# 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] (*Both authors contribute equally)
[1]National Yang Ming Chiao Tung University, [2]Netflix Eyeline Studios

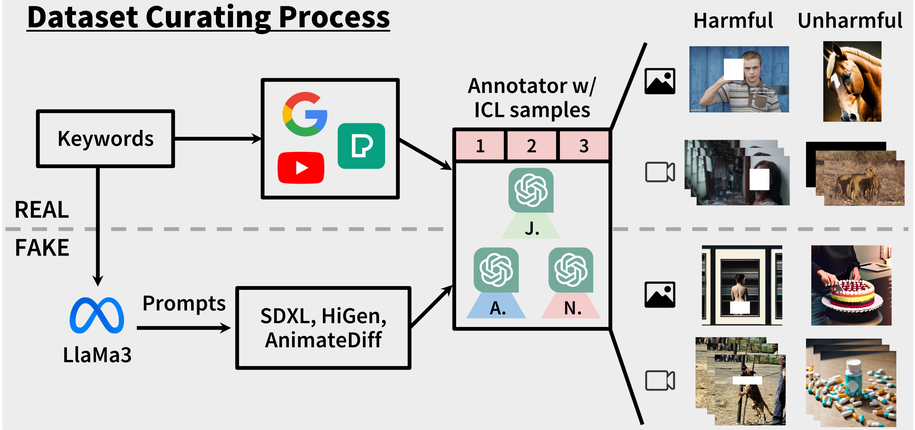
NATIONAL YANG MING CHIAO TUNG UNIVERSITY

EYELINE STUDIOS™ POWERED BY NETFLIX

NEURAL INFORMATION PROCESSING SYSTEMS

Project Page

## Dataset Curating Process



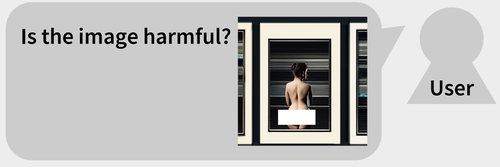

## Motivation: limitations on existing datasets

- Limited scopes of harmful contents (e.g., nudity, gun, knife)
- Comprise only real images, without synthesized images / videos
- Focus on detecting certain harmful objects, without considering the whole visual context.

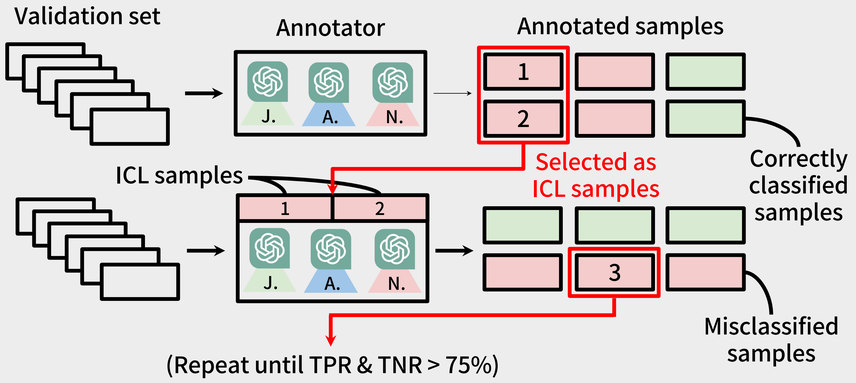
## Contribution

- A scalable multimodal harmful contents dataset, VHD11K
  - 10,000 images and 1,000 videos sourced from the Internet and 4 generative models.
  - Cover in total 10 categories with non-trivial definitions.
- A novel annotation framework for harmful content
  - Reformulate the annotation as multi-agent visual question answering problem.
  - Utilize 3 vision-language models to debate the harmfulness.
- Benchmarking on 8 existing harmfulness detectors with VHD11K
  - Explore their limitations on detecting harmful contents.
  - Demonstrate the effectiveness of VHD11K by improved performances of finetuned methods.

## Annotation Example




## In-Context Learning Samples Selection




## Harmfulness Detection Accuracies

### Pretrained models benchmarking

| | 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 | - | 49.80 | 84.80 | 67.30 | - |
| 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 | **70.40** | 45.20 | 97.00 | 71.10 | **70.70** |
| GPT-4V [22] (long) | 64.08 | 93.12 | **78.60** | - | 67.40 | 91.80 | **79.60** | - |

### Pretrained vs. prompt-tuned models on SMID & VHD11K

| | 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** |

## Distribution of Each Category