Project Title: Your Project Title Here

Team Members: Renee Zbizika
Emails: rzbizika@stanford.edu

1 Objective

Image-to-video (I2V) generation has made significant progress with models like AnimateDiff and CogVideoX Yang et al. (2024), but evaluating generated videos remains a major bottleneck. Current evaluations rely on handcrafted metrics and human preference studies, but overall lack a unified, differentiable, and human-aligned reward model. Existing learned reward models, such as VideoReward Liu et al. (2025), are designed for text-to-video tasks and are not directly applicable to I2V generation. This project proposes adapting the VideoReward architecture to the I2V setting by introducing an Image Consistency (IC) reward head and training on structured human annotations from the VBench benchmark Huang et al. (2024). By combining VBench's multi-dimensional evaluation suite with differentiable reward modeling, I aim to create a scalable, human-aligned, and generalizable reward model for I2V evaluation. This learned reward will serve as a robust alternative to handcrafted human evaluation, enabling better benchmarking, model selection, and potentially reward-guided generation.

2 Related Work

VideoReward Liu et al. (2025) introduces a learned reward model for text-to-video (T2V) generation. It uses a multi-head architecture to predict dimensions like visual quality, motion realism, and text alignment based on human preference data. Since VideoReward is trained specifically for video-text pairs, it cannot directly evaluate I2V generation, where alignment with a source image is a primary metric.

VBench Huang et al. (2024) is a comprehensive benchmark suite for evaluating video generation models across multiple dimensions, such as motion realism, perceptual quality, and image consistency. It has structured handcrafted metrics and large-scale human preference annotations for both T2V and I2V tasks. However, VBench doesn't offer a learned, generalizable reward model that can predict preferences without repeated collection of human feedback.

Recent I2V models like AnimateDiff and CogVideoX Yang et al. (2024) have pushed the quality of generated videos, but evaluation remains ad-hoc, relying on CLIP scores, frame difference metrics, or small human studies. These approaches often correlate poorly with true human judgments and lack the scalability needed for training or fine-tuning generative models.

My work addresses these gaps by adapting the VideoReward architecture to I2V tasks using VBench's human-labeled datasets. In doing so, I provide a novel, scalable, and human-aligned reward function for evaluating I2V models.

3 Technical Outline

Dataset: I will generate an I2V dataset from VBench's prompt and source image suites, selecting 2–3 strong I2V models to produce diverse video outputs. Each example will consist of a source image, multiple generated videos, and associated human preference annotations.

Reward Model Adaptation: Starting from the VideoReward architecture, I plan to retain the visual quality and motion quality heads but replace the text alignment head with a new image consistancy (IC) head. The model input will be adapted from (video, text prompt) to (video, source image) pairs. Training will use pairwise preference losses (e.g., Direct Preference Optimization or margin ranking loss).

Training: The reward model will be trained to predict human preferences on labeled pairs as specified above. We will apply regularization to prevent overfitting to specific prompts or models.

Evaluation: Success is defined by strong human alignment, competitive baseline performance, and robustness to distribution shift.

Alignment with Human Preferences

I will compute Spearman & Pearson correlation between learned reward scores and human preference labels to indicate either successful alignment or failure to capture human judgments.

Comparison to VBench Baselines

I will compare the ranking quality of the learned reward against VBench's handcrafted evaluation pipelines (CLIP similarity, frame-difference metrics, etc.), analyzing failure cases like reward hacking or misaligned rankings.

Robustness to Distribution Shift

I will test the reward model's ability to generalize onto unseen prompts, images, and I2V models. Performance degradation (like drop in correlation score) will be measured relative to in-distribution validation sets. A robust model should maintain reasonable performance across different data sources; sharp drops or erratic behavior would signal overfitting.

4 Team Contributions

- **Renee**: sole contributor

References

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, Yaohui Wang, Xinyuan Chen, Ying-Cong Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. 2024. VBench++: Comprehensive and Versatile Benchmark Suite for Video Generative Models. (2024). `https://doi.org/10.48550/ARXIV.2411.13503`

Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Wenyu Qin, Menghan Xia, Xintao Wang, Xiaohong Liu, Fei Yang, Pengfei Wan, Di Zhang, Kun Gai, Yujiu Yang, and Wanli Ouyang. 2025. Improving Video Generation with Human Feedback. (2025). `https://doi.org/10.48550/ARXIV.2501.13918`

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. 2024. CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer. (2024). `https://doi.org/10.48550/ARXIV.2408.06072`