotional marketing."* Similarly, P17 is *"willing to tell"* a shopping app their *"size, height, and weight"* for clothes selection, or a news app their preferences for content, as this data directly serves the app's explicit function. This willingness is strictly conditional on utility. P15 suggests that any information obscured (e.g., via mosaics) is *"always unimportant,"* implying a binary view where private information should not be compromised for utility, stating they would *"never trade-off privacy to post something."* This highlights that while some users engage in a granular cost-benefit analysis per information type, others adopt a blanket protective posture.

5 General Discussions

5.1 The Emergence and Prevalence of Inference Risks

Early works primarily focused on privacy risks unknown to users, such as location information hidden within photo metadata [21]. There are also many works on social media which focus on other people's privacy [11, 12] (e.g., bystanders [24]), the potential privacy disclosure [40] and surveillance [25]. Different from images [30], videos brought the additional dimension of temporality, which might bring additional privacy risks from inferring these videos. Regarding inferential privacy risks, while users gain some awareness after being explicitly informed, they often struggle to detect these risks independently. Such risks are not only prevalent in widespread social media platforms [13, 30] but can also raise during task completion with smart wearable glasses [10] and on numerous self-logging devices [6], potentially leading to diverse threats. Our risks of inference privacy primarily targets the owner itself, as the inferences, and the resulted user profiles are usually interconnected with the social media account [26], rather than on the bystanders shot, and we acknowledged that future work could explore the potential threats around inference even on a bystander recorded in the videos, especially under VLM-induced data aggregation from different sources [46], which could potentially profile the bystander.

Our work focused on the inference problems of videos, which service providers could use the temporal, and even causal relationships for inference. As evidenced by users, the videos uploaded on the social media are often edited, which means attackers may aggregate videos from different perspectives for inferring users' attributes [2]. This method, similar to self-consistency [47], may results in stronger inference capability, posing greater harms than inferring from pictures [30], warranting further attention.

Therefore, we advocate for improved privacy management methods. This call is partly driven by the inadequacy of current consent mechanisms [3]. These mechanisms fail to inform users about inferential risks, much less enable granular control over such privacy-sensitive data. We propose exploring automated, content-aware fine-grained control methods [44] and leveraging them to address

inferential privacy risks [59]. For instance, this could involve performing inference using on-device models, presenting the inferred information to users, and allowing them to edit it via an interface.

5.2 The Duality of Inference

We followed Staab et al. [41] for deciding the personal attributes of the inference, where different information has different chances of being inferred. In the realm of recommendations, inferred user profiles can contain diverse aspects beyond our current scope [48, 55], and may benefit users in the form of providing personalized service [3]. However, in our interview, most of the users mainly articulated the negative aspects of inferences rather than the positive aspects, citing it as a threat rather than benefits. This reflects participants' privacy calculus [28] that the benefits seems minimal in comparison to the negative consequences, especially on social medias. This could be because that participants held an opaque models about the data usage before, therefore cannot easily think out beneficial usage. We acknowledged the duality of this inference, and deemed the detailed trade-offs of this inference as the future work.

5.3 Accuracy of Inference and their Effects

Our interviews revealed users hold dichotomous perceptions of erroneous inferences from VLMs. One user group believes the model memorizes these errors, compelling them to make corrections to prevent future service degradation and viewing the inaccuracies as systemic flaws. In contrast, a second group dismisses incorrect inferences as inconsequential to their privacy and thus ignores them. Regardless of their mental model, users consistently reported that such errors induce confusion and cognitive load, fostering a perception of the system as "broken".

Fundamentally, VLM inferences are probabilistic estimations estimations based on limited cues and are thus inherently prone to inaccuracy and bias [41, 59]. These errors can cause harmful misrepresentations, such as stereotyping a user wearing a skirt as female. We argue that potential harm is not limited to accurate inference. Therefore, VLMs should refrain from making definitive but potentially incorrect claims. Instead, when inferential confidence is low, systems should be designed to express ambiguity or transparently solicit user clarification.

5.4 Complexity and Burden of User-Driven Privacy Mitigation

We found that users employ proactive content obfuscation methods to mitigate privacy risks. However, these user-driven mitigation strategies often prove insufficient. Many users propose simply avoiding the upload of "risky" videos altogether, which significantly compromises usability. For users, blurring individual objects within videos is cumbersome, especially given the current lack of robust video privacy editing tools. Prior anonymization techniques [44, 60] and approaches to inferential privacy [61] are predominantly image-based, failing to account for the dynamic nature of video and its associated challenges. Therefore, we advocate for improved methods that effectively balance user effort with user control and agency. This could involve direct user control or the adoption of automated method, where intelligent agents assist users by learning and applying their privacy preferences [57, 58].

5.5 Design Implications

Our findings reveal a significant discrepancy between the inferential capabilities of VLMs and user perception. We therefore propose the following design implications, spanning interface design, system architecture, and platform governance, to foster a trustworthy and user-centric VLM ecosystem.

Implication 1: Designing for inferential awareness. To mitigate user distress from unexpected and unsettlingly accurate inferences, systems must enhance inferential awareness. Instead of merely declaring what is inferred, interfaces should visually ground inferences in real-time by highlighting the specific visual evidence that contributes to a conclusion, such as cues indicating a location. This approach makes the inferential process transparent and scrutable to the user.

Implication 2: Empowering user agency with proactive mitigation tools. Our study shows that users perceive manual mitigation strategies as both ineffective and burdensome. To better empower users, systems should integrate proactive, AI-driven tools. Such tools would automatically detect and flag potential privacy risks in content prior to publication, offering users summarized assessments and one-click mitigation solutions, thereby reducing user burden [61].

Implication 3: Shifting the burden through systemic safeguards and accountable governance. Users strongly advocate for shifting the onus of privacy protection from individual effort to platform-level responsibility. This necessitates systemic safeguards, including architecturally enforced and verifiable data lifecycle controls to rebuild trust. Furthermore, platforms must replace blanket, one-time consent models with granular consent mechanisms that require separate user approval for distinct classes of data inference, thereby aligning with users' preferences for selective sharing [44].

5.6 Limitations

We acknowledge that our papers have three limitations. First, our recruitment is grounded in China, and the experiment involved solely Chinese participants with their experience primarily on Chinese platforms. Our recruitment is also biased towards university students. This may cause bias as Chinese users may create and upload different topics of videos on the social media platform. We envisioned a cross-cultural comparison of the inference risk as our future work. Second, although we provided vignettes for users and let them to reflect on their uploaded videos as examples, our qualitative study may subject to recall bias and desirability bias, which may not reflect the real opinions of participants. Thirdly, our experiment may not reflect the real capabilities of VLMs, as our experiment primarily involves interviews on participants about their mental models, strategies and expectations. Although the capability of VLMs are evolving, we believe the fundamental threats and risks are insightful for the future work.

Acknowledgments

This work was supported by the Natural Science Foundation of China under Grant No. 62472243 and 62132010.

References

[1] Adrian Aiordachioae and Radu-Daniel Vatavu. 2019. Life-tags: a smartglasses-based system for recording and abstracting life with tag clouds. *Proceedings of the ACM on human-computer interaction* 3, EICS (2019), 1–22.

[2] Torin Anderson and Shuo Niu. 2025. Making AI-Enhanced Videos: Analyzing Generative AI Use Cases in YouTube Content Creation. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–7.

[3] Sumit Asthana, Jane Im, Zhe Chen, and Nikola Banovic. 2024. " I know even if you don't tell me": Understanding Users' Privacy Preferences Regarding AI-based Inferences of Sensitive Information for Personalization. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–21.

[4] Michael Bailey, David Dittrich, Erin Kenneally, and Doug Maughan. 2012. The menlo report. *IEEE Security & Privacy* 10, 2 (2012), 71–75.

[5] Natã M Barbosa, Zhuohao Zhang, and Yang Wang. 2020. Do Privacy and Security Matter to Everyone? Quantifying and Clustering {User-Centric} Considerations About Smart Home Device Adoption. In *Sixteenth Symposium on Usable Privacy and Security (SOUPS 2020)*. 417–435.

[6] Matthew Barker-Canler, Daniel Gooch, Janet Van Der Linden, and Marian Petre. 2024. Flexible minimalist self-tracking to support individual reflection. *ACM Transactions on Computer-Human Interaction* 31, 3 (2024), 1–35.

[7] Tom L Beauchamp et al. 2008. The belmont report. *The Oxford textbook of clinical research ethics* (2008), 149–155.

[8] Divyanshu Bhardwaj, Alexander Ponticello, Shreya Tomar, Adrian Dabrowski, and Katharina Krombholz. 2024. In Focus, Out of Privacy: The Wearer's Perspective on the Privacy Dilemma of Camera Glasses. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–18.

[9] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712* (2023).

[10] Runze Cai, Nuwan Janaka, Hyeongcheol Kim, Yang Chen, Shengdong Zhao, Yun Huang, and David Hsu. 2025. AiGet: Transforming Everyday Moments into Hidden Knowledge Discovery with AI Assistance on Smart Glasses. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–26.

[11] Modesto Castrillón-Santana, Elena Sánchez-Nielsen, David Freire-Obregón, Oliverio J Santana, Daniel Hernández-Sosa, and Javier Lorenzo-Navarro. 2023. Evaluation of a Visual Question Answering Architecture for Pedestrian Attribute Recognition. In *International Conference on Computer Analysis of Images and Patterns*. Springer, 13–22.

[12] Xinhua Cheng, Mengxi Jia, Qian Wang, and Jian Zhang. 2022. A simple visual-textual baseline for pedestrian attribute recognition. *IEEE Transactions on Circuits and Systems for Video Technology* 32, 10 (2022), 6994–7004.

[13] SoeYoon Choi. 2023. Privacy literacy on social media: Its predictors and outcomes. *International Journal of Human–Computer Interaction* 39, 1 (2023), 217–232.

[14] Victoria Clarke and Virginia Braun. 2014. Thematic analysis. In *Encyclopedia of critical psychology*. Springer, 1947–1952.

[15] Tamara Denning, Zakariya Dehlawi, and Tadayoshi Kohno. 2014. In situ with bystanders of augmented reality glasses: Perspectives on recording and privacy-mediating technologies. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2377–2386.

[16] Vasisht Duddu and Antoine Boutet. 2022. Inferring sensitive attributes from model explanations. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 416–425.

[17] Motahhare Eslami, Sneha R Krishna Kumaran, Christian Sandvig, and Karrie Karahalios. 2018. Communicating algorithmic process in online behavioral advertising. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–13.

[18] Asbjørn Følstad, Cecilie Bertinussen Nordheim, and Cato Alexander Bjørkli. 2018. What makes users trust a chatbot for customer service? An exploratory interview study. In *Internet Science: 5th International Conference, INSCI 2018, St. Petersburg, Russia, October 24–26, 2018, Proceedings 5*. Springer, 194–208.

[19] Erin Griffiths, Salah Assana, and Kamin Whitehouse. 2018. Privacy-preserving image processing with binocular thermal cameras. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4 (2018), 1–25.

[20] David Harborth and Alisa Frik. 2021. Evaluating and redefining smartphone permissions with contextualized justifications for mobile augmented reality apps. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*. 513–534.

[21] Benjamin Henne and Matthew Smith. 2013. Awareness about photos on the web and how privacy-privacy-tradeoffs could help. In *Financial Cryptography and Data Security: FC 2013 Workshops, USEC and WAHC 2013, Okinawa, Japan, April 1, 2013, Revised Selected Papers 17*. Springer, 131–148.

[22] Yiqing Hua, Shuo Niu, Jie Cai, Lydia B Chilton, Hendrik Heuer, and Donghee Yvette Wohn. 2024. Generative AI in user-generated content. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–7.

[23] Jaeyeon Jung and Matthai Philipose. 2014. Courteous glass. In *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing: Adjunct publication*. 1307–1312.

[24] Marion Koelle, Swamy Ananthanarayan, Simon Czupalla, Wilko Heuten, and Susanne Boll. 2018. Your smart glasses' camera bothers me! exploring opt-in and opt-out gestures for privacy mediation. In *Proceedings of the 10th Nordic Conference on Human-Computer Interaction*. 473–481.

[25] Marion Koelle, Matthias Kranz, and Andreas Möller. 2015. Don't look at me that way! Understanding user attitudes towards data glasses usage. In *Proceedings of the 17th international conference on human-computer interaction with mobile devices and services*. 362–372.

[26] Anastasia Kozyreva, Philipp Lorenz-Spreen, Ralph Hertwig, Stephan Lewandowsky, and Stefan M Herzog. 2021. Public attitudes towards algorithmic personalization and use of personal data online: Evidence from Germany, Great Britain, and the United States. *Humanities and Social Sciences Communications* 8, 1 (2021), 1–11.

[27] Josephine Lau, Benjamin Zimmerman, and Florian Schaub. 2018. Alexa, are you listening? Privacy perceptions, concerns and privacy-seeking behaviors with smart speakers. *Proceedings of the ACM on human-computer interaction* 2, CSCW (2018), 1–31.

[28] Robert S Laufer and Maxine Wolfe. 1977. Privacy as a concept and a social issue: A multidimensional developmental theory. *Journal of social Issues* 33, 3 (1977), 22–42.

[29] Yu-li Liu, Wenjia Yan, Bo Hu, Zhuoyang Li, and Yik Ling Lai. 2022. Effects of personalization and source expertise on users' health beliefs and usage intention toward health chatbots: Evidence from an online experiment. *Digital Health* 8 (2022), 20552076221129718.

[30] Ying Ma, Shiquan Zhang, Dongju Yang, Zhanna Sarsenbayeva, Jarrod Knibbe, and Jorge Goncalves. 2025. Raising Awareness of Location Information Vulnerabilities in Social Media Photos using LLMs. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–14.

[31] Nathan Malkin, David Wagner, and Serge Egelman. 2022. Runtime permissions for privacy in proactive intelligent assistants. In *Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022)*. 633–651.

[32] Shagufta Mehnaz, Sayanton V Dibbo, Ehsanul Kabir, Ninghui Li, and Elisa Bertino. 2022. Are your sensitive attributes private? novel model inversion attribute inference attacks on classification models. In *31st USENIX Security Symposium (USENIX Security 22)*. 4579–4596.

[33] George R Milne, George Pettinico, Fatima M Hajjat, and Ereni Markos. 2017. Information sensitivity typology: Mapping the degree and type of risk consumers perceive in personal data sharing. *Journal of Consumer Affairs* 51, 1 (2017), 133–161.

[34] Robert Nimmo, Marios Constantinides, Ke Zhou, Daniele Quercia, and Simone Stumpf. 2024. User Characteristics in Explainable AI: The Rabbit Hole of Personalization?. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–13.

[35] Judith S Olson, Jonathan Grudin, and Eric Horvitz. 2005. A study of preferences for sharing and privacy. In *CHI'05 extended abstracts on Human factors in computing systems*. 1985–1988.

[36] Tribhuvanesh Orekondy, Mario Fritz, and Bernt Schiele. 2018. Connecting pixels to privacy and utility: Automatic redaction of private information in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 8466–8475.

[37] Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. 2017. Towards a visual privacy advisor: Understanding and predicting privacy risks in images. In *Proceedings of the IEEE international conference on computer vision*. 3686–3695.

[38] Emilee Rader, Samantha Hautea, and Anjali Munasinghe. 2020. " I Have a Narrow Thought Process": Constraints on Explanations Connecting Inferences and {Self-Perceptions}. In *Sixteenth Symposium on Usable Privacy and Security (SOUPS 2020)*. 457–488.

[39] Joel Reardon, Álvaro Feal, Primal Wijesekera, Amit Elazari Bar On, Narseo Vallina-Rodriguez, and Serge Egelman. 2019. 50 ways to leak your data: An exploration of apps' circumvention of the android permissions system. In *28th USENIX security symposium (USENIX security 19)*. 603–620.

[40] Cong Shi, Xiangyu Xu, Tianfang Zhang, Payton Walker, Yi Wu, Jian Liu, Nitesh Saxena, Yingying Chen, and Jiadi Yu. 2021. Face-Mic: inferring live speech and speaker identity via subtle facial dynamics captured by AR/VR motion sensors. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*. 478–490.

[41] Robin Staab, Mark Vero, Mislav Balunovic, and Martin Vechev. [n. d.]. Beyond Memorization: Violating Privacy via Inference with Large Language Models. In *The Twelfth International Conference on Learning Representations*.

[42] Julian Steil, Marion Koelle, Wilko Heuten, Susanne Boll, and Andreas Bulling. 2019. Privaceye: privacy-preserving head-mounted eye tracking using egocentric scene image and eye movement features. In *Proceedings of the 11th ACM symposium on eye tracking research & applications*. 1–10.

[43] Batuhan Tömekçe, Mark Vero, Robin Staab, and Martin Vechev. 2024. Private Attribute Inference from Images with Vision-Language Models. *arXiv preprint arXiv:2404.10618* (2024).
[44] Hari Venugopalan, Zainul Abi Din, Trevor Carpenter, Jason Lowe-Power, Samuel T King, and Zubair Shafiq. 2024. Aragorn: A privacy-enhancing system for mobile cameras. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 4 (2024), 1–31.
[45] M Vimalkumar, Sujeet Kumar Sharma, Jang Bahadur Singh, and Yogesh K Dwivedi. 2021. 'Okay google, what about my privacy?': User's privacy perceptions and acceptance of voice based digital assistants. *Computers in Human Behavior* 120 (2021), 106763.
[46] Xiao Wang, Jiandong Jin, Chenglong Li, Jin Tang, Cheng Zhang, and Wei Wang. 2024. Pedestrian attribute recognition via clip based prompt vision-language fusion. *IEEE Transactions on Circuits and Systems for Video Technology* (2024).
[47] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. [n. d.]. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *The Eleventh International Conference on Learning Representations*.
[48] Yifan Wang, Weizhi Ma, Min Zhang, Yiqun Liu, and Shaoping Ma. 2023. A survey on the fairness of recommender systems. *ACM Transactions on Information Systems* 41, 3 (2023), 1–43.
[49] Jeffrey Warshaw, Nina Taft, and Allison Woodruff. 2016. Intuitions, analytics, and killing ants: inference literacy of high school-educated adults in the {US}. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*. 271–285.
[50] Ben Weinshel, Miranda Wei, Mainack Mondal, Euirim Choi, Shawn Shan, Claire Dolin, Michelle L Mazurek, and Blase Ur. 2019. Oh, the places you've been! User reactions to longitudinal transparency about third-party web tracking and inferencing. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*. 149–166.
[51] Craig E Wills and Mihajlo Zeljkovic. 2011. A personalized approach to web privacy: awareness, attitudes and actions. *Information Management & Computer Security* 19, 1 (2011), 53–73.
[52] Zhenyu Wu, Haotao Wang, Zhaowen Wang, Hailin Jin, and Zhangyang Wang. 2020. Privacy-preserving deep action recognition: An adversarial learning framework and a new dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 4 (2020), 2126–2139.
[53] Yaxing Yao, Justin Reed Basdeo, Oriana Rosata Mcdonough, and Yang Wang. 2019. Privacy perceptions and designs of bystanders in smart homes. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–24.
[54] Yaxing Yao, Davide Lo Re, and Yang Wang. 2017. Folk models of online behavioral advertising. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 1957–1969.
[55] Jizhi Zhang, Keqin Bao, Yang Zhang, Wenjie Wang, Fuli Feng, and Xiangnan He. 2023. Is chatgpt fair for recommendation? evaluating fairness in large language model recommendation. In *Proceedings of the 17th ACM Conference on Recommender Systems*. 993–999.
[56] Shuning Zhang and Shixuan Li. 2024. " Confrontation or Acceptance": Understanding Novice Visual Artists' Perception towards AI-assisted Art Creation. *arXiv preprint arXiv:2410.14925* (2024).
[57] Shuning Zhang, Ying Ma, YongquanOwen' Hu, Ting Dang, Hong Jia, Xin Yi, and Hewu Li. 2025. From Patient Burdens to User Agency: Designing for Real-Time Protection Support in Online Health Consultations. *arXiv preprint arXiv:2508.00328* (2025).
[58] Shuning Zhang, Ying Ma, Xin Yi, and Hewu Li. 2025. Evaluating the Efficacy of Large Language Models for Generating Fine-Grained Visual Privacy Policies in Homes. *arXiv preprint arXiv:2508.00321* (2025).
[59] Shuning Zhang, Lyumanshan Ye, Xin Yi, Jingyu Tang, Bo Shui, Haobin Xing, Pengfei Liu, and Hewu Li. 2024. " Ghost of the past": identifying and resolving privacy leakage from LLM's memory through proactive user interaction. *arXiv preprint arXiv:2410.14931* (2024).
[60] Shuning Zhang, Xin Yi, Haobin Xing, Lyumanshan Ye, Yongquan Hu, and Hewu Li. 2024. Adanonymizer: Interactively Navigating and Balancing the Duality of Privacy and Output Performance in Human-LLM Interaction. *arXiv preprint arXiv:2410.15044* (2024).
[61] Ziyang Zhang, Chong Bao, Xiaokun Pan, Chia-Ming Chang, Takeo Igarashi, and Guofeng Zhang. 2025. Through the Lens of Privacy: Exploring Privacy Protection in Vision-Language Model Interactions on Smart Glasses. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–8.

A Ethics Considerations

We acknowledged that our paper may have ethical concerns. We meticulously designed the study and considered all aspects related to the ethics. We followed Menlo Report [4] and Belmont Report [7] in designing the experiment, and our study got the approval of our institution's Institutional Review Board (IRB). Our study aims to understand users' perceptions towards VLM's inference on users' videos uploaded to the social media, and the potential privacy implications. Our results provided implications for designing ethical systems and platforms, which could benefit end-users. Before the study, we provided users' the user consent, and got the acceptance before conducting the study. We informed participants that our study is completely voluntary, and participants could quit the study at any time without penalty, and did not need to provide any explanation. We compensated participants appropriately according to the local wage standard.

B Interview Script

Our interview protocol was designed to explore participants' perception with AI-generated private attribute inference, their mitigation strategies, and the effectiveness of their methods. The interview includes several sequentially progressive research questions. The original interview script was conducted in simplified Chinese to ensure natural communication with participants, and has been translated into English for presentation in this paper.

B.1 Research Questions

- Have you ever used a large model with visual capabilities? If so, have you ever felt that these models sometimes infer unexpected information from images? If so, can you give an example that left a strong impression on you?
- Do you think a large model can infer the aforementioned information (e.g., age range, gender) from videos you have previously filmed? If so, how do you think it makes these inferences? If not, why?
- Do you think if someone else filmed such a video (like the one you previously provided), it could infer the corresponding privacy attributes of the video creator (e.g., age range, gender)? If so, how would you infer this? If not, what conditions do you think are missing?
- Do you think this poses any negative risks to you? Specifically, what would you be concerned about?
- In what tasks or situations do you think the situation described above could specifically invade your privacy? Can you give an example that comes to your mind, or one that you think is the most relevant?
- In this example, what measures would you currently take to avoid such risks? Or, did you not realize these risks before? Can you describe the methods you are now taking?
- Do you think the methods you have taken are effective? Why?
- What methods do you think could be implemented in the future, such as interface-based or reminder box-based solutions, to avoid current privacy risks?

C User Consent

We are a research group from XXX University, investigating on users' perception of AI-generated private attribute inference. The inference is mainly based on visual data that users upload to the model, especially VLMs, which can be analyzed and used to infer personal information of users, leading to privacy risks. Our study's

focus is to understand your perception on the private inference of AI, your concerns of privacy risks caused by this, as well as your mitigation strategies and their effectiveness.

The interview would take approximately 30-60 minutes depending on its content, and would be audio recorded, and transcribed for academic analysis and publication. We would not use your material including the audio and transcribed text for any other usage than outlined above. Your participation is completely voluntary, and you has the right to withdraw at any time without penalty or explanation. Your data would be kept confidential and anonymized before processing. If you complete the experiment, you could get a compensation of 90RMB per hour.

If you have any other questions, you could contact XXXX (Email: XXXX) for further clarification.