A Multimodal AI Approach for Predicting Personal Privacy Preferences in Photo Sharing

Extended Abstract

Yuqi Niu
Shanghai Jiao Tong University
Shanghai, China
niuyuqi@sjtu.edu.cn

Kejia Zhang
University of Edinburgh
Edinburgh, United Kingdom
k.zhang-61@sms.ed.ac.uk

Weidong Qiu*
Shanghai Jiao Tong University
Shanghai, China
qiuwd@sjtu.edu.cn

Nadin Kökciyan
University of Edinburgh
Edinburgh, United Kingdom
nadin.kokciyan@ed.ac.uk

ABSTRACT

Social media users frequently share photos containing others' faces without explicit consent, raising significant privacy risks. Existing face privacy protection methods face a major limitation: they cannot reliably predict individuals' diverse privacy preferences, particularly with limited training data. To address this challenge, we analyze a large-scale public survey dataset, examining both the influence of contextual factors on privacy decisions and the variability of personalized privacy needs across individuals. Based on these insights, we propose a novel framework that leverages multimodal large language models to predict users' privacy preferences for public photo sharing using in-context learning. Our framework achieves strong performance on the dataset, reaching 84.45% accuracy and improving over baselines by up to 18.77%. Moreover, for 87.5% of 486 users, the model correctly predicts at least 75% of the 16 scenarios. These results demonstrate the potential of our approach to advance automated privacy protection on social media by incorporating individual preferences and enabling face-level privacy management.

KEYWORDS

Privacy; Photo-sharing; Social media

ACM Reference Format:
Yuqi Niu, Kejia Zhang, Weidong Qiu, and Nadin Kökciyan. 2026. A Multimodal AI Approach for Predicting Personal Privacy Preferences in Photo Sharing: Extended Abstract. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 3 pages. https://doi.org/10.65109/QNBC3731

1 INTRODUCTION

The widespread use of smart cameras and social media enables massive sharing of user-generated visual content. Platforms like Instagram, TikTok, and YouTube each host over 1.5 billion monthly active users [15]. Shared content often includes others' faces, and when privacy preferences differ without explicit consent, multi-user privacy conflicts (MPCs) may arise [16], potentially violating regulations such as GDPR [18]. Facial data exposure also raises risks of cyberbullying, identity theft, and re-identification [3, 17].

Prior work has proposed face protection mechanisms [13], yet most operate at the photo level and ignore individual privacy preferences [1, 4, 5, 14]. Face-level approaches based on recognition or fixed rules [7, 9, 10] often trade usability for protection. Moreover, existing designs often do not explicitly account for the mismatch between stated preferences and actual photo-sharing decisions, i.e., the privacy paradox. Consequently, automated tools for accurately predicting individual privacy preferences remain limited.

Recent studies [2, 6, 12, 20] show that privacy perceptions in photo sharing are shaped by both visual and contextual factors. Some Internet of Things (IoT) privacy solutions attempt to predict users' privacy decisions from contextual cues, but they typically require substantial historical data to train models [11] or construct experience bases [8]. Moreover, their reasoning processes are often insufficiently transparent to end-users. In photo sharing settings, computational approaches that work with limited data and jointly model visual and contextual factors are still lacking.

Our analysis of a real-world photo-sharing dataset [12] shows that privacy perceptions depend on contextual factors and vary across individuals. Self-reported privacy statements often do not align with actual decisions, and group-level patterns do not reliably predict individuals. Motivated by these findings, we propose a face-level privacy-preserving framework that predicts personalized privacy preferences using in-context learning [19] of multimodal large language models (MLLMs). The framework operates in two stages: constructing a user preference profile from stated preferences and a small set of past decisions, and predicting comfort for new scenarios using rule-based matching with MLLM fallback explanations. Experiments show that our method achieves 84.45% accuracy in a few-shot setting.

2 ANALYSIS OF PHOTO-SHARING SCENARIOS

To understand user privacy behavior in photo sharing, we conducted a manual consistency analysis on a real-world dataset [12]

*Corresponding author: qiuwd@sjtu.edu.cn.

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/QNBC3731

Table 1: Prediction performance under leave-one-out evaluation. Left: overall metrics averaged over all user–scenario predictions. Right: per-user results (number of users with x correct predictions out of 16 scenarios; percentages relative to 486 participants).

	Overall				Per-user				
Method	Acc	Prec	Recall	F1	≤7	8–11	12–13	14–15	16
SL-S	0.6568	0.3812	0.5324	0.4443	30 (6.17%)	304 (62.55%)	136 (27.98%)	16 (3.29%)	0 (0.00%)
SL-F	0.6830	0.4138	0.5519	0.4730	23 (4.73%)	267 (54.94%)	174 (35.80%)	22 (4.53%)	0 (0.00%)
SL-V	0.7814	0.6407	0.3453	0.4488	20 (4.12%)	118 (24.28%)	157 (32.30%)	173 (35.60%)	18 (3.70%)
PPF-S	0.6731	0.4107	0.6187	0.4934	49 (10.08%)	232 (47.74%)	136 (27.98%)	56 (11.52%)	13 (2.67%)
PPF-E	0.7604	0.5467	0.4117	0.4697	13 (2.67%)	168 (34.57%)	158 (32.51%)	122 (25.10%)	25 (5.14%)
PPF-H-D	**0.8445**	**0.7455**	0.6023	**0.6663**	**3 (0.62%)**	**58 (11.93%)**	162 (33.33%)	**219 (45.06%)**	**44 (9.05%)**
PPF-H-F	0.7776	0.5556	**0.6851**	0.6136	5 (1.03%)	135 (27.78%)	**198 (40.74%)**	128 (26.34%)	20 (4.12%)

containing photo content, sharing context, user comfort, and self-reported privacy attitudes by mapping privacy statements to structured factors and comparing them with scenario-level ratings.

Our analysis yields five insights: (1) Visual content plays a central role, with comfort shaped by facial expression, sensitivity, and visual prominence; (2) Preferences vary under the same context, indicating strong personalization needs; (3) Subjective importance and statistical significance may diverge because self-reported influences on privacy comfort do not always align with group-level statistical effects; (4) Privacy orientation does not reliably predict behavior, as stated attitudes and scenario ratings are often inconsistent; (5) Privacy statements are often ambiguous and too vague to reflect actual decision logic.

3 PRIVACY PREDICTION FRAMEWORK

Our findings expose a gap between declared preferences and actual decisions, motivating methods that integrate contextual, visual, and behavioral signals to predict user comfort. We now introduce a two-stage framework to build a privacy prediction framework.

In the first stage, we construct a personalized privacy preference profile (PP_u) using a user's explicit privacy statement ($\mathcal{P}_u$), privacy score ($\mathcal{P}s_u$), historical decisions (HE), and a consensus rule base (CRB) derived from factors widely reported in prior research. We begin by extracting privacy-related factors and overall preferences (always comfortable f_c or uncomfortable f_{un}) from $\mathcal{P}_u$ using MLLM forming an initial factor map ($F_{\mathcal{P}_u}$). We then analyze each scenario in HE with the MLLM to identify factor–comfort pairs and store them in the contextual experience base (CEB_u). To reduce coverage bias and spurious generalization, examples are sampled from CEB_u to ensure factor–value coverage. For each determinable factor in $F_{\mathcal{P}_u}$, we retrieve its value set and select examples matching each factor–value condition. An upper bound k limits the number of sampled examples. If coverage is insufficient, additional cases are randomly drawn from CEB_u. From the sampled cases, we derive candidate rules that associate factor combinations with comfort outcomes. Conceptually, this corresponds to identifying observed factor–value combinations and their associated comfort labels. A rule is added to PP_u only when its factor combination appears in more than τ examples with consistent comfort labels. These validated rules are then merged with CRB to form the final PP_u. For users with extreme privacy scores or strong global preferences in $\mathcal{P}_u$, sampled decisions are further used to verify general tendencies. If at least n of k samples share the same comfort outcome, a corresponding global rule ($rule_c$ or $rule_{un}$) is added.

In the second stage, we extract contextual factors for a target scenario using the MLLM. If PP_u contains applicable rules, they are triggered when scenario factors match their conditions, producing rule-based decisions. If any rule indicates an *uncomfortable* outcome, the final decision is *uncomfortable*; otherwise, the MLLM estimates comfort from the scenario factors.

4 EXPERIMENT

We evaluate our approach as a binary classification task on a real-world dataset [12] (as discussed in Section 2), comparing models with three levels of personalization. (1) *Scene-Level (SL)* models use no user information: **SL-S** uses the photo and description, **SL-F** uses contextual factors, and **SL-V** uses descriptive factors. (2) *Personalized Preference (PPF)* models use user inputs without rule validation: **PPF-S** relies on privacy statements ($\mathcal{P}_u$) and **PPF-E** on prior decisions (CEB_u). (3) Our *Hybrid framework (PPF-H)* combines rule validation with MLLM reasoning: **PPF-H-D** uses descriptive factors and **PPF-H-F** uses descriptive and determinable factors.

For personalized methods, we set $k = 6$, $\tau = 2$, and $n = 6$. Experiments are conducted using Qwen-2.5-VL 32B on NVIDIA H100 GPUs. Table 1 reports results from two perspectives: aggregate prediction metrics (left) and per-user correctness distributions (right). The per-user view shows how many users obtain different numbers of correct predictions out of 16 scenarios, reflecting personalization quality. Hybrid methods (PPF-H) consistently outperform baselines, with PPF-H-D achieving the best overall accuracy and the highest proportion of users with 12 or more correct predictions.

5 CONCLUSION

In this study, we present a novel method for automatically predicting users' personalised privacy preferences when photos containing them are publicly shared. Our framework leverages in-context learning and reasoning capabilities of open-source MLLMs to identify factors that influence privacy perceptions from photo and context descriptions. With only a small number of samples, it reconciles discrepancies between self-reported preferences and actual choices, producing predictions that more closely reflect real-world behavior. Our evaluation shows that the method achieves accurate predictions without requiring extensive training data, addressing a key limitation of existing approaches. This represents an important step toward safeguarding facial privacy and fostering a more privacy-respecting digital environment.

REFERENCES

[1] Rawan Alharbi, Mariam Tolba, Lucia C. Petito, Josiah Hester, and Nabil Alshurafa. 2019. To Mask or Not to Mask?: Balancing Privacy with Visual Confirmation Utility in Activity-Oriented Wearable Cameras. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3, Article 72 (2019), 29 pages. https://doi.org/10.1145/3351230

[2] Mary Jean Amon, Aaron Necaise, Nika Kartvelishvili, Aneka Williams, Yan Solihin, and Apu Kapadia. 2023. Modeling User Characteristics Associated with Interdependent Privacy Perceptions on Social Media. *ACM Trans. Comput.-Hum. Interact.* 30, 3, Article 40 (June 2023), 32 pages. https://doi.org/10.1145/3577014

[3] Mark Andrejevic and Neil Selwyn. 2020. Facial recognition technology and the end of privacy for good. Web page. https://lens.monash.edu/2020/01/23/1379547/facial-recognition-tech-and-the-end-of-privacy

[4] Gonul Ayci, Murat Sensoy, Arzucan Özgür, and Pinar Yolum. 2023. Uncertainty-Aware Personal Assistant for Making Personalized Privacy Decisions. *ACM Trans. Internet Techn.* 23, 1 (2023), 13:1–13:24. https://doi.org/10.1145/3561820

[5] Mariella Dimiccoli, Juan Marín, and Edison Thomaz. 2018. Mitigating Bystander Privacy Concerns in Egocentric Activity Recognition with Deep Learning and Intentional Image Degradation. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 4, Article 132 (2018), 18 pages. https://doi.org/10.1145/3161190

[6] Rakibul Hasan, Bennett I. Bertenthal, Kurt Hugenberg, and Apu Kapadia. 2021. Your Photo is so Funny that I don't Mind Violating Your Privacy by Sharing it: Effects of Individual Humor Styles on Online Photo-sharing Behaviors. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. ACM, New York, NY, USA, Article 556, 14 pages. https://doi.org/10.1145/3411764.3445258

[7] Panagiotis Ilia, Iasonas Polakis, Elias Athanasopoulos, Federico Maggi, and Sotiris Ioannidis. 2015. Face/off: Preventing privacy leakage from photos in social networks. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*. ACM, 781–792. https://doi.org/10.1145/2810103.2813603

[8] Nadin Kökciyan and Pinar Yolum. 2022. Taking Situation-Based Privacy Decisions: Privacy Assistants Working with Humans. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence, IJCAI 2022*. ijcai.org, Vienna, Austria, 703–709. https://doi.org/10.24963/IJCAI.2022/99

[9] Ang Li, Wei Du, and Qinghua Li. 2018. PoliteCamera: Respecting Strangers' Privacy in Mobile Photographing. In *Security and Privacy in Communication Networks: 14th International Conference, SecureComm 2018, Singapore, Singapore, August 8-10, 2018, Proceedings, Part I (Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, Vol. 254)*. Springer, 227–247. https://doi.org/10.1007/978-3-030-01701-9_13

[10] Fenghua Li, Zhe Sun, Ang Li, Ben Niu, Hui Li, and Guohong Cao. 2019. Hideme: Privacy-Preserving Photo Sharing on Social Networks. In *Proceedings of 2019 IEEE Conference on Computer Communications*. IEEE, 154–162. https://doi.org/10.1109/INFOCOM.2019.8737466

[11] Pardis Emami Naeini, Sruti Bhagavatula, Hana Habib, Martin Degeling, Lujo Bauer, Lorrie Faith Cranor, and Norman M. Sadeh. 2017. Privacy Expectations and Preferences in an IoT World. In *Proceedings of 13th Symposium on Usable Privacy and Security*. USENIX Association, Santa Clara, CA, USA, 399–412. https://www.usenix.org/conference/soups2017/technical-sessions/presentation/naeini

[12] Yuqi Niu, Nicole Meng-Schneider, Weidong Qiu, and Nadin Kokciyan. 2025. "I am not the primary focus" - Understanding the Perspectives of Bystanders in Photos Shared Online. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, CHI 2025*. ACM, Yokohama, Japan, 899:1–899:23. https://doi.org/10.1145/3706598.3713826

[13] Yasmeen Rashidi, Tousif Ahmed, Felicia Patel, Emily Fath, Apu Kapadia, Christena Nippert-Eng, and Norman Makoto Su. 2018. "You don't want to be the next meme": College Students' Workarounds to Manage Privacy in the Era of Pervasive Photography. In *Proceedings of the 14th Symposium on Usable Privacy and Security*. USENIX Association, 143–157. https://www.usenix.org/conference/soups2018/presentation/rashidi

[14] Anna Cinzia Squicciarini, Cornelia Caragea, and Rahul Balakavi. 2017. Toward Automated Online Photo Privacy. *ACM Trans. Web* 11, 1 (2017), 2:1–2:29. https://doi.org/10.1145/2983644

[15] Statista GmbH. 2025. Most popular social networks worldwide as of February 2025, by number of monthly active users. Web page. https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

[16] Jose M. Such, Joel Porter, Sören Preibusch, and Adam Joinson. 2017. Photo Privacy Conflicts in Social Media: A Large-Scale Empirical Study. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, Denver, CO, USA, 3821–3832. https://doi.org/10.1145/3025453.3025668

[17] Nishant Vishwamitra, Hongxin Hu, Feng Luo, and Long Cheng. 2021. Towards Understanding and Detecting Cyberbullying in Real-world Images. In *Proceedings of 28th Annual Network and Distributed System Security Symposium, NDSS 2021*. The Internet Society, virtually. https://www.ndss-symposium.org/ndss-paper/towards-understanding-and-detecting-cyberbullying-in-real-world-images/

[18] Paul Voigt and Axel Von dem Bussche. 2017. *The EU General Data Protection Regulation (GDPR): A Practical Guide*. Vol. 10. https://doi.org/10.1007/978-3-319-57959-7

[19] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent Abilities of Large Language Models. *Trans. Mach. Learn. Res.* 2022 (2022). https://openreview.net/forum?id=yzkSU5zdwD

[20] Yuhang Zhao, Yaxing Yao, Jiaru Fu, and Nihan Zhou. 2023. "If sighted people know, I should be able to know: " Privacy Perceptions of Bystanders with Visual Impairments around Camera-based Technology. In *Proceedings of the 32nd USENIX Security Symposium, USENIX Security 2023*. USENIX Association, Anaheim, CA, USA, 4661–4678.