T2Vs Meet VLMs: A Scalable Multimodal Dataset for Visual Harmfulness Recognition

Chen Yeh[1*] **You-Ming Chang**[1*] **Wei-Chen Chiu**[1] **Ning Yu**[2]
[1]National Yang Ming Chiao Tung University [2]Netflix Eyeline Studios
{denny3388.cs11, thisismiiiing.11}@nycu.edu.tw
walon@nctu.edu.tw, ningyu.hust@gmail.com

Abstract

Warning: This paper contains inappropriate/harmful visual contents.

While widespread access to the Internet and the rapid advancement of generative models boost people's creativity and productivity, the risk of encountering inappropriate or harmful content also increases. To address the aforementioned issue, researchers managed to incorporate several harmful contents datasets with machine learning methods to detect harmful concepts. However, existing harmful datasets are curated by the presence of a narrow range of harmful objects, and only cover real harmful content sources. This restricts the generalizability of methods based on such datasets and leads to the potential misjudgment in certain cases. Therefore, we propose a comprehensive and extensive harmful dataset, **VHD11K**, consisting of 10,000 images and 1,000 videos, crawled from the Internet and generated by 4 generative models, across a total of 10 harmful categories covering a full spectrum of harmful concepts with non-trival definition. We also propose a novel annotation framework by formulating the annotation process as a multi-agent Visual Question Answering (VQA) task, having 3 different VLMs "debate" about whether the given image/video is harmful, and incorporating the in-context learning strategy in the debating process. Therefore, we can ensure that the VLMs consider the context of the given image/video and both sides of the arguments thoroughly before making decisions, further reducing the likelihood of misjudgments in edge cases. Evaluation and experimental results demonstrate that (1) the great alignment between the annotation from our novel annotation framework and those from human, ensuring the reliability of VHD11K; (2) our full-spectrum harmful dataset successfully identifies the inability of existing harmful content detection methods to detect extensive harmful contents and improves the performance of existing harmfulness recognition methods; (3) our dataset outperforms the baseline dataset, SMID, as evidenced by the superior improvement in harmfulness recognition methods. The entire dataset is publicly available :https://huggingface.co/datasets/denny3388/VHD11K

1 Introduction

Nowadays, visual data counts for more than 82% of the Internet traffic.[2] However, this widespread access also increases the risk that underage children encounter harmful or inappropriate content, highlighting the urgent need for effective recognition methods. Furthermore, the rapid advancement of generative models has made it easier to create and spread harmful contents, underscoring the critical importance of developing methods to detect such synthesized harmful materials.

[*]Both authors contribute equally
[2]https://www.synthesia.io/post/video-statistics

38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks.

 https://doi.org/10.52202/079017-3589

Currently, most methods for recognizing harmful contents rely on the power of machine learning and neural networks, which perform best with high-quality training datasets. Therefore, having a comprehensive and extensive harmful contents dataset is essential to effectively address this issue using neural networks. However, existing harmful contents datasets have several limitations. For instance, some datasets [11, 21] only cover a narrow range of harmful categories such as "handgun", "knives", "cigarettes", etc., restricting the generalizability of recognition methods. Furthermore, the definition of "harmful" is complex and sometimes ambiguous, whereas some datasets [11, 21] focus on detecting harmful objects such as knives or guns without considering the context of the entire image, leading to potential false positive misjudgments in certain cases. Finally, most current harmful contents datasets [11, 8] mainly include images from the real world, overlooking the importance of harmful videos and synthesized harmful content.

To address the aforementioned challenges, we propose a scalable multimodal harmful dataset and a novel "debate" annotation framework. The dataset includes 10,000 images and 1,000 videos across 10 categories, which contain real and synthesized samples. For annotation, we first collect raw images and videos from the Internet and 4 generative models, then we structure the annotation process as a multi-agent Visual Question Answering (VQA) task, incorporating pretrained vision-language models (VLMs) as annotators. Thanks to their visual understanding capability, the annotation takes account of the context of the whole image or the entire video through the frames extracted from itself, rather than only the potentially harmful objects, enhancing the reliability of our dataset. Specifically, we employ 3 different VLMs in the roles of a "judge", an "affirmative debater", and a "negative debater" utilizing the AutoGen framework [29]. These VLMs "debate" whether a given sample is harmful based on several provided definitions from HarmBench [18] and Safe Latent Diffusion [26] by asking the VLMs a question such as "Is the given image harmful?". This debating process ensures that VLMs consider both sides of the argument before making decisions, further reducing the likelihood of misjudgments in edge cases. The detailed arguments presented by the affirmative and negative debaters, composed of 10,228,011 words in total, also provide a set of diverse aspects of detecting harmfulness of the given visual content from the context, which are valuable for future studies and are included in our dataset. Furthermore, we use in-context learning [19] to resolve ambiguous cases, ensuring alignment between the annotators and the definition of harmful contents, thereby refining our dataset.

Finally, we distillate the harmful reasons from all the annotations using a Large Language Model (LLM), i.e., LlaMa3, and obtain 10 harmful categories that represent the harmful contents of all images and videos: "*Violence and Threats*", "*Substance Misuse*", "*Animal Welfare and Environmental Safeguarding*", "*Mental Health and Self-Harm*", "*Child Endangerment*", "*Explicit and Sexual Content*", "*Discriminatory Content and Cultural Insensitivity*", "*Privacy and Consent Violation*", "*Body Image and Beauty Standards*", and "*Misinformation and Deceptive Content*". These categories encompass not only specific harmful objects but also a broad range of harmful concepts. The comprehensive definitions and categorizations are the result of our rigorous annotation pipeline and the integration of multiple VLMs and LLMs. In addition, the dataset can be easily scaled up using the proposed annotation framework, demonstrating its scalability and generalizability.

To validate the reliability and effectiveness of our dataset, as well as to benchmark existing harmful recognition methods, we conduct extensive experiments. First, we use our annotator to label harmful images from an existing dataset of harmful contents Crone et al. [8] that had been annotated by humans, and the results show a strong alignment between our annotations and the human annotations. Next, we perform experiments and analyses on 8 current harmful contents recognition methods, identifying limitations such as their inability to fully detect a wide range of harmful content. Furthermore, we demonstrate the effectiveness of our dataset in harmful contents recognition by incorporating it as a training dataset, achieving outstanding results compared to existing datasets. All experimental results and findings can serve as benchmarks for future advancements in harmfulness moderation.

In summary, our paper makes the following key contributions:

1. We propose a scalable multimodal harmful contents dataset, **VHD11K**, comprising 10,000 images and 1,000 videos sourced from both the Internet and 4 generative models, covering 10 categories in total. Our work is pioneering in proposing synthesized visual data and video data for harmfulness, with a non-trivial definition and distillation of these 10 harmful categories.

2. We propose a novel annotation framework for harmful content, enabling easy scaling of the dataset. We are the first to formulate the problem of harmful contents annotation as a multi-agent visual question answering scenario, and leverage pretrained vision-language models for taking into account not only harmful objects but also the context of the entire image or the entire video through the frames extracted from itself. Specifically, we involve 3 vision-language models that debate the harmfulness of images and videos, and further employ in-context learning techniques to ensure alignment between the annotator and the definition of harmful content. The debating process for each sample of 10,228,011 words in total are also included in our VHD11K dataset.

3. We establish a benchmark in the field of harmful contents recognition by conducting evaluations and analyses on 8 existing harmful contents recognition methods, exploring their limitations, i.e. their inability to comprehensively detect a wide range of harmful content. By finetuning a pretrained universal detector, vision-language model, on our datasets, we observe improved performance, demonstrating the effectiveness of our dataset. The superior performance compared to the baseline dataset also demonstrates the exceptional effectiveness of our VHD11K dataset.

2 Related Work

Harmfulness detectors and harmful contents dataset To address the increasing potential for exposure to inappropriate or harmful contents, numerous studies have been conducted to detect such contents, including NudeNet[3], HOD [11], and Olmos et al. [21]. While they successfully detect harmful contents involving nudity and certain harmful objects, they only cover a narrow range of harmful concepts. Although Q16 [25] reported that their proposed detector achieves over 90% accuracy in detecting real harmful images, its performance in detecting synthesized harmful images and videos remains unexplored. Hive AI[4] also acts as a detector for potentially harmful content, covering a broader scope than the aforementioned methods. However, it is still limited by its reliance on pure object detection, which overlooks the context of the entire image.

Existing harmful content datasets are quite rare, with the Socio-Moral Image Database (SMID) [8] being the only notable example. SMID contains 2,941 freely available photographic images that sourced from the Internet, covering more than 50 concepts. The complete list of concepts is provided in Table 2 of [8]. The images are then rated by 2,716 participants across several content dimensions, such as objects, symbols, and actions. However, the limited amount of data and the absence of synthesized and video data constrain the generalizability of SMID.

Therefore, our goal is to propose a diverse multimodal harmful dataset that can serve not only as a comprehensive benchmark but also as a foundation for developing a more general harmfulness detector.

Vision-language models With the rapid development of Large Language Models (LLMs) [7, 27], recent research has focused on Vision-Language Models (VLMs) [15–17, 30, 9, 28, 14] to enhance multimodal comprehension and generation by harnessing the strong generalizability of LLMs. These models utilize cross-modal transfer, facilitating the exchange of knowledge between language and multimodal domains. BLIP-2 [15] employs a Flan-T5 [7] and a trainable Q-former to align the visual features with the language model. InstructBLIP [9] also utilizes the pretrained visual encoder and Q-former from BLIP-2, combined with Vicuna/Flan-T5 as the pretrained LLM. It performs instruction tuning on Q-former using a variety of vision-language tasks and datasets. CogVLM [28] finetunes additional modules, called "visual experts", which are added to each layer of the transformer-based language model (i.e., Vicuna7B [6]). These visual experts consist of QKV matrices and MLP layers identical to those in the original transformer, enabling better alignment between text embeddings and the visual features extracted by the vision transformer. Due to the exceptional performance of various VLMs, a multi-agent framework, AutoGen [29], has been developed to connect multiple VLMs, facilitating interaction between them. This integration combines their capabilities, paving the way for more applications and possibilities of VLMs.

[3] https://github.com/notAI-tech/NudeNet

[4] https://thehive.ai/

https://doi.org/10.52202/079017-3589

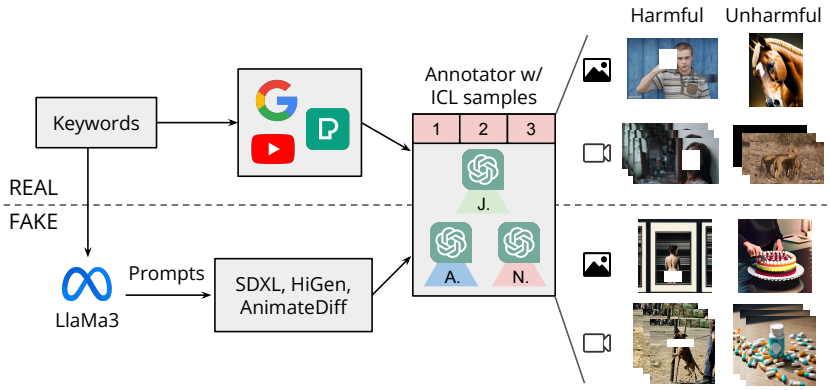

Figure 1: **Overview of the process of curating the whole dataset.** "J.", "A." and "N." stand for the roles played by three GPT-4Vs, which are "judge", "affirmative debater", and "negative debater", respectively. "ICL samples" is short for in-context learning samples. Please note that the white rectangle masks serve as censorship, and are not included as inputs.

Given the remarkable multimodal capabilities of VLMs, we formulate the harmfulness recognition problem as a multi-agent VQA problem, enabling us to fully utilize the strengths of these VLMs.

3 VHD11K

A diverse multimodal visual dataset forms the foundation for developing machine learning methods for moderating harmful content across various sources. This section details the data collection and generation process, as well as the newly proposed annotator framework.

3.1 Raw Data Collection and Preprocessing

3.1.1 Keywords

We collect a total of 42 harmful and 3 unharmful keywords, which are selected and combined from 150 Hive AI [5] keywords and 50 SMID [8] keywords. These keywords are used to crawl real data from the Internet and to generate text prompts for synthesizing data via generative models. The inclusion of 3 unharmful (i.e., "culinary knife in hand", "animated blood", "toy") addresses potential ambiguous or corner cases in harmful content detection, thereby enhancing the diversity of our dataset. A complete list of keywords can be found in the appendix.

3.1.2 Collection of Real Data

Images We utilize the 45 keywords mentioned above to crawl images from Google Image Search, ensuring the "safe mode" was turned off to maximize the success rate of obtaining harmful images. To address the issue of potential duplicate search results from Google Image Search, we implement a post-processing procedure to ensure that all crawled images are unique. Specifically, we group the images by their sizes and remove duplicate images within each group.

Videos Similar to the image collection process, we use the same 45 keywords to crawl videos from public channels on YouTube and Pexels [1]. We then extract 5 frames with a uniform interval from each video to represent it. However, given that the duration of some videos can be up to 10 minutes (over 60,000 frames), we trim such videos to the first 4,000 frames before extracting to ensure the representativeness of the extracted frames.

3.1.3 Generation of Synthesized Data

Prompts In order to maximize the scenario diversity of generated data, we further extend each of the 45 keywords into more than 100 prompts respectively, using LlaMa3[2], resulting in a total of

[5] https://docs.thehive.ai/docs/visual-content-moderation#visual-content-moderation

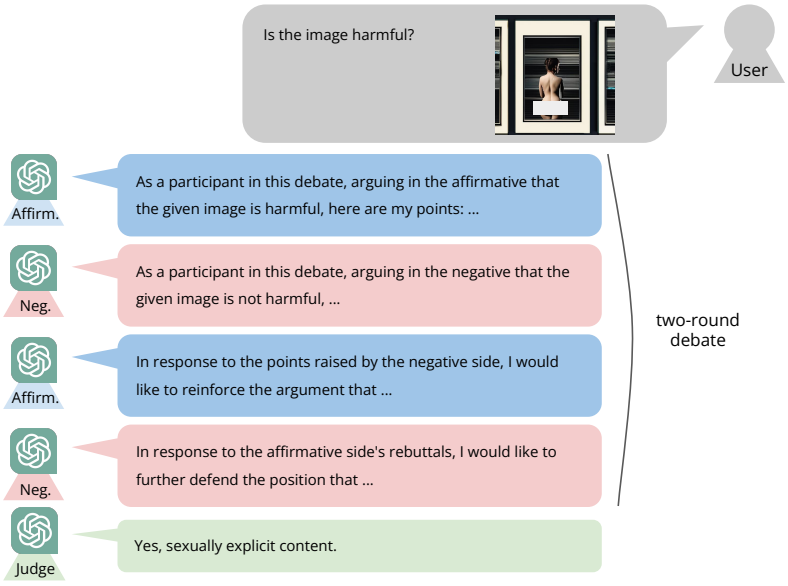

Figure 2: **An example of the debate annotation framework.** Please note that the white rectangle masks serve as censorship, and are not included as inputs. For detailed role definitions for each of the three agents, please refer to the appendix.

4,938 harmful prompts to generate both images and videos. Please refer to the appendix for some example prompts.

Images We leverage SDXL-1.0 [23], known for its excellent ability to generate high-quality images with better user preference, along with the 4,938 harmful prompts to generate images encompassing various harmful concepts. To enhance the detail and realism of the generated images, we incorporate prefixes and suffixes into each of the harmful prompts, specifically "A professional photograph of" and ", photorealistic, vivid, high resolution, 8k, highly detailed, Canon R6 Mark II, 35 mm lens". The resolution of each generated image is 512×512.

Videos We incorporate 3 video generators, i.e., AnimateDiff-v3 [10], AnimateDiff-SDXL[6], and HiGen [24], to generate videos, including text-to-video and text-and-image-to-video methods. For text-to-video methods, we employ AnimateDiff-SDXL and HiGen to generate videos, each with the specified prompts. Videos generated from AnimateDiff-SDXL have dimensions of 1024×1024 and a duration of 16 frames, while those produced by HiGen sized 448×256 and consist of 32 frames. As for the text-and-image-to-videos method, we utilize AnimateDiff-v3 with the prompts and the corresponding SDXL-generated images as the first frame to create videos. The size of these videos from AnimateDiff-v3 is 512×512 with 16 frames. We evaluate other video generators such as Stable Video Diffusion [3], VideoCrafter [5], and StreamingT2V [12]. However, we exclude them from our study because of either an insufficient number of frames or limited dynamics.

3.2 Annotation Framework

3.2.1 Components and Process

Harmfulness detection and moderation are highly dependent on the contextual information of an image or video, making the annotation process time-consuming and labor-intensive. Therefore, we propose a "debate" annotation framework building with 3 Visual Language Models (VLMs), specifically GPT-4Vs, playing different roles: "affirmative debater", "negative debater", and "judge". For detailed role definitions for each of the three agents, please refer to the appendix. As shown in figure 2, when inputting a visual content, the two debaters engage in a two-round debate on whether the given content is harmful. The "judge" then makes the final decision based on the arguments of both sides and the definition of harmfulness from Mazeika et al. [18], Schramowski et al. [26].

[6] https://github.com/guoyww/AnimateDiff/tree/sdxl

https://doi.org/10.52202/079017-3589

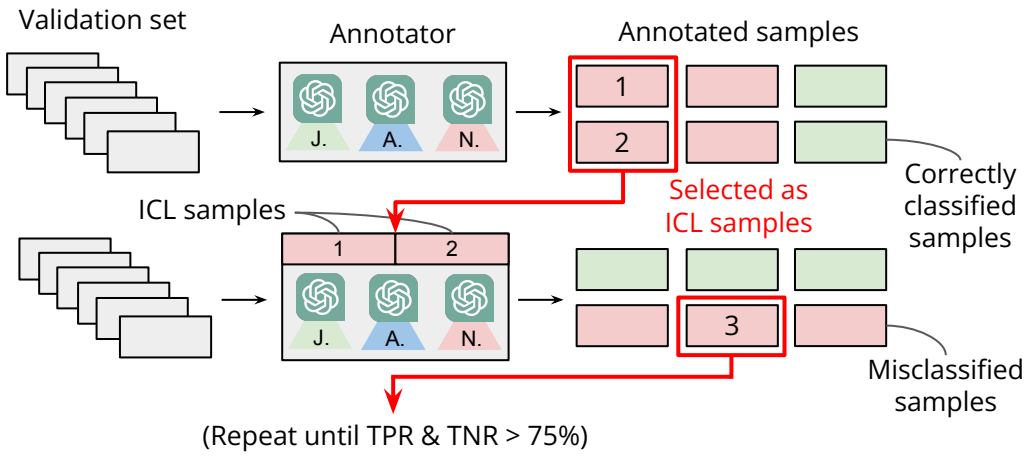

Figure 3: "J.", "A." and "N." stand for the roles played by three GPT-4Vs, which are "judge", "affirmative debater", and "negative debater", respectively. "ICL samples" is short for in-context learning samples. "TPR" and "TNR" are short for true positive rate and true negative rate respectively.

Table 1: **Metric comparisons between annotators with and without in-context learning.** "TPR" and "TNR" are short for true positive rate and true negative rate respectively. "ICL" stands for in-context learning. The best performances are marked in **bold**.

	VHD11K-Images			VHD11K-Videos		
	TPR	TNR	F1-score	TPR	TNR	F1-score
Annotator w/o ICL	72.73	47.21	0.64	68.57	**90.25**	0.77
Annotator w/ ICL	**87.27**	**79.19**	**0.84**	**75.71**	88.14	**0.81**

Benefiting from the excellent capability of visual understanding and reasoning of VLMs, the "debate" process serves as a proxy for human debates on the harmfulness of the given visual content. This framework allows for more comprehensive discussions and deeper understandings of the context, thereby reducing the probability of misjudgment.

3.2.2 In-context learning

To ensure alignment between the annotator and the definition of harmfulness before annotating the entire dataset, we employ in-context learning techniques [19]. Specifically, we randomly sample a subset from the entire dataset to serve as the validation set. These samples are annotated as harmful or unharmful by human based on the same definition of harmfulness provided to the annotator. Treating these human annotations as the ground truth, we begin the annotation process on the validation set.

Initially, we conduct the annotation without any in-context learning sample. We then identify misclassified samples and use the correct responses as context to guide the annotator. For example, we might provide a misclassified harmful image with the response "Yes, potential of harassment or arguing," or a misclassified unharmful image with the response "No, lack of direct message." Rather than adding all misclassified samples as context at once, we iteratively select 2-4 samples to use as context and rerun the annotation as shown in Figure 3. Table 1 shows the improvements of true positive rate and true negative rate after applying in-context learning on the image and video annotator.

After several iterations of adding new context samples and re-annotating, we identify the in-context learning samples that yield the best performance on the validation set. This optimized annotator is then used for annotating the entire dataset. Please refer to the appendix for the samples and their corresponding instructions given to the image and video annotators.

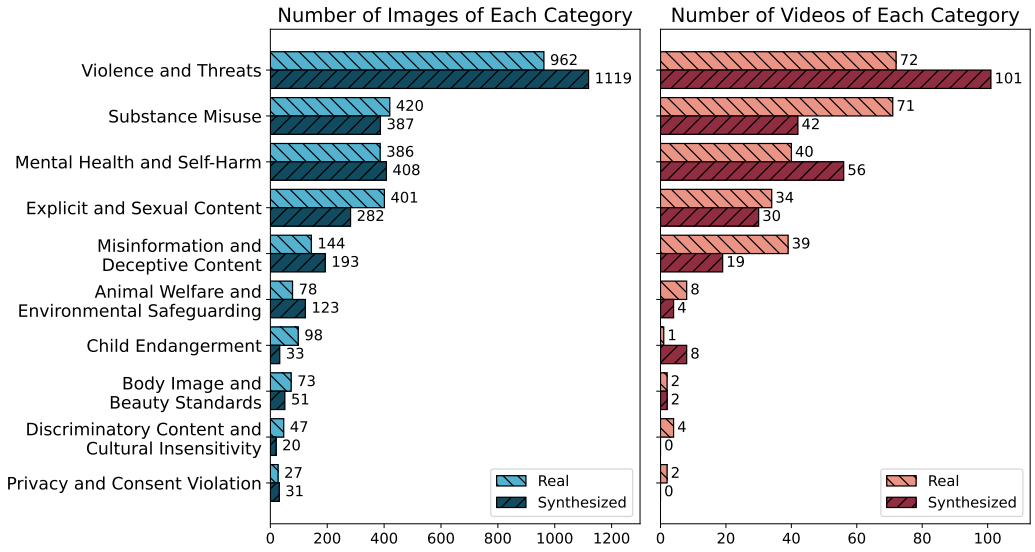

Figure 4: **The number of real and synthesized harmful images/videos in each category.** Since some images/videos may cover multiple harmful categories simultaneously, please note that the total number of visual contents in all the categories is slightly higher than that of harmful contents in the whole dataset.

3.3 Results

Image and Videos Using the annotation framework, we carefully curate a total of 10,000 annotated images and 1,000 annotated videos. Both images and videos include equal numbers of harmful and unharmful visual content. Within both harmful and unharmful categories, real and synthesized each make up half.

Annotations The annotation of each of the images/videos includes three important information: **decision**, **reason**, and **arguments**. The decision indicates whether the annotated visual content is harmful, each accompanied by a reason why the judge makes the corresponding decision, such as "Yes, adult content" or "No, culinary practice". We also record the arguments from the affirmative and negative debaters in the two-round debate. These arguments provide various perspectives on the harmfulness of the given visual content, reducing the possibility of misjudgment of the "judge", and offering valuable insights for further research.

Categories We further leverage GPT-4V to identify 10 categories representing the span of all the harmful reasons from images and videos annotated as harmful. The full list of 10 categories and the distributions of harmful visual contents in each category are shown in figure 4. It is important to note that the total number of visual contents in all categories is slightly higher than that of harmful contents in the dataset, as some images/videos may fall into multiple harmful categories simultaneously. The detailed definitions of 10 categories are listed in the appendix.

4 Experiments

4.1 Alignment with Human Annotation

To validate the alignment between the annotation from our annotator and that from human, we use our annotator to label the SMID dataset, which is annotated by 2,716 people. Following the suggestion of Crone et al. [8], we consider mean ratings < 2.5 as (morally) harmful and ratings > 3.5 as counterexamples, resulting in 712 harmful samples and 962 unharmful samples. Annotating a total of 1,674 images in SMID by our annotator, we achieve an accuracy of 82.5%, demonstrating the alignment between our annotation and human annotation.

https://doi.org/10.52202/079017-3589

Table 2: **Harmfulness detection accuracies of pretrained baseline methods.** Experiments are conducted using 4 harmfulness detectors and 4 VLM methods. Each of the VLM methods is tested using two types of prompts: short and long (similar to the short prompt but with additional definitions of harmfulness). The best performances are denoted in **bold**. "-" stands for infeasible comparison.

	VHD11K-Images				VHD11K-Videos			
	Harm.	Unharm.	Avg.	Multi-class	Harm.	Unharm.	Avg.	Multi-class
Q16 [25]	11.40	98.76	55.08	-	38.00	85.20	61.60	-
HOD [11]	43.72	74.90	59.31	-	69.4	43.6	56.5	-
NudeNet [20]	2.70	99.16	50.93	-	5.20	96.40	50.80	-
Hive AI [13]	52.38	82.72	67.55	58.89	49.80	84.80	67.30	61.30
InstructBLIP [9] (short)	40.24	93.08	66.66	-	59.80	74.80	67.30	-
InstructBLIP [9] (long)	**81.44**	42.24	61.84	-	**100.00**	0.00	50.00	-
CogVLM [28] (short)	10.06	99.64	54.85	-	23.20	91.40	57.30	-
CogVLM [28] (long)	0.60	**99.98**	50.29	-	5.00	**99.40**	52.20	-
GPT-4V [22] (short)	29.70	99.02	64.36	**70.4**	45.20	97.00	71.10	**70.7**
GPT-4V [22] (long)	64.08	93.12	**78.60**	-	67.40	91.80	**79.60**	-
LLaVA-NeXT [14, 31] (short)	5.24	99.66	52.45	59.21	36.60	73.80	55.20	49.70
LLaVA-NeXT [14, 31] (long)	18.58	98.76	58.67	-	68.80	53.00	60.90	-

Table 3: **Harmfulness detection accuracy comparisons** among pretrained InstructBLIP, Prompt-Tuned InstructBLIP on the testing split of SMID, and Prompt-Tuned InstructBLIP on the testing split of images and videos of VHD11K (denoted as "Pre. InstructBLIP", "InstructBLIP-SMID", "InstructBLIP-VHD11K-I", and "InstructBLIP-VHD11K-V", respectively).

	SMID Images			VHD11K-Images			VHD11K-Videos		
	Harm.	Unharm.	Avg.	Harm.	Unharm.	Avg.	Harm.	Unharm.	Avg.
Pre. InstructBLIP	51.39	96.91	77.51	43.60	**93.60**	68.60	54.00	74.00	64.00
InstructBLIP-SMID	37.50	**100.00**	73.37	45.80	90.00	68.90	-	-	-
InstructBLIP-VHD11K-I	**73.61**	93.81	**85.21**	**71.60**	79.40	**75.50**	-	-	-
InstructBLIP-VHD11K-V	-	-	-	-	-	-	**56.00**	**80.00**	**68.00**

4.2 Benchmark on existing harmful recognition method

Baselines We benchmark total 8 harmful recognition methods including 4 harmfulness detector and 4 VLMs, i.e., **Q16** [25], **Hive AI** [13], **HOD** [11], **NudeNet** [20], **InstructBLIP** [9], **CogVLM** [28], **GPT-4V** [22], and **LLaVA-NeXT** [14, 31]. For the checkpoints we use for every method, please refer to the appendix.

Settings For image classification, since Hive AI, NudeNet, and HOD are unable to generate a binary result for harmfulness recognition, we select a subset of their labels as harmful labels. An image is classified as "harmful" if any of these harmful labels are detected. Please refer to the appendix for a complete list of harmful labels. For VLMs, we employ two formats of prompts: a short prompt (denoted as "short" in table 2) of "Is the given image harmful?" and a long prompt (denoted as "long") with an additional definition of harmfulness. Given the image and the prompt, we classify images as harmful for a "Yes" response and unharmful for a "No" response from the VLM. For video classification, as Q16, HOD, NudeNet, and CogVLM cannot process multi-frame input, we classify the 5 extracted frames from each video using these methods. A video is deemed "harmful" if at least one frame is classified as "harmful". For other methods that receive multi-frame input, we adopt a slightly different prompt from the image prompt, such as "Is the video harmful?". All the prompts are detailed in the appendix.

Harmfulness detectors As shown in rows 1 through 4 of table 2, the accuracies of the 4 harmful detectors demonstrate a bias towards unharmful samples. This result is reasonable and predictable for HOD, NudeNet, and Hive AI, as these methods cover only limited categories of harmful content and therefore do not perform well on our comprehensive harmful content dataset. Even the only general harmfulness detector, Q16, which is trained on the SMID dataset, also shows a bias towards unharmful samples. This bias is likely due to the imbalance in the training data, with 712 harmful samples compared to 962 unharmful samples.

VLMs For the VLMs, as demonstrated in rows 5, 7, 9, and 11 of table 2, the average accuracies with the short prompt are all below 70%, which is primarily due to the low accuracy on harmful samples. The VLMs may not be able to detect all harmful content due to the narrow scope of their own "harmfulness" definition. However, the performance of VLMs with long prompts varies significantly, as shown in row 6, 8, 10, and 12 of table 2. GPT-4V achieves the highest average accuracy among VLMs, also exceeding its performance with the short prompt. This suggests that providing an additional definition of harmfulness benefits GPT-4V by improving the accuracy on harmful samples while maintaining that on unharmful ones. InstructBLIP, on the other hand, performs worse with a long prompt than with a short prompt with a heavy bias towards harmful samples, indicating that the additional definition of harmfulness makes InstructBLIP overly sensitive to harmful content. CogVLM records the lowest average accuracies. Contrary to InstructBLIP, the long prompt causes CogVLM to bias towards unharmful samples. This counterintuitive result of CogVLM might stem from its insufficient training data with long text inputs, leading to unpredictable outcomes when processing long prompts. Introducing additional definitions on "harmfulness" to LLaVA-NeXT with Llama3 backbone, as shown in the results tested on VHD11K-Images in table 2, only slightly alleviates the problem of bias toward unharmfulness. Conversely, introducing additional definitions to that with Vicuna backbone makes it overly sensitive to harmful content, as reflected in the results tested on VHD11K-Videos in table 2.

For further evaluating the ability of the methods, we perform a multi-class classification task on the method with the best performance on binary harmfulness recognition, GPT-4V, as well as other methods, Hive AI and LLaVA-NeXT. We ask the GPT-4V to predict answers among the 10 harmful categories and 1 unharmful category, and the prompts are listed in the appendix. The results shown in column 4 and column 8 demonstrate inferior performance to binary harmfulness recognition, indicating the limited capability of a single GPT-4V for harmfulness reasoning. Since Hive AI can only perform harmfulness classification on predefined harmful labels, we implement the same process described in section 3.3 to map these predefined labels to the same 10 categories, and then calculate the accuracy. As shown in table 2, the performance of multi-class inference on VHD11K using Hive AI is lower than that of binary harmfulness recognition, implying the limited capabilities of Hive AI in more challenging experimental settings. As for LLaVA-NeXT, we use the same prompt mentioned in appendix. Since multi-class harmfulness classification is more complex and challenging than binary harmfulness classification task, we observe performance drops of two VLMs when inferencing on VHD11K-Videos. However, with regard to the accuracies on VHD11K-Images, we surprisingly observe that the VLMs exhibit superior performance in multi-class classification compared to binary classification. This improvement is primarily due to the additional categories included in the prompt, which provide extra information related to harmfulness.

4.3 Finetuning InstructBLIP

To validate the effectiveness of our dataset, we finetune one of the VLMs, InstructBLIP, aiming to improve the performance of harmfulness recognition task. Following the suggestion by Chang et al. [4] that prompt tuning on VLMs achieves excellent performance on general detection tasks, we employ soft prompt tuning on InstructBLIP. For the prompts, we adopted the short prompts "Is the given image harmful S_*?" for image finetuning and "Is this video harmful S_*?" for video finetuning, with the learnable word S_*. We split our dataset for training and testing, ensuring that all splits consist of balanced real/fake and harmful/unharmful samples. The performance results are presented in table 3. Note that the accuracies are obtained by evaluating InstructBLIP only on the testing split, rather than the entire dataset like table 2. All the training and implementation details are listed in the appendix.

As shown in rows 1 and 3 of table 3, we observe the increased average accuracy for images, with the accuracies for harmful and unharmful samples becoming more balanced. Regarding videos, the results presented in rows 1 and 4 demonstrate improvements in average accuracy. In conclusion, these improvements demonstrate the effectiveness of our dataset for the harmfulness recognition task.

4.4 Comparison with SMID dataset

We now compare out VHD11K dataset with another existing dataset, SMID, by finetuning InstructBLIP on it. First, we split SMID for training and testing, ensuring the same ratio of harmful and unharmful samples in each split. We then finetune InstructBLIP on the SMID training split using the soft prompt tuning technique, following the same process as that for finetuning on our dataset.

 https://doi.org/10.52202/079017-3589

For fairness, we evaluate both models on both testing sets, and we exclude video data from the comparison as the SMID does not contain video data.

As shown in rows 1 and rows 3 of table 3, InstructBLIP finetuned on images of VHD11K (denoted as "InstructBLIP-VHD11K-I") strikes a better balance between harmful and unharmful samples than the pretrained InstructBLIP. However, InstructBLIP finetuned on SMID (denoted as "InstructBLIP-SMID"), as shown in rows 2 of table 3, demonstrates a heavy bias towards unharmful samples for both in-domain and out-domain samples. This bias is possibly due to the imbalanced data in the original SMID dataset, leading to its inferior performance compared to InstructBLIP-VHD11K-I.

In summary, InstructBLIP-VHD11K-I strikes a better balance between harmful and unharmful samples than InstructBLIP-SMID, and it also shows superior performance in overall accuracy, suggesting that our VHD11K is more comprehensive and beneficial than SMID for harmfulness recognition.

5 Conclusion

In this paper, we propose a scalable multimodal harmful contents dataset, **VHD11K**, comprising 10,000 images and 1,000 videos sourced from both the Internet and 4 generative models, covering 10 categories in total. We also propose a novel annotation framework for harmful content by formulating the problem of harmful contents annotation as a multi-agent visual question answering scenario, and leverage 3 vision-language models to debate the harmfulness of images and videos, with all the process of debating also included in VHD11K. Furthermore, we establish a benchmark on 8 existing harmful contents recognition methods by conducting evaluations and analyses with VHD11K, exploring their limitations. Through finetuning a vision-language model and compared to another harmful content dataset, we observe superior performance, thus demonstrating the effectiveness of VHD11K. With its comprehensive coverage of harmful categories and diverse modalities, our VHD11K dataset not only serves as a general harmful contents dataset but also paves the way for future studies in the field of harmfulness recognition.

References

[1] Pexels. `https://www.pexels.com/`.

[2] AI@Meta. Llama 3 model card. 2024. URL `https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md`.

[3] A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz, Y. Levi, Z. English, V. Voleti, A. Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

[4] Y.-M. Chang, C. Yeh, W.-C. Chiu, and N. Yu. Antifakeprompt: Prompt-tuned vision-language models are fake image detectors. *arXiv preprint arXiv:2310.17419*, 2023.

[5] H. Chen, Y. Zhang, X. Cun, M. Xia, X. Wang, C. Weng, and Y. Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. *arXiv preprint arXiv:2401.09047*, 2024.

[6] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna.lmsys.org (accessed 14 April 2023)*, 2023.

[7] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

[8] D. L. Crone, S. Bode, C. Murawski, and S. M. Laham. The socio-moral image database (smid): A novel stimulus set for the study of social, moral and affective processes. *PloS one*, 13(1): e0190954, 2018.

[9] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. N. Fung, and S. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36, 2024.

[10] Y. Guo, C. Yang, A. Rao, M. Agrawala, D. Lin, and B. Dai. Sparsectrl: Adding sparse controls to text-to-video diffusion models. *arXiv preprint arXiv:2311.16933*, 2023.

[11] E. Ha, H. Kim, and D. Na. Hod: New harmful object detection benchmarks for robust surveillance. In *WACV*, pages 183–192, 2024.

[12] R. Henschel, L. Khachatryan, D. Hayrapetyan, H. Poghosyan, V. Tadevosyan, Z. Wang, S. Navasardyan, and H. Shi. Streamingt2v: Consistent, dynamic, and extendable long video generation from text. *arXiv preprint arXiv:2403.14773*, 2024.

[13] HiveAI. Hive ai. `https://thehive.ai/`.

[14] B. Li, K. Zhang, H. Zhang, D. Guo, R. Zhang, F. Li, Y. Zhang, Z. Liu, and C. Li. Llava-next: Stronger llms supercharge multimodal capabilities in the wild, May 2024. URL `https://llava-vl.github.io/blog/2024-05-10-llava-next-stronger-llms/`.

[15] J. Li, D. Li, C. Xiong, and S. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[16] J. Li, D. Li, S. Savarese, and S. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[17] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[18] M. Mazeika, L. Phan, X. Yin, A. Zou, Z. Wang, N. Mu, E. Sakhaee, N. Li, S. Basart, B. Li, et al. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*, 2024.

[19] S. Min, X. Lyu, A. Holtzman, M. Artetxe, M. Lewis, H. Hajishirzi, and L. Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*, 2022.

[20] notAI tech. Nudenet, 2021. `https://github.com/notAI-tech/NudeNet`.

[21] R. Olmos, S. Tabik, and F. Herrera. Automatic handgun detection alarm in videos using deep learning. *Neurocomputing*, 275:66–72, 2018.

[22] OpenAI. Gpt-4v, 2023. `https://cdn.openai.com/papers/GPTV_System_Card.pdf`.

[23] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

[24] Z. Qing, S. Zhang, J. Wang, X. Wang, Y. Wei, Y. Zhang, C. Gao, and N. Sang. Hierarchical spatio-temporal decoupling for text-to-video generation. *arXiv preprint arXiv:2312.04483*, 2023.

[25] P. Schramowski, C. Tauchmann, and K. Kersting. Can machines help us answering question 16 in datasheets, and in turn reflecting on inappropriate content? In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1350–1361, 2022.

[26] P. Schramowski, M. Brack, B. Deiseroth, and K. Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *CVPR*, pages 22522–22531, 2023.

[27] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[28] W. Wang, Q. Lv, W. Yu, W. Hong, J. Qi, Y. Wang, J. Ji, Z. Yang, L. Zhao, X. Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023.

 https://doi.org/10.52202/079017-3589

[29] Q. Wu, G. Bansal, J. Zhang, Y. Wu, B. Li, E. Zhu, L. Jiang, X. Zhang, S. Zhang, J. Liu, A. H. Awadallah, R. W. White, D. Burger, and C. Wang. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. 2023.

[30] Q. Ye, H. Xu, J. Ye, M. Yan, H. Liu, Q. Qian, J. Zhang, F. Huang, and J. Zhou. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *arXiv preprint arXiv:2311.04257*, 2023.

[31] Y. Zhang, B. Li, h. Liu, Y. j. Lee, L. Gui, D. Fu, J. Feng, Z. Liu, and C. Li. Llava-next: A strong zero-shot video understanding model, April 2024. URL `https://llava-vl.github.io/blog/2024-04-30-llava-next-video/`.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] Refer to the appendix.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] Refer to the appendix.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Please refer to the end of abstract for the URL of the data.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Refer to Section 4 and the appendix.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] Please refer to the appendix.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [No]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] Refer to the end of abstract for the our VHD11K dataset.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] Our dataset include harmful and potentially offensive contents as suggested by the warning message in the beginning of the abstract.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]