# Exploring Foundation Models in Detecting Concerning Daily Functioning in Psychotherapeutic Context based on Images from Smart Home Devices

Yuang Fan*, Jingping Nie*, Xinghua Sun†, and Xiaofan Jiang*
*Columbia University, New York, New York, USA
†University of Washington, Seattle, Washington, USA

***Abstract*—The surge of cyber-physical systems (CPS) and smart home devices (e.g., vacuum robots and pet cameras) equipped in U.S. households opens up the potential to screen the day-to-day functioning of individuals in the smart home environment and to provide precautionary assistance to individuals who may need psychotherapies. Meanwhile, recent advances in foundation models (FMs) enable the large language models (LLMs) and vision-language models (VLMs) to have strong reasoning capabilities even in complex scenarios. In this paper, we investigate the integration of FMs and photos taken from the perspective of vacuum robots to screen the behaviors indicative of mental state that need further attention from therapists and health care practitioners. Specifically, we explore the possibility of using VLMs and LLM-based reasoners to accurately detect two of the most concerning behaviors at home: smoking- and drinking-alone. Compared to existing methods based on object detection, we demonstrate that the integration of LLMs and VLMs can significantly enhance detection accuracy, especially in complex home environments with ambiguous patterns and distinguishing `concerning` events from `benign` events. We showcase the potential of employing FMs in CPS to discern nuanced insights into day-to-day functioning behaviors in the psychotherapeutic contexts.**

***Index Terms*—Applications of Foundation Models, Image Reasoning, Cyber-physical Systems**

## I. Introduction

The surge of cyber-physical systems (CPS) and smart home devices equipped in U.S. households opens up the potential to screen the day-to-day functioning of individuals in the smart home environment and provide precautionary assistance to individuals living alone who may need psychotherapies [1]. As claimed by therapists, **smoking- and drinking-alone** in the home environment are two of the most common activities that need further attention from mental health care practitioners [2].

Systems such as [1] have proposed objection-detection-based techniques leveraging pictures taken by cameras equipped on vacuum robots to identify all occurrences of smoking- and drinking-alone activities in the one-person household and provide precautionary intervention. However, not all smoking- and drinking-alone activities need further attention if conducted in moderation. For example, an individual having a single glass of wine at the dining table is practically harmless (defined as **`benign`** in this work), and a person binge drinking a bottle of hard liquor lying on the ground needs immediate attention (defined as **`concerning`** in this work). To improve the experience for the at-home precautionary self-screening, it is necessary to distinguish and exempt those smoking- or drinking-alone behaviors that do not indicate any mental health concerns. This is where pure object-detection-based systems appear limited as they fail to extract contextual information needed to assess such an activity's severity.

Meanwhile, recent advances in foundation models (FMs) have enabled large language models (LLMs) and vision-language models (VLMs) to have strong reasoning capabilities, even in complex scenarios [3]. This motivates us to investigate the potential of incorporating such FMs in the pipeline of precautionary self-screening in home environments, leveraging the existing smart home devices. We explore using VLMs and LLM-based reasoners to accurately decide whether a smoking- or drinking-alone activity is actually **`concerning`** or **`benign`**, and integrate it as part of the home robot module of the smart home system proposed in [1].

To achieve our goal, we design and experiment 4 `Image Analyzer` modules (see Section V) to distinguish an actual **`concerning`** smoking- or drinking-alone event from *(i)* **`benign`** smoking- or drinking-alone event or *(ii)* no smoking- or drinking-alone event detected. We select LLaVA-1.5 [4] as the VLM and GPT-4 [5] as the LLM reasoner to be integrated into the `Image Analyzers`. We collect 10,767 images from the perspective of vacuum machines, with 10 subjects acting out various scenarios reflecting different intensities of smoking- and drinking-alone activities in a diverse set of home environments. We demonstrate that the integration of LLMs and VLMs into the precautionary screening pipeline can distinguish the actually **`concerning`** smoking- or drinking-alone events from the **`benign`** events in a complex home environment. Evaluated on a test subset of the images, the best-performing `Image Analyzer` shows a 73% accuracy in identifying an actual **`concerning`** event, an improvement of at least 13% compared to the pure objection-detection-based method proposed by [1] that flag the presence of alcohol containers and flocculation smoke, which includes both **`concerning`** and **`benign`** smoking- or drinking-alone event. We demonstrate the value of FMs in CPS to gain a detailed understanding of daily functioning behaviors within the psychotherapeutic context, which can be conveniently integrated into the existing smart home systems.

979-8-3503-6345-6/24/$31.00 ©2024 IEEE
DOI 10.1109/FMSys62467.2024.00012

## II. Related Works

Monitoring a person's smoking and drinking behavior is crucial in psychotherapies due to the substantial impact these substances can have on mental health and treatment outcomes [6]. Researchers have suggested that psychological distress is significantly more frequent in groups that smoked for more than 7 consecutive days (53.3% vs 31.2%) and for individuals who consumed at least 1 drink of alcohol in the last 30 days (42.4% vs 28.6%) [7]. By incorporating monitoring, therapists can adapt treatment plans to match the patient's specific situations, providing the foundational motivation for our work. Wearable devices like SPIDER+ [8] have been developed to collect hints of mental status. Systems such as aiMSE [9] incorporated multimodal processing algorithms to automatically administer mental status examinations online. These works make solid attempts to automate the techniques used by mental health practitioners during therapeutic sessions, but does not provide analysis of individuals' daily activities in home, especially for individuals living in one-person households as there are no other people to check on them.

A list of 37 dimensions of day-to-day functioning has been proposed in [2] to infer the mental status of a person. Additionally, a system for self-screening the day-to-day functioning of individuals in the home environment is proposed in [1], leveraging the existing sensors on the common smart home devices. In particular, as mentioned in Section I, [1] demoed a non-intrusive and privacy-aware system that uses a vacuum robot with cameras to detect and validate the occurrence of smoking- and drinking-alone behaviors, providing the structure of a system which will be further discussed in Section IV.

With the popularity of LLMs like GPT-4 [5] and Llama 2 [10], researchers have leveraged their immense prior knowledge trained on colossal amounts of corpus to create large multimodal models (LMMs) that enable artificial intelligence (AI) chat models to process information beyond text. Promising advancements have been made especially in the vision tasks, with VLMs such as LLaVA [4] and MiniGPT-4 [11] demonstrating the ability to generate conceivable responses to user prompts with regard to supplemented images. Additionally, works like LLaVA-Plus [12] explored the ability of VLMs and LLMs to learn and use general-purpose vision-language pre-trained models as tools, enabling new scenarios of LMM workflows. On the other hand, few works have been able to establish specific integration of VLMs into existing activity detection and monitoring systems, which are predominantly based on objection-detection or motion-detection methods. Yet, works like LLaVA-Grounding [13] and InstructDET [14] reveal the capability of VLMs in extracting a considerable level of information and context from images when given specific instructions about the artifact of interest, showing comparable performance to pure objection-detection methods.

## III. Challenge

Determining whether an image exhibits **`concerning` behaviors of smoking- or drinking-alone** described in Section I

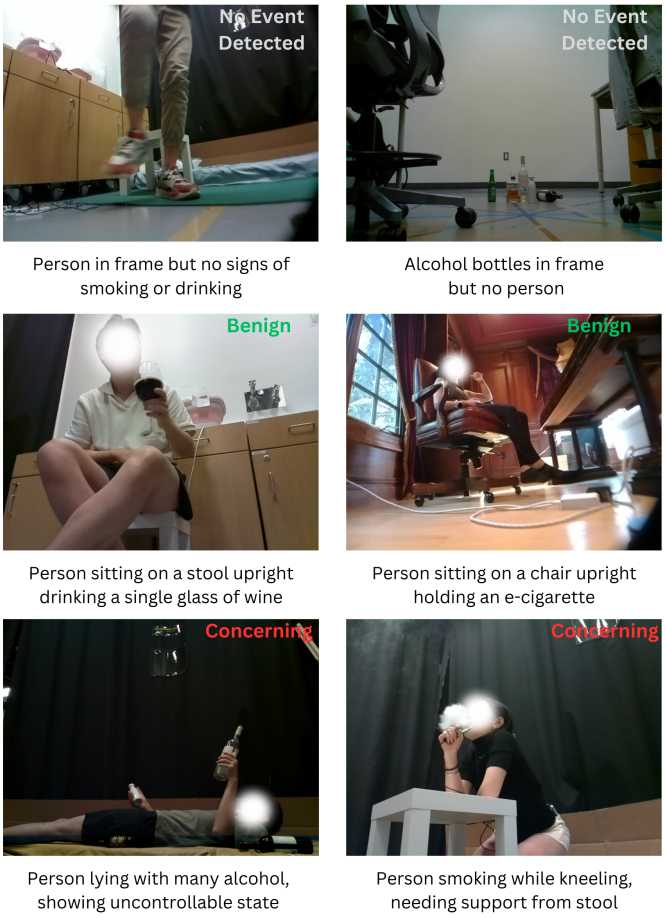


Fig. 1: Annotated picture examples illustrating **`concerning`** and **`benign`** behavior.

involves examining if the image illustrates a person smoking- or drinking-alone, and given the occurrence of such an event, reasoning if the illustrated behavior raises an alert that needs further attention (**`concerning`**). The object-detection-based system is an adequate approach for the primary task of detecting the occurrence of a smoking- or drinking-alone event as it can identify specific artifacts in the image that are necessary for such an event (e.g., smoke, alcohol containers). Still, detecting these artifacts is intrinsically challenging, as the smoke, cigarette, and alcohol containers are comparably small in size in the complex home environment. Identifying traces of smoke is especially difficult as the smoke might dilute at the moment that the home robot takes the pictures.

Moreover, the object-detection-based system is limited in analyzing whether such an event is inherently **`concerning`**, for which the decision relies on retrieving hints from factors such as consumption amount, posture, and apparent mental state and making a logical inference. Having said that, making such a decision is non-trivial even for humans. Given the lack of a universal standard to separate a **`concerning`** smoking- or drinking-alone behavior from a **`benign`** smoking- or drinking-alone behavior defined in [1], we propose to define for the scope of this paper, based on intuition and common

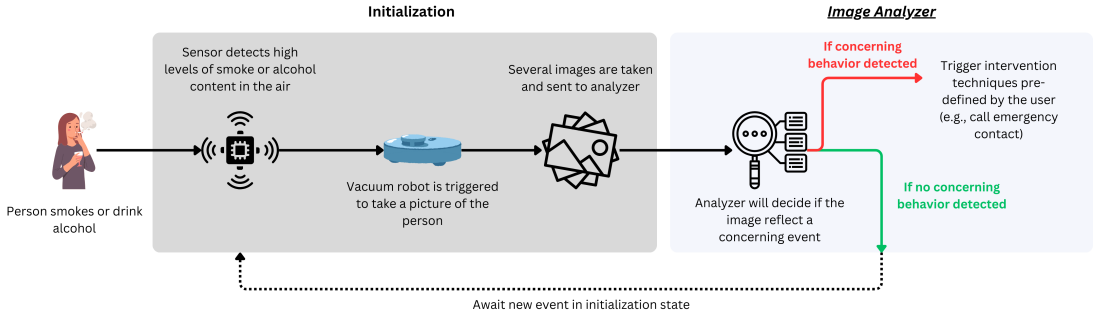


Fig. 2: The system architecture incorporating Initialization stage proposed in [1] and innovating the `Image Analyzer`.

sense, a **`concerning`** smoking- or drinking-alone behavior as "*a person smoking or drinking in excess, expressing apparent volatile mental state, or showing impaired control of body*". To better illustrate the definition, several annotated example images are provided in Figure 1.

## IV. System Architecture

Detecting **`concerning`** behaviors of smoking- and drinking- alone in the home environment involves several steps as depicted in Figure 2. Unlike public and commercial environments where continuous monitoring with surveillance cameras is common, the home environment usually lacks these devices due to privacy concerns. Moreover, even if there are always-on surveillance cameras equipped, people might not feel comfortable or secure and not express their natural behaviors. This poses a challenge for us to rely on purely vision-based detection methods. Acknowledging the privacy concerns, we incorporate the system proposed in [1] which combines the always-on low-fidelity sensors to pick up hints of a **`concerning`** event and validate the event with temporarily-activated cameras equipped on vacuum robots. The images are then passed to the proposed `Image Analyzer` to decide whether they reflect a **`concerning`** behavior and trigger a predefined intervention if necessary.

### A. Initialization

Similar to [1], the entry point of our system is the low-fidelity sensors commonly present in smart home devices, MQ3 (alcohol), and MQ9 (flammable gas) sensors. If a threshold reading of alcohol or smoke particles as defined in [1] is detected by these sensors, an alert will be triggered and a command will be sent to the vacuum robot to validate if there is a **`concerning`** event of smoking or alcohol consumption.

Upon receiving a command to validate a **`concerning`** event, the vacuum robot will navigate around the household to search for a person through LiDAR and cameras, which are commonly installed on some recent robot vacuums such as Roborock S7 MaxV Plus and other home robots, such as Amazon Astro. The 2D range LiDAR is a promising privacy-aware sensing modality for human detection and tracking. Trying to minimize the time of using privacy-invasive sensor modality, camera, the home robot will follow the user in the single-occupant room using 2D LiDAR tracking that uses simultaneous localization and mapping (SLAM), background subtraction, and Robot Operating System (ROS). Once a person is found, the robot will proceed to take several pictures of the person from different angles to ensure that sufficient context is captured. These pictures will then be sent to the `Image Analyzer` on the server and the robot will navigate back to the charging stating until receiving another command.

TABLE I: Content of Specialized Dataset

| Scenarios | Number of images |
|---|---|
| Smoking | 4689 |
| Alcohol Consumption | 2347 |
| Combination | 3731 |
| **Total** | 10767 |

### B. Image Analyzer

The `Image Analyzer` is implemented with a combination of VLM and LLM reasoner, which provides a decision on whether the images provide enough evidence for a **`concerning`** behavior of smoking- or drinking-alone. To realize a robust `Image Analyzer` is non-trivial with the challenges listed in Section III and the approaches depicted in Section V. Section VI illustrates the different VLM-LLM-based microbenchmarks and evaluations for the `Image Analyzer`. If no images show hints of **`concerning`** behavior, then the detection pipeline will be suspended for a short time for the sensor readings to dissipate. If the `Image Analyzer` decides that any of the input images exhibit **`concerning`** behaviors, it will create an alert leading to either re-validation or intervention based on the specific adaptations of the system (e.g., call emergency contact).

## V. Method

This section first presents the dataset collected and describes 4 `Image Analyzers` that was experimented on. These methods will include: (*i*) baseline objection detection model; (*ii*) high-level VLM; (*iii*) break-down VLM with expert rules; and (*iv*) break-down VLM with LLM-based reasoner.

### A. Dataset

Given the lack of existing datasets with smoking- and drinking-alone images taken from an upward-angled camera view, a specialized dataset containing 10767 images are collected from the angle of vacuum robots to fine-tune and evaluate the models proposed in this paper. 10 subjects are recruited, approved by the Institutional Review Board (IRB),

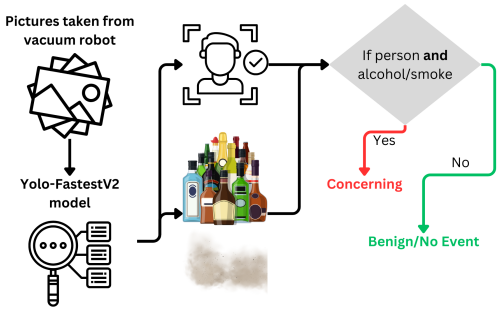


Fig. 3: Object detection method using Yolo-FastestV2.

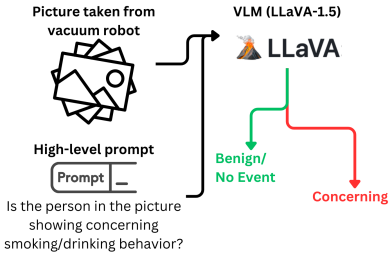


Fig. 4: High-level prompt VLM (LLaVA-1.5) method.

to perform/simulate scenarios related to smoking, alcohol consumption, and a combination of the two in 7 different environments (at the subjects' apartments and in the lab). As illustrated in Table I, there are 4689 images capturing scenarios of smoking which may contain people, flocculation smoke, and cigarettes/e-cigarettes; 2347 images capturing scenarios of alcohol consumption which may contain people, drink-ware, alcohol bottles, and other types of alcohol containers; and 3731 images capturing a combination of smoking and drinking events. The images are human-labeled for the ground truth with 3 variables: if the image contains flocculation smoke, if the image contains any alcohol containers, and if the image reflects concerning smoking- or drinking-alone behavior. As outlined in Section I, not all smoking- or drinking-alone behaviors are concerning, hence the dataset aims to capture this premise and reflect both **`concerning`** and **`benign`** events of smoking- or drinking-alone as well as images with no signs of smoking- or drinking-alone. This data is then randomly split into a development set (training and validation) of 9767 images and a testing set of 1000 images.

### B. Object Detection

The object-detection-based approach proposed in [1], which marks all smoking- and drinking-alone events as **`concerning`**, is leveraged as the baseline approach and evaluated under the definition for **`concerning`** behaviors defined in Section III as the ground truth. Particularly, a Yolo-FastestV2 [15] model is fine-tuned for 300 epochs using a batch size of 64 on images containing various types of alcohol bottles and flocculation smoke. If the object detection model detects any occurrences of alcohol bottles or smoke in the image along with the presence of a person, the `Image Analyzer` will then classify this image as **`concerning`**.

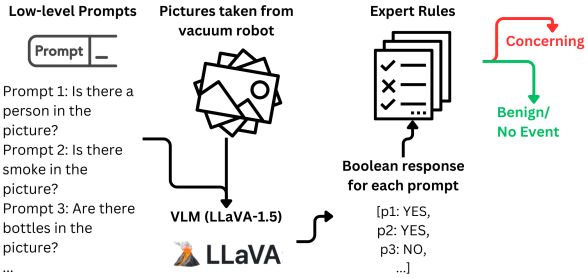


Fig. 5: Break-down VLM (LLaVA-1.5) with expert rules method.

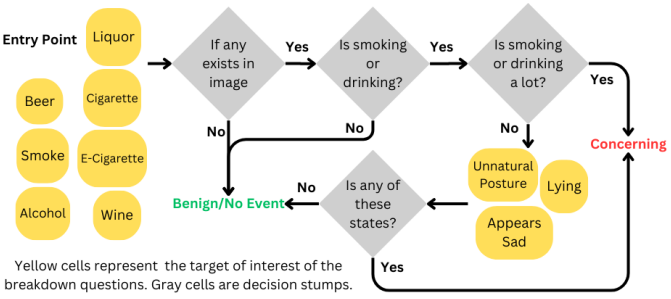


Fig. 6: Expert-rule-based classifier.

### C. High-level VLM

An end-to-end high-level VLM approach is experimented with as the `Image Analyzer` as depicted in Figure 4. LLaVA-1.5 is chosen as the VLM in this work due to its state-of-the-art performance on over 11 benchmarks [4]. Along with the picture, a high-level prompt is passed into LLaVA asking *"Is the person in the picture smoking or drinking alone and exhibiting concerning behavior?"* . The output of LLaVA are: (*i*) a YES or NO string as the answer to the high-level prompt and (*ii*) a short description of the image. The YES or NO string is directly used as the boolean indicator of whether the image displays a **`concerning`** drinking- or smoking-alone behavior and the descriptions are saved for evaluation purposes.

### D. Break-down VLM with Expert Rules

Although LLaVA can provide an answer to any question, it is often inaccurate in reasoning tasks that need multiple inferencing steps [3]. This applies to the use case where detecting concerning behavior involves considerations of many factors and is inherently ambiguous. To tackle this problem, the high-level prompt is broken down into many low-level prompts, each addressing a specific element of the picture. For example, separate questions are created for: (*i*) the presence of smoke, (*ii*) the presence of cigarettes, (*iii*) the presence of beer bottles, (*iv*) the presence of liquor bottles, and (*iv*) the posture of the person. After passing these prompts along the image to LLaVA, an array of boolean YES or NO strings and an array of analyses are generated as the return values for each specific low-level prompt. Then, the boolean array is fed into an expert-rule-based classifier illustrated in Figure 6 that will identify the image as either **`concerning`** or **`benign`**.

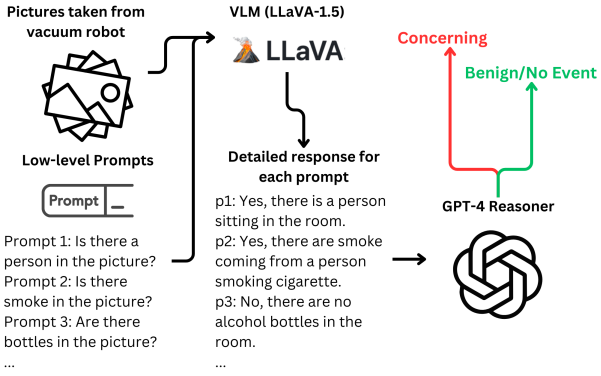


Fig. 7: Break-down VLM (LLaVA-1.5) with LLM reasoner (GPT-4) method.

### E. Break-down VLM with LLM Reasoner

An LLM reasoner takes advantage of the rich knowledge base and strong reasoning ability of LLMs pre-trained on colossal amounts of data [3]. As such, the expert-rule-based classifier is replaced by an LLM reasoner to extract meaningful contexts from the break-down responses returned by the VLM and decide whether the responses are **`concerning`**. We have prompted a GPT-4 reasoner with: (*i*) the break-down VLM responses including both YES/NO responses and short analyses described in Section V-D as well as the questions asked in the prompts, and (*ii*) a few examples of **`concerning`** and **`benign`** scenarios. The output of GPT-4 reasoner will be a final YES/NO decision and a brief reason of whether the image shows **`concerning`** smoking- or drinking-alone behaviors.

## VI. Evaluation and Microbenchmarks

To reinforce the motivation for using VLMs and highlight their raw potential of extracting information from images, a micro-benchmark is first conducted to purely compare the object-detection performance of the fine-tuned Yolo-FastestV2 with plain LLaVA-1.5 in detecting alcohol containers and flocculation smoke. Applying both methods to the test set of 1000 images from the dataset mentioned in Section V-A, we record whether the Yolo-FastestV2 model detected alcohol containers or smoke for each image, and the LLaVA-1.5 model is given an image along with two prompts: (*i*) *"Are there alcohol containers in the picture?"* and (*i*) *"Are there smoke in this picture?")* and returns YES or NO. As shown in Table II, the plain LLaVA-1.5 model outperforms the fine-tuned Yolo-FastestV2 model for both alcohol container and smoke detection by more than 10% accuracy. We acknowledge that there might be better object detection models such as YOLOv8 [16]. However, with a 90.2% accuracy in detecting alcohol containers and a 77% accuracy in identifying smoke, LLaVA-1.5 proves to be a competitive alternative to specialized object-detection models in our application scenario.

The 4 `Image Analyzers` proposed in Section V are evaluated on the test set with 1000 images and checked against the human-labeled ground truths for **`concerning`** smoking- or drinking-alone behaviors defined in Section III, illustrated in Table III.

TABLE II: Comparison of smoke/alcohol container detection performance of Yolo-FastestV2 with LLaVA-1.5.

| Models | `Alcohol Container Accuracy` | `Smoke Accuracy` |
|---|---|---|
| **Yolo-FastestV2** | 79.63% | 65.21% |
| **LLaVA-1.5** | 90.20% | 77.36% |

The pure object-detection-based `Image Analyzer` produced a baseline performance of 59.67% accuracy in identifying **`concerning`** behaviors in the image. Note that this object-detection-based `Image Analyzer` will also completely miss pictures captured in between smoke exhales so that no flocculation smoke is actually present in the image.

The high-level end-to-end VLM based `Image Analyzer` using LLaVA-1.5 shows comparable accuracy of 62%. Although end-to-end VLMs are not exceptionally good at collecting contextual information that requires multiple inferencing steps, inspecting the descriptions returned along with the YES or NO string reveals that even the high-level VLM `Image Analyzer` can use multiple pieces of evidence related to smoking or drinking alone activities without direct observation of bottles or smoke. For instance, in a picture taken between exhales when the smoke has already faded and the cigarette held in hand is partly concealed, the high-level VLM `Image Analyzer` is able to identify this situation as smoking by pointing out that *"a man is holding a cigarette up to his mouth"*. This shows that VLM-based `Image Analyzer` poses a clear advantage when a **`concerning`** event is not shown by evident traces of specific objects but instead is hinted by the relative posture of the person in the frame.

The break-down VLM `Image Analyzer` with an expert-rule-based classifier demonstrates an evident improvement in performance of 73% accuracy in detecting **`concerning`** behaviors of smoking- or drinking-alone. The increase in performance is expected as a result of fine-tuned low-level prompts that aim to extract specific traces from various factors that may reflect a **`concerning`** behavior. This `Image Analyzer` also produced a strong recall score of 0.822, leading to a more balanced F1-score of 0.809.

The break-down VLM `Image Analyzer` with LLM reasoner shows comparable performance compared to break-down VLM `Image Analyzer` with an expert-rule-based classifier, attaining an accuracy score of 71.33%. It also produces the highest F1-score of 0.821 as a direct result of its exceptionally good recall score of 0.947, reflecting the GPT-4 reasoner's tendency to mark an image as showing **`concerning`** behavior for the slightest hint generated by the LLaVA-1.5 responses while maintaining a reasonable false positive rate as indicated by the precision score of 0.724.

## VII. Discussion and Future Work

Using VLM-based `Image Analyzers` has shown a clear advantage in distinguishing **`concerning`** behaviors from **`benign`** behaviors as compared to the pure object-detection-based method. Although the break-down VLM with expert rule `Image Analyzer` is shown to have the best performance

TABLE III: Performance of `Image Analyzers` in classifying **`concerning`** behavior.

| Proposed Analyzers | `Accuracy` | `Precision` | `Recall` | `F1-score` |
|---|---|---|---|---|
| **Object Detection** | 59.67% | 0.601 | 0.767 | 0.674 |
| **High-level VLM** | 62.0% | 0.867 | 0.534 | 0.661 |
| **Break-down VLM & Expert Rules** | 73.0% | 0.795 | 0.822 | 0.809 |
| **Break-down VLM & LLM Reasoner** | 71.33% | 0.724 | 0.947 | 0.821 |

listed in Table III, the break-down VLM with LLM (GPT-4) reasoner, while demonstrating comparable performance, also retains the ability to process unexpected information that is not explicitly defined in the prompts. Contextual information about the quantitative and descriptive aspects of an event cannot be coded into a limited set of expert rules but can be picked up by LLM reasoners, demonstrating their flexibility and potential for tackling complex problems in the real world.

We see the potential in using VLM-based `Image Analyzers` to screen other behaviors at home that need further attention from therapists or health care. We will enrich the dataset by collecting images related to more potentially **`concerning`** activities and labeling the ground truth through voting with multiple annotators. We plan to incorporate LLM reasoners into the precautionary screening pipeline to enable time-series analysis to detect **`concerning`** behaviors. For example, another LLM reasoner can record the analysis of the break-down VLM with LLM reasoner and report trends in day-to-day functioning. With the momentum of AI acceptance in the general public led by the popularity of OpenAI products, we envision a boom in demand for VLM and LLM solutions for tasks like precautionary in-home screening for their superior capability in automating precautionary procedures and improving the living environment.

## VIII. Conclusion

Incorporating the system proposed in [1], which leverages smart home devices and home robots to screen daily functioning and provide precautionary psychotherapeutic interventions in one-person household, we prompt FMs (VLMs and LLMs) as an intelligent `Image Analyzer` to detect two of the most concerning behaviors at home, smoking- and drinking-alone, using photos captured by vacuum robot. We improve the definition of concerning smoking- and drinking-alone event proposed by [1], and design and experiment 4 `Image Analyzer` methods. We demonstrate that break-down VLM (LLaVA-1.5) with expert-rule-based classifier `Image Analyzer`, with an accuracy of 73%, performs the best in identifying **`concerning`** smoking- or drinking-alone behaviors and distinguishing them from **`benign`** smoking- or drinking-alone ones in the home environment from images taken by ground-level vacuum robots in our experiment setup. Meanwhile, the break-down VLM with LLM (GPT-4) reasoner `Image Analyzer` has comparable performance, and considering the flexibility of tolerance of LLM reasoner, this method has the potential to be integrated into complex real-world in-home precautionary daily functioning screening systems in the psychotherapeutic context.


## Acknowledgment

This research was partially supported by the National Science Foundation under Grant Number CNS-1943396. The views and conclusions contained here are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of Columbia University, NSF, or the U.S. Government or any of its agencies.


## References


[1] J. Nie, M. Zhao, S. Xia, X. Sun, H. Shao, Y. Fan, M. Preindl, and X. Jiang, "Ai therapist for daily functioning assessment and intervention using smart home devices," in *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*, SenSys '22, (New York, NY, USA), p. 764–765, Association for Computing Machinery, 2023.

[2] J. Nie, H. Shao, M. Zhao, S. Xia, M. Preindl, and X. Jiang, "Conversational ai therapist for daily function screening in home environments," in *Proceedings of the 1st ACM International Workshop on Intelligent Acoustic Systems and Applications*, pp. 31–36, 2022.

[3] H. You, R. Sun, Z. Wang, L. Chen, G. Wang, H. A. Ayyubi, K.-W. Chang, and S.-F. Chang, "Idealgpt: Iteratively decomposing vision and language reasoning via large language models," 2023.

[4] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," 2023.

[5] OpenAI(2023), "Gpt-4 technical report," 2023.

[6] N. Choi and D. DiNitto, "Drinking, smoking, and psychological distress in middle and late life," *Aging & Mental Health*, vol. 15, pp. 720 – 731, 2011.

[7] V. R. Ferreira, T. V. Jardim, A. L. L. Sousa, B. M. C. Rosa, and P. C. V. Jardim, "Smoking, alcohol consumption and mental health: Data from the brazilian study of cardiovascular risks in adolescents (erica)," *Addictive Behaviors Reports*, vol. 9, p. 100147, 2019.

[8] J. Nie, Y. Liu, Y. Hu, Y. Wang, S. Xia, M. Preindl, and X. Jiang, "Spiders+: A light-weight, wireless, and low-cost glasses-based wearable platform for emotion sensing and bio-signal acquisition," *Pervasive and Mobile Computing*, vol. 75, p. 101424, 2021.

[9] Y. Liu, S. Xia, J. Nie, P. Wei, Z. Shu, J. A. Chang, and X. Jiang, "aimse: Toward an ai-based online mental status examination," *IEEE Pervasive Computing*, vol. 21, no. 4, pp. 46–54, 2022.

[10] M. GenAI, "Llama 2: Open foundation and fine-tuned chat models," 2023.

[11] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "Minigpt-4: Enhancing vision-language understanding with advanced large language models," 2023.

[12] S. Liu, H. Cheng, H. Liu, H. Zhang, F. Li, T. Ren, X. Zou, J. Yang, H. Su, J. Zhu, L. Zhang, J. Gao, and C. Li, "Llava-plus: Learning to use tools for creating multimodal agents," 2023.

[13] H. Zhang, H. Li, F. Li, T. Ren, X. Zou, S. Liu, S. Huang, J. Gao, L. Zhang, C. Li, and J. Yang, "Llava-grounding: Grounded visual chat with large multimodal models," 2023.

[14] R. Dang, J. Feng, H. Zhang, C. Ge, L. Song, L. Gong, C. Liu, Q. Chen, F. Zhu, R. Zhao, and Y. Song, "Instructdet: Diversifying referring object detection with generalized instructions," 2023.

[15] Y.-S. Poon, C.-C. Lin, Y.-H. Liu, and C.-P. Fan, "Yolo-based deep learning design for in-cabin monitoring system with fisheye-lens camera," in *2022 IEEE International Conference on Consumer Electronics (ICCE)*, pp. 1–4, IEEE, 2022.

[16] Ultralytics, "Yolov8." https://github.com/ultralytics/ultralytics, 2023.