Low-Latency Privacy-Aware Robot Behavior guided by Automatically Generated Text Datasets

Yuta Irisawa[1], Tomoaki Yamazaki[1], Seiya Ito[2], Shuhei Kurita[3],
Ryota Akasaka[4], Masaki Onishi[5], Kouzou Ohara[1], Ken Sakurada[6]

***Abstract*— Humans typically avert their gaze when faced with situations involving another person's privacy, and humanoid robots should exhibit similar behaviors. Various approaches exist for privacy recognition, including an image privacy recognition model and a Large Vision-Language Model (LVLM). The former relies on datasets of labeled images, which raise ethical concerns, while the latter requires more time to recognize images accurately, making real-time responses difficult. To this end, we propose a method of automatically constructing the LLM Privacy Text Dataset (LPT Dataset), a privacy-related text dataset with privacy indicators, and a method of recognizing whether observing a scene violates privacy without ethically sensitive training images. In constructing the LPT Dataset, which consists of both private and public scenes, we use an LLM to define privacy indicators and generate texts scored for each indicator. Our model recognizes whether a given image is private or public by retrieving texts with privacy scores similar to the image in a multi-modal feature space. In our experiments, we evaluated the performance of our model on three image privacy datasets and a realistic experiment with a humanoid robot in terms of accuracy and responsibility. The experiments show that our approach identifies the private image as accurately as the highly tuned LVLM without delay.**

I. INTRODUCTION

Recent advancements in interactive humanoid robots are becoming increasingly integrated into our daily lives. Designed with human-like features, these robots are expected to exhibit natural, human-like behaviors, including consideration of the privacy of others. When confronted with privacy-sensitive situations, such as when someone changes clothes or enters a password, humans instinctively avert their gaze. Likewise, humanoid robots must demonstrate privacy awareness by responding appropriately in real time. Our preliminary survey (see Section III) indicates that individuals experience discomfort when being observed by robots in such contexts. For this purpose the robots are required to identify whether the image captured by the camera contains content that may infringe privacy without delay.

[1]Yuta Irisawa, Tomoaki Yamazaki, Kouzou Ohara are with Aoyama Gakuin University {yuta.irisawa@dslabo.org, yamazaki@it.aoyama.ac.jp, ohara@it.aoyama.ac.jp}
[2]Seiya Ito is with National Institute of Information and Communications Technology (NICT) seiya.ito@nict.go.jp
[3]Shuhei Kurita is with National Institute of Informatics (NII) skurita@nii.ac.jp
[4]Ryota Akasaka is with Osaka University akasaka@elsi.osaka-u.ac.jp
[5]Masaki Onishi is with National Institute of Advanced Industrial Science and Technology (AIST) onishi-masaki@aist.go.jp
[6]Ken Sakurada is with Kyoto University sakurada@i.kyoto-u.ac.jp

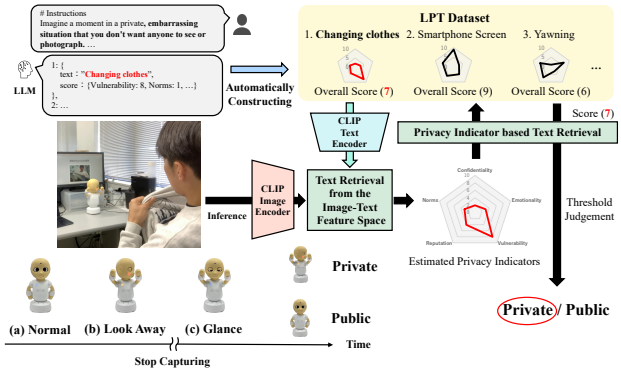


Fig. 1. Our methodology for recognizing a private image consists of two phases: 1) the automatic construction of a privacy-related text dataset with multiple privacy indicators, the LPT Dataset, and 2) image privacy recognition that uses texts through a two-stage image-text retrieval process. We integrate the proposed method with robots to realize a look-away action (b) in situations where it might violate an individual's privacy.

The primary approach to privacy recognition involves the development of image privacy recognition models [1]–[7]. Conventional methods develop these models with labeled image datasets [7]–[9]. However, the collection and annotation of private images raises significant ethical concerns that cannot be overlooked. A promising alternative is to leverage a Large Vision-Language Model (LVLM), which facilitates zero-shot recognition by leveraging extensive image-text pair datasets, eliminating the need for direct training on private images. However, this approach often involves complex reasoning processes, which leads to substantial computational costs that render it unsuitable for real-time applications.

To enable a robot to perform privacy-aware behavior with low latency, we achieve image privacy recognition without relying on ethically sensitive images. Specifically, we fully utilize a Large Language Model (LLM) and a Vision-Language Model (VLM) and propose a methodology consisting of two phases: 1) the automatic construction of the LLM Privacy Text Dataset (LPT Dataset) and 2) the image privacy recognition model based solely on textual data. An overview of this study is shown in Figure 1. In the first phase, we use an LLM to generate textual descriptions related to privacy, each scored from 0 to 10 across multiple privacy indicators (e.g., "norms" and "reputation"). In the second phase, our model predicts the privacy score of an input image, representing how private an image (or a text) context is, through a two-stage text retrieval process that references multiple sentences rather than just one to improve robustness.

In the first stage, it retrieves texts from the LPT Dataset that are similar to the image in the vision language feature space and estimates the scores for each indicator. In the second stage, these scores are used to retrieve similar texts, and the privacy score is computed from their overall scores.

For evaluation, we assess the performance of the recognition model based on the LPT Dataset and the privacy-aware behavior of the robot equipped with the recognition model. The performance of the model is tested on PrivacyAlert [9], VISPR [7], and the CommU Privacy Image Dataset (CPI Dataset), which contains human action images captured by the robot. Finally, we conduct a user study with 20 subjects to evaluate the model's effectiveness in real-world scenarios.

The contributions of this study are summarized as follows:

- We propose a method that enables humanoid robots to recognize privacy-sensitive situations and respond appropriately.
- We introduce a low-latency and robust approach to image privacy recognition using only automatically generated privacy-labeled text, eliminating the need for ethically sensitive image datasets.

II. RELATED WORK

This study proposes an approach to image privacy recognition without using ethically sensitive images. Our approach leverages a VLM and a text dataset automatically generated by an LLM, allowing interactive humanoid robots to exhibit privacy-aware behaviors. To position our work in the broader research landscape, we review relevant studies on human-robot interaction, image privacy recognition, and advances in foundation models such as LLMs or LVLMs.

A. AI for Human-Aware Robotics

As robots increasingly interact with humans, natural communication has become essential [10]. Technologies such as cloud computing and computer vision enable robots to recognize gestures and improve interaction [11], [12]. In addition, natural language processing and speech recognition enhance verbal communication, while reinforcement learning supports autonomous navigation and task execution [13].

To achieve more natural interactions, robots that can adjust gaze, nod, or blink in response to conversation cues have been developed [14]. Advanced systems integrate image recognition with LLMs to interpret group conversations and provide context-aware support, such as serving drinks [15]. These developments enhance the social acceptability and usability of humanoid robots.

Despite these advances, data-driven approaches raise ethical concerns, particularly regarding privacy and bias [16]. AI models trained on human-generated data introduce the risks of surveillance and unintended privacy violations. Ensuring privacy-aware behavior in humanoid robots is essential for their ethical deployment.

B. Image Privacy Recognition

The notion of privacy is ambiguous, and the criteria for its violation vary across jurisdictions. However, in most countries, capturing images with a camera is widely recognized as an act that raises concerns about potential privacy violations. For example, "information collection" is one category in Solove's taxonomy of actions that pose risks to privacy [17]. Furthermore, in Japan, unauthorized photography has been recognized as a potential infringement of privacy, even in cases related to Google Street View, when it encroaches on an individual's private life (Google Street View Case, Fukuoka High Court Judgment, 13 July 2012, Japan).

Existing image privacy recognition methods classify images as private or public based on labeled datasets [18]. Notable examples of such datasets include PicAlert [8], which annotates images from Flickr, and VISPR [7], which categorizes images using privacy attributes derived from legal regulations. PrivacyAlert [9] refines VISPR with recent social media data, while cross-cultural studies highlight differences in privacy perception [19]. However, the collection of ethically sensitive images for these datasets raises concerns [20].

Earlier privacy recognition models relied on hand-crafted features and text annotations [1]. More recent methods incorporate convolutional neural networks [2] and bidirectional encoder representations from transformers for text-based feature extraction [3]. Some approaches integrate scene context and object information [4], while others apply topic modeling to improve interpretability [21]. More recently, the query processing capabilities of LVLMs for images have improved, enabling the exploration of methods for privacy awareness. However, achieving high-accuracy recognition with LVLMs requires approximately three seconds per image, making them impractical for real-time applications.

C. Foundation Models

Foundation models (e.g., LLMs and LVLMs) can be applied to general-purpose tasks using knowledge obtained from vast amounts of training data and can generate highly accurate responses even without a task-specific dataset. LLMs are text generation models capable of producing diverse sentences by varying the input, known as a "prompt." Prompt-based techniques such as in-context learning [22] and Chain of Thought (CoT) [23] further enhance their ability to generate text that aligns with user intent. Leveraging this property, LLMs have been utilized for synthetic dataset generation in specific domains [24], [25]. In this study, we extend this approach to generate privacy-related textual data, thereby eliminating the need for ethically sensitive image datasets in privacy detection.

LVLMs, such as GPT-4, have been developed as multimodal generation models that extend LLMs to tasks such as image captioning [26] and text-to-image generation [27]. Contrastive Language-Image Pretraining (CLIP) [28] is a representative model of VLMs, achieving strong performance in zero-shot classification by mapping images and texts into a shared feature space using contrastive learning. This capability allows CLIP to generalize across diverse vision-language tasks without task-specific fine-tuning. Beyond general applications, CLIP has been explored for ethical assessments of images. Jeong et al. [29] developed a model that identifies immoral images without requiring direct training on new

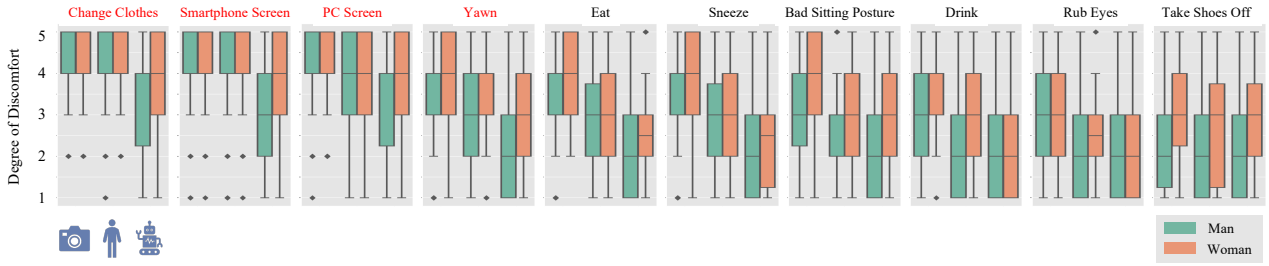


Fig. 2. Questionnaire results: Discomfort ratings for 10 types of actions and information were collected on a 5-point scale across three conditions—photography, human gaze, and robotic observation—separately for male and female participants. Types marked in red are subject to a user study with the robot in Section V-C.

datasets, while Park et al. [30] extended this approach to detect and correct moral issues in AI-generated images. Privacy-focused VLMs have also been proposed for visual question answering tasks [20].

Most VLM-based approaches to image privacy recognition still rely on sensitive images that cannot be collected and annotated in an ethical manner. Our study addresses this limitation by leveraging VLMs to recognize the image privacy solely on textual data, eliminating the need for direct exposure to private images. Unlike previous work [29], we do not train a classification model; instead, we estimate the privacy score of an image using a memory-based approach. In general, privacy recognition requires the ability to handle detailed expressions in images and text, even in unknown scenes. This often makes classification overly sensitive to local samples near the boundary and outliers. In contrast, the memory-based approach is more robust in such cases by leveraging multiple examples that have been retrieved.

III. Preliminary Survey on Robot Observation

To enable a robot to properly perform privacy-aware behavior according to a given situation, it is essential to understand how uncomfortable people feel when their own actions and personal information are observed by others. The degree of discomfort depends on the context, the observer, and the nature of the subject. However, this aspect has not been well investigated so far. Thus, in this study, we first conducted a survey through a crowdsourcing service. Specifically, we asked 200 workers to rate their degree of discomfort on a 5-point scale for 10 types of actions and information across three types of observation: photography, human gaze, and robots. In this survey, the specific form and recognition mechanism of the robot were not provided.

As shown in Figure 2, photography was the most uncomfortable observation type, while robot observation was relatively less uncomfortable. However, in highly private situations, such as changing clothes or viewing a smartphone screen, strong discomfort was also reported even when observed by the robot. In addition, being observed by robots tended to cause discomfort, particularly among female participants.

These results clarify the impact of robot observation on privacy perceptions and emphasize the need for context-aware privacy considerations. This study examines scenarios, such as changing clothes and viewing smartphone screens, where discomfort was particularly high, to evaluate the acceptability of the robot's observational behavior.

IV. Methodology

In this study, we propose a robotic system capable of behaving in a privacy-sensitive manner by leveraging an ethically sound textual dataset. To this end, we construct a comprehensive textual dataset and introduce a model that classifies an image as private or public based solely on textual descriptions, thereby eliminating the dependence on privacy-sensitive visual data. The proposed dataset encompasses a diverse range of privacy-related scenarios while autonomously defining privacy indicators. Furthermore, to ensure efficient and robust recognition for practical applications, the proposed model references multiple texts obtained through two-stage image-text retrievals, where retrieval methods are pretrained to effectively find neighbors in inference. For our retrieval-based method, we employed k-Nearest Neighbors (kNN) primarily for its efficiency in adding and removing privacy-sensitive texts.

In this section, we explain three key components in our methodology: the construction procedure of the LPT Dataset, image-text retrieval-based image privacy recognition, and the integration mechanism of the model into CommU.

A. Construction of the LLM Privacy Text Dataset

To recognize privacy-sensitive situations, the content must include descriptions of private actions or objects. However, constructing such a dataset manually is highly labor intensive and poses significant challenges due to the inherent subjectivity of privacy perceptions, which vary across individuals, cultures, and temporal contexts. To address this issue, we leverage an LLM to automatically generate the LPT Dataset, an ethically sound textual dataset for privacy recognition.

Figure 3 illustrates the construction process of the LPT Dataset. The LPT Dataset conceptualizes privacy based on whether the subjects being observed or photographed experience discomfort, reflecting human sensitivity to privacy intrusions. Since privacy concerns vary in both contextual scope (e.g., sensitive personal data vs. bodily exposure) and severity (e.g., yawning vs. changing clothes), we adopt a graded privacy representation rather than a binary classification. Each textual entry is assigned a privacy score ranging

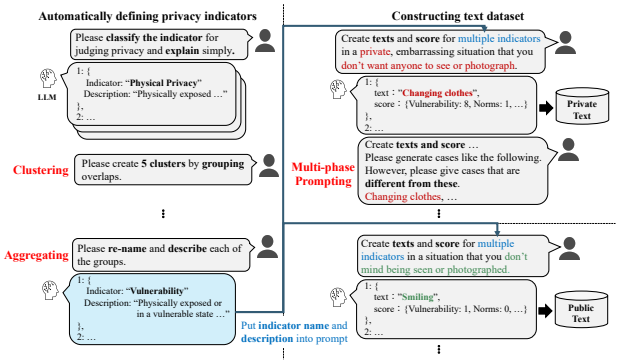


Fig. 3. Constructing the LPT Dataset: This process consists of two key steps—automatically defining multiple privacy indicators and constructing a text dataset that considers both private and public cases.

from 0 to 10 across multiple privacy indicators, encapsulating different dimensions of privacy violations. These indicators are dynamically adjusted to reflect contemporary privacy perceptions inferred from the LLM's training data. In order to construct an excellent dataset, we implement three schemes from the perspectives of reliability, diversity, and balance.

To enhance the dataset's reliability, privacy indicators undergo iterative refinement through clustering and aggregation. Since a single generation cycle may introduce inconsistencies, multiple iterations are conducted to ensure that privacy attributes remain robust and well-defined. This process enables the LLM to generate structured descriptions of private and public scenarios, promoting dataset diversity and ensuring adaptability to evolving privacy standards and temporal variations.

Since repeated prompts we assign impose inherent token limitations of LLMs, a prompt without any device may lead to make a scenario bias and produce overly homogeneous outputs. To maintain the diversity of scenarios, we employ a multi-phase prompting strategy, where previously generated text is used as conversational context. This strategy guides the LLM to mitigate redundancy while preserving contextual coherence, ensuring adaptability to evolving privacy standards and temporal variations.

Furthermore, to ensure dataset balance, we generate both private and public content using distinct prompts. Since specifying "public" often yields descriptions predominantly focused on outdoor environments, we introduce two separate prompts: indoor and outdoor. This distinction enhances the diversity of public content, allowing privacy recognition models to better distinguish various public contexts.

B. Training-free Image Privacy Recognition via Text

To realize an image privacy recognition model using only textual data with privacy indicators, it is crucial to effectively link the privacy context of a given image with private scenes described in texts. To this end, we propose CLIP-driven Text-based Image Privacy Recognition (CLIP-TIPR), a retrieval-based image privacy recognition model that identifies texts with privacy scores similar to the image in the CLIP's multi-modal feature space. Our model determines whether an image is private through a two-stage retrieval process: the first stage utilizes CLIP's feature space, while the second stage incorporates privacy indicators. If an LLM defines only one indicator, our model performs only the first retrieval-based on text features. We refer to our models as CLIP-TIPR (indicator-based) and CLIP-TIPR (text-based), respectively.

In this paper, we adopt the regression based on vanilla kNN for the two image-text retrievals. Here, we describe the nearest neighbor search problem we aim to solve. Note that kNN can reduce the computation time during inference by adopting acceleration algorithms such as approximate nearest neighbor search and kd-trees.

Let $\mathscr{T} = \{T_1, T_2, \ldots, T_N\}$ be the set of all texts in the LPT Dataset, as described in Section IV-A, and N be the total number of texts. Each text T_n is associated with multiple privacy indicators, which is represented by a vector $\boldsymbol{s}_n$ summarizing their respective scores as follows:

$$\boldsymbol{s}_n = (s_n^1, s_n^2, \ldots, s_n^M) \tag{1}$$

where M represents the total number of privacy indicators, and s_n^m denotes the m-th privacy indicator score for text T_n. The image encoder and text encoder in CLIP, mapping an image I and a text T into a shared D-dimensional feature space, are denoted as $f_I(I)$ and $f_T(T)$, respectively.

In the first phase, the model retrieves texts from the LPT Dataset similar to the input image in the multi-modal feature space, and then computes the scores of the image from the indicator scores assigned to the retrieved texts. The problem of selecting the indices $\mathscr{I}$ of the k_I texts that best match an image I in CLIP's feature space is formulated as follows:

$$\begin{aligned} \mathscr{I}_I &= \underset{\mathscr{I}}{\operatorname{argmin}} \sum_{i \in \mathscr{I}} d(f_I(I), f_T(T_i)) \\ &\text{s.t. } \mathscr{I} \subseteq \{1, \ldots, N\}, |\mathscr{I}| = k_I \end{aligned} \tag{2}$$

where $d(\cdot,\cdot)$ denotes the distance function.

Since each of the k_I selected texts is associated with a privacy indicator vector, their average serves as the predicted indicator scores for the image I. The resulting predicted score vector $\hat{\boldsymbol{s}}(I)$ is defined as follows:

$$\hat{\boldsymbol{s}}(I) = \frac{1}{k_I} \sum_{i \in \mathscr{I}_I} \boldsymbol{s}_i \tag{3}$$

Similarly, the same procedure can be applied to the overall score, yielding the predicted privacy score for the image I given by:

$$\hat{s}^{OA}(I) = \frac{1}{k_I} \sum_{i \in \mathscr{I}_I} s_i^{OA} \tag{4}$$

Furthermore, we apply kNN again in the second stage to obtain a privacy score from the indicator scores. Specifically, we compute the distance between the predicted score vector $\hat{\boldsymbol{s}}(I)$ and all score vectors $\{\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_N\}$ in the text dataset. The problem of selecting the indices $\mathscr{J}$ of the $k_{\boldsymbol{s}}$ texts that best match an image I using indicator scores is as follows:

$$\begin{aligned} \mathscr{J}_{\hat{\boldsymbol{s}}(I)} &= \underset{\mathscr{J}}{\operatorname{argmin}} \sum_{j \in \mathscr{J}} d(\hat{\boldsymbol{s}}(I), \boldsymbol{s}_j) \\ &\text{s.t. } \mathscr{J} \subseteq \{1, \ldots, N\}, |\mathscr{J}| = k_{\boldsymbol{s}} \end{aligned} \tag{5}$$

where $d(\cdot,\cdot)$ denotes the distance function.

Finally, the privacy score $\hat{S}(I)$ of the image I is computed by averaging the overall scores assigned to texts as follows:

$$\hat{S}(I) = \frac{1}{k_{\boldsymbol{s}}} \sum_{j \in \mathscr{I}_{\hat{\boldsymbol{s}}(I)}} s_j^{OA} \tag{6}$$

The proposed procedure allows a multi-dimensional and stepwise evaluation of image privacy by referencing multiple texts scored for each indicator similar to the image.

C. Integrating Image Privacy Recognition into Robots

In this section, we describe the integration of the image privacy recognition method described in Section IV-B into CommU (Vstone Co. Ltd.), a humanoid robot.

CommU captures environmental images using a forehead-mounted camera and stores them as buffered images for internal processing. These images are transmitted to a server on the same network via socket communication, where they are processed by an image privacy recognition function. Due to the extremely limited computing power of CommU's embedded computer, this function runs on the server to estimate the privacy score from the image.

If the computed privacy score exceeds a predefined threshold, CommU mimics human behavior by considering the gaze of others and performs a look-away action, as illustrated in Figure 1(b). During this action, the camera temporarily stops capturing images for a fixed duration (2 seconds) before resuming image capture.

This reassessment process simulates the human tendency to glance at the environment again in uncomfortable situations. During reassessment, CommU observes the environment with its arms raised and eyes half-open while gradually returning its body orientation and posture to its original position (Figure 1(c)). At this stage, CommU captures a new image and transmits it to the server for re-evaluation, following the same procedure as before. If the recomputed privacy score still exceeds the threshold, CommU repeats the look-away action. In contrast, if the score falls below the threshold, CommU returns to its original posture (Figure 1(a)), indicating a reduction in privacy concerns.

On the server, the image privacy recognition function operates continuously, receiving and processing images sent from CommU in real time. To ensure real-time performance, the model is preloaded into memory, allowing immediate inference on incoming images. Through this process, CommU detects privacy violations in its environment using the image privacy recognition model, enabling it to exhibit behavior that naturally considers human gaze and privacy awareness.

V. EXPERIMENTS

A. Building the LPT Dataset

We constructed the LPT Dataset following the procedure described in Section IV-A, assigning scores ranging from 0 to 10 to multiple privacy indicators and their overall score. To facilitate the comparative analysis of datasets, we constructed a dataset scored from 0 to 10 based on a single abstract privacy indicator: whether one would feel embarrassed if seen or photographed by others. All datasets were generated under the same conditions using GPT-4o as the LLM, with a `temperature` of 0.9, a `presence_penalty` of 0.0, and a `frequency_penalty` of 0.3.

TABLE I
DISTRIBUTION OF PRIVACY INDICATORS FOR GENERATED TEXTS

Generated Indicator	Mean±Std
Emotional Privacy and Sensitivity	2.17 ± 2.29
Physical Privacy and Vulnerability	0.87 ± 1.55
Social Norms and Embarrassment	3.25 ± 3.13
Confidentiality and Information Security	0.41 ± 1.48
Reputation and Identity Concerns	1.84 ± 2.12
Overall Score	3.39 ± 2.83

In the construction of the LPT Dataset, the indicators and their meanings generated are as follows:

Emotional Privacy and Sensitivity (*EMO*) Indicator of emotional exposure or psychological burden.

Physical Privacy and Vulnerability (*VULN*) Indicator related to the state of physical unprotection or invasion of physical privacy of an individual.

Confidentiality and Information Security (*CONF*) Indicator related to the risk of leakage of confidential information, personal information, etc.

Social Norms and Embarrassment (*NORMS*) Indicator of situations involving behavior or embarrassment that deviate from social norms.

Reputation and Identity Concerns (*REP*) Indicator related to the likelihood that an individual's behavior or situation will affect his or her reputation, identity, etc.

For simplicity, each indicator will hereafter be denoted by its abbreviation in parentheses. Note that the overall score is denoted as *OA*.

Table I shows the distribution of privacy indicators for generated texts. The *NORMS* and *OA* have a wide range of values, while the values for *VULN* and *CONF* tend to be concentrated around 0.

Table II presents examples of generated text and its privacy indicator scores in the LPT Dataset. The top example "Having your pants accidentally torn in a very public place" receives a score of 8 for *VULN*, 7 for *REP*, and 9 for *NORMS*, with an overall score of 10. This suggests an intuitive alignment between the sentence context and the degree of privacy violation. In contrast, the bottom example "Photographed yawning widely in a serious meeting" has a *NORMS* score of 6 and an overall score of 6. Because this situation may be perceived as a privacy violation in certain contexts, its overall score is close to the boundary between private and public, which appears reasonable.

B. Evaluation of Image Privacy Recognition

1) Dataset: To evaluate image privacy recognition, we utilized the existing evaluation image datasets PrivacyAlert and VISPR, as well as the newly constructed CommU Privacy Image Dataset (CPI Dataset). PrivacyAlert is a dataset labeled with binary classifications (private or public) based on the presence or absence of privacy-related elements.

TABLE II

EXAMPLES OF LPT DATASET TEXT AND SCORES WHERE VALUES OF 6 OR HIGHER IN **BOLD** AND 8 OR HIGHER IN **RED**.

Text	*CONF*	*EMO*	*VULN*	*REP*	*NORMS*	*OA*
Having your pants accidentally torn in a very public place.	0	**6**	**8**	**7**	**9**	**10**
Having your phone sent a personal message to a group chat by mistake.	5	**7**	0	**6**	**8**	**9**
Photographed yawning widely in a serious meeting.	0	2	1	3	**6**	**6**

VISPR is a multi-label dataset that includes 67 types of privacy-related elements along with a "safe" label, making a total of 68 labels. In this study, we classified images as private if they contained at least one privacy-related element, while those labeled only as "safe" were categorized as public.

To evaluate the ability of robots to recognize privacy in real-world environments, we constructed the CPI Dataset. This dataset comprises images of human actions captured by CommU in real-world settings. Based on the preliminary survey described in Section III, we focused on actions that were rated as particularly discomforting: changing clothes, viewing PC or smartphone screens, and sloppy behavior (e.g., yawning). Images were captured before, during, and after these actions, and five annotators assessed whether they contained information or actions that they would not want others to see if they were in the same situation. The final label was determined by majority vote, adopting a label if at least four out of five annotators agreed. Additionally, we ensured that the number of private and public images was balanced for each action type to facilitate the interpretation of recognition accuracy. The distribution of private/public images in each dataset is as follows: PrivacyAlert (489/1,281), VISPR (4,967/3,000), and CPI Dataset (75/75).

2) Comparison Methods: To verify the effectiveness of the proposed method, we conducted a comparison using two datasets described in Section II-B. Specifically, we evaluated the methods based on the LPT Dataset with multiple indicators and a dataset with a single abstract indicator. Within the LPT Dataset, we compared a text-based retrieval using only the overall score indicator and each individual indicator separately, as well as an indicator-based retrieval that considers all multiple indicators. Furthermore, we explored an approach utilizing an LVLM, GPT-4o mini, to classify images as private or public. The LVLM-based methods included LVLM-simple, which directly predicts labels, and LVLM-cot, which first outputs the reasoning process before determining the final label by using the CoT technique.

In addition, we conducted ablation studies on both the text-based retrieval and the indicator-based retrieval. Specifically, regarding the text-based retrieval, we compared our kNN-based approach to one following Jeong et al. [29], which trains a linear classifier based on the prompt type (private/public) of the generated texts, referred to as ZSVCIP[1] in our experiment. Regarding the indicator-based retrieval, we compared our classification method, which applies thresholding based on overall, to voting based on prompt type.

For the proposed method, we utilized three CLIP models[2]. We employed kNN for both image-text retrievals, with $k = 7$, using Euclidean distance as the distance function, and set a threshold of 5 on the privacy score ranging from 0 to 10 to distinguish between private and public.

3) Evaluation Metrics: We used the macro F1-score as an evaluation metric, averaging it across all CLIP models for methods without LVLM. We also measured the processing time on an M2 MacBook Air (Apple Inc.), which was used for the actual implementation. For the proposed method, we measured only the image recognition model's processing time, excluding image transfer. In contrast, for LVLM, the processing time was measured from request to response, including image transfer.

4) Results: Table III presents the evaluation results using the image dataset. In terms of accuracy, CLIP-TIPR (indicator-based) consistently outperformed both the text-based method trained with a single abstract indicator and LVLM-simple across all datasets. Furthermore, CLIP-TIPR (indicator-based) achieved higher F1-scores in two out of three datasets compared to LVLM-cot and ZSVCIP [29]. Especially in VISPR, images are labeled as private if any private attribute is present, meaning that even those containing only a human face tend to be classified as private. Since our study assumes human interaction, this labeling difference led to lower scores for our method. Additionally, as shown in Table IV, text-based inference on the CPI Dataset for each individual indicator resulted in an F1-score below 0.4 for most indicators, highlighting the difficulty of recognition. However, *NORMS* achieved significantly higher accuracy, even surpassing the F1-score of the indicator-based approach. Nevertheless, selecting the best indicator is not straightforward without first reviewing the results. Furthermore, in the second-phase indicator-based retrieval, the voting approach resulted in a significantly lower performance compared to thresholding. These findings indicate that CLIP-TIPR (indicator-based) effectively integrates multiple indicators rather than relying on a single one, enabling stable and high recognition accuracy.

Regarding processing time evaluation, as shown in Table III, methods using LVLM required at least two seconds, even in the shortest cases, while LVLM-cot, which achieves higher accuracy, required more computations and took over three seconds. In contrast, both ZSVCIP [29] and CLIP-TIPR (text-based) demonstrated significantly lower processing latency than CLIP-TIPR (indicator-based). However, all

[1]ZSVCIP does not work on the overall score of the LPT Dataset. Therefore, we defined labels based on the prompt type instead.

[2]https://huggingface.co/openai/{clip-vit-base-patch32,clip-vit-base-patch16,clip-vit-large-patch14}

TABLE III

COMPARISON OF QUANTITATIVE EVALUATION RESULTS. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD** AND THE SECOND BEST IN UNDERLINE.

Method	Reference text	macro F1-score ↑			Process Time (s) ↓
		CPI Dataset	PrivacyAlert	VISPR	
ZSVCIP [29]	LPT Dataset (binary)	0.500	0.543	**0.525**	0.059
CLIP-TIPR (text-based)	Single Abstract Indicator	0.440	0.577	0.306	**0.058**
CLIP-TIPR (indicator-based)	LPT Dataset	**0.680**	0.582	0.461	0.074
LVLM-simple	-	0.417	0.249	0.437	2.040
LVLM-cot	-	0.644	**0.706**	0.415	3.198

TABLE IV

ABLATION RESULTS FOR DIFFERENT METHODS ON THE LPT DATASET

Method	Indicator	CPI Dataset
CLIP-TIPR (text-based)	*EMO*	0.355
	VULN	0.376
	CONF	0.343
	NORMS	**0.694**
	REP	0.385
	OA	0.626
CLIP-TIPR (indicator-based, Voting)	ALL	0.554
CLIP-TIPR (indicator-based, Thresholding)	ALL	0.680

methods completed processing in under 0.1 seconds. These results indicate that the proposed method enables low-latency privacy recognition.

C. User Study on Privacy-Aware Robot Behavior

1) Experimental Setup: In this study, we evaluated both the performance of a recognition model trained on an image privacy dataset and its effectiveness when integrated into the humanoid robot CommU. To this end, we conducted an experiment involving 20 male participants in their twenties. The evaluation targeted behaviors previously identified as uncomfortable when photographed, namely changing clothes, viewing a PC or smartphone screen, and yawning. Each participant enacted a scenario corresponding to one of these behaviors for approximately 30 seconds.

The robot's responses were assessed based on three criteria: 1) the accuracy of gaze aversion, 2) the accuracy of gaze maintenance, and 3) the immediacy of response. The two accuracy metrics were rated on a five-point Likert scale from 1 (low) to 5 (high), with their harmonic mean serving as the overall evaluation metric. Immediacy was similarly rated on a five-point scale, with a score of 0 assigned if the robot continuously averted its gaze or failed to respond.

For comparative analysis, we examined four methods: a text-based approach trained using a single abstract indicator, the proposed indicator-based method, LVLM-cot alone, and a hybrid approach in which the proposed method initially identifies privacy-sensitive situations, followed by final verification via LVLM. This study is predicated on the notion that human responses can be classified into two types: spinal reflexes that bypass the cerebrum and conscious responses involving cognitive processing. The latter category can be further subdivided into immediate intuitive judgments and deliberative decisions. To approximate human-like behavior, we employed our proposed method for faster recognition while leveraging LVLM, which is computationally slower but may be effective in deliberative scenarios. In this hybrid approach, LVLM is used only after the proposed method identifies an image as private, allowing for a more efficient decision-making process.

2) System Evaluation and User Feedback: The evaluation results of this study indicate that the proposed method, as well as its combination with LVLM-cot, received the highest ratings among the evaluation items and exhibited consistent trends across all evaluation metrics (Figure 4). Regarding the text-based method, while it demonstrated a high accuracy in not averting gaze (True Negative Rate, TN Rate), its accuracy in averting gaze (True Positive Rate, TP Rate) was low. Consequently, the harmonic mean of these rates fell below 3, leading to an overall negative evaluation. As with the previous case, LVLM-cot alone also tended to have a higher TN Rate than TP Rate, with its harmonic mean reaching 3.2, resulting in a relatively positive evaluation. Its reaction immediacy rating was only 1.8, strongly suggesting that participants found it noticeably slow to respond.

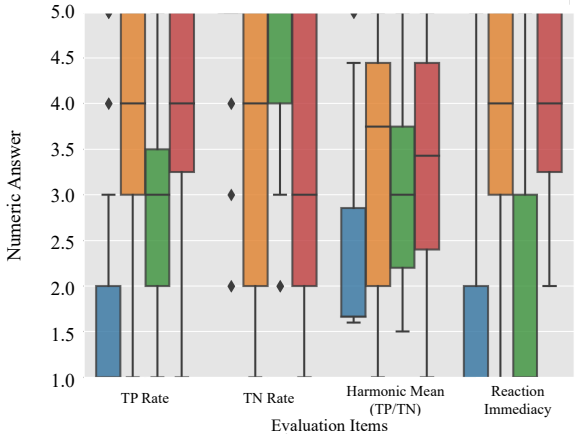
On the other hand, the proposed method achieved a significantly higher rate of gaze aversion, with a harmonic mean of 3.3, compared to the text-based method in a t-test ($p < 0.01$). Additionally, its responsiveness rating exceeded 3.5, demonstrating a clear improvement over LVLM-cot ($p < 0.01$), indicating that participants perceived it as more responsive.

No significant differences were observed between the proposed method and its combination with LVLM in the present evaluation. Similarly, qualitative feedback in the open-ended responses did not highlight any notable advantages of their combination. While this difference was not evident in the current evaluation, using alternative evaluation criteria might yield further insights.

Additionally, in a post-experiment survey, participants were asked to rate, on a five-point scale, whether they felt their privacy was protected by the robot's ability to suspend recording in response to human movements. The results showed that over 95% of participants rated their experience as 4 or higher, indicating that the proposed method is also effective from a privacy protection perspective.

VI. CONCLUSIONS

We propose a framework for a humanoid robot to exhibit privacy-aware behavior using an image privacy recognition



	TP Rate	TN Rate	Harmonic Mean (TP/TN)	Reaction Immediacy
Text-based	1.98±1.38	4.67±0.79	2.48±1.11	1.02±1.71
Indicator-based	3.72±1.31	3.25±1.42	3.35±1.34	3.63±1.78
LVLM-cot	2.73±1.24	4.40±1.01	3.16±1.10	1.85±1.65
Indicator-based × LVLM-cot	3.81±1.13	3.30±1.45	3.41±1.27	3.76±1.62

Fig. 4. User Feedback Results: Our indicator-based method and its combination with LVLM-cot received high ratings in terms of action accuracy and reaction immediacy.

model. The robot exhibited privacy-aware behavior, such as looking away when a human engages in actions they prefer not to be observed by others. To achieve this, we introduced a low-latency privacy-aware method that eliminates the need for ethically sensitive image privacy datasets.

We incorporated CLIP-TIPR, a retrieval-based model designed to operate solely on textual data, as the robot's image privacy recognition. While CLIP-TIPR enables privacy-aware behavior, integrating it with an LVLM could help the robot balance reflexive responses with deliberate privacy considerations, making its behavior more human-like.

Since privacy perception varies not only across different time periods but also individuals, developing a personalized privacy awareness model tailored to each user is desirable. This requires constructing an interactive learning mechanism that adapts to individual user preferences over time.

ACKNOWLEDGMENT

This work was supported by JST, PRESTO Grant Number JPMJPR22C4, Japan.

REFERENCES

[1] S. Zerr, S. Siersdorfer, J. Hare, and E. Demidova, "Privacy-Aware Image Classification and Search," in *SIGIR*, 2012, pp. 35–44.

[2] A. Tonge and C. Caragea, "Image Privacy Prediction Using Deep Neural Networks," *ACM Trans. Web*, pp. 1–32, 2020.

[3] C. Zhao and C. Caragea, "Knowledge Distillation with BERT for Image Tag-Based Privacy Prediction," in *RANLP*, 2021, pp. 1616–1625.

[4] A. Tonge and C. Caragea, "Dynamic Deep Multi-modal Fusion for Image Privacy Prediction," in *WWW*, 2019, pp. 1829–1840.

[5] C. Zhao and C. Caragea, "Deep Gated Multi-modal Fusion for Image Privacy Prediction," *ACM Trans. Web*, vol. 17, no. 34, pp. 1–24, 2023.

[6] G. Yang, J. Cao, Z. Chen, J. Guo, and J. Li, "Graph-based Neural Networkds for Explainable Image Privacy Inference," *Pattern Recognition*, vol. 105, no. 107360, 2020.

[7] T. Orekondy, M. Fritz, and B. Schiele, "Connecting Pixels to Privacy and Utility: Automatic Redaction of Private Information in Images," in *CVPR*, 2022, pp. 8466–8475.

[8] S. Zerr, S. Siersdorfer, and J. Hare, "PicAlert!: A System for Privacy-aware Image Classification and Retrieval," in *CIKM*, 2012, pp. 2710–2712.

[9] C. Zhao, J. Mangat, S. Koujalgi, A. Squicciarini, and C. Caragea, "PrivacyAlert: A Dataset for Image Privacy Prediction," in *ICWSM*, 2022, pp. 1352–1361.

[10] B. Obrenovic, X. Gu, G. Wang, D. Godinic, and I. Jakhongirov, "Generative AI and Human–Robot Interaction: Implications and Future Agenda for Business, Society and Ethics," *AI & SOCIETY*, pp. 1–14, 2024.

[11] M. Andronie, G. Lăzăroiu, O. L. Karabolevski, R. Ştefănescu, I. Hurloiu, A. Dijmărescu, and I. Dijmărescu, "Remote Big Data Management Tools, Sensing and Computing Technologies, and Visual Perception and Environment Mapping Algorithms in the Internet of Robotic Things," *Electronics*, vol. 12, no. 1, p. 22, 2022.

[12] H. Liu and L. Wang, "Gesture Recognition for Human-Robot Collaboration: A Review Author Links Open Overlay Panel," *International Journal of Industrial Ergonomics*, vol. 68, pp. 355–367, 2018.

[13] T. Zhang and H. Mo, "Reinforcement Learning for Robot Research: A Comprehensive Review and Open Issues," *International Journal of Advanced Robotic Systems*, vol. 18, no. 3, 2021.

[14] S. Noguchi, Y. Nakamura, and Y. Okadome, "Development of a Attentive Listening Robot Using the Motion Prediction Based on Surrogate Data," *HCI International 2024 Posters*, pp. 387–394, 2024.

[15] D. Tanneberg, F. Ocker, S. Hasler, J. Deigmoeller, A. Belardinelli, C. Wang, H. Wersing, B. Sendhoff, and M. Gienger, "To Help or Not to Help: LLM-based Attentive Support for Human-Robot Group Interactions," *arXiv preprint, arXiv:2403.12533*, 2024.

[16] R. Wullenkord and F. Eyssel, "Societal and Ethical Issues in HRI," *Current Robotics Reports*, vol. 1, no. 3, pp. 85–96, 2020.

[17] Y. Zong, O. Bohdal, T. Yu, Y. Yang, and T. Hospedales, "Safety Fine-Tuning at (Almost) No Cost: A Baseline for Vision Large Language Models," *arXiv preprint arXiv:2402.02207*, 2024.

[18] G. Zhang, B. Liu, T. Zhu, A. Zhou, and W. Zhou, "Visual Privacy Attacks and Defenses in Deep Learning: A Survey," *Artificial Intelligence Review*, vol. 55, no. 6, pp. 4347–4401, 2022.

[19] A. Xu, Z. Zhou, K. Miyazaki, R. Yoshikawa, S. Hosio, and K. Yatani, "DIPA2: An Image Dataset with Cross-cultural Privacy Perception Annotations," in *Proc. ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 7, no. 4, 2024, pp. 1–30.

[20] L. Samson, N. Barazani, S. Ghebreab, and Y. M. Asano, "Privacy-Aware Visual Language Models," *arXiv preprint arXiv:2405.17423*, 2024.

[21] G. Ayci, A. Özgür, M. Şensoy, and P. Yolum, "PEAK: Explainable Privacy Assistant through Automated Knowledge Extraction," *arXiv preprint arXiv:2301.02079*, 2023.

[22] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, *et al.*, "Language Models are Few-Shot Learners," in *NeurIPS*, 2020, pp. 1877–1901.

[23] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, *et al.*, "Chain-of-thought Prompting Elicits Reasoning in Large Language Models," in *NeurIPS*, 2022, pp. 24 824–24 837.

[24] B. Chintagunta, N. Katariya, X. Amatriain, and A. Kannan, "Medically Aware GPT-3 as a Data Generator for Medical Dialogue Summarization," in *MLHC*, 2021, pp. 354–372.

[25] C. Zheng, S. Sabour, J. Wen, Z. Zhang, and M. Huang, "AugESC: Dialogue Augmentation with Large Language Models for Emotional Support Conversation," in *Findings of the Association for Computational Linguistics: ACL*, 2023, pp. 1552–1568.

[26] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and Tell: A Neural Image Caption Generator," in *CVPR*, 2015, pp. 3156–3164.

[27] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models," in *ICML*, 2021, pp. 16 784–16 804.

[28] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning Transferable Visual Models from Natural Language Supervision," in *PMLR*, 2021, pp. 8748–8763.

[29] Y. Jeong, S. Park, S. Moon, and J. Kim, "Zero-shot Visual Commonsense Immorality Prediction," *arXiv preprint, arXiv:2211.05521*, 2022.

[30] S. Park, S. Moon, and J. Kim, "Judge, Localize, and Edit: Ensuring Visual Commonsense Morality for Text-to-Image Generation," *arXiv preprint, arXiv:2212.03507*, 2022.