Project Page

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



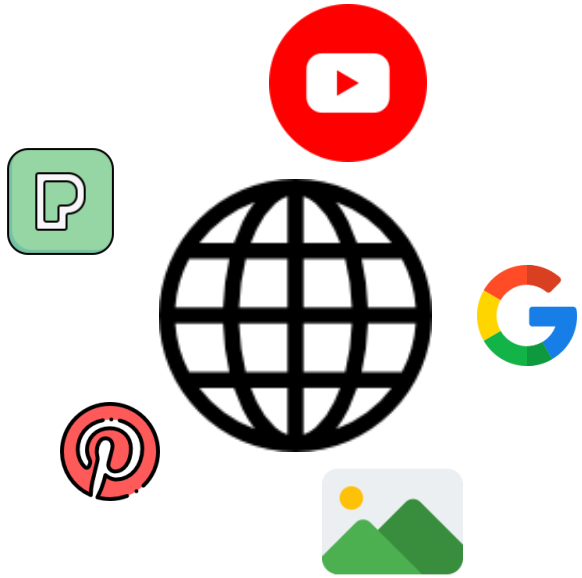

Chen Yeh[1*] You-Ming Chang[1*] Wei-Chen Chiu[1] Ning Yu[2]

[1]National Yang Ming Chiao Tung University, [2]Netflix Eyeline Studios

(*Both authors contribute equally)

Introduction: Background

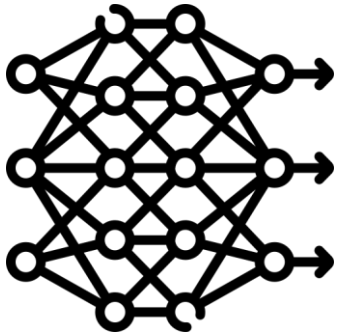
- Visual data (images/videos) counts for more than 82% of the Internet traffic.
- Increases the risk that underage children encounter “harmful” or “inappropriate” contents.



Introduction: Harmful Contents Detection

- Researchers detect harmful content by machine learning methods.
- The performance relies on the quality of Dataset

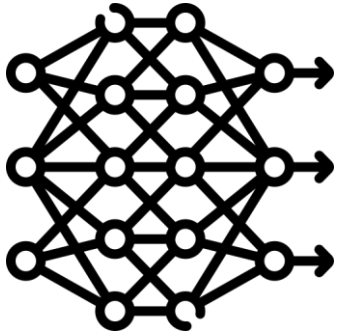
Training



Harmful

Unharmful

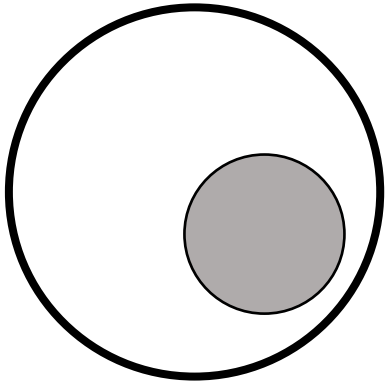
Testing



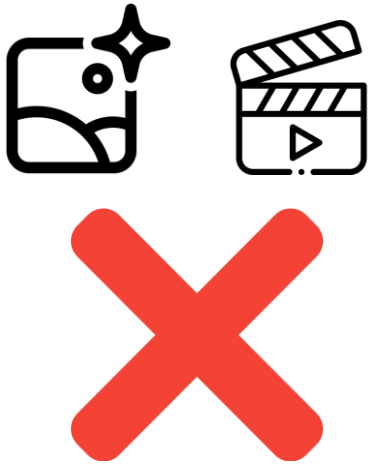
Harmful or Unharmful

Introduction: Issues of Existing Datasets

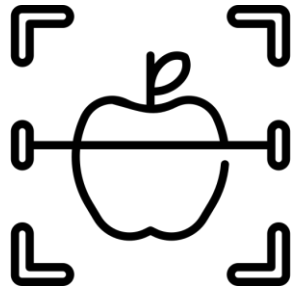
1. Only contains a limited scope of harmful contents (e.g., nudity, gun, knife)
2. Only contain real images, not including synthesized images / videos
3. Only contain pure object detection, ignoring context



Limited scope



synthesized images / videos not included

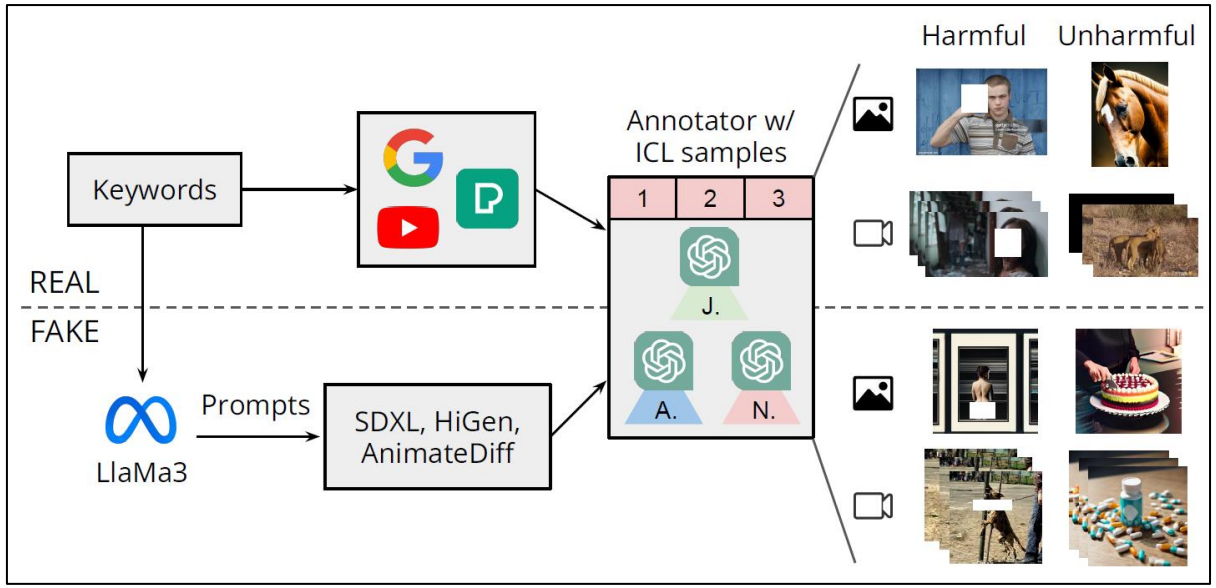


Pure object detection

Introduction: Our VHD11K

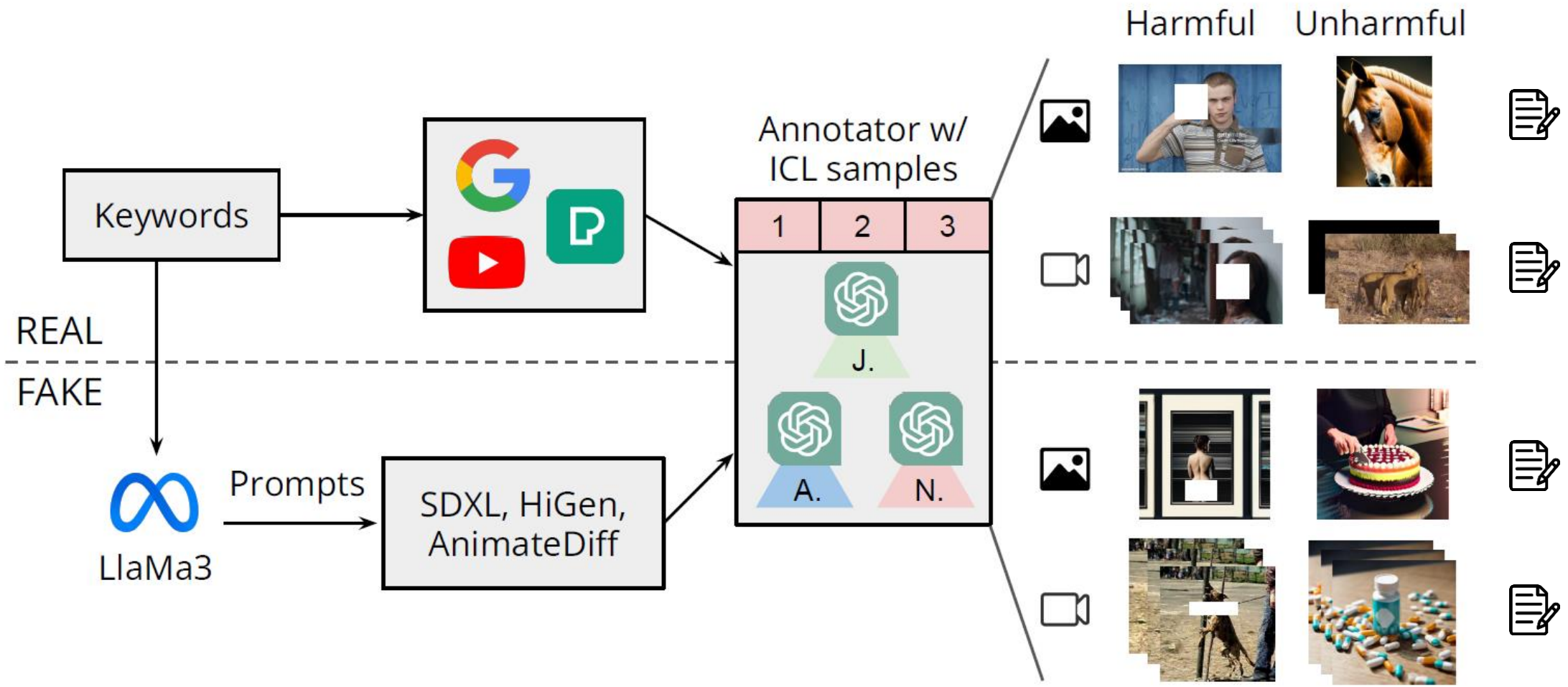
1. Only contains a limited scope of harmful contents (e.g., nudity, gun, knife)
2. Only contain real images, not including synthesized images / videos
3. Only contain pure object detection, ignoring context

Visual Harmful Dataset 11K
(VHD11K)

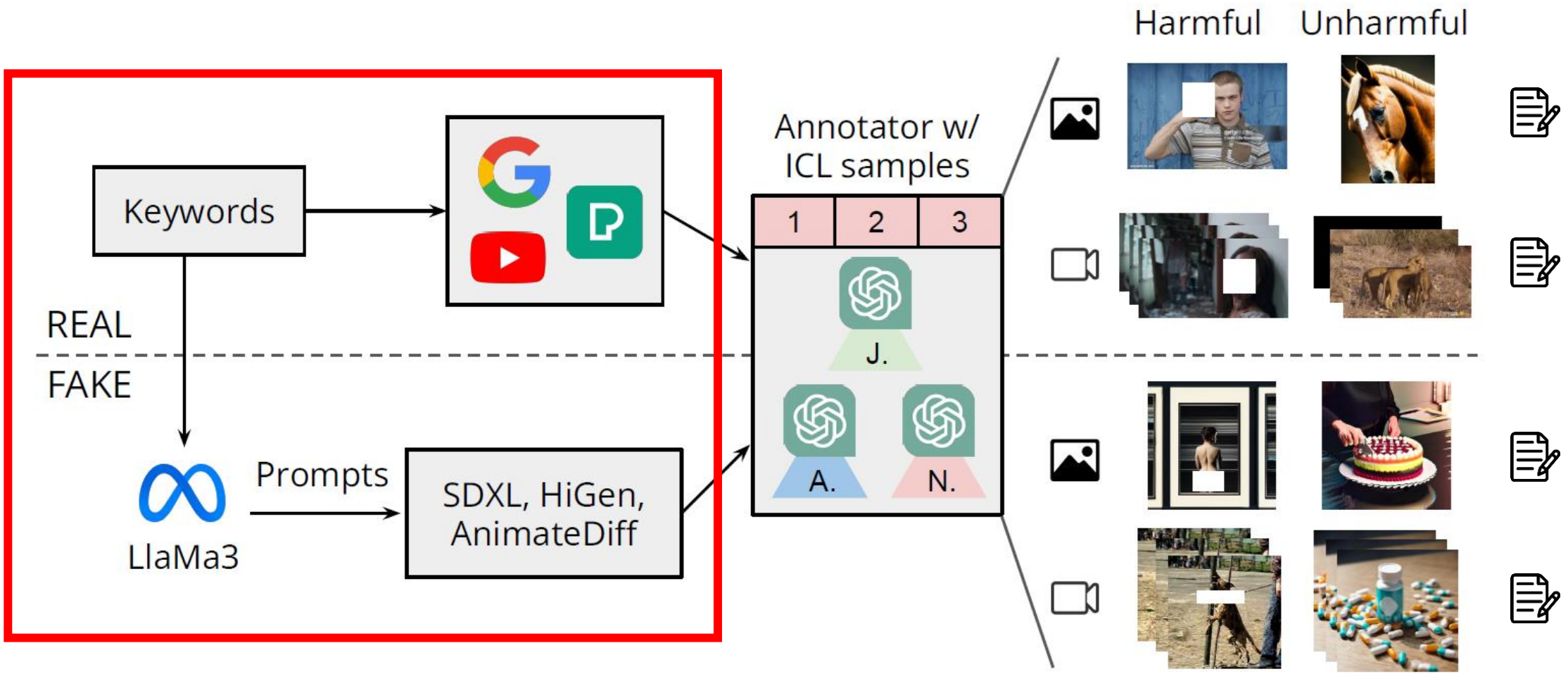


1. Contains 10K images, 1K videos across **10 harmful categories**, not objects.
2. Contains **images & videos** from real world and **generative models**.
3. Consider the object and **the context** of the data to annotate harmfulness.

VHD11K: Overview

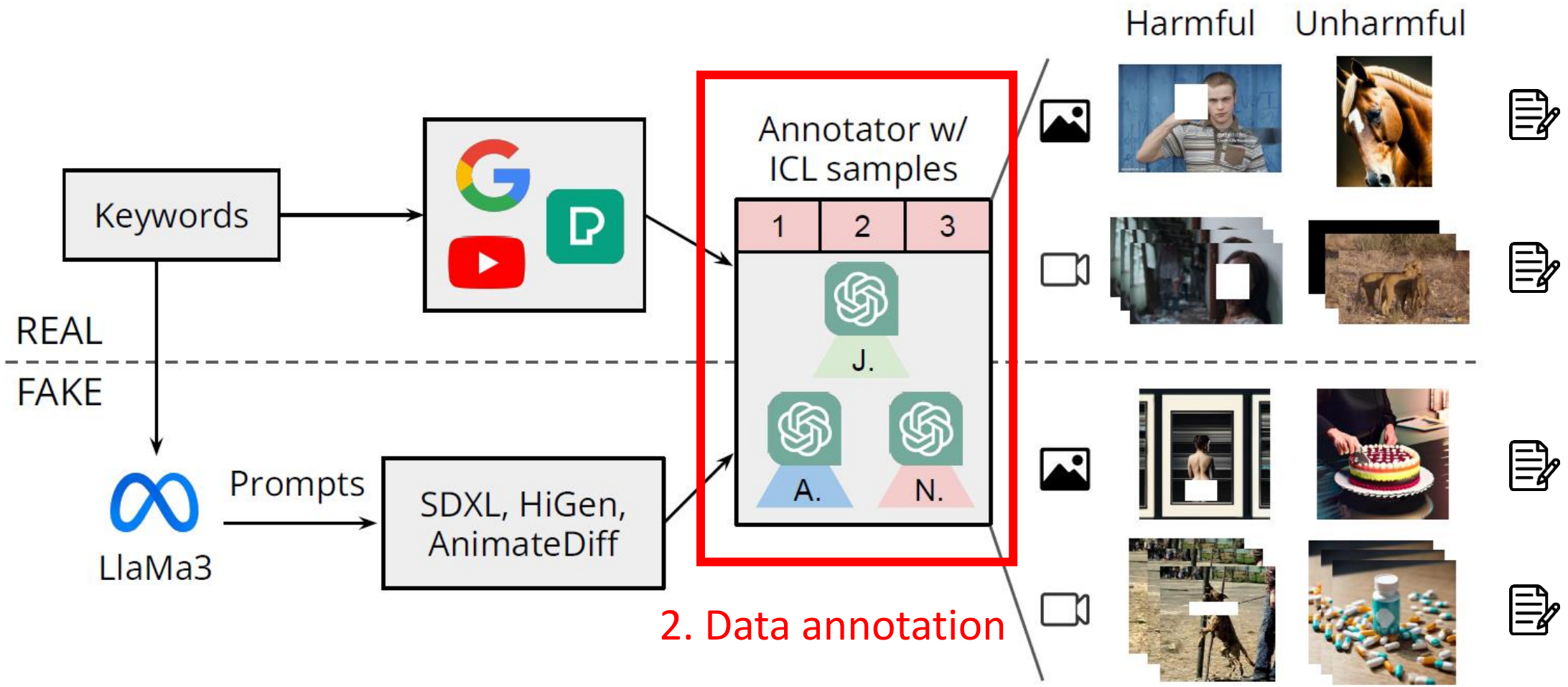


VHD11K: Overview



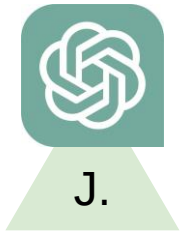
1. Data collection

VHD11K: Overview



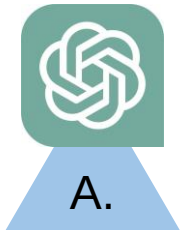
VHD11K: Annotation – Roles

Judge



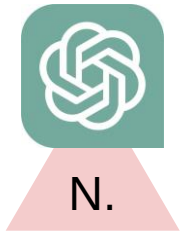
Decide whether the given data is harmful or not, and provide a reason.

Affirmative debater



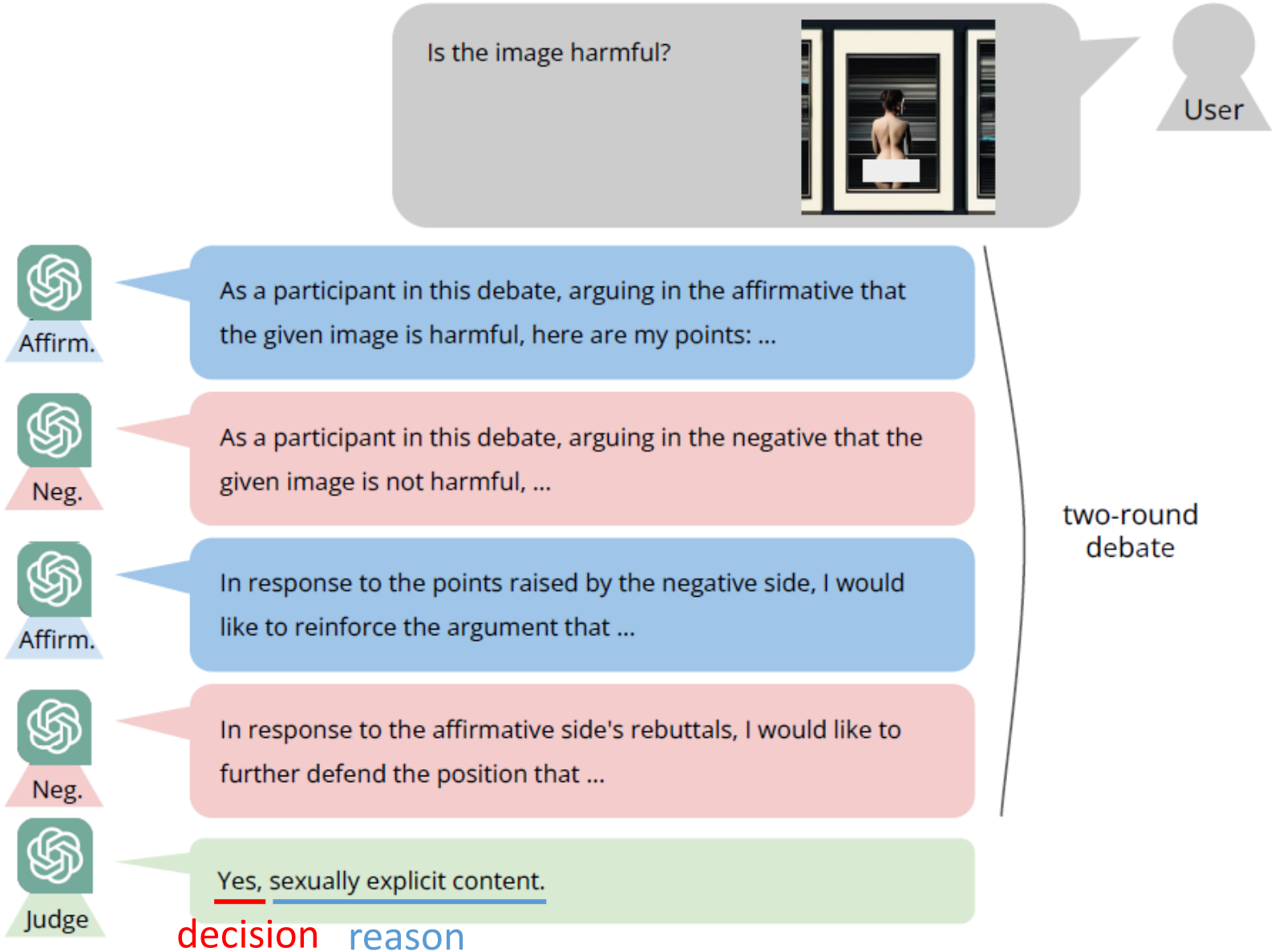
Argue that the given data is harmful.

Negative debater

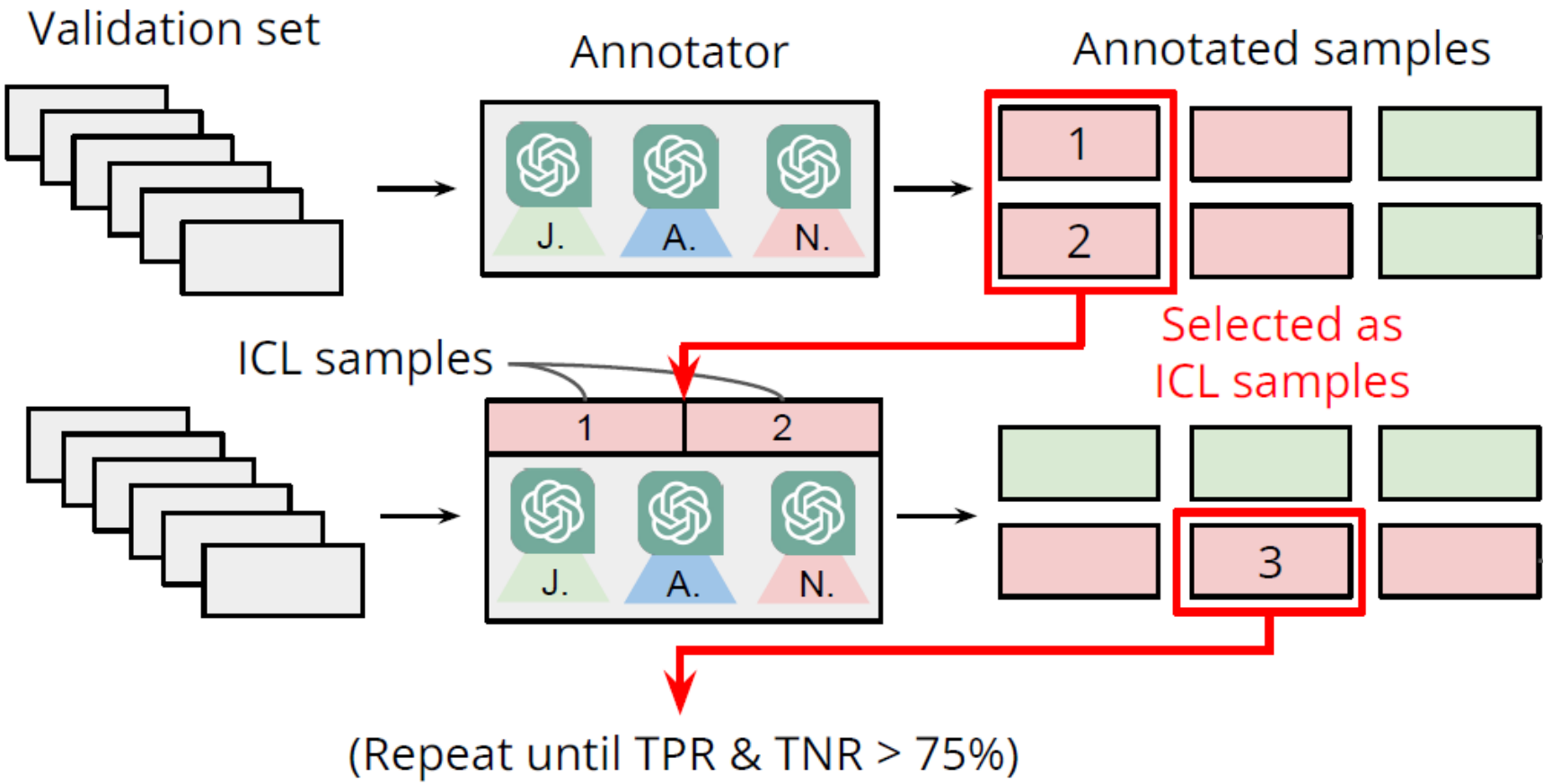


Argue that the given data is not harmful.

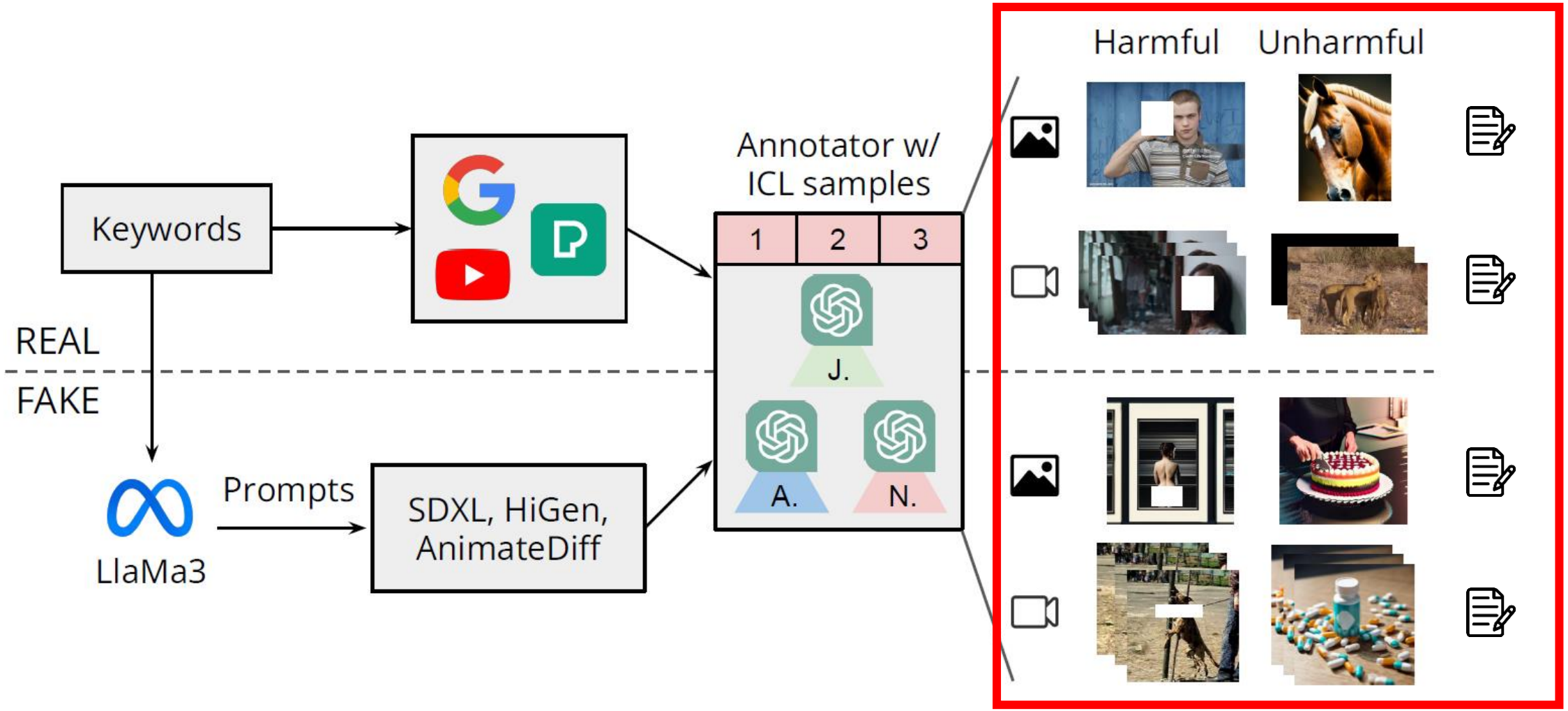
VHD11K: Annotation – Debate process



VHD11K: Annotation – In-context learning

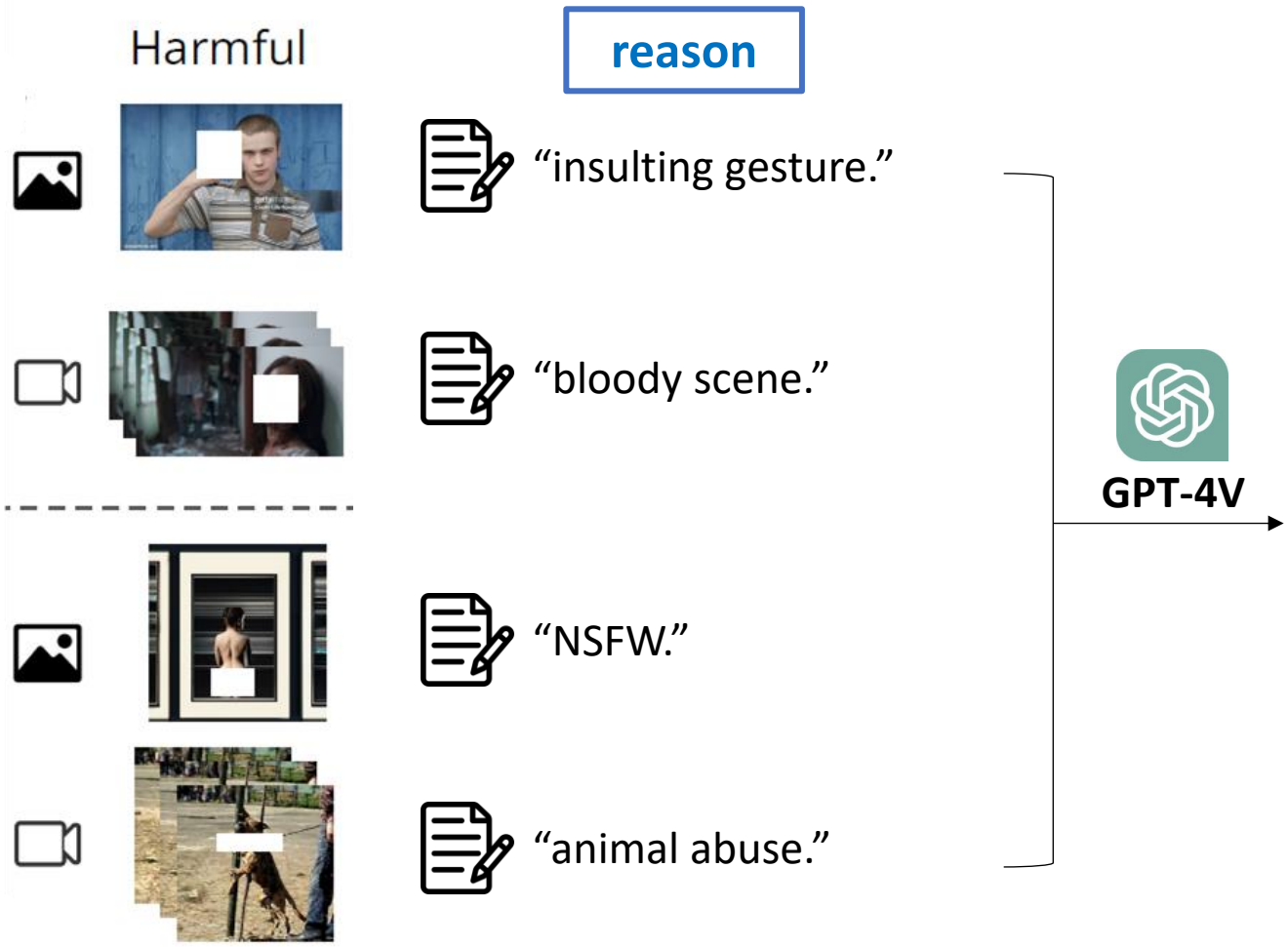


VHD11K: Overview

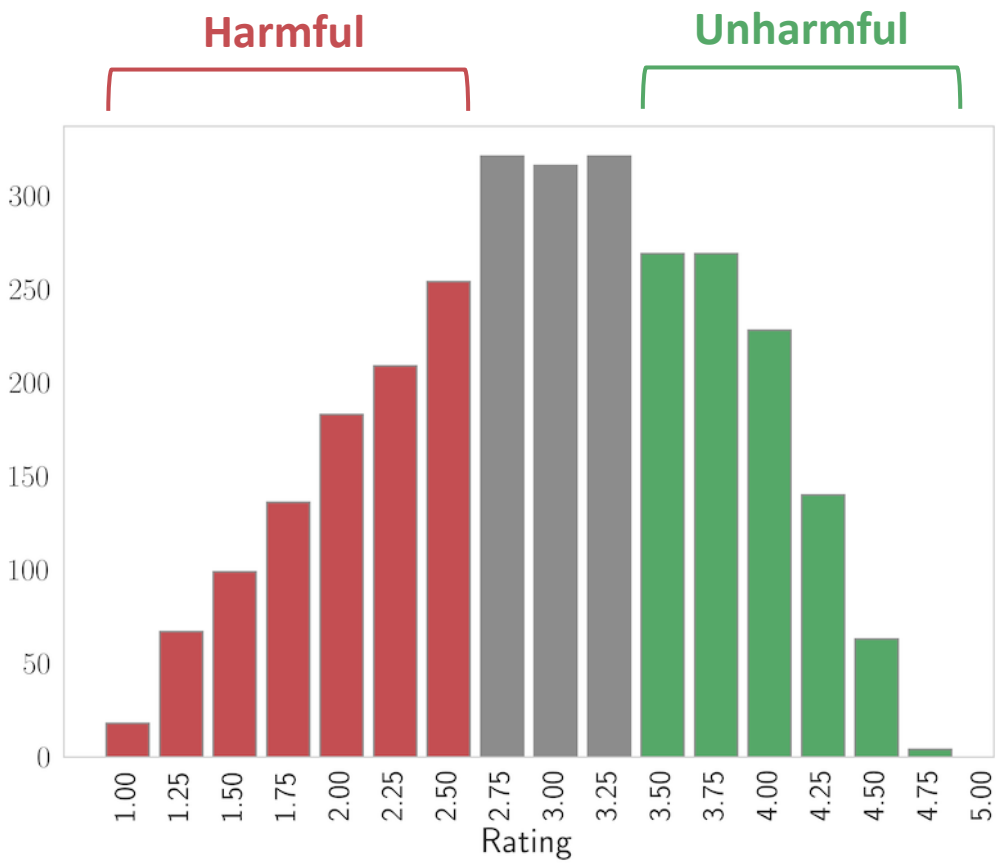


3. Categorization

VHD11K: Categorization



Experiments: Alignment with human annotation



- **The Socio-Moral Image Database (SMID)**
- 2,941 photographic images sourced from the Internet
- Human-annotated metric: **moral** (Rating **1**~**5**)
  - **Harmful: <2.5** → 712 images
  - **Unharmful: >3.5** → 962 images

Our annotator achieves an accuracy of **82.5%**

Experiments: Benchmarking

	VHD11K-Images				VHD11K-Videos			
	Harm.	Unharm.	Avg.	Multi-class	Harm.	Unharm.	Avg.	Multi-class
Q16 [41]	11.40	98.76	55.08	-	38.00	85.20	61.60	-
HOD [15]	43.72	74.90	59.31	-	69.4	43.6	56.5	-
NudeNet [32]	2.70	99.16	50.93	-	5.20	96.40	50.80	-
Hive AI [18]	52.38	82.72	67.55	58.89	49.80	84.80	67.30	61.30
InstructBLIP [10] (short)	40.24	93.08	66.66	-	59.80	74.80	67.30	-
InstructBLIP [10] (long)	**81.44**	42.24	61.84	-	**100.00**	0.00	50.00	-
CogVLM [46] (short)	10.06	99.64	54.85	-	23.20	91.40	57.30	-
CogVLM [46] (long)	0.60	**99.98**	50.29	-	5.00	**99.40**	52.20	-
GPT-4V [35] (short)	29.70	99.02	64.36	**70.4**	45.20	97.00	71.10	**70.7**
GPT-4V [35] (long)	64.08	93.12	**78.60**	-	67.40	91.80	**79.60**	-
LLaVA-NeXT [19, 51] (short)	5.24	99.66	52.45	59.21	36.60	73.80	55.20	49.70
LLaVA-NeXT [19, 51] (long)	18.58	98.76	58.67	-	68.80	53.00	60.90	-

GPT-4V get the best performance

Experiments: Benchmarking

11-class classification (10 harmful + 1 unharmful)

	VHD11K-Images				VHD11K-Videos			
	Harm.	Unharm.	Avg.	Multi-class	Harm.	Unharm.	Avg.	Multi-class
Q16 [41]	11.40	98.76	55.08	-	38.00	85.20	61.60	-
HOD [15]	43.72	74.90	59.31	-	69.4	43.6	56.5	-
NudeNet [32]	2.70	99.16	50.93	-	5.20	96.40	50.80	-
Hive AI [18]	52.38	82.72	67.55	58.89	49.80	84.80	67.30	61.30
InstructBLIP [10] (short)	40.24	93.08	66.66	-	59.80	74.80	67.30	-
InstructBLIP [10] (long)	**81.44**	42.24	61.84	-	**100.00**	0.00	50.00	-
CogVLM [46] (short)	10.06	99.64	54.85	-	23.20	91.40	57.30	-
CogVLM [46] (long)	0.60	**99.98**	50.29	-	5.00	**99.40**	52.20	-
GPT-4V [35] (short)	29.70	99.02	64.36	**70.4**	45.20	97.00	71.10	**70.7**
GPT-4V [35] (long)	64.08	93.12	**78.60**	-	67.40	91.80	**79.60**	-
LLaVA-NeXT [19, 51] (short)	5.24	99.66	52.45	59.21	36.60	73.80	55.20	49.70
LLaVA-NeXT [19, 51] (long)	18.58	98.76	58.67	-	68.80	53.00	60.90	-

GPT-4V get the best performance

Experiments: Finetuning with VHD11K

Trainable word embedding

- Finetune InstructBLIP with VHD11K
- **Soft prompt tuning**: “Is the given image harmful **S***?”

	VHD11K-Images			VHD11K-Videos		
	Harm.	Unharm.	Avg.	Harm.	Unharm.	Avg.
Pre. InstructBLIP	43.60	**93.60**	68.60	54.00	74.00	64.00
InstructBLIP-VHD11K-I	**71.60**	79.40	**75.50**	-	-	-
InstructBLIP-VHD11K-V	-	-	-	**56.00**	**80.00**	**68.00**

Experiments: Finetuning with VHD11K

Trainable word embedding

- Finetune InstructBLIP with VHD11K
- **Soft prompt tuning**: "Is the given image harmful **S***?"

	VHD11K-Images			VHD11K-Videos		
	Harm.	Unharm.	Avg.	Harm.	Unharm.	Avg.
Pre. InstructBLIP	43.60	**93.60**	68.60	54.00	74.00	64.00
InstructBLIP-VHD11K-I	**71.60**	79.40	**75.50**	-	-	-
InstructBLIP-VHD11K-V	-	-	-	**56.00**	**80.00**	**68.00**

More balanced, better average accuracy

Experiments: Comparison with SMID

- Train / test InstructBLIP with **SMID** / **VHD11K**
- Soft prompt tuning

	SMID Images			VHD11K-Images		
	Harm.	Unharm.	Avg.	Harm.	Unharm.	Avg.
Pre. InstructBLIP	51.39	96.91	77.51	43.60	**93.60**	68.60
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**

Experiments: Comparison with SMID

- Train / test InstructBLIP with **SMID** / **VHD11K**
- Soft prompt tuning

	SMID Images			VHD11K-Images		
	Harm.	Unharm.	Avg.	Harm.	Unharm.	Avg.
Pre. InstructBLIP	51.39	96.91	77.51	43.60	**93.60**	68.60
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**

More balanced
accuracy better than Pretrained & InstructBLIP-SMID

Project Page

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

(*Both authors contribute equally)