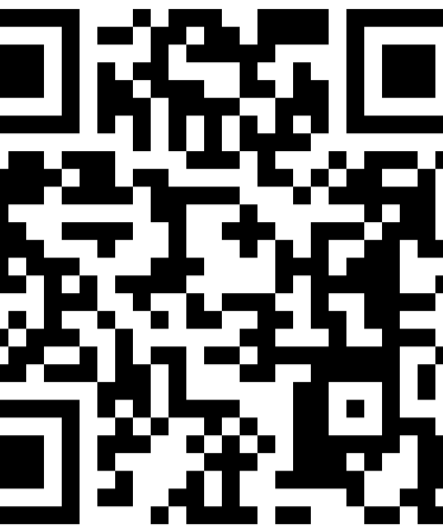


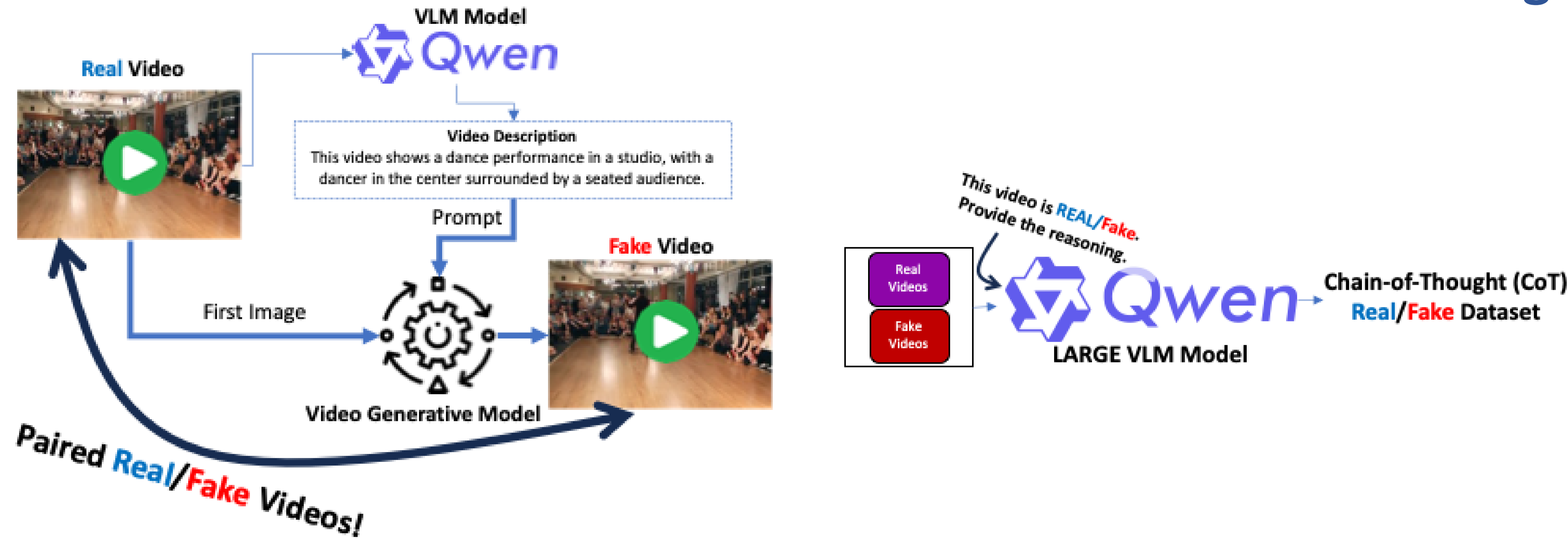
VidGuard-R1: AI-Generated Video Detection and Explanation via Reasoning MLLMs and RL

Kyoungjun Park¹, Yifan Yang², Juheon Yi², Shicheng Zheng², Muhammad Muaz¹, Yifei Shen², Dongqi Han², Caihua Shan², Lili Qiu^{1,2}

¹The University of Texas at Austin, ²Microsoft Research

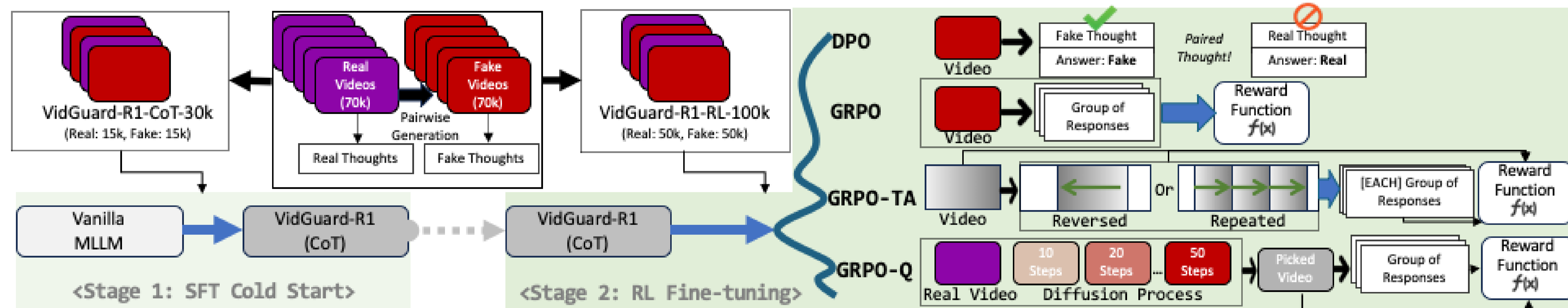


Paired Video Generation with Contextual and Format Matching



- Construct 140k videos in contextual pairs:
 - Real (InternVid, ActivityNet) & Fake (HunyuanVideo-I2V, CogVideoX-5B)
- Standardize video format: 49 frames, 8 FPS, 720p, and YUV420p.

Fine-tuning Pipeline with SFT + RL



1. SFT Cold Start: learn chain of thought (CoT) reasoning.
2. RL Fine-tuning
 - **DPO** (Direct Preference Optimization), **GRPO** (Group Relative Policy Optimization)

✓ GRPO-TA (Temporal Artifacts)

- Step 1: Randomly select a video segment using Gaussian distribution.
- Step 2: Apply temporal edits: reverse, repeat, or drop segments.

✓ GRPO-Q (Quality)

- Adjust diffusion steps to control generated video quality during diffusion.
- Use generated intermediate videos for scoring as reward function in GRPO.

Experimental Results

We evaluate across various baselines, including CNNs, Transformers, and MLLMs.

Our Dataset

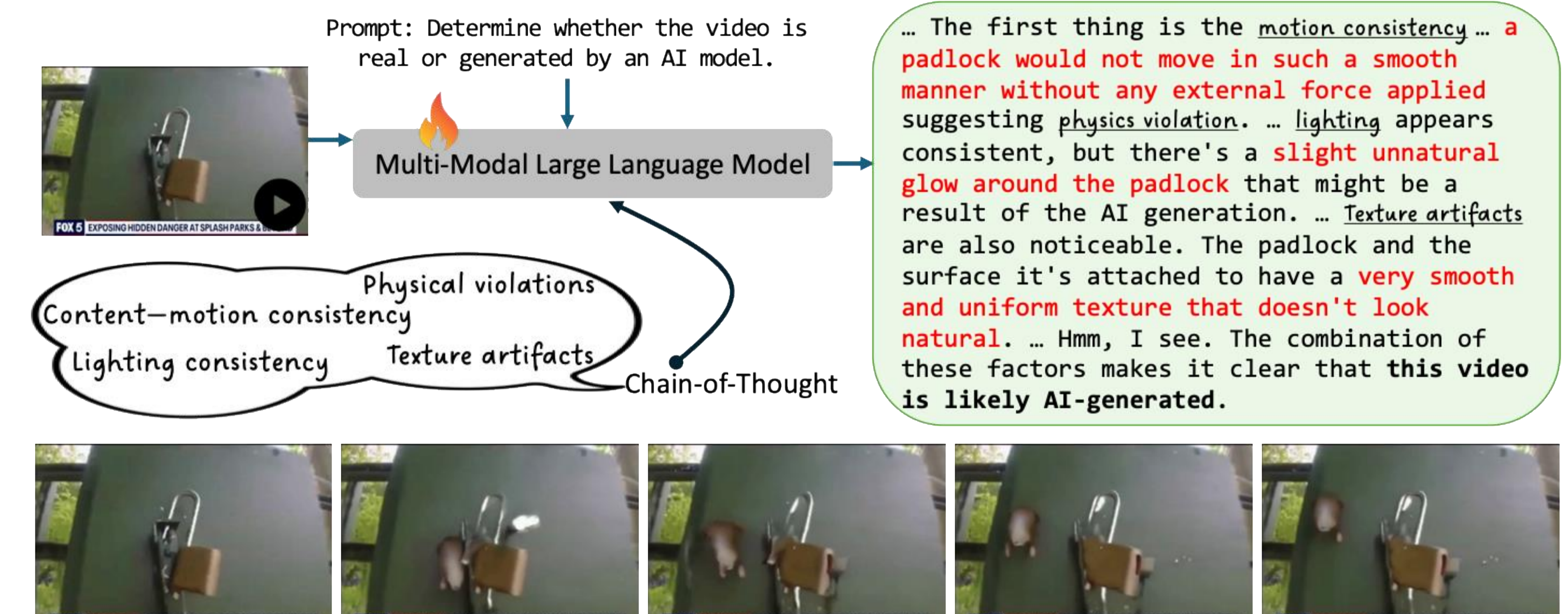
| Method | Type | CogVideoX | HunyuanVideo |
|-------------------------------|------|--------------|--------------|
| SlowFast | CNN | 77.87 | 77.03 |
| I3D | CNN | 64.78 | 62.13 |
| TRN | CNN | 68.73 | 69.87 |
| UniFormer V2 | TF | 73.95 | 71.92 |
| TimeSformer | TF | 78.53 | 74.55 |
| VideoSwin | TF | 76.81 | 79.71 |
| MViT V2 | TF | 58.38 | 53.91 |
| Qwen2.5-VL-7B | MLLM | 50.95 | 52.83 |
| GPT-4.1 mini | MLLM | 54.95 | 55.31 |
| VidGuard-R1 (CoT) | MLLM | 66.18 | 63.19 |
| VidGuard-R1 (DPO) | MLLM | 79.13 | 80.88 |
| VidGuard-R1 (GRPO, Zero-shot) | MLLM | 81.30 | 81.90 |
| VidGuard-R1 (GRPO-TA) | MLLM | 82.17 | 83.72 |
| VidGuard-R1 (GRPO-Q) | MLLM | 84.32 | 86.17 |

Benchmark: GenVidBench

| Method | Type | MuseV | SVD | CogVideo | Mora | HD-VG | Mean |
|-------------------------------|------|--------------|--------------|--------------|--------------|--------------|--------------|
| SlowFast | CNN | 12.25 | 12.68 | 38.34 | 45.93 | 93.63 | 41.66 |
| I3D | CNN | 8.15 | 8.29 | 60.11 | 59.24 | 93.99 | 49.23 |
| TRN | CNN | 38.92 | 26.64 | 91.34 | 93.98 | 93.97 | 71.26 |
| UniFormer V2 | TF | 20.05 | 14.81 | 45.21 | 99.21 | 96.89 | 57.55 |
| TimeSformer | TF | 73.14 | 20.17 | 74.80 | 39.40 | 92.32 | 64.28 |
| VideoSwin | TF | 62.29 | 8.01 | 91.82 | 45.83 | 99.29 | 67.27 |
| MViT V2 | TF | 76.34 | 98.29 | 47.50 | 96.62 | 97.58 | 79.90 |
| Qwen2.5-VL-7B | MLLM | 25.86 | 27.06 | 68.51 | 43.26 | 71.15 | 47.30 |
| GPT-4.1 mini | MLLM | 26.07 | 33.78 | 94.07 | 57.19 | 87.64 | 59.62 |
| VidGuard-R1 (CoT) | MLLM | 36.52 | 16.02 | 99.35 | 76.94 | 99.94 | 66.09 |
| VidGuard-R1 (GRPO, Zero-shot) | MLLM | 97.24 | 96.59 | 99.88 | 99.93 | 88.14 | 96.37 |
| VidGuard-R1 (GRPO) | MLLM | 97.38 | 94.98 | 99.90 | 99.99 | 95.46 | 97.53 |

- ✓ SFT with CoT data enables the MLLM to better understand and reason, surpassing baseline MLLM performance.
- ✓ GRPO fine-tuning further improves accuracy beyond standard GRPO results.

Qualitative Results



AI Video (Dreamvideo)



Let me think about this. When evaluating whether this video is real or AI-generated, I'll consider the factors of **motion consistency**, **lighting consistency**, **texture artifacts**, and **physics violations**.

1. **Motion Consistency**: The astronaut riding the horse appears to be in motion, but the movement looks **somewhat unnatural**. The horse's legs and the astronaut's body seem to move in a way that doesn't quite align with real-life physics. The motion looks **slightly jerky and inconsistent**, which is a common sign of AI-generated content.
2. **Lighting Consistency**: The lighting in the video seems to be consistent with what one might expect in a real outdoor setting. However, the lighting on the astronaut and the horse is **somewhat flat and lacks the depth and variation** that would be present in a real scene.
3. **Texture Artifacts**: The textures of the astronaut's suit and the horse's coat appear **somewhat artificial**. There are no visible details that would suggest a real-life material.
4. **Physics Violations**: The most striking aspect is the **physics violation**. An astronaut in a spacesuit would not be able to ride a horse in a real-world scenario.

Answer: Generated!