Project Title: Learning Image-to-Video Consistency Rewards

Team Members: Agnes Liang, Renee Zbizika
Emails: agliang@stanford.edu, rzbizika@stanford.edu

1 Experiments

Setup: Dataset: We use a 300-example subset of the VBench++ dataset Huang et al. (2024) with videos from VideoCrafter1.0. Each sample includes a source image, two generated videos, and a human-labeled image-to-video consistency score. **Method:** Our baseline is a supervised MLP using frozen image and mean-pooled video features. Our main model adapts VideoReward Liu et al. (2025) by replacing text with the reference image, sharing a ViT encoder across image and video frames, and introducing an Image Consistency (IC) head with attention-based fusion. For the milestone, we implemented and validated this model, showing it learns meaningful consistency rewards.

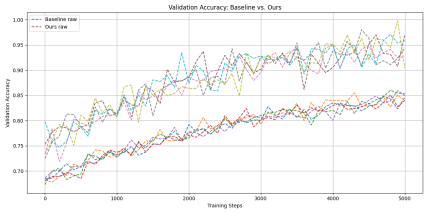

Figure 1: Visualization of results.

Initial Results: Our results show that our model learns more accurate signals than the MLP baseline, supporting our hypothesis that adapting VideoReward with an image consistency head would be effective for reward modeling. Our higher accuracy of the suggests that incorporating transformer-based fusion and learned attention enables better modeling of image-video relationships, albeit with increased sensitivity to training dynamics.

2 Changes to Research Hypothesis or Objective

Original Hypothesis: A reward model trained on VBench human preference data can outperform handcrafted metrics for evaluating I2V model outputs.

Revised Hypothesis: No changes. Our initial experiments support the hypothesis that human-aligned reward learning is feasible for I2V tasks.

Justification: Results on the subset confirm the feasibility of training a differentiable, human-aligned reward model. We found no need to revise the objective. Wang et al. (2023)

3 Next Steps

1. **Further Experimentation:** Expand training data by running VBench on low-quality videos from non-SOTA I2V models to better represent failure cases and improve contrastive learning.
2. **Algorithm Refinement:**
   - Incorporate the IC-Reward model into an open-source I2V generator (e.g., Open-Sora) as a reward function for reward-guided fine-tuning.
   - Extend the current supervised loss to include contrastive or pairwise preference learning, aligning more closely with human-annotated relative judgments.
3. **Evaluation & Analysis:** Explore adding a transformer branch for text prompts, enabling multi-modal reward learning using both image and text inputs from VBench.

Potential Challenges: Integrating the reward model into an existing I2V pipeline may require extensive architectural modifications. Additionally, collecting sufficient low-quality video samples and ensuring reliable human-aligned labels for contrastive learning can be time-consuming and labor-intensive. Multimodal training introduces additional complexity in aligning visual and textual modalities effectively.

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. *arXiv preprint arXiv:2411.13503* (2024).

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

Cong Wang, Jiaxi Gu, Panwen Hu, Songcen Xu, Hang Xu, and Xiaodan Liang. 2023. DreamVideo: High-Fidelity Image-to-Video Generation with Image Retention and Text Guidance. (2023). https://doi.org/10.48550/ARXIV.2312.03018