# Task-Aware Minimal Disclosure for Privacy-Preserving GenAI Photo Uploads

Goun Bae
*Clemson University*

Jinkyung Katie Park
*Clemson University*

## 1 Introduction

Generative AI (GenAI) systems increasingly support image-based interactions as Vision-Language Models (VLMs) advance [7, 8]. While useful for complex situational queries [8, 11], photo-based GenAI interactions heighten privacy risks, as uploaded images may be stored, leaked, or reused [1, 3, 4]. VLMs can infer private attributes, such as location or socioeconomic status [12, 13, 15]. Such risks may arise not only from attributes that are sensitive or identifying on their own, but also through the composition of seemingly benign visual attributes with other attributes or auxiliary evidence, which may enable sensitive inference, profiling, or identification [5, 14]. Yet, users often underestimate these implicit inferences [10].

Despite this, privacy support for image-based GenAI interactions remains underexplored. Existing visual privacy tools often obscure visibly sensitive regions (e.g., faces or PII) [2, 6, 9], but may leave contextual background cues exposed to VLM inference. Motivated by these limitations, we explore task-aware minimal disclosure: helping users share only the visual information needed for a specific GenAI photo query. We designed a prototype that locally analyzes an uploaded image and its accompanying prompt, surfaces potential privacy risks, including such compositional risks, and generates a minimized image for user review before submission to an external GenAI service. We ask: **How do users perceive, discover, and act on task-aware minimal-disclosure support during GenAI photo-upload interactions?** Through a preliminary survey with 28 GenAI users and an iterative formative usability study with 7 participants, we identify initial design priorities and early usability challenges that can inform future task-aware privacy tools for GenAI photo upload. Our findings highlight design tensions around making implicit inference risks visible, motivating use before upload, and avoiding feedback that feels either overly reassuring or invasive.


*USENIX Symposium on Usable Privacy and Security (SOUPS)* 2026.
August 23–26, 2026, Hannover, Germany.

## 2 Prototype Design and Formative Study

**2.1 Initial Survey** A survey with 28 GenAI users[1] revealed two design priorities. First, privacy support should be low-friction. Although 58% worried about photo uploads, 61% of them still uploaded original images, with 82% citing the hassle as the reason. Second, users need help recognizing implicit AI inferences. While 83% identified direct privacy cues, only 28% recognized contextual cues such as background objects. These gaps motivated our prototype, which embeds automated privacy support in the chat upload flow and makes potential privacy risks, including contextual inferences, visible before upload.

**2.2 Task-Aware Privacy Prototype Design** Based on these priorities, we implemented the prototype in a simulated chat-based GenAI interface (Figure 1). To avoid sending original photos to external GenAI servers, the tool runs locally and generates a minimized image before upload. Users can open *Privacy Mode* from the uploaded image thumbnail. Once activated, an on-device pipeline[2] analyzes the user's prompt and image to isolate task-relevant visual elements from the background. The tool then presents a two-panel review screen: the original image with hoverable warning icons for visual elements associated with potential privacy risks, including

[1] Survey topics included photo-upload experience, privacy concerns, protective actions and barriers, privacy-relevant visual cues, and understanding of photo processing/storage. Respondents were 18-44 years old; 66.7% identified as female, 29.2% as male, and 4.2% as non-binary/third gender.

[2] The local pipeline: FastVLM (element detection), Qwen3-4B (prompt analysis and task-relevance reasoning), and Florence-2-base (segmentation).

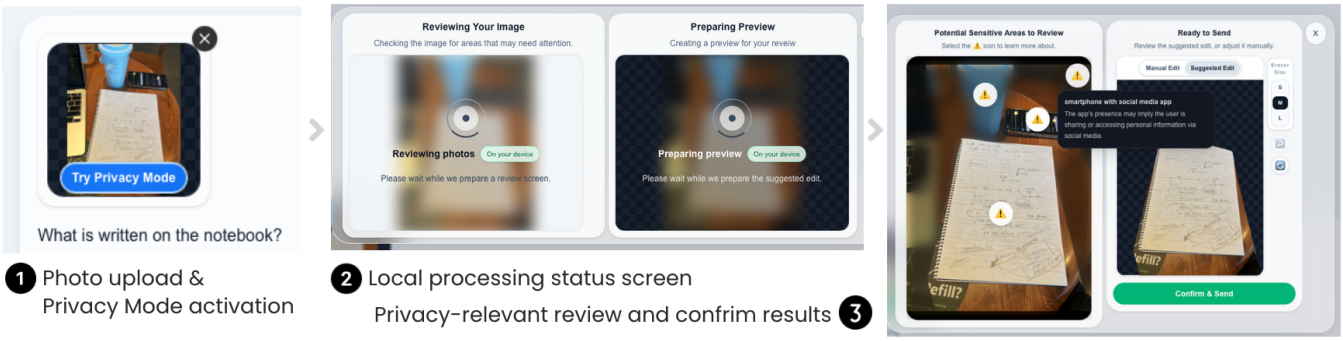



Figure 1: (1) Users activate *Privacy Mode* from the uploaded image thumbnail. (2) The image is processed locally. (3) The final view presents a suggested version for the user's query, alongside hoverable inference warnings on the original image.

compositional risks, and a minimal-disclosure version that removes unnecessary context. Users can accept the suggestion, adjust it manually, or continue with the original image.

**2.3 Formative Usability Study** The study was approved by Clemson University's IRB; participants completed informed consent before the session, and all tasks used researcher-provided photos (not participants' personal images). We conducted a remote scenario-based usability study via Zoom with 7 participants[3] to iteratively refine the prototype. Participants completed two photo-upload tasks: checking whether a sprouted carrot was edible and organizing photographed handwritten notes. They first used the simulated interface naturally without intervention, then tried *Privacy Mode* while thinking aloud, followed by an interview exploring their expectations, perceived usefulness, and understanding of the privacy features. Sessions lasted about 30 minutes on average.

## 3 Formative Findings & Next Steps

**3.1 Visibility alone did not motivate use because users expected privacy tools to target obvious sensitive content.** Our design iterations suggested that participants' initial non-use of *Privacy Mode* reflected not only discoverability, but also perceived need. The first three participants overlooked the original entry point next to the send button in a chat input, so we moved it onto the uploaded image thumbnail to make it more salient and contextually tied to the photo. After this change, the remaining four participants noticed it, but only two clicked it, suggesting that visibility alone was not sufficient to motivate use.

A key barrier appeared to be a mismatch in privacy expectations. Participants generally expected photo privacy tools to redact visibly sensitive regions. Because the photos in our study lacked such obvious triggers, they perceived no immediate privacy issue and felt little hesitation uploading them as-is. In particular, they did not initially consider the extent to which AI could infer sensitive information from an uploaded photo. As a result, the value of our tool was not apparent before participants entered it. These findings suggest that task-aware privacy tools should make their purpose clear early enough to support informed privacy decisions, not just be visible.

**3.2 Inference feedback raised awareness but could create false reassurance or feel invasive.** The inference panel helped participants recognize privacy considerations in a visual context that they had not initially considered. However, they also pointed to potential unintended effects of inference feedback. First, it could create a sense of reassurance when flagged elements appeared individually acceptable to share. While reviewing the inference panel, P2 noted that some flagged elements seemed acceptable to share, suggesting that users may treat the tool's feedback as a completed privacy check, while risks may still emerge from combined cues within an image or accumulated disclosures across repeated uploads [5, 14]. Second, inference feedback could make the tool feel invasive: P6 found the tool "creepy," noting that the system seemed to inspect the image too closely, even after understanding that the processing occurred locally. These reactions suggest that inference feedback should be framed carefully so that it supports privacy reflection without appearing as either a definitive privacy clearance or an overly invasive inspection.

**3.3 Next Steps** Our findings suggest that task-aware minimal-disclosure tools must solve two design problems: helping users understand why minimization is needed, and helping them evaluate whether the minimized image still supports their task. We plan to refine the prototype based on these insights and conduct a larger study to evaluate users' acceptance and sharing decisions in GenAI interactions.

[3] 5 female, 2 male; ages 25-34. Six had GenAI photo-upload experience; one never uploaded due to privacy. Sessions were screen-recorded.

## References


[1] Chen, T., Li, P., Zhou, K., Chen, T., and Wei, H. Unveiling privacy risks in multi-modal large language models: Task-specific vulnerabilities and mitigation challenges. In *Findings of the Association for Computational Linguistics: ACL 2025* (2025), pp. 4573–4586.

[2] Hasan, R., Li, Y., Hassan, E., Caine, K., Crandall, D. J., Hoyle, R., and Kapadia, A. Can privacy be satisfying? on improving viewer satisfaction for privacy-enhanced photos using aesthetic transforms. In *Proceedings of the 2019 CHI conference on human factors in computing systems* (2019), pp. 1–13.

[3] Kelley, P. G., Cornejo, C., Hayes, L., Jin, E. S., Sedley, A., Thomas, K., Yang, Y., and Woodruff, A. " there will be less privacy, of course": How and why people in 10 countries expect {AI} will affect privacy in the future. In *Nineteenth Symposium on Usable Privacy and Security (SOUPS 2023)* (2023), pp. 579–603.

[4] Lee, H.-P., Yang, Y.-J., Von Davier, T. S., Forlizzi, J., and Das, S. Deepfakes, phrenology, surveillance, and more! a taxonomy of ai privacy risks. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (2024), pp. 1–19.

[5] Liu, F., Zhang, Y., Huang, X., Peng, Y., Li, X., Wang, L., Shen, Y., Duan, R., Qin, S., Jia, X., et al. The eye of sherlock holmes: Uncovering user private attribute profiling via vision-language model agentic framework. In *Proceedings of the 33rd ACM International Conference on Multimedia* (2025), pp. 4875–4883.

[6] Monteiro, K., Wu, Y., and Das, S. Imago obscura: An image privacy ai co-pilot to enable identification and mitigation of risks. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology* (2025), pp. 1–26.

[7] OpenAI. ChatGPT can now see, hear, and speak. https://openai.com/index/chatgpt-can-now-see-hear-and-speak/, Sept. 2023. Accessed: 2026-05-16.

[8] OpenAI. Thinking with images. https://openai.com/index/thinking-with-images/, Apr. 2025. Accessed: 2026-05-16.

[9] Orekondy, T., Fritz, M., and Schiele, B. 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* (2018), pp. 8466–8475.

[10] Qia Wang, S., Peddinti, S. T., Taft, N., and Feamster, N. Beyond pii: How users attempt to estimate and mitigate implicit llm inference. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems* (2026), pp. 1–17.

[11] Sharma, T., Tseng, Y.-Y., Zhang, L., Ide, A., Mack, K. A., Findlater, L., Gurari, D., and Wang, Y. "before, i asked my mom, now i ask chatgpt": Visual privacy management with generative ai for blind and low-vision people. In *Proceedings of the 27th International ACM SIGACCESS Conference on Computers and Accessibility* (2025), pp. 1–14.

[12] Staab, R., Vero, M., Balunovic, M., and Vechev, M. Beyond memorization: Violating privacy via inference with large language models. In *International Conference on Learning Representations* (2024), vol. 2024, pp. 33832–33878.

[13] Tömekçe, B., Vero, M., Staab, R., and Vechev, M. Private attribute inference from images with vision-language models, 2024. Advances in Neural Information Processing Systems.

[14] Tsaprazlis, E., Feng, T., Ramakrishna, A., Karimireddy, S. P., Gupta, R., and Narayanan, S. Rethinking visual privacy: A compositional privacy risk framework for severity assessment with vlms. *arXiv preprint arXiv:2603.21573* (2026).

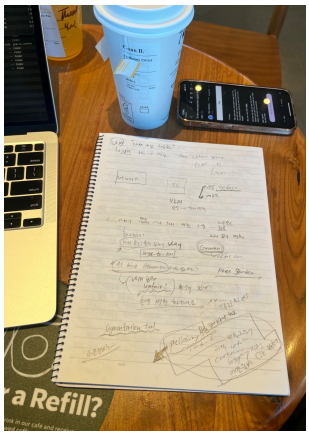
[15] Wang, X., Yi, X., Xie, X., and Jia, J. Specify privacy yourself: Assessing inference-time personalized privacy preservation ability of large vision-language models. In *Proceedings of the 33rd ACM International Conference on Multimedia* (2025), pp. 12304–12313.


## A Study Procedure

Each session consisted of four parts: (1) introduction and consent, (2) a natural-use photo-upload task without intervention, (3) a guided task in which participants used *Privacy Mode* while thinking aloud, and (4) a post-study interview. Participants accessed the prototype through a temporary study-specific URL. Sessions were conducted remotely over Zoom with screen sharing enabled, allowing the researchers to observe participants' interactions with the prototype and their think-aloud comments.

Participants completed two researcher-prepared photo-upload scenarios, each consisting of a study photo and a short contextual description. Scenario order was counterbalanced across participants. For each scenario, participants read the context, wrote their own prompt for the provided photo, and completed the task in the prototype. Table 1 summarizes these two scenarios.

Table 1: Photo-upload scenarios used in the formative study.

| **Sprouted carrot** | **Handwritten notes** |
|---|---|
|  |  |
| Participants were asked to imagine that they took a carrot out of the refrigerator, noticed that it had sprouted, and wanted to know whether it was still safe to eat. | Participants were asked to imagine that they had taken handwritten notes during a meeting and wanted help transcribing and organizing the notes. |

## B Prototype Interface Used in the Study

Figure 2 illustrates the three-step workflow of our simulated chat-based GenAI prototype. In the first stage, participants uploaded one of the study photos and wrote their own prompt. After a photo was uploaded, a *Try Privacy Mode* button appeared over the image thumbnail, allowing participants to open the privacy-support tool within the normal upload flow (Figure 2 (a)).

After participants activated *Privacy Mode*, the interface

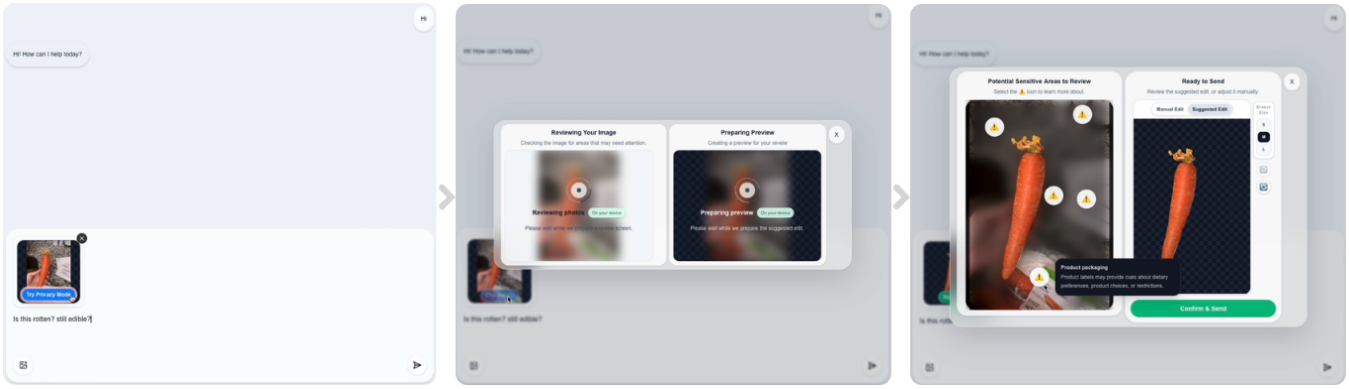


Figure 2: Full-screen screenshots of the prototype interface used. (a) Photo upload and *Try Privacy Mode* activation. (b) Local processing screen shown while the prototype reviewed the image and prepared a preview. (c) Two-panel review screen with privacy-relevant warnings, a suggested minimal-disclosure output, and manual editing options.

showed a processing screen while the prototype reviewed the image and prepared a suggested preview. This screen communicated that the image was being processed locally and gave users an idea of the upcoming step (Figure 2 (b)).

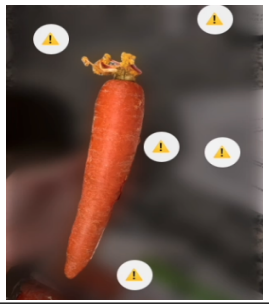
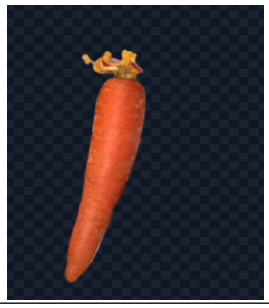
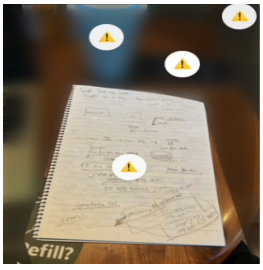
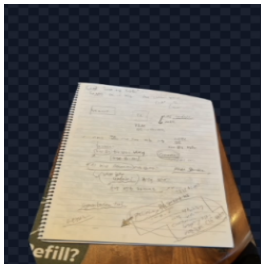
The final stage presented a two-panel review interface. The left panel showed the uploaded image with warning icons placed over visual cues that the prototype identified as potentially privacy-relevant. When participants hovered over an icon, a short explanation appeared describing why that visual element might raise a privacy concern.

The right panel showed the suggested minimal-disclosure output, which aimed to preserve information needed for the user's prompt while reducing unnecessary visual disclosure. The interface also provided a manual editing option so participants could further adjust the suggested output if it omitted relevant content or if they wanted to hide additional details (Figure 2 (c)).

## C Prototype-Generated Privacy Cues

Table 2 summarizes the privacy cues generated by the prototype. These cues were shown to participants as possible privacy-relevant visual details.

Table 2: Study stimuli, prototype-generated privacy cues, and minimal-disclosure outputs.

| Annotated image | Visual cue | Potential privacy implication | Minimal-disclosure output |
|---|---|---|---|
|  | Medication containers | May indicate the presence of health-related items or routines. |  |
| | Mail | May expose names, addresses, sender information, or account-related details. | |
| | Kitchen setup | May support inferences about diet, routines, or living environment. | |
| | Product packaging | Product labels may provide cues about dietary preferences, product choices, or restrictions. | |
| | Granite countertop | Home materials may support inferences the living environment or interior preference. | |
|  | Text on coffee cup | May reveal a name, order information, or other personal identifier. |  |
| | Smartphone screen | May reveal app content, including messages, webpages, or account-related information. | |
| | Handwritten notes | May contain personal, academic, or work-related information. | |
| | Background setting | May support inferences about routine, location, or socioeconomic status. | |