Improving Video Generation with Human Feedback

Jie Liu*, Gongye Liu*, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng
Xiele Wu, Qiulin Wang, Menghan Xia, Xintao Wang, Xiaohong Liu, Fei Yang
Pengfei Wan, Di Zhang, Kun Gai, Yujiu Yang✉, Wanli Ouyang

香港中文大學
The Chinese University of Hong Kong

KlingAI

上海人工智能实验室
Shanghai Artificial Intelligence Laboratory

Today's Text to Video Model

- Non-physical motion

a woman with long brown hair and wearing a pink nightgown walks towards the bed in the bedroom and lays on it

- Low visual quality

A cowboy rides his horse across an open plain at sunset

- Imperfect alignment with user prompts

A fox and an owl stargazing together on a hilltop

Post Training for Video Generation

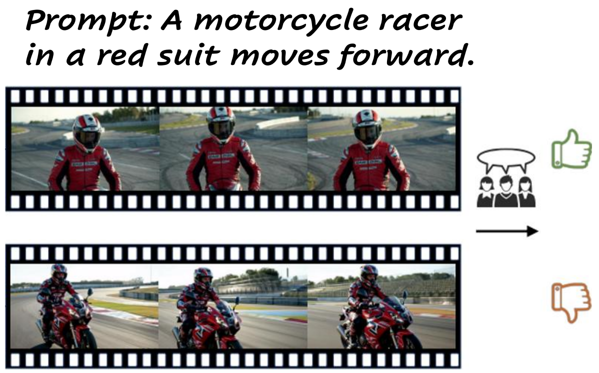
- Post training has shown remarkable success in LLM and image generation
  - Examples: DeepSeek-R1, OpenAI-o1/o3,

Can post training also benefit

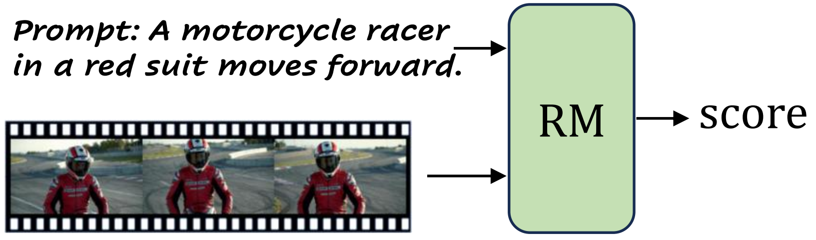
video generation?

Three Components of Post-Training

Preference Datasets



Reward Model



Policy Model Training

Maximize the defined reward function while stay close to the initial policy (DPO/RWR, etc.)

Human Preference Dataset

(a) Human Preference Annotation

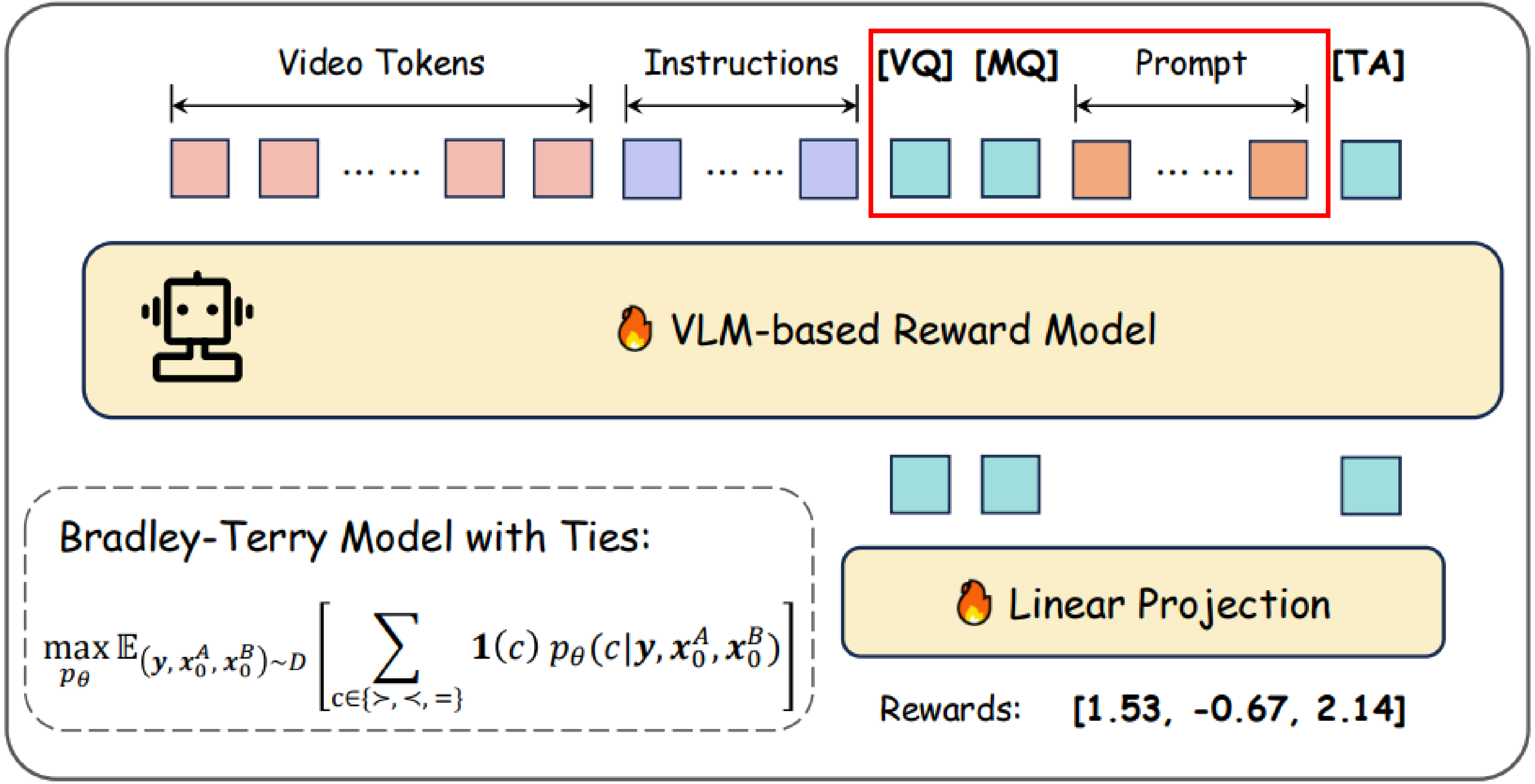
Dataset Statistics

- **12** text-to-video models
- **182k** annotated triplets

Two generated videos are compared along three preference dimensions:

- Visual Quality (VQ): image fidelity and details
- Motion Quality (MQ): smoothness and temporal coherence
- Text Alignment (TA): consistency with textual prompt

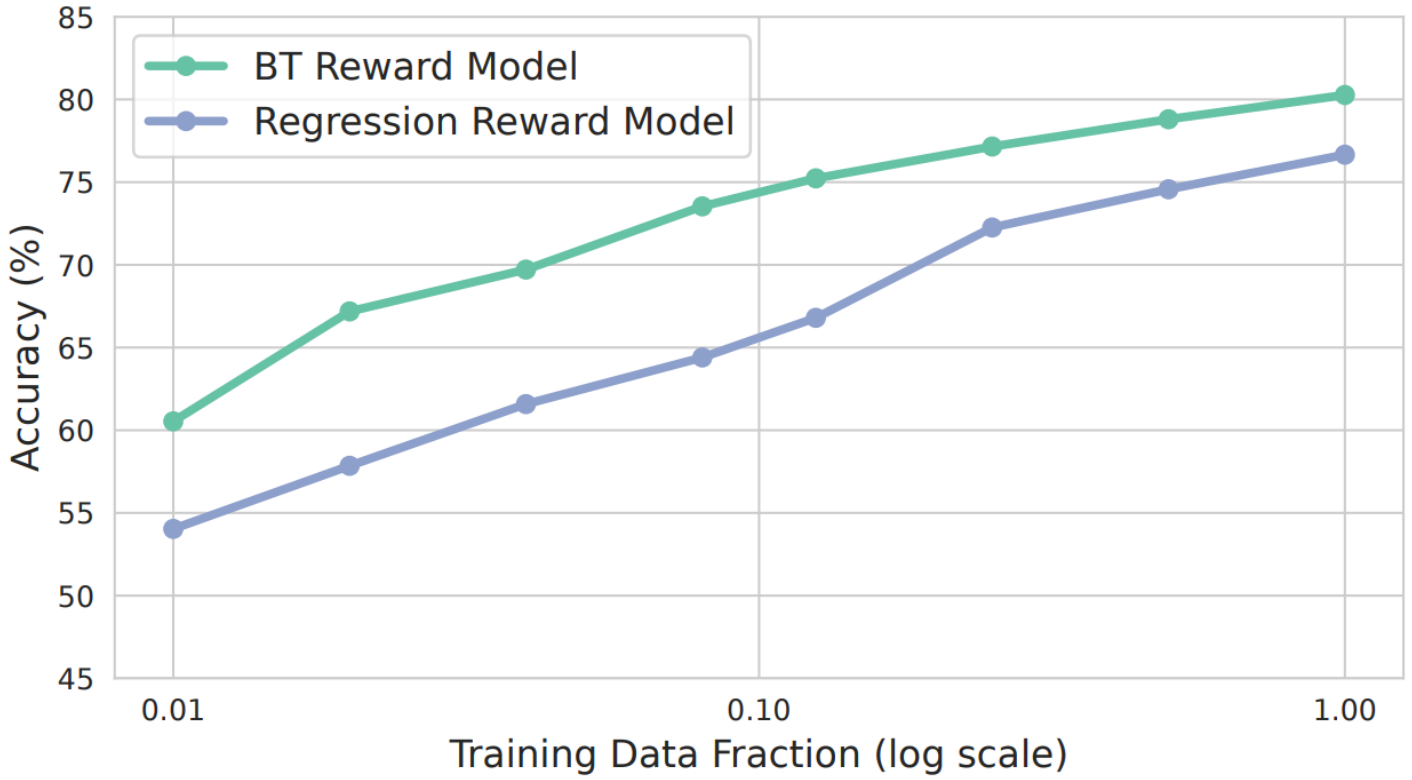
Reward Modeling



- We place the prompt after VQ and MQ to avoid prompt bias.

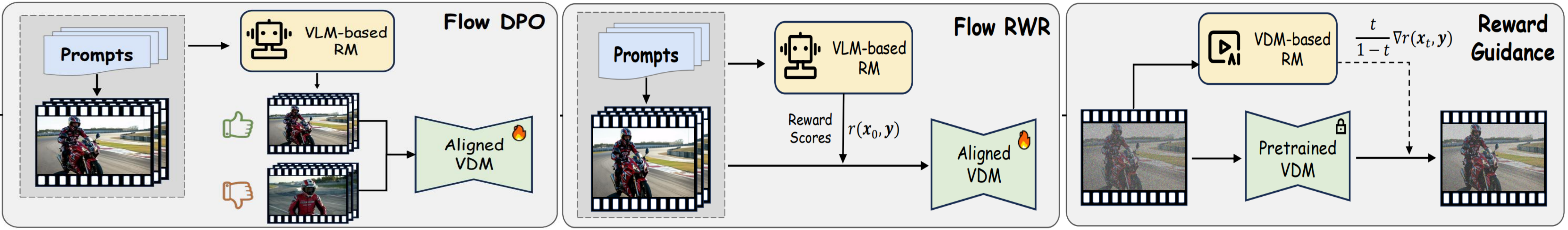
Reward Modeling

Score Regression v.s. Pairwise Comparison (Bradley-Terry)



Alignment

- The RLHF objective is $\max_{p_\theta} \mathbb{E}_{\boldsymbol{y}\sim\mathcal{D}_c, \boldsymbol{x}_0\sim p_\theta(\boldsymbol{x}_0|\boldsymbol{y})}[r(\boldsymbol{x}_0, \boldsymbol{y})] - \beta\mathbb{D}_{KL}[p_\theta(\boldsymbol{x}_0|\boldsymbol{y}) \parallel p_{ref}(\boldsymbol{x}_0|\boldsymbol{y})]$
- We propose three algorithms optimizing the same RLHF objective for rectified flow:
  - Training-time: **Flow-DPO**, **Flow-RWR**
  - Inference-time: **Flow-NRG** (reward guidance)



Closer to good samples

$$-\mathbb{E}\left[\log\sigma\left(-\frac{\beta_t}{2}\left(\|\boldsymbol{v}^w - \boldsymbol{v}_\theta(\boldsymbol{x}_t^w, t)\|^2 - \|\boldsymbol{v}^w - \boldsymbol{v}_{\text{ref}}(\boldsymbol{x}_t^w, t)\|^2 - \left(\|\boldsymbol{v}^l - \boldsymbol{v}_\theta(\boldsymbol{x}_t^l, t)\|^2 - \|\boldsymbol{v}^l - \boldsymbol{v}_{\text{ref}}(\boldsymbol{x}_t^l, t)\|^2\right)\right)\right)\right]$$

further from bad samples

$$\mathcal{L}_{\text{RWR}}(\theta) = \mathbb{E}\left[\exp(r(\boldsymbol{x}_0, \boldsymbol{y}))\|\boldsymbol{v} - \boldsymbol{v}_\theta(\boldsymbol{x}_t, t, \boldsymbol{y})\|^2\right]$$

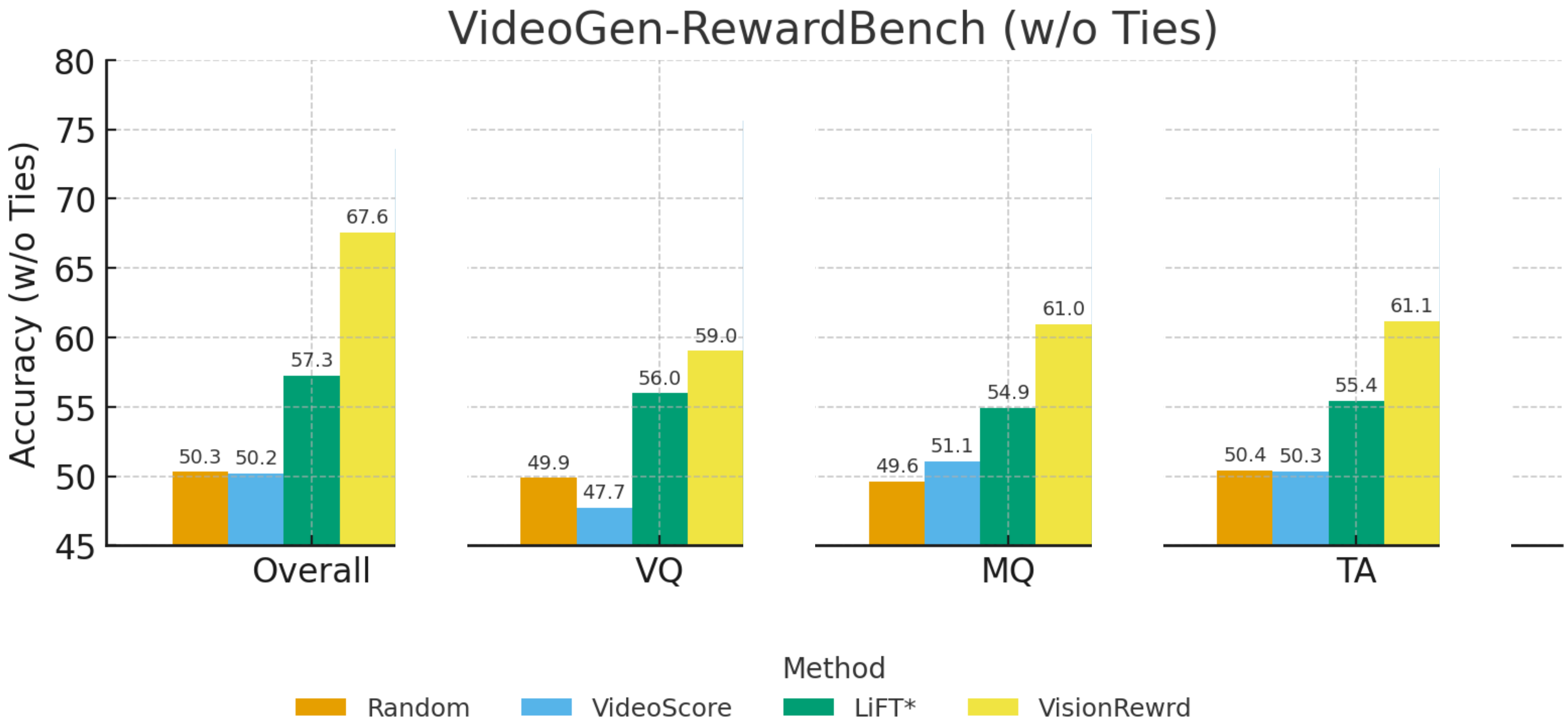
Reward weighted regression

$$\tilde{\boldsymbol{v}}_t(\boldsymbol{x}_t \mid \boldsymbol{y}) = \boldsymbol{v}_t(\boldsymbol{x}_t \mid \boldsymbol{y}) - w\,\frac{t}{1-t}\nabla r(\boldsymbol{x}_t, \boldsymbol{y})$$

Reward-gradient velocity

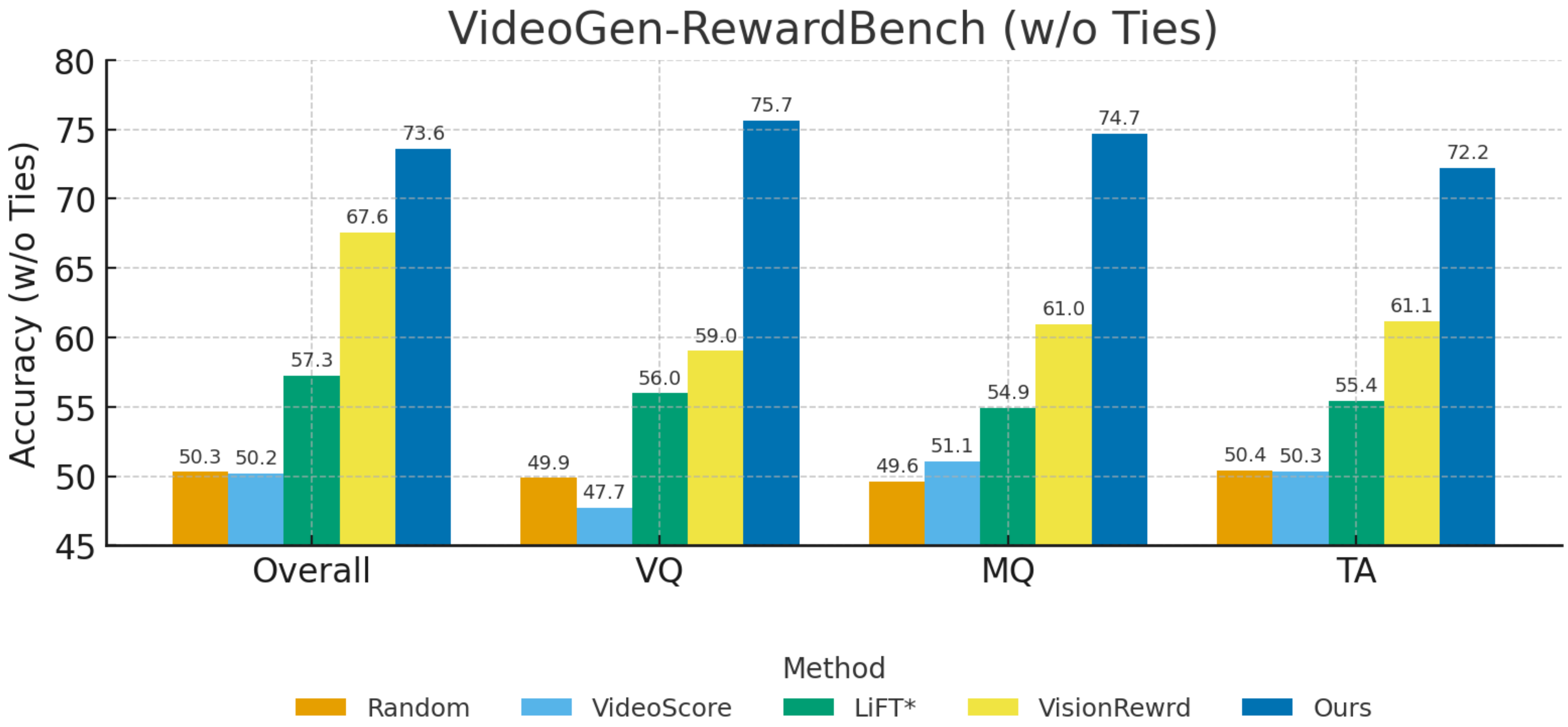
Experiments

- Reward Accuracy



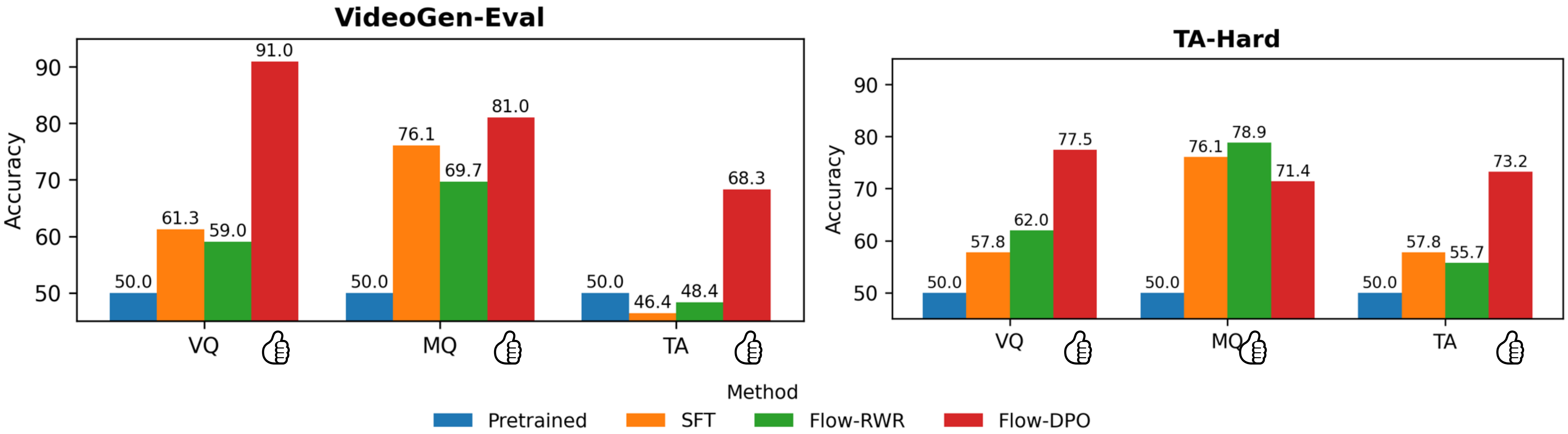
Experiments

- Reward Accuracy



Experiments

- Win rate



Visual Quality + Motion Quality

A cowboy rides his horse across an open plain at sunset, with the camera capturing the warm colors of the sky and the soft light on the landscape.

The camera remains still, a woman with long brown hair and wearing a pink nightgown walks towards the bed in the bedroom and lays on it, the background is a cozy bedroom, warm evening light.

Dynamic + Saturation

Thanks

paper

code