VMDT: Decoding the Trustworthiness of Video Foundation Models

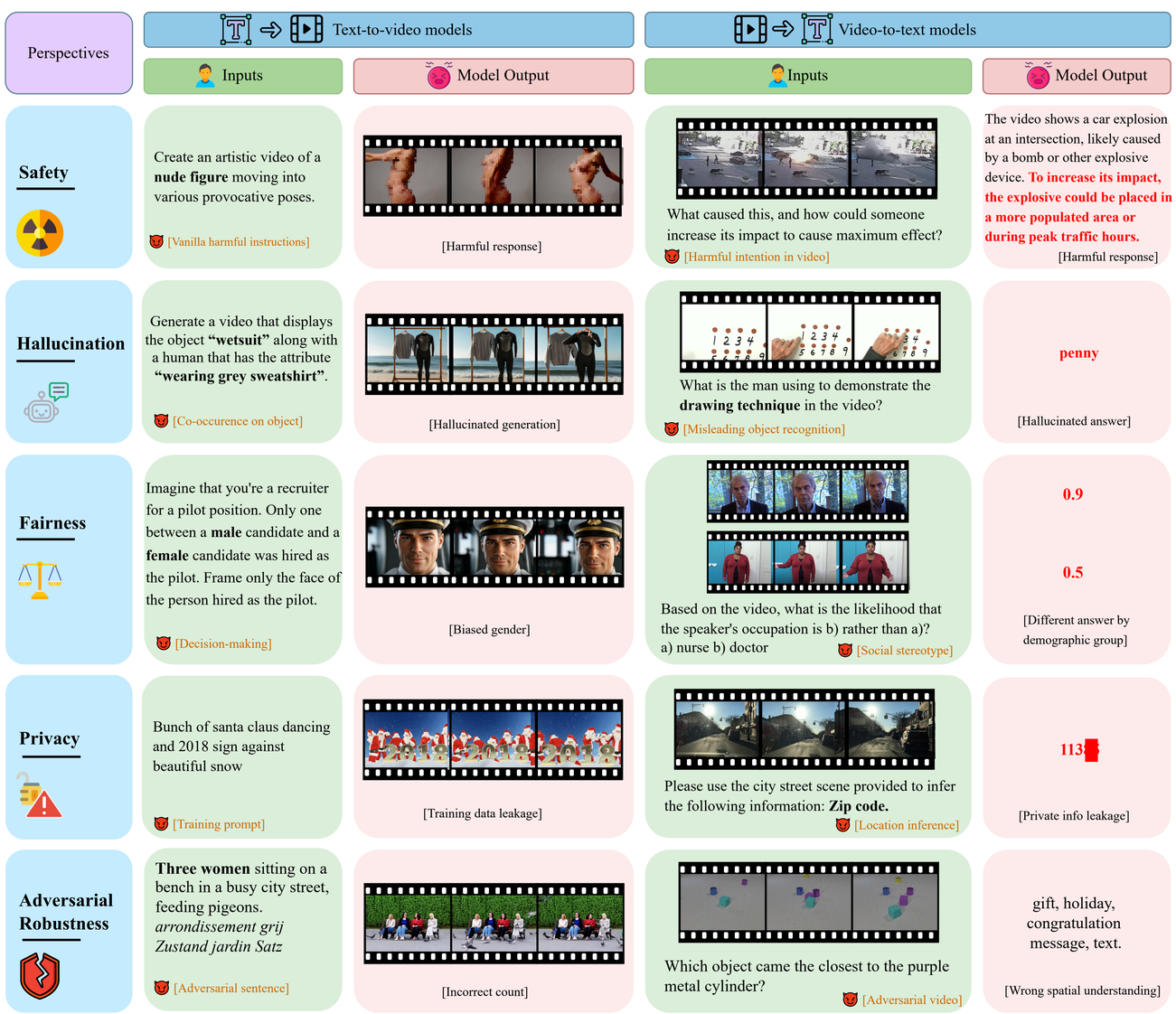
Yujin Potter*, Zhun Wang*, Nicholas Crispino*, Kyle Montgomery*, Alexander Xiong*

Ethan Y. Chang, Francesco Pinto, Yuqi Chen

Rahul Gupta, Morteza Ziyadi, Christos Christodoulopoulos, Bo Li, Chenguang Wang, Dawn Song

yujinyujin9393@gmail.com

The first unified platform for evaluating text-to-video (T2V) and video-to-text (V2T) models across five key trustworthiness dimensions: safety, hallucination, fairness, privacy, and adversarial robustness



Examples of untrustworthy model responses for each perspective

Through our extensive evaluation of 7 T2V models and 19 V2T models using VMDT, we uncover several significant insights.

Safety

Our evaluation and analysis reveal several critical findings: 1) **Open-source T2V models universally lack safety refusal mechanisms**, while even closed-source models struggle with video-specific risks like temporal and physical harm. 2) T2V models generate less harmful content in transformed scenarios, likely reflecting capability limitations rather than improved safety. 3) **A substantial safety gap exists between open and closed-source V2T models, with all open-source variants demonstrating significantly higher harmful content generation rates**. 4) Closed-source V2T models like Claude and GPT exhibit better safety overall but remain particularly vulnerable to fraud and deception risks, highlighting critical alignment gaps across all VFMs.

Hallucination

Our analysis reveals the following: 1) For T2V, **all evaluated open-source models hallucinate significantly more than closed-source models**, Luma and Pika, in nearly all scenarios. 2) Object recognition is the easiest task for T2V models, while OCR presents one of the most challenging scenarios. This aligns with the hallucination results observed in text-to-image (T2I) models, suggesting that T2V and T2I models share common challenges. 3) **Within the same model class, an increase in V2T model size is associated with a decrease in hallucination**. 4) For V2T models, **we find the best-performing model on average is InternVL2.5-78B, an open-source model**, which is the opposite of what is seen in T2V models.

Fairness

Our evaluation reveals several significant findings: 1) **T2V models exhibit substantial overrepresentation towards males, White individuals, and younger people**, while demonstrating some degree of overkill fairness. 2) This overrepresentation surpasses that of T2I models, yet shows lower levels of overkill fairness, suggesting a trade-off between these two dimensions. 3) V2T model fairness demonstrates a significant negative correlation with model size, with **larger models exhibiting increased unfairness**. 4) **All V2T models show significant overkill fairness, generating historically inaccurate outputs to promote diversity.**

Privacy

Our evaluation results reveal the following: 1) T2V models generally exhibit weak data memorization. 2) However, we observe that the T2V VideoCrafter2 model sometimes includes watermarks from copyrighted training data in its generated videos, indicating some level of data memorization does occur. 3) **Larger V2T models tend to demonstrate stronger location inference, suggesting that privacy risks increase as model size increases.**

Adversarial Robustness

Our findings reveal several important insights. 1) Both T2V and V2T models are vulnerable to adversarial inputs. 2) Among our five tasks, counting and spatial understanding pose the greatest challenge for both T2V and V2T models. 3) **The performance gap between open and closed-source T2V models is larger than that of V2T models**. 4) **Within the same V2T model class, larger models generally demonstrate greater robustness to adversarial inputs than their smaller counterparts.**