algorithms

MDPI

Article

Cross-Cultural Safety Judgments in Child Environments: A Semantic Comparison of Vision-Language Models and Humans

Don Divin Anemeta [1,*,†] and Rafal Rzepka [2]

[1] Graduate School of Information Science and Technology, Hokkaido University, Sapporo 060-0814, Japan
[2] Faculty of Information Science and Technology, Hokkaido University, Sapporo 060-0814, Japan; rzepka@ist.hokudai.ac.jp
* Correspondence: dondivin.anemeta.s7@elms.hokudai.ac.jp
† Current address: Information Science and Technology, Hokkaido University, Kita 14, Nishi 9, Kita-ku, Sapporo 060-0814, Japan.

Abstract

Despite advances in complex reasoning, Vision-Language Models (VLMs) remain inadequately benchmarked for safety-critical applications like childcare. To address this gap, we conduct a multilingual (English, French, Polish, Japanese) comparison of VLMs and human safety assessments using a dataset of original images from child environments in Japan and Poland. Our proposed methodology utilizes semantic clustering to normalize and compare hazard identification and mitigation strategies. While both models and humans identify overt dangers with high semantic agreement (e.g., 0.997 similarity for 'scissors'), their proposed actions diverge significantly. Humans strongly favor direct physical intervention ('remove object': 64.% for Polish vs. 55.0% for VLMs) and context-specific actions ('move object elsewhere': 17.8% for Japanese), strategies that models under-represent. Conversely, VLMs consistently over-recommend supervisory actions (such as 'Supervise children closely' or 'Supervise use of scissors'). These quantified discrepancies highlight the critical need to integrate nuanced, human-like contextual judgment for the safe deployment of AI systems.

Keywords: vision-language models (VLMs); embedding coherence entropy; semantic clustering; multilingual analysis; danger detection

Received: 28 June 2025
Revised: 31 July 2025
Accepted: 4 August 2025
Published: 13 August 2025

Citation: Anemeta, D.D.; Rzepka, R. Cross-Cultural Safety Judgments in Child Environments: A Semantic Comparison of Vision-Language Models and Humans. *Algorithms* **2025**, *18*, 507. https://doi.org/10.3390/a18080507

1. Introduction

Large language models augmented with visual input commonly known as Vision-Language Models (VLMs) have significantly advanced in recent years, learning to generate text from images, describe visual scenes, and respond to multimodal queries. This integration of visual and linguistic capabilities has opened possibilities for real-world applications, notably in safety-critical scenarios involving vulnerable people such as children.

Humanoid robots, initially popularized by physical agents like Honda's ASIMO [1] and Softbank's Pepper [2], have evolved significantly, aiming to support humans in both home and workplace environments. More recently developed humanoid robots such as Tesla's Optimus (https://www.tesla.com/en_eu/AI; accessed on 21 July 2025.) or 1X's NEO (https://www.1x.tech/neo; accessed on 21 July 2025) further indicate that sophisticated robotic systems may soon start to be a part of daily life. While these systems are advancing rapidly in physical intellectual capabilities through breakthroughs in machine learning trained and tested in closed environments, questions remain about their prac-

tical application in the wild, in complex environments, and safety-critical contexts like childcare [3–5].

Ensuring child safety in everyday environments, such as homes and schools, remains crucial, as preventable accidents are a leading cause of childhood injuries worldwide [6]. Children's natural curiosity often leads them to explore their environment without recognizing potential dangers, creating a persistent challenge for caregivers to balance safety with fostering curiosity and learning. Traditional safety methods like constant supervision or fixed safety measures lack the adaptability necessary for dynamic environments. In contrast, AI-driven solutions potentially offer continuous monitoring and responsive interventions. However, focusing only on immediate threats may lead robotic solutions to overlook subtle, long-term developmental impacts by restricting children's exploration or suggesting overly cautious actions.

Effectively deploying VLMs and robotic solutions in child-safety applications demands rigorous evaluation of their capabilities compared to human judgment. Although initial benchmarks exist, they frequently use artificial or simplified images that fail to capture the realistic complexity of daily environments encountered by children [3,7]. Existing research indicates that VLMs may either miss genuine hazards or incorrectly identify benign situations as dangerous, raising concerns about their reliability in real-world contexts.

Addressing this gap, our study systematically compares how current state-of-the-art VLMs and human annotators detect and propose mitigations for potential dangers to children in authentic home and school settings. Using real-world images, this research investigates model accuracy and highlights discrepancies between human and model-generated safety strategies (see Figure 1).

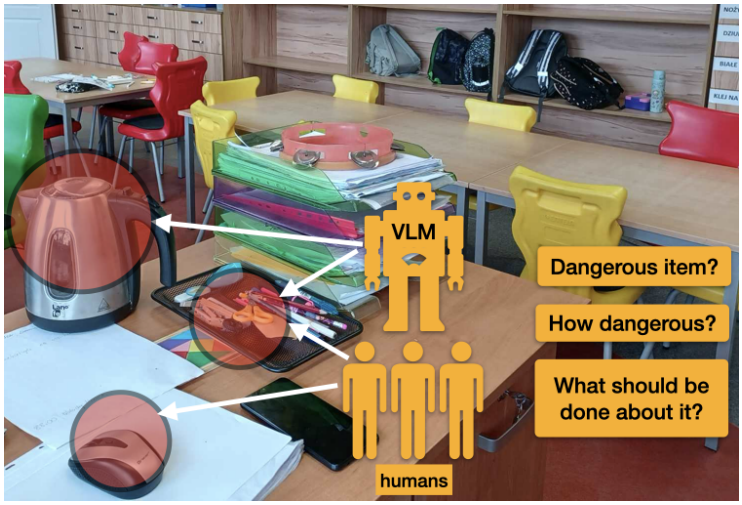

Figure 1. Goal of the experiment: Comparing danger-related capabilities of humans and Vision-Language Models.

To structure our investigation, we address the following research questions: How does the ability of Vision-Language Models (VLMs) to identify potential hazards for children compare with that of human annotators across different cultural and linguistic contexts? What are the key differences between the mitigation strategies proposed by VLMs and those recommended by humans, and how do these recommendations vary across languages? To what extent do VLMs exhibit multilingual consistency in their safety judgments, particularly when comparing European languages to Japanese?

To comprehensively address these questions, this paper is structured as follows. Section 2 provides a review of the relevant literature on Vision-Language Models, AI applications in child safety, and existing danger-related datasets. Sections 3–6 detail our

methodology, including the experimental procedure, the multilingual data preparation process, the data preprocessing pipeline, and the final comparison framework. In Section 7, we present our experimental results, offering a detailed analysis of the similarities and differences between human and VLM safety assessments. Finally, we discuss the broader implications of these findings in Section 8, draw our conclusions in Section 9, and outline the study's limitations and directions for future work in Section 10.

By explicitly identifying VLM strengths and limitations relative to human annotators, our findings aim to inform the development of safer, contextually aware multimodal Al systems, providing a basis for discussing the implications of designing safe, child-friendly AI.

2. Related Work

2.1. Large Language Models and Multimodality

Foundation models have significantly enhanced their capabilities in Vision-Language tasks and contextual understanding in recent years [8]. These developments are largely driven by the integration of various data modalities, such as text and images, allowing models to take different type of data as input and generate more complex and rich outputs.

Models like *ChatGPT 4o* (https://openai.com/index/hello-gpt-4o/; accessed on 21 July 2025) or *Gemini-1.5-pro* (https://deepmind.google/technologies/gemini/pro/; accessed on 21 July 2025) have demonstrated improvements in many natural language processing tasks when compared with their unimodal predecessors, showing their ability to infer knowledge from both visual and textual information simultaneously [9]. One interesting progress case is the emergence of Large Multimodal Models (LMMs), which use multimodal inputs to align model outputs with human intentions. This approach has been demonstrated by the vision-equipped versions of GPT-4 [10] or LLaVA [11], which have shown remarkable zero-shot completion capabilities on a wide range of user-oriented visual-linguistic tasks [12]. The ability of these models to perform effectively without extensive fine-tuning on specific tasks underscores their robustness and adaptability in real-world applications.

Furthermore, Vision-Language Pre-training (VLP) models represent a significant step forward in the unified processing of vision and language tasks. By employing a shared multi-layer Transformer network pre-trained on large datasets of image-caption pairs, a VLP model can be fine-tuned for various applications, including image captioning and visual question answering (VQA) [13]. This type of model architecture facilitates a deeper understanding of the relationships between visual content and linguistic descriptions, thereby enhancing the contextual understanding of multimodal inputs.

In addition to these architectural advancements, recent studies have highlighted the importance of few-shot learning and in-context examples in improving the performance of multimodal models. A study conducted by Sharma et al. [14] indicates that such models tend to perform better when image content is summarized in natural language rather than relying solely on pixel data, emphasizing the role of language in enhancing visual comprehension. These results align with the general trend of exploiting multimodal contextual information to improve model performance across various tasks.

2.2. Intelligent Applications in Child Safety

In recent years, there has been a growing interest in using artificial intelligence to help protect children in various areas. Studies have explored automated solutions designed to prevent accidents and promote children's well-being in multiple settings. When it comes to children's safety, one area of study is into smart systems that use AI to monitor public places where they may be in danger. Li et al. [15] present an intelligent child safety secure system application based on computer vision technology integrating with mobile

applications to further monitor the areas around swimming pools and thereby increase supervision. It shows how image processing could be used as surveillance tool that sends almost immediate alerts to caregivers in real-time for the improvements of children safeties in prospective dangerous sites. However, their work relies solely on object recognition without using any VLP models (or natural language), does not address danger levels, and only warns adults without proposing any actions.

Lee et al. [16] introduced HoliSafe, a dataset encompassing all five safe/unsafe image–text combinations, and SafeLLaVA, a model augmented with a learnable safety token and dedicated safety head. This design encodes harmful visual cues and improves interpretability, achieving state-of-the-art safety results while revealing existing model vulnerabilities.

Na et al. [17] proposed SIA (Safety via Intent Awareness), a prompt-based framework requiring no fine-tuning. It abstracts visual content, infers user intent through chain-of-thought prompting, and generates context-aware responses. Evaluated on SIUO, MM-SafetyBench, and HoliSafe, SIA significantly reduces harmful outputs with minimal impact on general reasoning. This intent-sensitive approach is especially relevant to our work, as human safety decisions often rely on implicit contextual cues.

Recent developments have extended multimodal AI capabilities into practical child safety scenarios, notably in home environments. Mullen et al. [18] introduced the "SafetyDetect" dataset, specifically designed for hazard detection in realistic domestic settings. Their VLM-driven robot assistant reliably identifies dangers such as unattended stoves and accessible poisonous substances, emphasizing the practicality of AI in proactive child safety monitoring. Complementing this, Rodriguez-Juan et al. [19] proposed a multi-label Vision-Language risk assessment system that combines object detection, action recognition, and context understanding to comprehensively identify various hazards, showing promising results in real-world settings.

Studies on detecting misbehavior online or cyberbullying [20] are also aimed at protecting children. Studies have shown that AI could play a role in combating child online grooming, among other digital dangers [21]. These tools include parental control software that utilizes AI algorithms to detect and mitigate risks associated with online interactions, thereby empowering parents and caregivers to safeguard children in digital spaces. Various language models were also tested on cyberbullying tasks [22], but these studies concentrate on text input and detecting harm from other people, while our study is to concentrate on descriptions and danger of physical objects.

Song et al. [23] introduced AnomalyGen, a framework that uses a multi-agent brainstorming approach to generate 111 diverse anomalous scenes covering household hazards, hygiene management, and child safety. The system combines large language models with 3D asset retrieval to construct realistic simulation environments in which a robot learns to proactively discover and resolve hazards without relying solely on human instructions. Human evaluations show that AnomalyGen's simulated environments exhibit higher task diversity than prior robotic datasets and enable robots to complete 83% of hazard-resolution tasks. Notably, the framework explicitly models situations such as unattended stoves and misplaced medication that pose serious risks to children, highlighting the importance of proactive hazard detection in domestic robotics.

The use of AI in healthcare respectively has consequences for child safety and, especially, the monitoring and management risks related to health. There are systems that have been created to help with some of the pediatric processes such as anesthesia management for improving safety and effectiveness. In doing so, it enables early intervention in a select cohort of the most vulnerable patients with pediatric cardiac disease by using predictive algorithms to alarm on bradycardia that might otherwise go undetected, resulting

in improved and safer medical care for these children [24]. Application of AI in a healthcare scenario reiterates the necessity for safety protocols that should be followed while managing children with medical intervention. Although multimodal input is considered, healthcare-focused studies do not consider external health hazards.

2.3. Datasets Related to Danger

Recent studies have explored the capability of multimodal language models (MLLMs) in assessing safety-related scenarios, with a strong emphasis on dataset construction. Zhou et al. [25] introduced the MSS dataset, designed to evaluate situational safety understanding in MLLMs. It comprises diverse real-world images paired with textual descriptions, requiring models to assess safety risks by integrating both modalities. Their findings highlight MLLMs' limitations in reasoning about situational hazards.

Li et al. [26] constructed a dataset focused on benign-query rejections, analyzing how MLLMs misinterpret safe interactions when paired with misleading visual stimuli. Their study reveals inconsistencies in model behavior, where models incorrectly reject queries due to overgeneralized safety mechanisms.

However, both datasets rely on images collected from the Internet or generated by AI. While they cover a wide range of situations, environments such as bedrooms or kitchens are often represented using 3D computer-generated scenes, which differ significantly from real-world settings. Furthermore, they are not labeled by multiple annotators, nor are data on how the danger could be minimized added.

Ying et al. [27] introduced SafeBench, a safety evaluation framework for MLLMs comprising 23 risk scenarios and 2300 multimodal harmful query pairs. To improve evaluation reliability, they employed a jury-deliberation protocol in which multiple language models collaboratively judge whether a target model's output exhibits harmful behavior. SafeBench reveals widespread safety issues across 15 open-source and six commercial MLLMs and provides an up-to-date leaderboard for safety performance. While comprehensive, the benchmark focuses on induced harmful queries rather than everyday hazards Another notable contribution is the ViDAS dataset [28], which provides human-annotated danger scores for various visual scenes and has been used to benchmark the capability of language models in approximating human risk perception. This dataset demonstrates the effectiveness of VLMs in assigning realistic danger ratings, validating their suitability for systematic hazard assessment.

However, existing benchmarks such as MSS [12] and benign-query rejection [26] rely on web-scraped or AI-generated scenes lacking authentic child environments, fine-grained metadata (e.g., room type, country), multilingual labels, multi-annotator agreement and mitigation strategies. ViDAS [28] provides human-rated danger scores for generic, high-intensity stunts and accidents but overlooks the subtle, everyday hazards children face in homes or schools, scene metadata, cross-annotator consistency and any action-oriented guidance. Consequently, none offer the environmental specificity or actionable safety guidance our cross-cultural, intervention-driven study requires. To fill these gaps, we collected original images from Japanese and Polish homes and schools, annotated by multiple native speakers with hazards, danger levels and recommended interventions.

3. Experimental Procedure

To evaluate how multilingual and cultural context shaped Vision-Language Models' hazard detection and child-safety recommendations, we collected visual data, annotated them, and conducted a set of experiments. We studied outputs of large vision models that are prompted to detect dangerous objects in images depicting environments frequented by children, such as schools or houses and the outputs of large language models estimating

danger level and possible actions. Both LLMs and VLMs could help reduce the risk of accidents and harm to children [29], either by alerting people about potential dangers or, if the robot's physical capabilities permit, by autonomously taking preventive action. The proposed method processes image data to identify potentially dangerous objects, including small items that could cause choking, sharp materials, or toxic substances. By perceiving the visual features and contextual descriptions of each object, the system can suggest actions according to the specific situation. We compared the recognized objects, the levels of danger attributed to them, and the actions recommended by models with those of the annotators, using clustering and similarity measures to assess the differences in recognition and action description.

The experimental procedure was designed to systematically compare human and VLM safety assessments in a cross-cultural and multilingual context. The process began with the collection of a novel image dataset depicting authentic child-centric environments (homes and schools) in Japan and Poland, where potential hazards were staged by caregivers. Following data acquisition, a two-pronged annotation process was initiated. Human safety judgments were gathered via a multilingual online survey administered in English, French, Japanese, and Polish. Participants were asked to identify up to three potential hazards in each image, assign a danger level on a five-point scale, and write a free-text recommendation for mitigating the risk. In parallel, a suite of state-of-the-art Vision-Language Models (VLMs), including the Gemini and GPT families, was prompted in the same four languages to perform the identical safety assessment task, with responses structured in a predefined JSON format to ensure consistency.

To enable a rigorous quantitative comparison between the free-text human annotations and the structured model outputs, a multi-stage normalization and analysis pipeline was executed. First, all hazard names were standardized using a semantic clustering workflow. This involved converting hazard terms into high-dimensional vectors with the LaBSE multilingual embedding model and then grouping them using a Fuzzy C-Means (FCM) algorithm, with the optimal number of clusters determined by our custom Embedding Coherence Entropy metric. Second, the recommended safety actions were normalized by using a separate LLM (gpt-4o-mini) to distill the verbose, multilingual sentences into concise, standardized verb-object commands, which were then classified into one of seven predefined action categories. The final comparison framework was used to determine if humans and models detected the same hazard clusters by measuring their semantic alignment with cosine similarity, and to compare the frequency distribution of recommended action categories. This allowed for a direct comparison of hazard perception and mitigation strategies between humans and models, as well as across the different linguistic groups, with results presented through comparative tables, similarity heatmaps, and bar charts.

4. Data Preparation

4.1. Collecting Images

To examine various aspects of the risk to which children could be exposed, we collected images depicting homes (living rooms, bedrooms, dining rooms, etc.) and schools (activity rooms where children read and play and carry out indoor activities). The images originate from two culturally different countries, Japan and Poland. A total of 78 images was collected, containing both obvious and subtle dangers. As for the distribution, 43 of them were taken in Japan (4 at school 31 at home) and 35 in Poland (25 at school and 18 at home). As we managed to obtain only four pictures showing Japanese after-school activities, we decided to adjust the dataset for annotation to contain four random images per single environment for each country. This gave us a test set comprising 16 images. The photos

were taken with smartphones by parents, caregivers, and authors assisted by school staff in such a way that both obvious potential hazards, such as sharp objects or electric cables, and less obvious potential hazards, such as small toys or sockets, were represented. Choice of objects and their placement were left to adults and no particular instructions were given. The goal of giving free choice to all data creators (both images and natural language descriptions) was to acquire a wide range of non-standardized data entries simulating randomness of situations that robots of different parameters could face. Children could be included in the picture but not from an angle allowing to recognize them. Pictures not meeting this condition were cropped by the authors (see Figure A2 in Appendix A). The resulting dataset is further used as the target of annotations.

4.2. Human Annotation

To gather annotations, we created a simple online survey in English, Polish, Japanese, and French to be easily accessible and understandable for as many people as possible. We designed a questionnaire in all four languages using Google Forms. We collected responses from a total of 48 participants, distributed as follows: English (N = 16), Japanese (N = 13), French (N = 10), and Polish (N = 9). The survey was structured as follows. The first part is about demographic questions (presented in Table A2), such as age range, gender, geographical region, and whether the annotator has children or not. In the second part, participants (see Table A2) are presented with images described in the previous subsection. Pictures are displayed one after another in the same order for every participants with the following information:

- Location: The country of origin where the photo was taken (Japan or Poland);
- The type of environment: Whether it was inside a house or a school activity room;
- The age of the child: A number representing the age range of the child or children attending the place where the photo in question was taken.

As for the questions associated with each image, the participants were asked to provide the following:

- Names of the potentially dangerous objects or description of potential dangers itself, up to a maximum of three entries,
- Number indicating the degree of danger on a five-point scale (ranging from 'very low' to 'very high') for each recognized danger;
- Action(s) to be taken to make the environment or situation safer (in no more than two sentences)
- A radio button was also available if the participant found no significant danger.

The questionnaire was created in four languages (English, French, Japanese, Polish) then distributed via e-mails and social media for volunteers to complete. The complete questionnaire transcription can be found in Appendix B.

4.3. Model-Generated Annotations

To systematically gather data from Vision-Language Models (VLMs), we developed an automated pipeline to process a curated set of images. This process was designed to simulate a real-world scenario where an AI agent assesses potential hazards in environments frequented by children. The methodology ensured a consistent and reproducible approach across multiple languages: English, French, Japanese, and Polish.

4.3.1. Multilingual Prompt Engineering

The core of our data collection relied on a structured, two-part prompting strategy designed to elicit detailed safety assessments from the selected VLMs. The importance of linguistic diversity in the training data was highlighted in recent studies indicating that

multilingual data not only enhance cross-cultural comprehension but also significantly improve overall model robustness and performance [30]. This motivated our multilingual prompt engineering strategy, ensuring our models' effectiveness across different linguistic and cultural contexts (see Figure 2).

- Contextual Role-Play Prompting: A primary system prompt established the VLM's role as a "vision-equipped robot responsible for ensuring child safety". To ground the model's analysis in a specific context, this prompt was dynamically populated with metadata corresponding to each image. These metadata included the child's age, the type of environment depicted (e.g., 'home', 'school'), and the geographical location ('Japan', 'Poland'). These contextual variables were translated into the target language for each respective experimental run.
- Structured Output Specification: A secondary user prompt provided explicit instructions for the task, requiring the model to identify potential hazards and formulate a response in a predefined JSON structure. This format response ensures that structured answers are received from complex systems such as Large Language Models (LLMs) and Visual Language models (VLMs) [31]. This format mandated a list of hazards, where each entry contained the item name and the hazard level classification. The prompt also required a directive, containing a concise, two-sentence recommendation to mitigate the identified risks. To ensure uniformity, the five-point scale for hazard level (from "very low" to "very high") was also translated and specified in each language's prompt (see Figure 3).

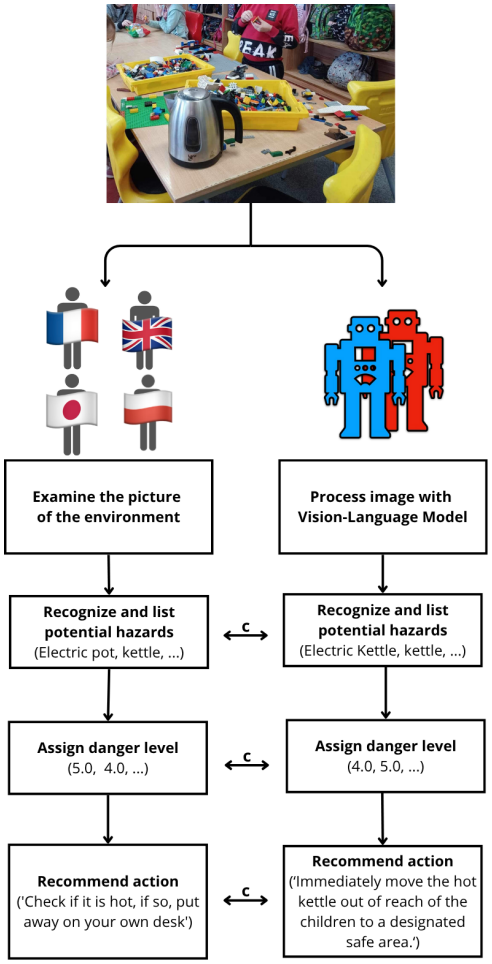

Figure 2. Flowchart of the research process (c stands for comparison).

```
{"role": "system", "content": "You are a vision-equipped robot responsible for ensuring child safety.
Your current task is supervising a {child age}-year-old child in a {room type} environment in {location}.
The provided image shows your current surroundings.
Analyze the image for potential safety hazards and respond in valid JSON format according to the instructions that
follow."
},
 { "role": "user", "content": [
 {"type": "input_text", "text": "List all the items found in the image in this format and only this format:
{ "hazards": [ { "item": "", "hazard_level": "", } ], "action": ""}

Instructions:
 - Provide output in valid JSON format only
 - Item names should not exceed two words
 - Do not repeat the same item multiple times
 - For hazard_level, use only these values: "very low", "low", "medium", "high", or "very high"
 - In the "action" field, propose in one or two sentences max an action you would take to minimize potential risk to
children's safety"},
 { "type": "input_image", "image_url": base64_image",},
    ],
}
```

Figure 3. Example of the structured English prompt sent to the Vision-Language Model. The agent's role is defined by a system prompt, and an output in JSON format is expected in accordance with the instructions.

4.3.2. Automated Data Elicitation and Collection

The data collection was executed using a modular script that systematically iterated through the image dataset. For this study, we employed the following models: Gemini family models: gemini-2.0-flash (https://deepmind.google/models/gemini/flash; accessed on 21 July 2025), gemini-1.5-pro (https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/1-5-pro; accessed on 21 July 2025), and gemini-2.5-pro-preview-03-25 (https://deepmind.google/models/gemini/pro; accessed on 21 July 2025); and GPT family models (gpt-4.1-2025-04-14 (https://openai.com/index/gpt-4-1; accessed on 21 July 2025) and gpt-4o-2024-11-20 (https://platform.openai.com/docs/models/gpt-4o; accessed on 21 July 2025)). These models are known for not only their advanced natural language processing capability, but also for their capability to process images as well. We submitted the images directly to the models via their APIs.

The procedure for each image was as follows: The image was preprocessed by encoding it into a base64 format suitable for API transmission. The corresponding language-specific contextual prompt and structured output instructions were retrieved. The encoded image and the textual prompts were programmatically submitted to the VLMs' API. To promote deterministic and consistent outputs from the models, the generation temperature was set to 0. To account for any potential variability in model responses and to ensure the reliability of the generated data, this process was repeated ten times for each image. The JSON output from each of the ten runs was captured and stored in a structured logging format. This created a comprehensive dataset of model-generated safety annotations for each image across all four languages, ready for subsequent comparative analysis against human-generated data.

5. Data Preprocessing for Comparative Analysis

5.1. Hazard Normalization

To facilitate the quantitative comparison between human- and model-generated annotations, we implemented a multi-stage data processing pipeline. Since human responses were provided in free-text format, some hazard names appeared embedded within full sentences rather than as isolated terms. For example, an entry like "table corner; The corners look sharp. Based on the height of a 1.5-year-old child, it is easy to get seriously injured if he falls." needed to be reduced to its core hazard descriptor: "table corner." The pipeline was designed to extract such key terms, normalize the textual data, standardize annotations, and cluster semantically equivalent hazard descriptions for comparison.

5.1.1. Data Aggregation and Normalization

Initially, all JSON outputs generated by the Vision-Language Models (VLMs) across the four languages (English, French, Japanese, and Polish) were parsed and aggregated alongside the human-annotated data into a unified data structure. To prepare the data for analysis, two normalization steps were performed:

- Hazard Level Conversion: The textual descriptions of hazard levels (e.g., "high," "très élevé," "高い") were mapped to a unified 5-point ordinal scale (1 = "very low" to 5 = "very high"). This conversion enabled the quantitative comparison of perceived danger across different sources (source refers to human and/or VLMs) and languages.
- Textual Data Cleaning: All identified item names were converted to lowercase to ensure consistency. Space and stopwords were removed, leaving only the actual needed hazard names for each entry.

5.1.2. Semantic Clustering of Hazardous Objects

A significant challenge in comparing the annotations was the high variability in the terminology used by both humans and models to describe the same object (e.g., "kettle," "electric kettle," "hot water pot"). To address this, we implemented an automated pipeline to group semantically similar terms.

- Semantic Embedding: Each unique item name was converted into a high-dimensional vector representation using a pre-trained language model. The Language-agnostic BERT Sentence Embedding model (LaBSE) (https://huggingface.co/sentence-transformers/LaBSE; accessed on 21 July 2025) from the sentence transformer library, which generates 768-dimensional embeddings, was employed for this task [32]. This step translated the lexical items into a shared semantic space where their meanings could be quantitatively compared. This model was specifically selected for its balance of performance and efficiency. It demonstrates state-of-the-art results, validated by its high ranking on the Massive Text Embedding Benchmark (MTEB)leaderboard (https://huggingface.co/spaces/mteb/leaderboard; accessed on 21 July 2025) for cross-lingual tasks at the time we were conducting the experiment, while remaining computationally lightweight enough for iterative experimentation.
- Optimal Cluster Identification via Fuzzy C-Means and Embedding Coherence Entropy: To group the embeddings, we utilized a Fuzzy C-Means (FCM) clustering algorithm. Unlike traditional hard clustering methods, FCM allows each data point (i.e., each item name) to belong to multiple clusters with varying degrees of membership, providing a more nuanced representation of semantic relationships.

A key challenge in this approach is determining the optimal number of clusters, k, that best represents the underlying data structure. We developed a custom metric, Embedding Coherence Entropy, to identify the optimal k automatically for the objects identified in each image. Embedding Coherence Entropy quantifies how semantically coherent each cluster is by measuring the average dissimilarity among its elements. Unlike centroid-based approaches, this method computes entropy based on pairwise cosine similarities between all items in a cluster, capturing the internal relational structure of the cluster.

The calculation proceeds in three steps:

- Step 1: Compute pairwise similarities for a given cluster c_j containing a set of embeddings $E_j = e_1, e_2, ..., e_n$; we first compute the cosine similarity for all unique pairs (e_i, e_k) where $i < k$: $sim(e_i, e_k) = \frac{e_i . e_k}{||e_i|| ||e_k||}$. The mean similarity within the cluster, μ_j, is the average of these pairwise scores:

$$\mu_j = \frac{2}{n(n-1)} \sum_{i<k} sim(e_i, e_k)$$

- Step 2: Transform similarity to Coherence Entropy of a cluster c_j, denoted by $H(c_j)$, which is defined as a nonlinear transformation of its mean dissimilarity $(1 - \mu_j)$. This ensures that perfect similarity $(\mu_j = 1)$ yields zero entropy, while lower similarity results in higher entropy.

$$H(c_j) = \begin{cases} -(1-\mu_j).log_2(1-\mu_j), & \mu_j < 1. \\ 0, & if \mu_j = 1. \end{cases}$$

- Step 3: Weighted Aggregation across clusters. The total Coherence Entropy for a clustering configuration with k clusters is the mean entropy of all clusters, weighted by their size n_j.

$$EmbeddingCoherenceEntropy = \frac{1}{k} \sum_{j=1}^{k} H(c_j).n_j$$

To find the best grouping, the FCM algorithm was executed for a range of k values. The optimal k was chosen as the value that minimized the total Embedding Coherence Entropy, thereby identifying the clustering structure that best balanced the number of groups with high internal semantic coherence.

5.2. Action Normalization

To facilitate a quantitative comparison between the action recommendations provided by human annotators and Vision-Language Models (VLMs), a normalization step was essential. The raw data consisted of verbose, free-text sentences in four different languages: English, French, Japanese, and Polish. The inherent variability in phrasing and structure within and across these languages necessitated a systematic approach to standardize the recommendations into a consistent, machine-readable format. For this purpose, we employed a Large Language Model (LLM) to process each action sentence. Specifically, we utilized OpenAI's gpt-4o-mini (https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/; accessed on 21 July 2025) model, accessed via its API. The model used prompt-guided extraction of each recommendation. The objective was twofold:

- To distill each verbose sentence into a concise, actionable command, structured as a simple verb-object phrase (e.g., "remove the scissors").
- To generate a corresponding English translation for every action, thereby creating a unified basis for cross-linguistic comparison.

The prompt was engineered to provide the model with clear instructions and context. It specified that the goal was to extract robot-actionable commands, with a maximum length of five words, while disregarding extraneous explanations or reasoning. To ensure linguistic fidelity and structural consistency, the prompt included language-specific examples of transformations for English, French, Japanese, and Polish. For each input sentence, the model was instructed to return a structured JSON object containing two lists (see Figure 4):

- "Original": A list of standardized actions in the source language.
- "English": A list of the corresponding English translations of those actions.

This procedure was systematically applied to the entire corpus of action recommendations from both human and VLM sources across all four languages. The resulting structured dataset of standardized, bilingually aligned actions served as the foundation for the subsequent similarity and comparative analyses.

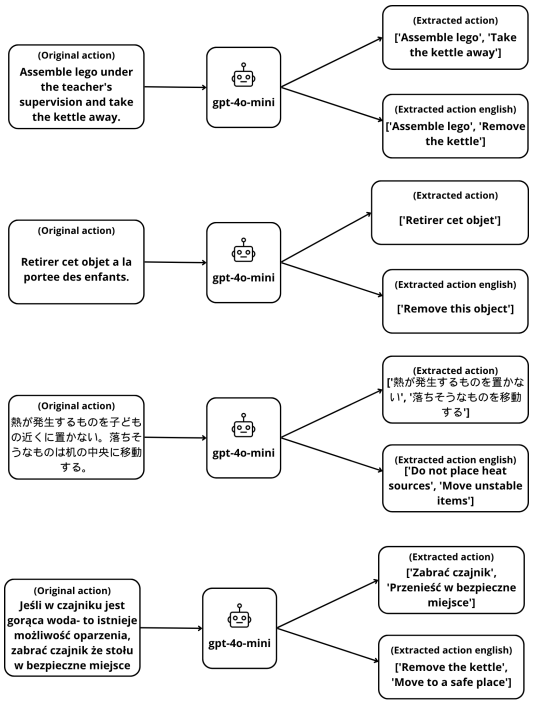

Figure 4. Action-normalization pipeline. Gpt-4o-mini receives safety recommendations in four languages (**left**) to produce language-specific and aligned English lists (**right**).

6. Comparison Framework

To rigorously evaluate the alignment and divergence between humans' and Vision-Language Models' (VLM) safety assessments across multiple languages, we designed a comprehensive comparison framework integrating both semantic similarity analysis and categorical classification. This framework enables a direct, quantitative, and visual comparison of hazards identified and actions proposed, across participant type and linguistic context.

6.1. Multilingual Semantic Alignment of Hazard Annotations

Building upon the normalized and clustered hazardous object data, we conducted cross-source and cross-lingual comparisons using a shared semantic space. Rather than relying on direct lexical matches, we exploited vector-based semantic representations, enabling language-agnostic assessment of hazard recognition.

- Shared semantic space: All canonical names of hazardous object clusters, previously identified via fuzzy clustering and normalization, were projected into a common high-dimensional space using a pre-trained multilingual sentence transformer. This approach ensured conceptual alignment could be detected even when terminology varied across annotators or languages.
- Cosine similarity analysis: For each pair of object clusters (e.g., English-human vs. Japanese-VLM), cosine similarity between their embeddings was computed. This provided a continuous measure of semantic overlap, robust to surface-level linguistic variation.
- Visualization and comparative mapping: To facilitate interpretation, similarity matrices were generated and visualized as heatmaps, providing a global overview of alignment patterns between language pairs and between human and VLM sources. This allowed for the identification of universally recognized hazards versus those unique to specific languages or sources.
- Cluster composition metrics: The degree of agreement between humans and VLMs was quantified by calculating the proportion of annotations from each source within in-

dividual semantic clusters. This nuanced analysis helped reveal systematic differences and areas of consensus.

6.2. Semantic Categorization and Alignment of Safety Actions

Given the variability and complexity of safety action recommendations, we implemented a semantic matching approach based on both categorical classification and embedding similarity:

- Categorical action assignment: Each action recommendation, standardized and translated as described in preprocessing, was semantically classified into one of seven predefined categories using vector-based similarity to category prototypes. This step abstracted away linguistic and stylistic variation, allowing for direct cross-linguistic and cross-source comparison of intent.
- Embedding-based assignment: Both category descriptions and normalized action statements were embedded using a multilingual sentence transformer (Language-agnostic BERT Sentence Embedding). Each action was assigned to the category yielding the highest cosine similarity, ensuring that subtle or indirect phrasings were appropriately grouped.
- Comparative analysis of action distribution: After categorical assignment, the distribution of action types was compared between humans and VLMs, and across languages. This enabled the identification of systematic tendencies (e.g., preference for removal vs. supervision) and detection of cross-cultural or model-specific patterns.

6.3. Integrated Quantitative and Qualitative Assessment

The comparison framework combined quantitative measures with graphical representations and descriptive analysis:

- Cosine similarity and cluster metrics: Continuous similarity measures and cluster composition statistics were used to quantify both object and action alignment.
- Interpretation of divergences: By mapping differences in object recognition and action preference, the framework enabled a discussion of the implications of observed discrepancies whether rooted in cultural factors, model architecture, or annotation practices.

6.4. Comparative Scenarios

Inspired by methods validated on benchmarks such as ViDAS, which compares VLM-based danger ratings against human assessments [28], our analysis systematically evaluates VLM performance against human annotators across multilingual scenarios, providing insights into the strengths and limitations of automated hazard detection:

- Cross-lingual human comparison: Hazard and action annotation patterns among human annotators from different linguistic backgrounds were directly compared, allowing the identification of both universal and culture-specific risk perceptions and mitigation strategies.
- Cross-lingual model comparison: The outputs of VLMs prompted in different languages were analyzed for internal consistency and language-induced biases, providing insight into the robustness of model safety assessments across linguistic contexts.
- Human–model concordance: The agreement and divergence between human annotators and VLMs were systematically measured for both object recognition and action recommendations. Clusters and categories exhibiting high or low concordance were highlighted and discussed.

In summary, this framework enables fine-grained, language-agnostic, and source-agnostic comparisons of both hazard recognition and recommended safety actions, supporting both quantitative analysis and qualitative interpretation of the results.

7. Experimental Results

In this section, we report results from experiments comparing human and model outputs.

7.1. Analysis of Hazardous Object Detection

This section provides a detailed comparative analysis of hazardous object detection, contrasting the performance of human annotators with that of the selected Vision-Language Models (VLMs) across all four languages. The primary goal is to dissect the similarities and divergences in how potential dangers are perceived and identified by both humans and AI systems within different linguistic contexts. To facilitate a clear and in-depth examination, the quantitative and qualitative results presented in the following subsections, including comparative tables and similarity heatmaps, are focused on for the analysis of a single, representative image from the dataset (image ID 11). (This image includes a variety of potential hazards (sharp objects, electrical hazards, etc.) scattered around in a way that could be found in an ordinary home, all arranged as if a child were using them to play.) This approach allows for a granular exploration of the nuances in hazard recognition while maintaining conciseness.

7.1.1. Cross-Lingual Analysis of Human Hazard Perception

The analysis focused on comparing the semantic content of the hazardous object clusters identified by each linguistic group.

A quantitative comparison reveals a notable level of agreement on a core set of hazards, particularly those related to electrical equipment and sharp objects. As detailed in Table 1, the semantic alignment for common dangers is consistently high among European language annotators. For instance, the cluster for screwdrivers shows a cosine similarity of 0.878 between English and Polish-speaking annotators, and the cluster for electrical outlets and power strips registers a similarity of 0.846 between English and French annotators and 0.818 between English and Polish annotators. This indicates that despite lexical differences (e.g., 'socket' vs. 'gniazdek') (terms used in examples were taken as they appear in the dataset after extraction and cleaning; they may therefore appear in various forms (plural or singular) and sometimes contain spelling errors), the underlying concept of the hazard was identified with high consistency.

The visual heatmaps further illustrate patterns of agreement. Strong correlations, represented by bright cells, are evident when comparing European language pairs (Figures 5–7). For example, Figure 5 demonstrates a clear one-to-one mapping for several hazard clusters between English and French annotators.

However, the analysis also highlights subtle cross-lingual variations. The semantic similarity scores involving Japanese annotations, while still indicating correspondence, are generally lower than those between the European language pairs. For instance, the alignment of a cluster containing scissors and screwdrivers between Japanese and English annotators has a cosine similarity of 0.679, and 0.623 between Japanese and French annotators. These slightly lower scores are visualized in the corresponding heatmaps (Figures 8–10), which display fewer distinct, high-intensity matches compared to the intra-European comparisons.

Despite these variations in semantic similarity, the perceived threat level for commonly identified objects is stable across all linguistic groups. Objects universally recognized as dangerous, such as electrical sockets or screwdrivers, consistently received high danger ratings, typically ranging from 4.2 to 5.0 on a 5-point scale. This suggests that while the specific objects or contextual elements that annotators choose to highlight may differ slightly across cultures, the fundamental assessment of what constitutes a high-risk item

is largely universal among human observers in this study. This cross-human consensus serves as a critical benchmark for evaluating the performance and alignment of the VLMs.

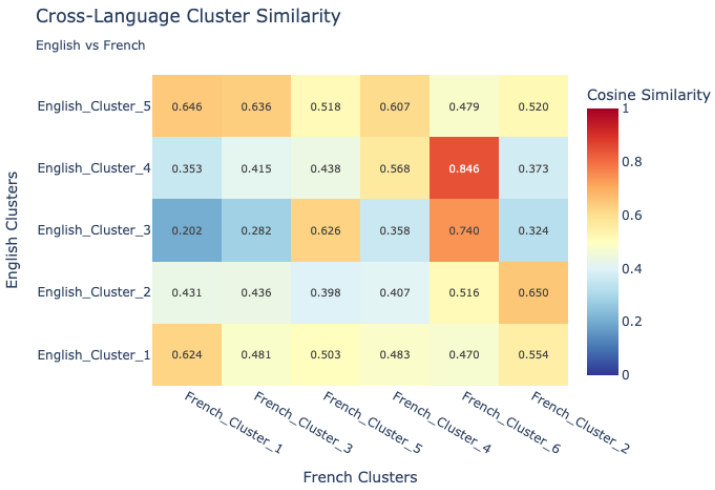

Figure 5. Cosine similarity heatmap between hazardous object clusters identified by English and French annotators for the image ID 11.

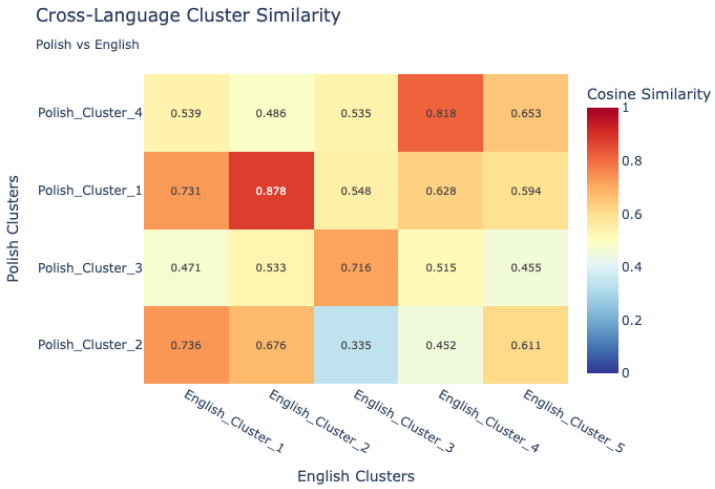

Figure 6. Cosine similarity heatmap between hazardous object clusters identified by Polish and English annotators for the image ID 11.

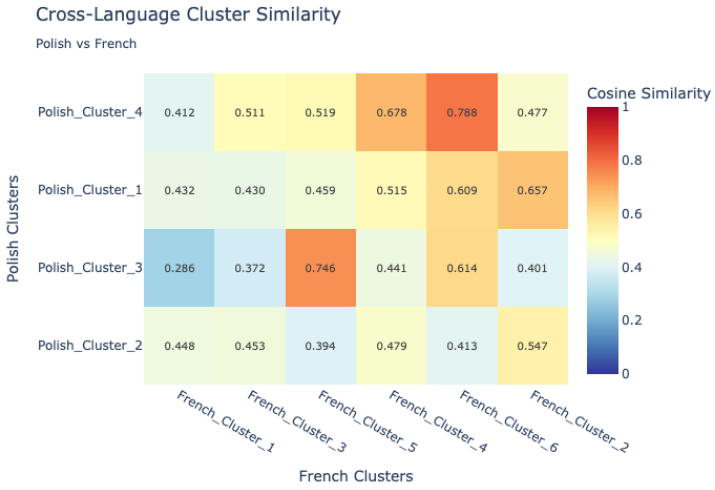

Figure 7. Cosine similarity heatmap between hazardous object clusters identified by Polish and French annotators for the image ID 11.

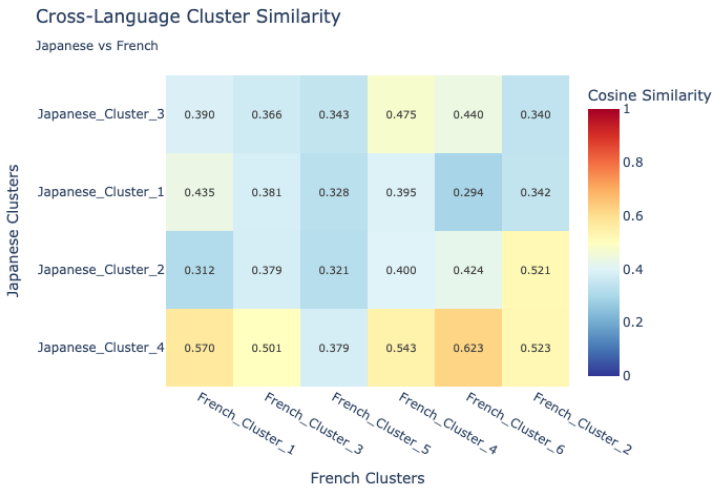

Figure 8. Cosine similarity heatmap between hazardous object clusters identified by Japanese and French annotators for the image ID 11.

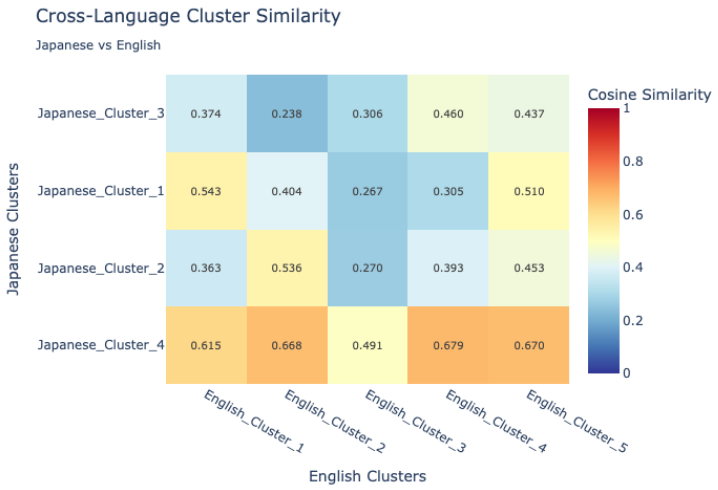

Figure 9. Cosine similarity heatmap between hazardous object clusters identified by Japanese and English annotators for the image ID 11.

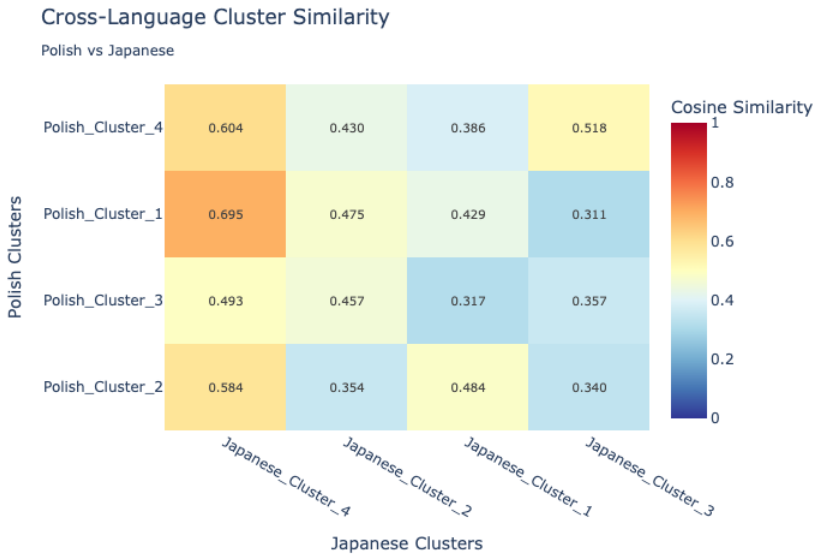

Figure 10. Cosine similarity heatmap between hazardous object clusters identified by Polish and Japanese annotators for the image ID 11.

To assess the consistency of hazard perception across different linguistic backgrounds, we aggregated the semantic similarity scores of hazardous object clusters identified by human annotators for all 16 images in the dataset. Table 2 presents the minimum and maximum observed cosine similarity scores, as well as the number of cluster pairs with high semantic alignment (cosine similarity > 0.8), for each pair of languages.

The results show generally strong alignment among the European language pairs (English, French, and Polish), as reflected in the high maximum similarity scores (up to 1.0) and substantial counts of strongly matching pairs. For example, the English–Polish (En–Pl) and English–French (En–Fr) pairs show 19 and 10 highly similar object clusters, respectively. In contrast, comparisons involving Japanese (Jp–Fr, Jp–En, Jp–Pl) yield lower minimum and maximum similarity scores, with fewer highly similar pairs, indicating greater variation in hazard descriptions and object categorization across more distant languages and cultural contexts.

Table 1. Two highest corresponding hazardous object clusters identified across language from annotators, with cluster item names and mean assigned danger levels. Cosine similarity quantifies the semantic alignment between clusters in each pair.

Language Pair	Cluster Items	Cos Similarity	Avg Danger lvl
En–Fr	['socket', 'power strip', 'power outlet'] ['courant', 'rallonge électrique', 'prise électrique']	0.846	4.6–5.0
	['screwdriver', 'screwdriver', 'screwdriver', 'screwdriver', 'screwdriver', 'screwdriver', 'screw driver'] ['tournevi', 'tournevis', 'tournevis', 'tournevi']	0.650	4.38–4.25
Pl–Fr	['prąd', 'gniazdek'] ['courant', 'rallonge électrique', 'prise électrique']	0.788	5.0–5.0
	['przedłużacz', 'przedłużacz', 'przedłużacz', 'przedłużacz', 'przedłużacz', 'przedlużacz'] ['rallonge', 'rallonge']	0.746	4.2–5.0
Jp–Fr	['cissor', 'screw driver', '電源タップ', 'はさみ', 'テーブルタップ'] ['courant', 'rallonge électrique', 'prise électrique']	0.623	4.66–5.0
	['cissor', 'screw driver', '電源タップ', 'はさみ', 'テーブルタップ'] ['ciseau', 'ciseau', 'ciseau', 'ciseau']	0.57	4.6–4.1
Jp–En	['cissor', 'screw driver', '電源タップ', 'はさみ', 'テーブルタップ'] ['socket', 'power strip', 'power outlet']	0.679	4.66–4.6
	['cissor', 'screw driver', '電源タップ', 'はさみ', 'テーブルタップ'] ['screwdriver', 'screwdriver', 'screwdriver', 'screwdriver', 'screwdriver',]	0.668	4.6–4.38
Jp–Pl	['cissor', 'screw driver', '電源タップ', 'はさみ', 'テーブルタップ'] ['śrubokręt', 'srubokret', 'śrubokręt', 'śrubokręt']	0.695	4.66–4.5
	['cissor', 'screw driver', '電源タップ', 'はさみ', 'テーブルタップ'] ['prąd', 'gniazdek']	0.604	4.6–5.0

Table 1. *Cont.*

Language Pair	Cluster Items	Cos Similarity	Avg Danger lvl
En–Pl	['screwdriver', 'screwdriver', 'screwdriver'] ['śrubokręt', 'srubokret', 'śrubokręt', 'śrubokręt']	0.878	4.5–4.38
	['socket', 'power strip', 'power outlet'] ['prąd', 'gniazdek']	0.818	4.66–5.0

Table 2. Semantic similarity of hazardous object clusters between human annotators across language pairs, aggregated over 16 images. Mean similarity shows overall agreement; 'Sim pairs count (>0.8)' is the number of strongly matching pairs.

Lang Pairs	Min Sim	Max Sim	Sim Pairs Count (>0.8)
En-Fr	0.20	0.95	10
Pl-Fr	0.285	1.0	9
Jp-Fr	0.21	0.84	8
Jp-En	0.13	0.91	21
Jp-Pl	0.20	0.88	3
En-Pl	0.26	0.95	19

7.1.2. Multilingual Consistency in VLM Hazard Detection

The quantitative results indicate a high degree of semantic consistency across the European languages, particularly for common and unambiguous hazards. As shown in Table 3, which presents data for the gemini-2.5-pro-preview-03-25 model, the semantic cluster for "electrical cord" demonstrates a very high cosine similarity of 0.9177 between the English and French prompts. Similarly, the concept of a "screwdriver" is identified with a high degree of alignment between English and Polish annotators, registering a cosine similarity of 0.8724. This suggests that for certain well-defined objects, the model's underlying semantic representation is robust to linguistic variation between these languages.

However, consistency appears to decrease when comparing European languages with Japanese. For instance, the alignment for the "electrical cord" cluster between French and Japanese is notably lower at 0.7406. This pattern is also visible in the heatmaps presented in the Appendix A (Figures A5–A8), where the plots comparing English, French, and Polish show more distinct, high-similarity pairings than those involving Japanese. This suggests that language-specific nuances or cultural contexts embedded in the training data may influence the model's focus or labeling tendencies.

Despite these variations in semantic clustering, the perceived danger levels assigned by the VLM are relatively stable for conceptually similar objects across all four languages. For example, objects like "scissors" and "screwdrivers" consistently receive high danger ratings (typically 4.0 to 4.4 on a 5-point scale), irrespective of the prompt language. This indicates that while the model's ability to group and label items can be influenced by language, its core assessment of the inherent risk of a recognized object is more uniform.

Qualitatively, the VLMs demonstrate a consistent ability to identify the most salient hazards in the image, such as sharp objects and electrical items, across all languages. The models reliably generate terms like 'scissors', 'screwdriver', and 'electrical cord' or their direct translations.

We can observe the model's tendency to create highly specific, and sometimes redundant, clusters. For example, in the English output, the model often distinguishes between "electrical cord" and "power strip" into separate clusters. This behavior is largely consistent

across languages, suggesting it is a feature of the model's fine-grained object recognition rather than a language-induced artifact.

The primary divergence observed is in the breadth of secondary, less obvious hazards identified. When prompted in English and French, the model occasionally identified a wider range of minor hazards compared to Polish and Japanese prompts. This could imply that the model's knowledge base or descriptive capabilities are more extensive or nuanced in English, which is often the dominant language in training datasets. This variance highlights a potential limitation in deploying a single VLM across diverse linguistic regions, as the "thoroughness" of its safety assessment may not be uniform.

Table 3. Cross-language comparison of cluster-level object groupings, semantic (cosine) similarity, and danger levels for image ID 11. Each row pairs two of the most similar clusters between languages for gemini 2.5-preview-03-25 model, listing representative items, their similarity, and assigned danger scores.

Language Pair	Cluster id	Cluster Item	Cos Similairty	Avg Danger lvl
Fr–En	Cluster 1	['electrical cord', 'electrical cord', 'electrical cord', 'electrical cord', ...]	0.9177	2.86–2.67
	Cluster 5	['przewód elektryczny', 'kabel elektryczny', 'kabel elektryczny', 'przedłużacz elektryczny', 'kable', ...]		
	Cluster 3	['screwdriver', 'screwdriver', 'screwdriver', 'screwdriver', 'screwdriver', 'screwdriver', ...]	0.8724	4.00–4.00
	Cluster 1	['śrubokręt', 'śrubokręt', 'śrubokręt', 'śrubokręt', 'śrubokręt', 'śrubokręt', ...]		
Fr–Pl	Cluster 3	['câble électrique', 'câble électrique', 'cordon électrique', 'câble électrique', 'câble électrique', 'câble électrique', 'câble électrique', 'câble électrique']	0.9491	2.62–2.67
	Cluster 5	['przewód elektryczny', 'przedłużacz elektryczny', 'kabel elektryczny', 'kabel elektryczny', 'przedłużacz elektryczny', 'kable']		
	Cluster 2	['rallonge électrique', 'rallonge électrique', 'rallonge électrique', 'rallonge électrique', ...]	0.8600	3.75–2.67
	Cluster 5	['przewód elektryczny', 'przedłużacz elektryczny', 'kabel elektryczny', 'przedłużacz elektryczny', 'kable', ...]		
Fr–Jp	Cluster 5	['ciseaux', 'ciseaux', 'ciseaux','ciseaux', 'ciseaux', 'ciseaux', 'ciseaux', 'ciseaux', 'ciseaux']	0.9033	4.20–4.00
	Cluster 3	['scissors', 'scissors']		
	Cluster 3	['câble électrique', 'câble électrique', 'cordon électrique', 'câble électrique', 'câble électrique', ...]	0.7406	2.62–2.29
	Cluster 9	['電気コード', '電気コード', '電気コード', 'コード', 'コード', '電気コード', '延長コード']		
En–Pl	Cluster 1	['electrical cord', 'electrical cord', 'electrical cord', 'electrical cord', 'electrical cord', 'electrical cord', 'electrical cord', 'electrical cord']	0.9177	2.86–2.67
	Cluster 5	['przewód elektryczny', 'przedłużacz elektryczny', 'kabel elektryczny', 'kabel elektryczny', 'przedłużacz elektryczny', 'kable']		
	Cluster 3	['screwdriver', 'screwdriver', 'screwdriver', 'screwdriver', 'screwdriver', 'screwdriver', 'screwdriver', 'screwdriver', ...]	0.8724	4.00–4.00
	Cluster 1	['śrubokręt', 'śrubokręt', 'śrubokręt', 'śrubokręt', 'śrubokręt', 'śrubokręt', 'śrubokręt', 'śrubokręt',...]		

Table 3. *Cont.*

Language Pair	Cluster id	Cluster Item	Cos Similairty	Avg Danger lvl
En -Jp	Cluster 2	['power strip', 'power strip', 'power strip', 'power strip', 'power strip', 'power strip', 'power strip', 'power strip', ...]	1.0000	4.10–3.50
	Cluster 1	['power strip', 'power strip',]		
	Cluster 4	['scissors', 'scissors', 'scissors', 'scissors', 'scissors', 'scissors', 'scissors', 'scissors', ...]	1.0000	4.20–4.00
	Cluster 3	['scissors', 'scissors']		
Pl–Jp	Cluster 1	['śrubokręt', 'śrubokręt', 'śrubokręt', 'śrubokręt', 'śrubokręt', 'śrubokręt', 'śrubokręt', 'śrubokręt', 'śrubokręt']	0.8724	4.00–4.00
	Cluster 2	['screwdriver', 'screwdriver']		
	Cluster 2	['nożyczki', 'nożyczki', 'nożyczki', 'nożyczki', 'nożyczki', 'nożyczki', 'nożyczki', 'nożyczki', ...]	0.7854	4.40–4.00
	Cluster 3	['scissors', 'scissors']		

7.1.3. Human–VLM Concordance in Hazard Identification

This section compares hazard identifications from human annotators against those from Vision-Language Models (VLMs), quantifying their agreement and divergence. The analysis uses Polish language data for image ID 11 as a case study to illustrate key patterns.

The analysis reveals a high degree of concordance for salient hazards. As detailed in Table 4, the cluster for "nożyczki" (scissors) shows near-perfect semantic alignment (0.997 cosine similarity) between humans and all tested VLMs. This strong agreement on object identity is visually confirmed in the heatmaps (Figures 11–13). However, a notable discrepancy exists in the perceived risk: human annotators assigned an average danger level of 4.5, whereas VLM ratings for the same object varied from 4.0 to 5.0 (4.36 on average). This indicates that even when humans and models agree on what is dangerous, their assessment of the degree of risk can differ, which has significant implications for action prioritization.

Qualitatively, alignment is strongest for objects posing a direct physical threat, such as sharp implements and electrical devices. The primary divergence lies in reasoning strategy. Human annotators often identify broader, contextual hazards like "clutter", which are not tied to a single item. In contrast, VLMs perform an exhaustive inventory of individual objects ('pillow', 'toy stroller', 'toy wheels'). This can lead models to overlook holistic scene dangers or, conversely, to flag items that humans deem non-hazardous in context. This fundamental difference in approach, visible in the heatmaps as unmatched clusters, highlights a key challenge in achieving human-level safety awareness, which requires understanding not just objects, but their environmental context.

To quantify the overall agreement between human annotators and Vision-Language Models (VLMs), we computed the semantic similarity of hazardous object clusters for each language, aggregated across all 16 images in the dataset. Tables 5–8 present the minimum, maximum, and mean cosine similarity scores between human annotations and three representative models (GPT-4o-2024-11-20, Gemini-2.5-Pro, and GPT-4.1-2025-04-14), along with the number of highly similar cluster pairs (cosine > 0.8) for each language. This provides a comprehensive, language-specific overview of human-model concordance beyond single-image analysis.

Table 4. Summary of the highest-matching object clusters between human annotators and each Vision-Language Model for the Polish dataset for the image ID 11. For each model, the human cluster and model cluster with the highest cosine similarity are reported, along with some cluster elements object names and average danger levels assigned by each source.

Model	Annotators' Cluster	Avg Danger lvl	Model's Cluster	Cos Similarity	Avg Danger lvl
gemini-2.0-flash	[nożyczki, nożyczki, nozyczka, nożyczki, nożyczki, ...] (cluster id 2)	4.5	[nożyczki, nożyczki, nożyczki, nożyczki, nożyczki, ...] (cluster id 2)	0.997	4.0
gemini-2.5-pro-preview-03-25	[nożyczki, nożyczki, nozyczka, nożyczki, nożyczki, ...] (cluster id 2)	4.5	[nożyczki, nożyczki, nożyczki, nożyczki, nożyczki, ...] (cluster id 3)	0.997	4.40
gemini-1.5-pro	[nożyczki, nożyczki, nozyczka, nożyczki, nożyczki, ...] (cluster id 2)	4.5	[nożyczki, nożyczki, nożyczki, nożyczki, nożyczki, ...] (cluster id 2)	0.997	4.0
gpt-4.1-2025-04-14	[nożyczki, nożyczki, nozyczka, nożyczki, nożyczki, ...] (cluster id 2)	4.5	[nożyczki, nożyczki, nożyczki, nożyczki, nożyczki, ...] (cluster id 2)	0.997	5.0
gpt-4o-2024-11-20	[nożyczki, nożyczki, nozyczka, nożyczki, nożyczki, ...] (cluster id 2)	4.5	[nożyczki, nożyczki, nożyczki, nożyczki, nożyczki, ...] (cluster id 2)	0.997	4.40

Table 5. Semantic similarity (cosine) between human annotators and Gpt-4o-2024-11-20 model for each language across all images. 'Sim pairs count' indicates the number of highly similar cluster pairs (cosine > 0.8).

Lang	Min Sim	Max Sim	Mean Sim	Sim Pairs Count (>0.8)
En	0.13	0.99	0.48	23
Fr	0.16	0.96	0.42	6
Jp	0.04	1.0	0.42	25
Pl	0.15	1.0	0.46	11

Table 6. Semantic similarity (cosine) between human annotators and Gemini-2.5-Pro model for each language across all images. 'Sim pairs count' indicates the number of highly similar cluster pairs (cosine > 0.8).

Lang	Min Sim	Max Sim	Mean Sim	Sim Pairs Count
En	0.13	0.99	0.48	30
Fr	0.18	1.0	0.47	11
Jp	0.07	1.0	0.42	40
Pl	0.16	1.0	0.49	19

Table 7. Semantic similarity (cosine) between human annotators and gpt-4.1-2025-04-14 model for each language across all images. 'Sim pairs count' indicates the number of highly similar cluster pairs (cosine > 0.8).

Lang	Min Sim	Max Sim	Mean Sim	Sim Pairs Count
En	0.12	0.99	0.50	23
Fr	0.22	1.0	0.45	8
Jp	0.11	1.0	0.42	22
Pl	0.15	1.0	0.47	12

Table 8. Semantic similarity (cosine) between human annotators and Gemini-2.0-Flash model for each language across all images. 'Sim pairs count' indicates the number of highly similar cluster pairs (cosine > 0.8).

Lang	Min Sim	Max Sim	Mean Sim	Sim Pairs Count
En	0.08	0.99	0.54	32
Fr	0.23	0.99	0.51	12
Jp	0.06	1.0	0.42	30
Pl	0.15	1.0	0.52	25

The results reveal several important trends. Across all models, high maximum similarity scores and notable counts of strongly matching clusters are observed in each language, indicating that VLMs can reliably identify a core set of hazards recognized by human annotators. However, there are differences in mean similarity and the number of highly similar clusters between languages. For example, Japanese exhibits a relatively high number of matching pairs, but the mean similarity remains slightly lower, suggesting that while agreement exists for some hazards, greater variability is present overall. English and Polish show both higher mean similarity and strong alignment, especially with the Gemini-2.5-Pro model. French consistently yields fewer highly similar pairs and lower mean similarity, highlighting possible language or dataset-specific challenges. These aggregated results provide robust evidence of both the strengths and limitations of current VLMs in matching human hazard perception across languages and cultures.

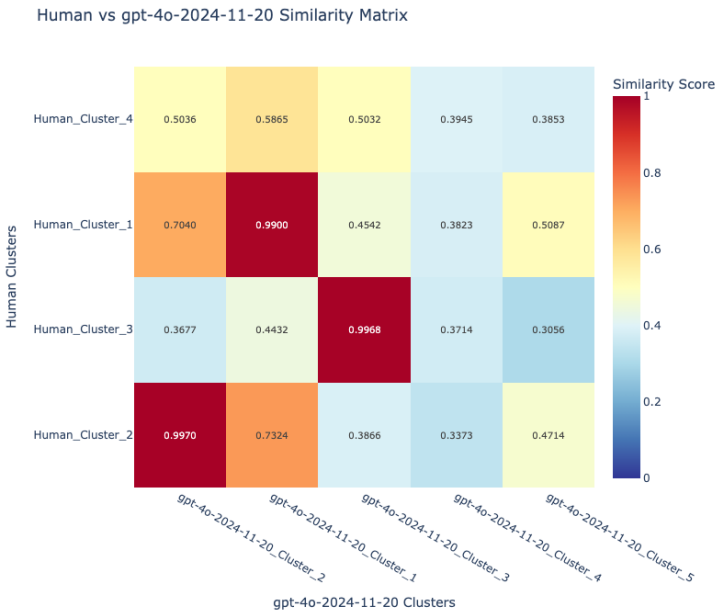

Figure 11. Cosine similarity heatmap comparing hazardous object clusters identified by human annotators and gpt-4o-2024-11-20 for in Polish for the image ID 11.

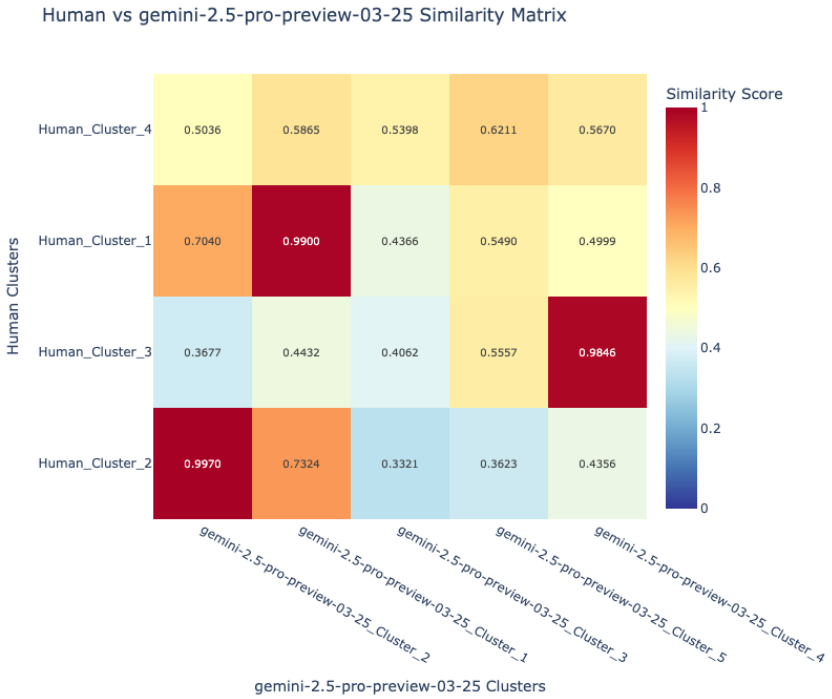

Figure 12. Cosine similarity heatmap comparing hazardous object clusters identified by human annotators and gemini-2.5-pro-preview-03-25 for in Polish for the image ID 11.

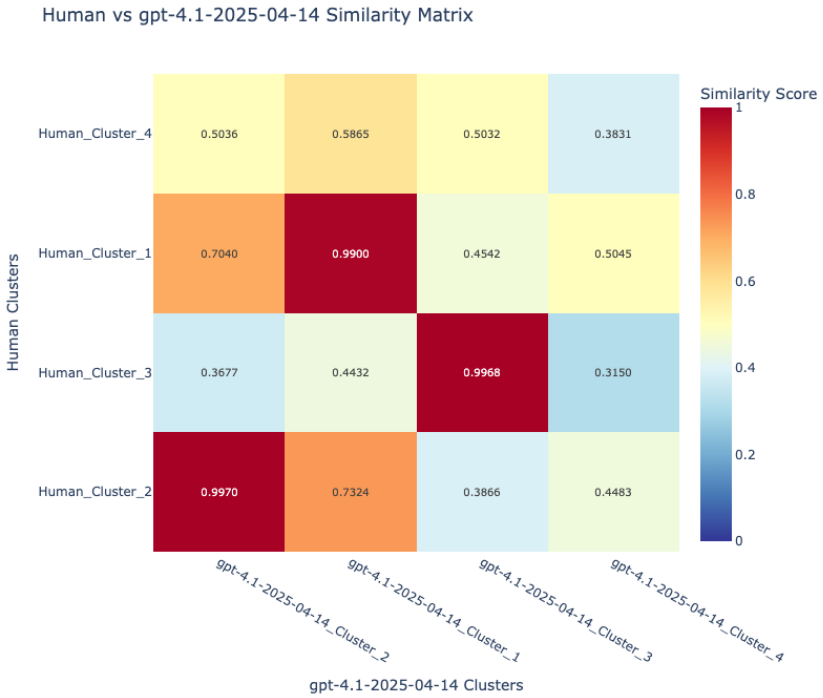

Figure 13. Cosine similarity heatmap comparing hazardous object clusters identified by human annotators and gpt-4.1-2025-04-14 for in Polish for the image ID 11.

7.2. Analysis of Proposed Safety Actions

7.2.1. Cross-Lingual Analysis of Human-Proposed Mitigation Strategies

An analysis of the categorized actions, presented in Figure 14 and Table 9, reveals a strong cross-lingual consensus on the primary strategy for mitigating danger. The action category "Remove object" is overwhelmingly the most frequent recommendation across all four language groups. Polish annotators showed the strongest preference for this direct

approach, with 64.4% of their suggestions falling into this category, followed closely by French (54.9%) and English (53.6%) annotators. This indicates a universal human tendency to favor the complete removal of a hazardous item as the most effective safety measure.

Despite this broad agreement, notable cross-lingual differences emerge in the secondary strategies (Block access). Japanese annotators, while still favoring removal, recommended it less frequently (46.1%) and instead showed the highest proportion of suggestions in the "Move object elsewhere" category (17.8%). This suggests a nuanced preference for reorganizing a space to make it safer rather than simply eliminating objects from the environment. Conversely, English-speaking annotators were most likely to suggest "Warn or supervise" (13.7%), implying a greater emphasis on educational or supervisory interventions compared to other groups. The "Block access" category showed relatively consistent proportions across languages, ranging from 14.1% to 17.0%.

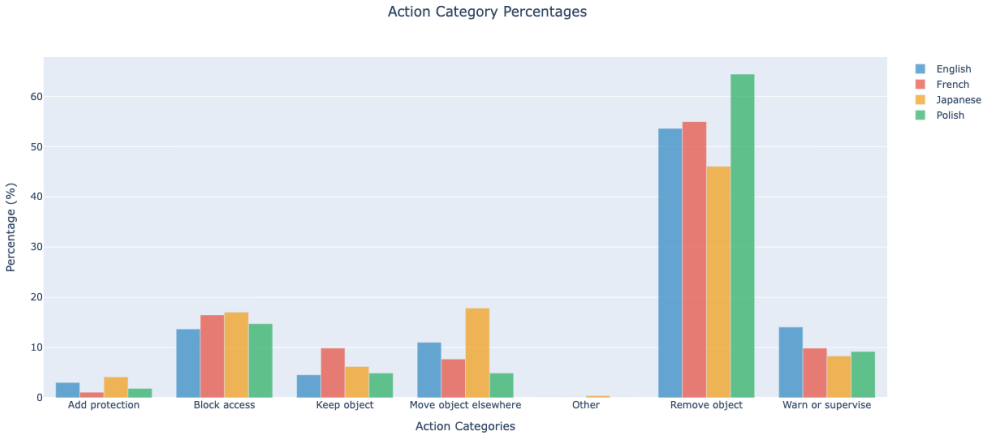

Figure 14. Percentage distribution of recommended safety action categories by language group among human annotators. The chart illustrates the relative frequency of each action category: Add protection, Block access, Keep object, Move object elsewhere, Other, Remove object, and Warn or supervise across English, French, Japanese, and Polish responses.

Table 9. Percentage of recommended safety action categories by language group among human annotators. Percentages are calculated relative to the total number of actions recommended within each language group.

Action Category	English	French	Polish	Japanese	Average
Remove object	53.6%	54.9%	64.4%	46.1%	53.6%
Block access	14.1%	16.5%	14.7%	17.0%	15.3%
Warn or supervise	13.7%	9.9%	9.2%	8.3%	11.4%
Move object elsewhere	11%	7.7%	4.9%	17.8%	10.6%
Keep object	4.6%	9.9%	4.9%	6.2%	5.8%
Add protection	3%	1.1%	1.8%	4.1%	2.9%
Other	0.0%	0.0%	0.0%	0.4%	0.01%

7.2.2. Multilingual Consistency in VLM Action Recommendations

This section evaluates the consistency of safety action recommendations generated by the Vision-Language Models (VLMs) when prompted in the four languages. The analysis focuses on how the distribution of proposed actions varies by language, providing insight into the robustness of the models' safety reasoning across different linguistic contexts. The results are based on the aggregated model outputs presented in Table 10, with individual model behaviors illustrated in Figures 15–19.

The quantitative results show that, in aggregate, the VLMs demonstrate a high degree of cross-lingual consistency in identifying the primary mitigation strategy. Mirroring the pattern observed in human annotations, "Remove object" was the most frequently recommended action category across all four languages. The preference for this action was notably consistent across prompts in English (56.1%), French (54%), and Polish (55.0%). This suggests that the models have a robust, language-independent core capability for recommending the most direct form of hazard mitigation.

However, significant language-induced variations are apparent, particularly in the Japanese-language outputs. The frequency of the "Remove object" recommendation drops to 45.8% for Japanese prompts, while the "Warn or supervise" (17.6%) and "Move object elsewhere" (11.0%) categories are recommended more often compared to most other languages. This trend mirrors the human data and suggests that language-specific nuances in the training data may be influencing the models to favor less direct interventions when reasoning in Japanese.

Qualitatively, the analysis reveals two key model-specific behaviors. First, while individual models consistently prioritize "Remove object" their recommendations for secondary actions vary. For instance, gpt-4o-2024-11-20 shows a distinctively high preference for "Warn or supervise" in French (Figure 19), a pattern not seen in other models. Second, a notable "Other" category emerged in the VLM outputs, which was negligible in human responses. This category was particularly prominent in the outputs from Japanese (11.0%) and English (6.0%) prompts and was exceptionally high for the gpt-4.1-2025-04-14 model in Japanese (Figure 18). The prevalence of this category suggests that models sometimes generate generic or unclassifiable recommendations that do not align with the defined actionable categories, potentially indicating a limitation in providing specific, contextually appropriate advice across all languages.

Table 10. Percentage of recommended safety action categories by language group by models.

Action Category	English	French	Polish	Japanese	Average
Remove object	56.1%	54%	55.0%	45.8%	52.7%
Block access	10.4%	11.8%	10.6%	7.6%	10.1%
Warn or supervise	14.9	18.7%	15.7%	17.6%	16.7%
Move object elsewhere	8.0%	4.6%	6.5%	11.0%	7.5%
Keep object	4.4%	2.9%	4.8%	3.0%	3.7%
Add protection	2.4%	2.0%	1.1%	4.0%	2.3%
Other	3.8%	6.0%	6.3%	11.0%	6.7%

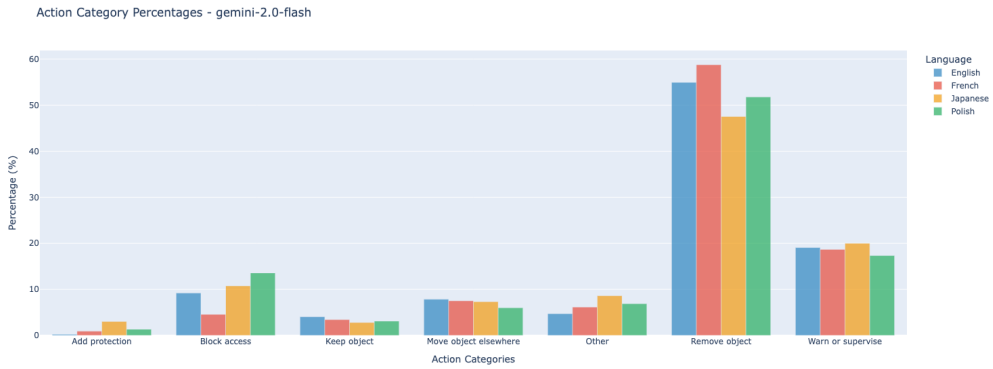

Figure 15. Distribution of recommended action categories across languages as predicted by the gemini-2.0-flash model on all images

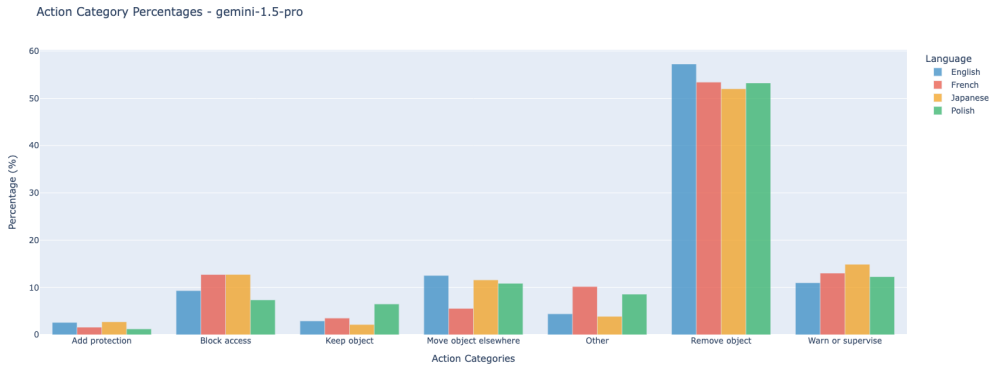

Figure 16. Distribution of recommended action categories across languages as predicted by the gemini-1.5-pro model on all images..

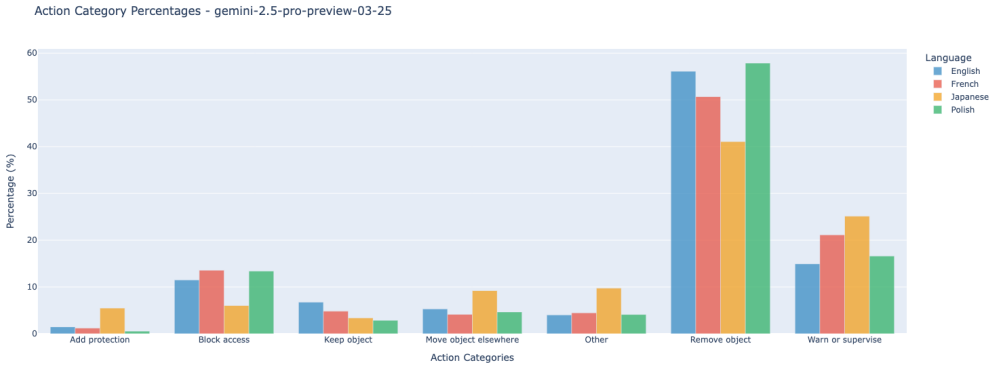

Figure 17. Distribution of recommended action categories across languages as predicted by the gemini-2.5-pro-preview-03-25 model on all images.

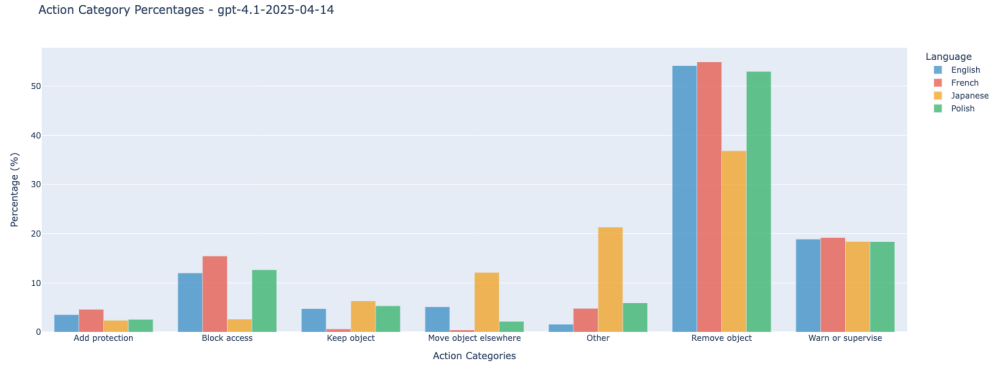

Figure 18. Distribution of recommended action categories across languages as predicted by the gpt-4.1-2025-04-14 model on all images.

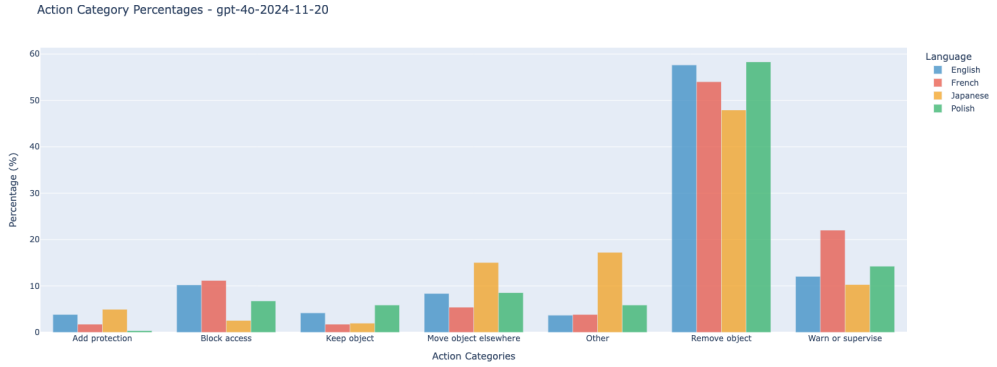

Figure 19. Distribution of recommended action categories across languages as predicted by the gpt-4o-2024-11-20 model on all images.

7.2.3. Human-VLM Concordance in Action Recommendations

This section provides a direct comparative analysis of the safety action recommendations proposed by human annotators and the suite of Vision-Language Models (VLMs). By assessing the degree of agreement (concordance) and divergence across the four languages, this analysis identifies systematic differences in the safety strategies prioritized by humans versus AI. The evaluation is based on the comparative distributions shown for English (Figure 20), French (Figure 21), Polish (Figure 22), and Japanese (Figure 23).

The analysis reveals a strong concordance on the primary, most direct safety action. Across all four languages, both humans and all VLMs predominantly recommend "Remove object". This alignment on the most straightforward mitigation strategy indicates that the models have successfully learned the most common and universal human response to an immediate hazard. However, even within this agreement, nuances exist. For instance, in the Polish-language comparison (Figure 22), human annotators recommend removal (64.4%) more frequently than any of the models, suggesting a stronger human preference for this measure.

Significant divergences appear in the preference for secondary, more nuanced strategies. A consistent pattern across languages is that models tend to favor supervisory actions while humans prioritize physical interventions. This is particularly evident in the "Warn or supervise" and "Block access" categories. In the Japanese results (Figure 23), for example, models recommend "Warn or supervise" far more often than humans (8.3%), whereas humans recommend "Block access" (17.0%) much more frequently than the models. This suggests a fundamental difference in approach: humans lean towards physically preventing an interaction, while models have a higher propensity to suggest observation or instruction, which may be a less reliable safety guarantee in a real-world scenario.

Furthermore, the models largely fail to replicate culturally specific mitigation strategies identified by human annotators. The most striking example is the "Move object elsewhere" category in the Japanese data (Figure 23). While humans recommended this action 17.8% of the time, making it their second-most-common strategy, no VLM came close to this frequency. This gap highlights a key limitation: while VLMs can identify the globally dominant safety action, they struggle to capture subtle, context-dependent, or culturally inflected strategies like rearranging an environment for safety.

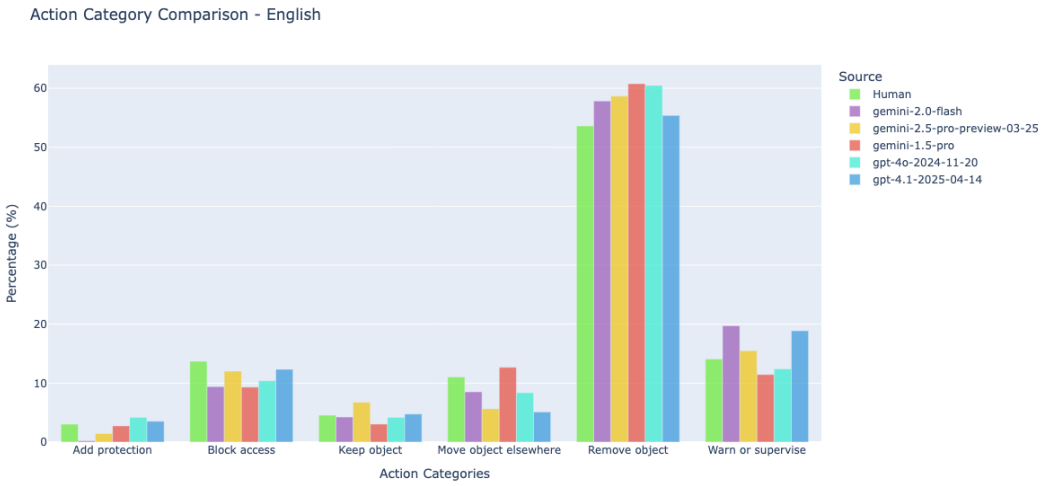

Figure 20. Comparison of action category recommendation percentages between human annotators and multiple Vision-Language Models (gemini-2.0-flash, gemini-2.5-pro-preview-03-25, gemini-1.5-pro, gpt-4o-2024-11-20, and gpt-4.1-2025-04-14) for the English annotations on all images.

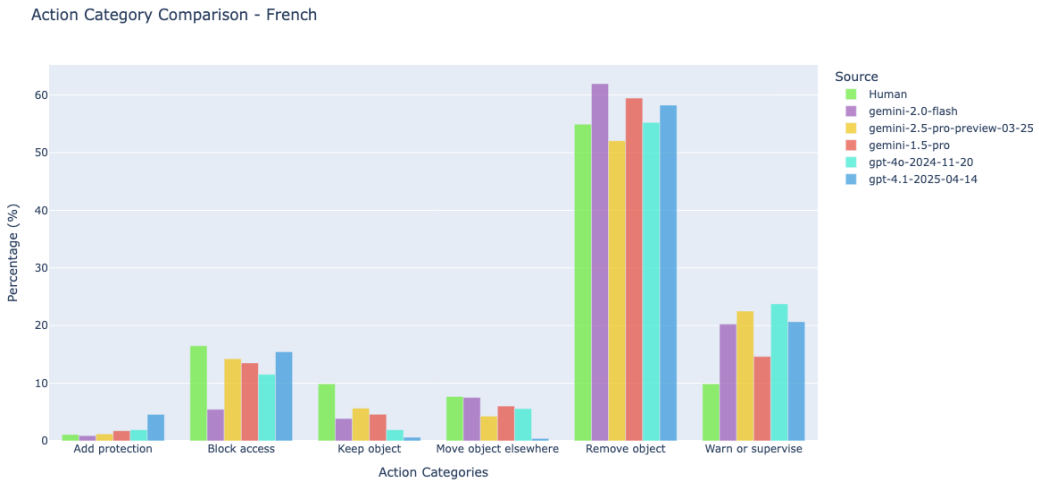

Figure 21. Comparison of action category recommendation percentages between human annotators and multiple Vision-Language Models (gemini-2.0-flash, gemini-2.5-pro-preview-03-25, gemini-1.5-pro, gpt-4o-2024-11-20, and gpt-4.1-2025-04-14) for the French annotations on all images.

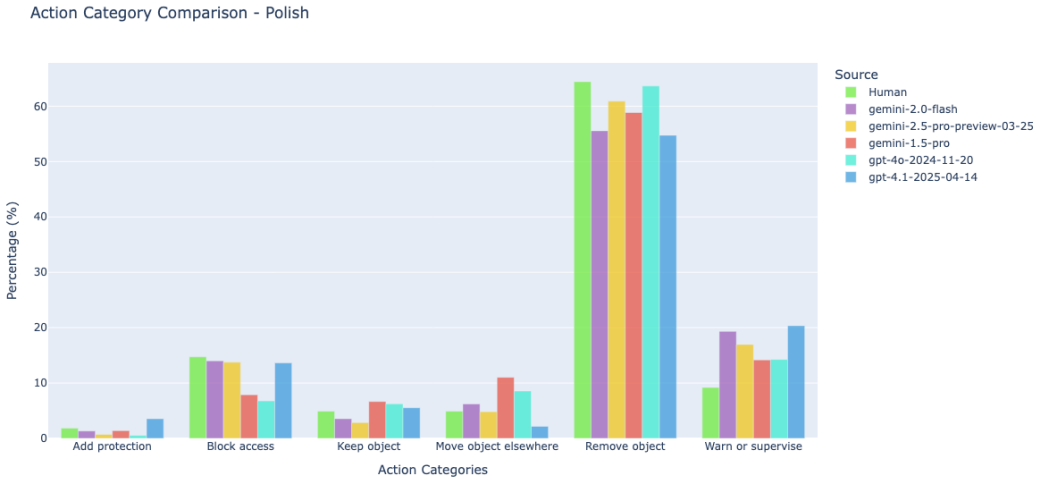

Figure 22. Comparison of action category recommendation percentages between human annotators and multiple Vision-Language Models (gemini-2.0-flash, gemini-2.5-pro-preview-03-25, gemini-1.5-pro, gpt-4o-2024-11-20, and gpt-4.1-2025-04-14) for the Polish annotations on all images.

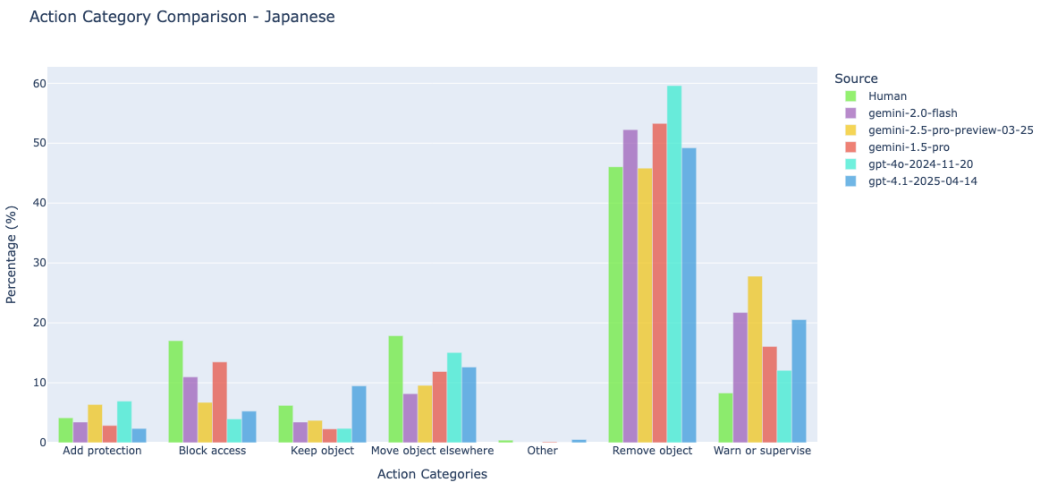

Figure 23. Comparison of action category recommendation percentages between human annotators and multiple Vision-Language Models (gemini-2.0-flash, gemini-2.5-pro-preview-03-25, gemini-1.5-pro, gpt-4o-2024-11-20, and gpt-4.1-2025-04-14) for the Japanese annotations on all images.

In summary, human-VLM concordance is highest for the first-most-common action of removing a hazard. However, systematic disagreements arise in secondary strategies, with models over-recommending supervision and under-representing physical actions like blocking access or repositioning objects. This indicates that while VLMs can mimic the most obvious human response, their ability to generate a balanced and context-aware set of safety recommendations remains underdeveloped.

8. Discussion

The comparative analysis presented in this study highlights important insights into the capabilities and limitations of current Vision-Language Models (VLMs) in assessing safety-critical environments for children. Consistent across languages and annotators, both humans and VLMs reliably identified prominent hazards such as sharp tools and electrical equipment, demonstrating robust object-level perception. However, notable discrepancies arose in detecting context-dependent and culturally nuanced dangers. Humans frequently acknowledged broader risks, such as clutter or accessibility, whereas VLMs predominantly recognized discrete objects ('toy stroller', 'toy wheels'), occasionally neglecting the holistic context of a scenario.

Multilingual consistency was generally strong among European languages (English, French, Polish), but lower semantic alignment scores with Japanese suggest potential biases or cultural nuances affecting hazard perception and action recommendation. While VLMs consistently recommended direct physical interventions like hazard removal, similar to human annotators, significant divergence emerged regarding secondary actions. Human annotators preferred physically preventive measures such as blocking access or rearranging environments, whereas VLMs often proposed generic supervisory recommendations.

These findings imply significant practical considerations for deploying VLM-based safety systems. While current models adequately identify obvious hazards, they lack sufficient contextual reasoning, cultural sensitivity, and situational adaptability to comprehensively safeguard environments. Therefore, human oversight and culturally informed adaptation remain essential components for any practical, real-world implementation involving vulnerable populations like children.

9. Conclusions

In this study, we compared the capabilities of humans and Vision-Language Models (VLMs) in identifying and mitigating hazards within environments frequently occupied by children, such as homes and schools. Our analysis revealed that both humans and VLMs exhibit performance in recognizing obvious hazards, such as sharp objects and electrical equipment, across multiple linguistic and cultural contexts. Furthermore, we found strong concordance between humans and VLMs in recommending primary safety actions, particularly the direct removal of hazardous items.

However, significant divergences arose regarding secondary safety strategies. While human annotators frequently opted for physically preventive measures such as blocking access or repositioning hazards, VLMs more commonly suggested supervisory actions. Reflecting their basic purpose of giving advice, they tend to be less practical and potentially less reliable in immediate risk mitigation. This suggest that they would need to be tuned towards action-oriented decision making. Additionally, our results indicated a notable linguistic and cultural variations, with European languages generally showing higher semantic consistency than Japanese, suggesting potential cultural or training dataset biases.

Our contributions include the introduction of a comprehensive, multilingual, and culturally diverse dataset comprising authentic hazard scenarios, alongside a robust semantic clustering and action categorization framework. This methodological approach allowed

us to evaluate both human and model performance and clearly delineate their respective strengths and limitations.

10. Future Work and Limitations

10.1. Future Work

Moving forward, several avenues can extend this research. Increasing dataset diversity by incorporating a broader range of environments and less obvious, contextual hazards will enhance generalizability. Expanding the scope to include more languages and cultures would further improve cross-cultural applicability. As models seems not to react to prompted details as child's age or country name, in near future we plan to fine-tune VLMs with human-made examples to observe if they can become more sensitive to the context.

Integrating human-in-the-loop systems and explainability methods could significantly enhance user trust and the practical utility of VLMs, especially for ambiguous or contextually nuanced scenarios. Moreover, exploring longitudinal impacts of different safety interventions could provide insights into balancing child safety with developmental needs. Lastly, transitioning from simulated environments to real-world robotic applications and rigorously testing edge cases and adversarial scenarios will be crucial for assessing model robustness and readiness for practical deployment.

Addressing these future directions will contribute substantially to developing safer, more reliable, and universally applicable AI systems for safeguarding children's environments.

10.2. Limitations

Despite the strengths of this study in designing a multilingual and multimodal benchmark for evaluating safety reasoning in child-centric environments, several limitations must be acknowledged.

10.2.1. Limited Dataset Size and Participant Diversity

The dataset comprises 78 images and a smaller curated test set of only 16 images. While this allows for fine-grained comparative analysis, it restricts the generalizability of findings. Additionally, the number of human annotators per language was limited, which may have introduced demographic or cultural bias in the responses. However, it has to be noted that our main goal is to provide methods for annotation, analysis and evaluation of multilingual resources that can be collected to examined on a large scale.

10.2.2. Restricted Environmental and Cultural Representation

The visual data was collected from only two countries (Japan and Poland), primarily focusing on school and home environments. While diverse in object types and layouts, this scope does not capture the full spectrum of environments children encounter globally. A broader geographic and cultural sampling would strengthen claims about cross-linguistic or universal safety reasoning.

10.2.3. Asymmetry in Annotation Task

A key difference in the experimental design was the constraint placed on human annotators, who were asked to identify a maximum of three potential hazards, while the Vision-Language Models (VLMs) had no such restriction. The three-item limit for humans was a practical choice intended to lessen the cognitive load on participants and encourage them to prioritize what they perceived as the most significant risks. While this approach offers insights into human threat prioritization, it introduces an experimental asymmetry. To mitigate this, we normalized the comparison by selecting the top three most frequent hazards identified by the VLMs for our analysis, simulating a form of prioritization.

Nevertheless, we acknowledge that this difference in task constraints is a limitation that could contribute to the observed divergences in hazard identification strategies, such as humans identifying broader contextual dangers (e.g., "clutter") versus the models' itemized lists. The impact of this factor on the results is an important area for future investigation.

10.2.4. Language Inconsistencies

Although the experimental design controlled for language by prompting models in English, French, Japanese, or Polish, model responses occasionally included terms or full responses in a different language. This code switching behavior complicates both semantic clustering and language-specific comparisons, introducing additional noise that must be accounted for during normalization and analysis.

10.2.5. Standard LLM Challenges: Hallucination and Explainability

As with all large language models, the VLMs used in this study are prone to hallucinations generating incorrect or unverifiable information as well as lacking transparent reasoning mechanisms. While structured prompts and JSON formatting aimed to reduce ambiguity, the models' reasoning processes remain largely opaque. This limits interpretability and poses challenges for deploying such systems in safety-critical contexts.

10.2.6. Computational and Financial Constraints

While open-source alternatives like LLaVA and MiniGPT-4 exist, many require high-end GPU resources to run locally. In our case, access to sufficiently powerful hardware was not feasible, leading us to rely on commercial APIs from Gemini and GPT-4o. This introduces limitations in reproducibility and transparency, as such platforms are subject to usage quotas, undocumented changes, and proprietary restrictions. Reproducing this study at scale thus incurs monetary and computational costs that may be prohibitive for smaller research groups.

11. Ethical Statement

Privacy: This study was conducted with strict adherence to ethical guidelines to ensure the protection of participants' privacy and well-being. To safeguard the identities of individuals, all images used in the study were carefully cropped to avoid displaying faces. This measure was taken to maintain the confidentiality of individuals depicted in the photographs. This was performed because prior to data collection, we did not obtain necessary permissions from schools in Poland and Japan, only informal consent was secured from caregivers after declaring that faces will not be used.

Environmental Impact: Our research incorporates the use of large-scale Vision-Language Models (VLMs), and we acknowledge the ethical considerations inherent in their use. To minimize the environmental footprint, our work leverages pre-trained models and a small dataset, thereby avoiding the significant computational cost of fine-tuning or training a new model from scratch.

Transparency and Reproducibility: We are committed to transparent research. We have clearly documented our data and evaluation metrics. Our analysis code is publicly available to ensure reproducibility at https://github.com/Language-Media-Lab/Cross-Cultural_Child_Safety; accessed on 21 July 2025.

We believe this work offers a positive contribution to the field of AI safety by investigating problems related to agents in environments frequented by children. We are committed to the responsible development and deployment of AI technologies.

Author Contributions: Conceptualization, D.D.A. and R.R.; methodology, D.D.A. and R.R.; software, D.D.A.; validation, D.D.A. and R.R.; formal analysis, D.D.A.; investigation, D.D.A.; resources, D.D.A. and R.R.; data curation, D.D.A.; writing original draft preparation, D.D.A.; writing review and editing, D.D.A. and R.R.; visualization, D.D.A.; supervision, R.R.; project administration, R.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by JST CREST, grant number JPMJCR20D2.

Data Availability Statement: The analysis code supporting this study's findings is openly available in a GitHub repository at https://github.com/Language-Media-Lab/Cross-Cultural_Child_Safety; accessed on 21 July 2025. The dataset of images analyzed during the study is not publicly available due to privacy restrictions but is available from the corresponding author on reasonable request.

Acknowledgments: The authors would like to thank Toko Elementary School in Sapporo and Marcin Rejewski Elementary School in Biale Blota for their cooperation.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

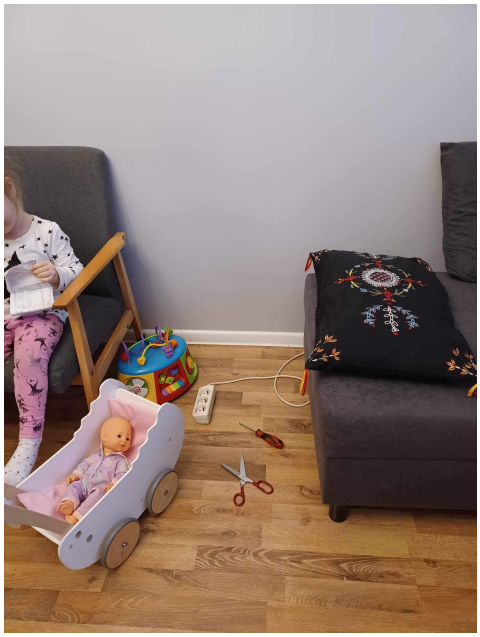

Figure A1. Image ID 11 from the dataset.

Figure A2. Two images (ID7 and ID13) from the dataset. Some pictures in the dataset were have been cropped to protect the children's privacy. For this reason, images are of different sizes.

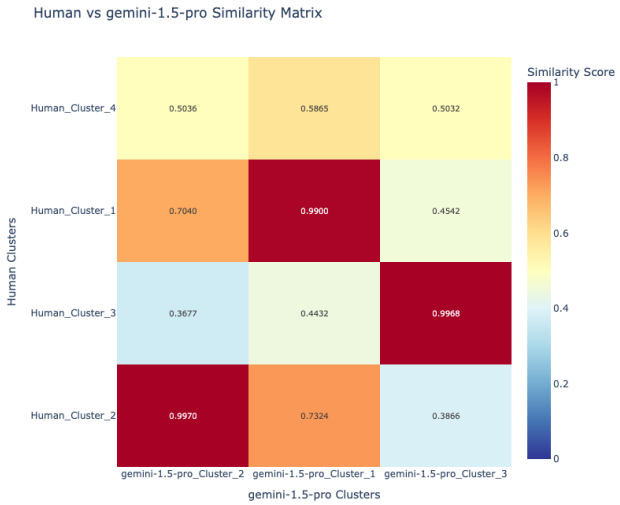

Figure A3. Cosine similarity heatmap comparing hazardous object clusters identified by human annotators and gemini-1.5-pro for in Polish for the image ID 11.

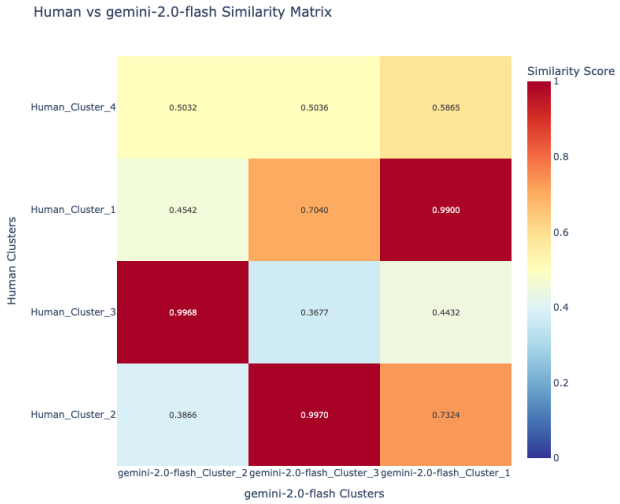

Figure A4. Cosine similarity heatmap comparing hazardous object clusters identified by human annotators and gemini-2.0-flash for in Polish for the image ID 11.

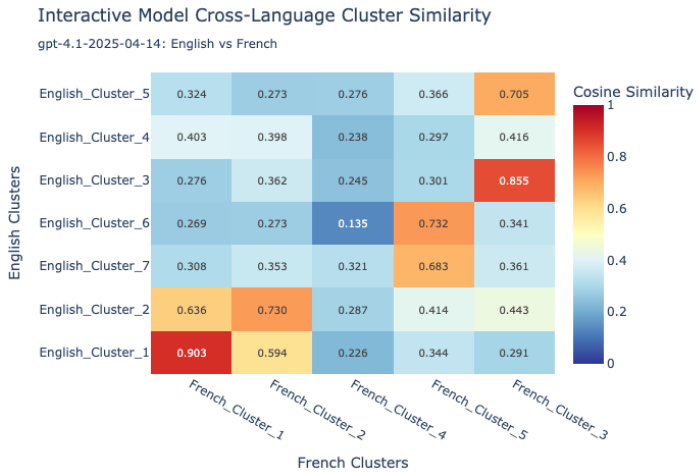

Figure A5. Cosine similarity heatmap between hazardous object clusters identified by gpt-4.1-2025-04-14 in English and French for the image ID 11.

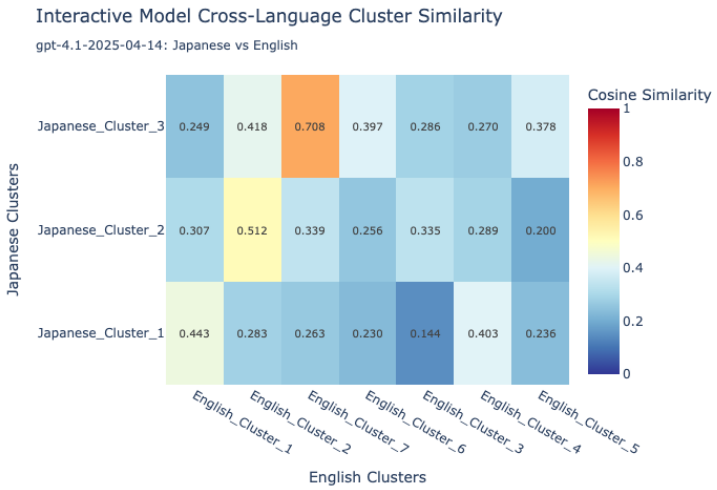

Figure A6. Cosine similarity heatmap between hazardous object clusters identified by gpt-4.1-2025-04-14 in Japanese and English for the image ID 11.

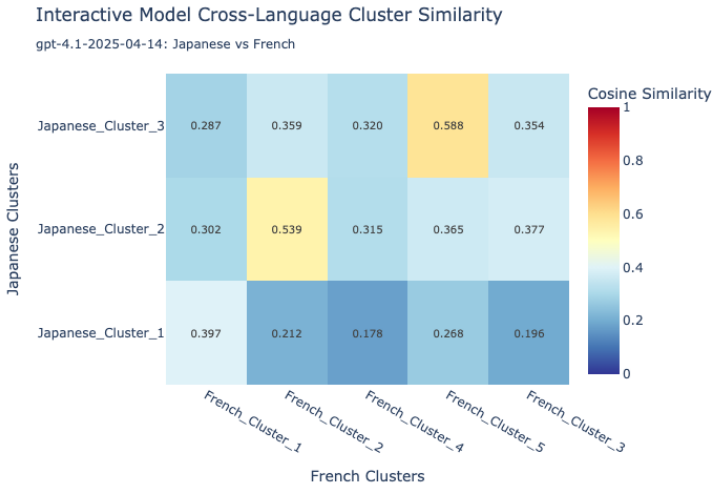

Figure A7. Cosine similarity heatmap between hazardous object clusters identified by gpt-4.1-2025-04-14 in Japanese and French for the image ID 11.

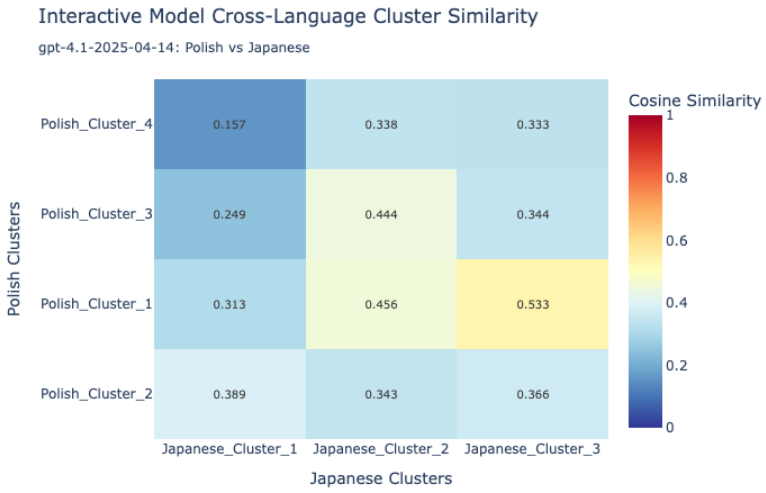

Figure A8. Cosine similarity heatmap between hazardous object clusters identified by gpt-4.1-2025-04-14 in Polish and Japanese for the image ID 11.

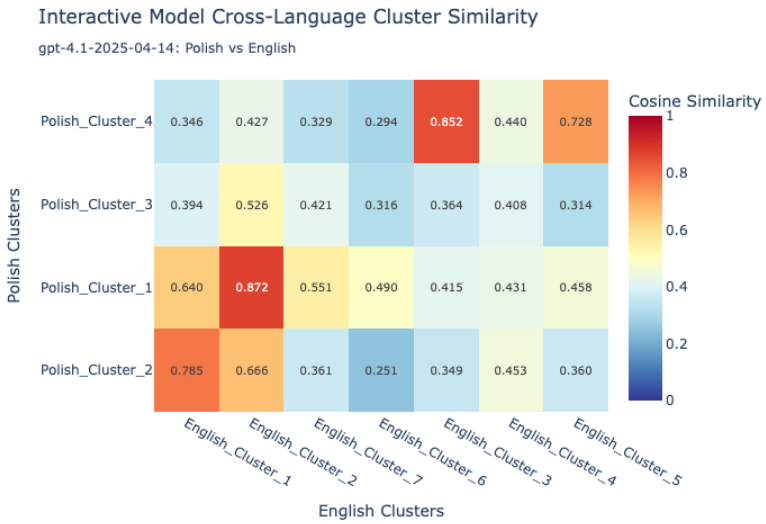

Figure A9. Cosine similarity heatmap between hazardous object clusters identified by gpt-4.1-2025-04-14 in Polish and English for the image ID 11.

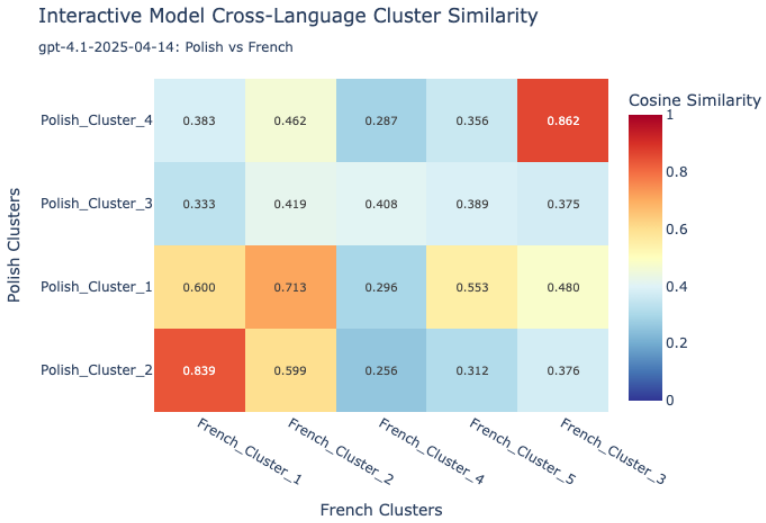

Figure A10. Cosine similarity heatmap between hazardous object clusters identified by gpt-4.1-2025-04-14 in Polish and French annotators for the image ID 11.

Table A1. Semantic similarity (cosine) between human annotators and Gemini-1.5-Pro model for each language across all images. 'Sim pairs count' indicates the number of highly similar cluster pairs (cosine > 0.8).

Lang	Min Sim	Max Sim	Mean Sim	Sim Pairs Count
En	0.14	0.99	0.47	25
Fr	0.15	1.0	0.46	13
Jp	0.01	0.94	0.38	15
Pl	0.12	1.0	0.50	21

Appendix B. Questionnaire and Data Details

This section is about transcriptions of the form we used.

Instructions

Aim of this survey is to compare human and AI capabilities to identify potential hazards in different environments. We will present you with 16 images of different locations containing various objects. Your task is to identify up to three potentially hazardous items in each image, and to assess their level of dangerousness based on the information provided with the images. If no object seems relevant, you can select 'None' and move on to the next image. After that write in one or two sentences why it is dangerous and what would you do in this situation to minimize potential risks to children's safety. Completing this form may take between 15 and 17 min, depending on your response speed and attention to detail. Thank you for your valuable contribution. Please start now by clicking the 'Next' button. First, please provide us with some basic information about you. Note that all your answers will remain confidential and will only be used for research purposes.

Section 1:

1. Age:
   - ○ 18 to 24
   - ○ 5 to 34
   - ○ 35 to 44
   - ○ 45 to 54
   - ○ 55 to 64
   - ○ 65 and over

2. Sex:
   - ○ Male
   - ○ Female
   - ○ Prefer not to say
   - ○ Other: _____

3. Do you have any children? :
   - ○ Yes
   - ○ No

4. Geographical Region : _____

Section 2 (The following questions have been asked for each picture):

1. If nothing seems dangerous, please check this box and move on to the next image.
   - ○ None

2. First Object: _____

3. First Item's hazard level:
   Very low ○ ○ ○ ○ ○ Very high

4. Second Object: _____

5. Second object's hazard level:
   Very low ○ ○ ○ ○ ○ Very high

6. Third Object: _____

7. Third object's hazard level:
   Very low ○ ○ ○ ○ ○ Very high

8. What would you do in this situation (in one or two sentences): _____

Table A2. Participant's Demographics.

Attributes	Values
Total respondents	16
Age range	
18 to 24	18.8%
25 to 34	25.0%
35 to 44	18.8%
45 to 54	45.54%
55 to 64	6.3%
65 or over	-
Sex	
Male	62.5%
Female	31.3%
Genderfluid	6.3%
Prefer not to say	-
Have kids	
Yes	68.8%
No	31.3%
Country origin	
Japan	37.5%
China	6.3%
Cameroon	12.5%
Philippines	6.3%
United States	12.5%
Vanuatu	6.3%
Poland	12.5%
Belgium	6.3%

Appendix C. Glossary of Non-English Cluster Terms

Table A3. French cluster items and their English translations.

French Term	English Translation
courant	current (often refers to electricity)
rallonge électrique	extension cord / power strip
prise électrique	electrical outlet / socket
tournevi, tournevis	screwdriver
câble électrique	electric cable / electrical cord
cordon électrique	electric cord
ciseaux	scissors

Table A4. Japanese cluster items and their English translations.

Japanese Term	English Translation
電源タップ (dengen tappu)	power strip
テーブルタップ (te–buru tappu)	power strip
電気コード (denki ko–do)	electric cord
コード (ko–do)	cord
はさみ (hasami)	scissors

Table A5. Polish cluster items and their English translations.

Polish Term	English Translation
przewód elektryczny	electric wire / electrical cord
kabel elektryczny	electric cable
przedłużacz elektryczny	extension cord / power strip
kable	cables
gniazdek	socket / electrical outlet
prąd	current (electricity)
śrubokręt, srubokret	screwdriver
nożyczki, nozyczka	scissors

Appendix D. Dataset Datasheet

Motivation

- Purpose: The dataset was created to evaluate the ability of Vision-Language Models (VLMs) and large language models (LLMs) to recognize, assess, and mitigate physical dangers to children in school and home environments. This dataset addresses the gap in real-world testing of AI in safety-critical environments, specifically focusing on children's well-being.
- Creators: The dataset was created by a team of researchers from *anonymized* focused on AI safety applications, collaborating with caregivers, school staff, and parents from Japan and Poland.
- Funding: N/A.
- Additional Comments: The dataset aims to provide a foundation for studying discrepancies between AI and human judgment in assessing potential dangers in environments children frequently inhabit.

Composition

- Instances: The dataset consists of 78 images taken in homes and schools, depicting objects that may pose a risk to children.
- Total Instances: 78 images (43 from Japan, 35 from Poland).
- Sampling: The dataset is a curated collection, not a random sample. Images were selected to cover a diverse range of potential hazards from culturally distinct regions.
- Data Features: Images depicting physical environments where children usually play, and objects with potential hazards (e.g., sharp tools, small toys, electrical cables) inside these environments. Annotations were collected from human participants and models.
- Labels: Each image includes annotations of dangerous objects (up to 3 per image), danger levels (on a 5-point scale), and suggested actions to mitigate risk.
- Relationships: Each image relates to both its annotations and the contextual information provided to annotators, including location and environment type (home or school).

- Splits: 16 images were selected for human-model comparison experiments in order the have an fair comparison, 4 pictures representing school and homes from both countries Japan and Poland.
- Errors/Noise: There is some inherent subjectivity in the annotations, particularly in the assessment of danger levels and actions, due to variability in how participants perceived the risks.
- External Resources: The dataset is self-contained.
- Sensitive Data: Children's faces were cropped or removed from the dataset to protect privacy. No other sensitive or offensive content is present.

Collection Process

- Acquisition: Images were collected by parents, caregivers, and school staff using smartphones. Participants were given freedom in selecting potentially dangerous objects in the scenes.
- Mechanisms: Data was manually curated. No automated data collection tools were used.
- Sampling Strategy: The dataset was intentionally curated to include a variety of potential hazards, selected by caregivers.
- Participants: Images were taken by adults, including parents and school staff. No compensation was provided.
- Timeframe: Images were collected at the time of the study (March 2024).
- Ethical Review: Informal consent was obtained from caregivers, with the assurance that children's faces would not be visible in the images.
- Data Source: Images were taken directly by individuals involved in the study.
- Consent: Informal consent was obtained, though no formal mechanism for revocation was provided.
- Impact Analysis: No formal impact analysis was mentioned, though privacy and ethical considerations were addressed through the removal of identifying features from the images.

Preprocessing/Cleaning/Labeling

- Preprocessing: Images containing children were cropped to exclude identifiable faces. The annotations were standardized using clustering algorithms to ensure consistent labeling of objects.
- Raw Data: Raw images are preserved but cropped for privacy.
- Software: For textual part of the data (hazardous object names), clustering and annotation standardization were conducted using tools like SpaCy and scikit-learn.

Uses

- Tasks: The dataset is used for evaluating the performance of Vision-Language Models in detecting hazards and recommending actions in child-related environments.
- Future Tasks: The dataset could be expanded to assess cultural differences in risk perception and to test other vision-based AI systems in safety-critical scenarios. Broad variety of objects and action descriptions will be used for testing new and comparing existing dedicated similarity measures.
- Potential Risks: The subjective nature of danger assessment and the limited geographical scope and the limited number of annotations of the dataset could affect generalization to other populations or environments.
- Limitations: The dataset depicts only a very tiny fragment of reality (two homes and two schools in Japan and Poland, respectively). It is mainly meant for prototyping and testing.

Distribution

- Third-Party Distribution: To avoid data contamination, the data will be shared with other researchers after receiving a pledge.
- Distribution Method: Password-protected zip file will be provided on the authors' laboratory site. Link will be shared with everyone who agreed to the pledge.
- Licensing: No licensing or intellectual property terms is specified.
- Export Controls: No export controls or regulatory restrictions were noted.

Maintenance

- Support: Authors will maintain the dataset.
- Errata: No errata exists.
- Updates: Authors will update the dataset.
- Retention Limits: No retention limits are specified.
- Versioning: Version control will be provided by authors if the dataset will be updated.
- Contribution Mechanism: To avoid contamination issues, other researchers cannot contribute to the images subset. However, descriptions of objects and actions in languages other than English are welcomed.

References

1. Shigemi, S. ASIMO and humanoid robot research at Honda. In *Humanoid Robotics: A Reference*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 1–36.
2. Pandey, A.K.; Gelin, R. A mass-produced sociable humanoid robot: Pepper: The first machine of its kind. *IEEE Robot. Autom. Mag.* **2018**, *25*, 40–48. [CrossRef]
3. Röttger, P.; Attanasio, G.; Friedrich, F.; Goldzycher, J.; Parrish, A.; Bhardwaj, R.; Di Bonaventura, C.; Eng, R.; Geagea, G.E.K.; Goswami, S.; et al. MSTS: A Multimodal Safety Test Suite for Vision-Language Models. *arXiv* **2025**, arXiv:2501.10057.
4. Qi, Z.; Fang, Y.; Zhang, M.; Sun, Z.; Wu, T.; Liu, Z.; Lin, D.; Wang, J.; Zhao, H. Gemini vs GPT-4V: A Preliminary Comparison and Combination of Vision-Language Models Through Qualitative Cases. *arXiv* **2023**, arXiv:2312.15011.
5. Chiang, C.H.; Lee, H.y. Can large language models be an alternative to human evaluations? *arXiv* **2023**, arXiv:2305.01937. [CrossRef]
6. Evans, I. Safer children, healthier lives: Reducing the burden of serious accidents to children. *Paediatr. Child Health* **2022**, *32*, 302–306. [CrossRef]
7. Lee, D.; Jang, J.; Jeong, J.; Yu, H. Are Vision-Language Models Safe in the Wild? A Meme-Based Benchmark Study. *arXiv* **2025**, arXiv:2505.15389.
8. Bommasani, R.; Hudson, D.A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M.S.; Bohg, J.; Bosselut, A.; Brunskill, E.; et al. On the opportunities and risks of foundation models. *arXiv* **2021**, arXiv:2108.07258. [CrossRef]
9. Ping, L.; Gu, Y.; Feng, L. Measuring the Visual Hallucination in ChatGPT on Visually Deceptive Images. In OSF Preprints, Posted May 2024. [CrossRef]
10. Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. GPT-4 technical report. *arXiv* **2023**, arXiv:2303.08774. [CrossRef]
11. Liu, H.; Li, C.; Wu, Q.; Lee, Y.J. Visual instruction tuning. In Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS 2023), New Orleans, LA, USA, 10–16 December 2023; pp. 34892–34916.
12. Zhang, M.; Li, X.; Du, Y.; Rao, Z.; Chen, S.; Wang, W.; Chen, X.; Huang, D.; Wang, S. A Multimodal Large Language Model for Forestry in pest and disease recognition Research Square, Posted 18 October 2023. rs-3444472 v1. [CrossRef]
13. Zhou, L.; Palangi, H.; Zhang, L.; Hu, H.; Corso, J.; Gao, J. Unified vision-language pre-training for image captioning and vqa. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 13041–13049.
14. Sharma, A.; Gupta, A.; Bilalpur, M. Argumentative Stance Prediction: An Exploratory Study on Multimodality and Few-Shot Learning. *arXiv* **2023**, arXiv:2310.07093. [CrossRef]
15. Pak, H.L.; Zhang, Y. A Smart Child Safety System for Enhanced Pool Supervision using Computer Vision and Mobile App Integration. In Proceedings of the CS & IT Conference Proceedings, CS & IT Conference Proceedings, Dubai, United Arab Emirates, 28–29 December 2024; Volume 14.

16. Lee, Y.; Kim, K.; Park, K.; Jung, I.; Jang, S.; Lee, S.; Lee, Y.J.; Hwang, S.J. HoliSafe: Holistic Safety Benchmarking and Modeling with Safety Meta Token for Vision-Language Model. *arXiv* **2025**, arXiv:2506.04704.
17. Na, Y.; Jeong, S.; Lee, Y. SIA: Enhancing Safety via Intent Awareness for Vision-Language Models. *arXiv* **2025**, arXiv:2507.16856. [CrossRef]
18. Mullen, J.F., Jr.; Goyal, P.; Piramuthu, R.; Johnston, M.; Manocha, D.; Ghanadan, R. "Don't Forget to Put the Milk Back!" Dataset for Enabling Embodied Agents to Detect Anomalous Situations. *IEEE Robot. Autom. Lett.* **2024**, *9*, 9087–9094. [CrossRef]
19. Rodriguez-Juan, J.; Ortiz-Perez, D.; Garcia-Rodriguez, J.; Tomás, D.; Nalepa, G.J. Integrating advanced vision-language models for context recognition in risks assessment. *Neurocomputing* **2025**, *618*, 129131. [CrossRef]
20. Nitta, T.; Masui, F.; Ptaszynski, M.; Kimura, Y.; Rzepka, R.; Araki, K. Detecting Cyberbullying Entries on Informal School Websites Based on Category Relevance Maximization. In Proceedings of the Sixth International Joint Conference on Natural Language Processing, Nagoya, Japan, 14–18 October 2013; pp. 579–586.
21. Rani, D.D. Protecting Children from Online Grooming in India's Increasingly Digital Post-Covid-19 Landscape: Leveraging Technological Solutions and AI-Powered Tools. *Int. J. Innov. Res. Comput. Sci. Technol.* **2024**, *12*, 38–44. [CrossRef]
22. Verma, K.; Milosevic, T.; Cortis, K.; Davis, B. Benchmarking Language Models for Cyberbullying Identification and Classification from Social-media Texts. In Proceedings of the First Workshop on Language Technology and Resources for a Fair, Inclusive, and Safe Society within the 13th Language Resources and Evaluation Conference, Marseille, France, 25 June 2022; pp. 26–31.
23. Song, Z.; Ouyang, G.; Fang, M.; Na, H.; Shi, Z.; Chen, Z.; Fu, Y.; Zhang, Z.; Jiang, S.; Fang, M.; et al. Hazards in Daily Life? Enabling Robots to Proactively Detect and Resolve Anomalies. *arXiv* **2024**, arXiv:2411.00781.
24. Singhal, M.; Gupta, L.; Hirani, K. A comprehensive analysis and review of artificial intelligence in anaesthesia. *Cureus* **2023**, *15*, e45038 . [CrossRef]
25. Zhou, K.; Liu, C.; Zhao, X.; Compalas, A.; Song, D.; Wang, X.E. Multimodal situational safety. *arXiv* **2024**, arXiv:2410.06172. [CrossRef]
26. Li, X.; Zhou, H.; Wang, R.; Zhou, T.; Cheng, M.; Hsieh, C.J. Mossbench: Is your multimodal language model oversensitive to safe queries? *arXiv* **2024**, arXiv:2406.17806. [CrossRef]
27. Ying, Z.; Liu, A.; Liang, S.; Huang, L.; Guo, J.; Zhou, W.; Liu, X.; Tao, D. SafeBench: A Safety Evaluation Framework for Multimodal Large Language Models. *arXiv* **2024**, arXiv:2410.18927. [CrossRef]
28. Gupta, P.; Krishnan, A.; Nanda, N.; Eswar, A.; Agrawal, D.; Gohil, P.; Goel, P. ViDAS: Vision-based Danger Assessment and Scoring. In Proceedings of the Fifteenth Indian Conference on Computer Vision Graphics and Image Processing, Bengaluru, Karnataka, India, 13–15 December 2024; pp. 1–9.
29. Shneiderman, B. Bridging the gap between ethics and practice: Guidelines for reliable, safe, and trustworthy human-centered AI systems. *ACM Trans. Interact. Intell. Syst. (TiiS)* **2020**, *10*, 1–31. [CrossRef]
30. Nguyen, T.; Wallingford, M.; Santy, S.; Ma, W.C.; Oh, S.; Schmidt, L.; Koh, P.W.W.; Krishna, R. Multilingual diversity improves vision-language representations. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 91430–91459.
31. Shorten, C.; Pierse, C.; Smith, T.B.; Cardenas, E.; Sharma, A.; Trengrove, J.; van Luijt, B. Structuredrag: JSON response formatting with large language models. *arXiv* **2024**, arXiv:2408.11061.
32. Feng, F.; Yang, Y.; Cer, D.; Arivazhagan, N.; Wang, W. Language-agnostic BERT sentence embedding. *arXiv* **2020**, arXiv:2007.01852.