# Adapting Text-to-Image Models for Video Generation

Samir Agarwala, Hong Ju Jeon, Jared Watrous

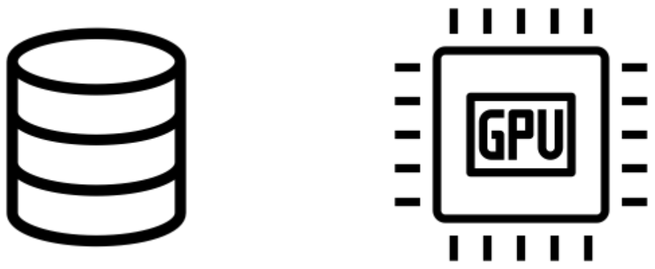
Department of Computer Science, Stanford University

## Introduction

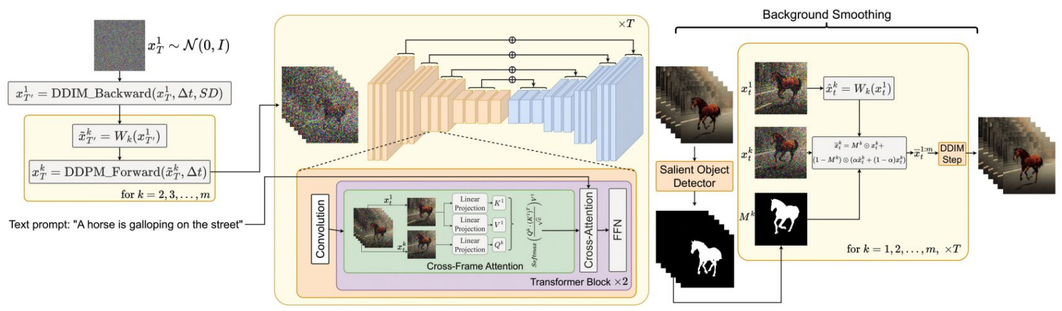
- Significant progress has been made in text-to-image models
- Training video generative models remains challenging
- **Leverage text-to-image (T2I) models to learn a text-to-video (T2V) generative model?**
  - *Input: Text Prompt*
  - *Output: Generated Video*

## Related Work

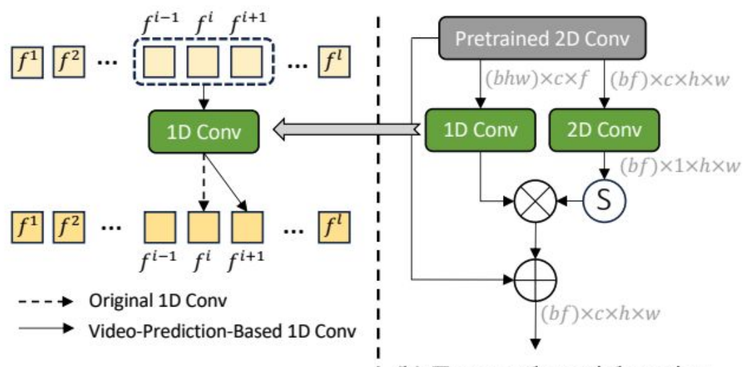
- *Unscalable*: Learning generative models from video datasets [1]



- *Unrealistic*: Zero-shot adaptations of T2I models [2]



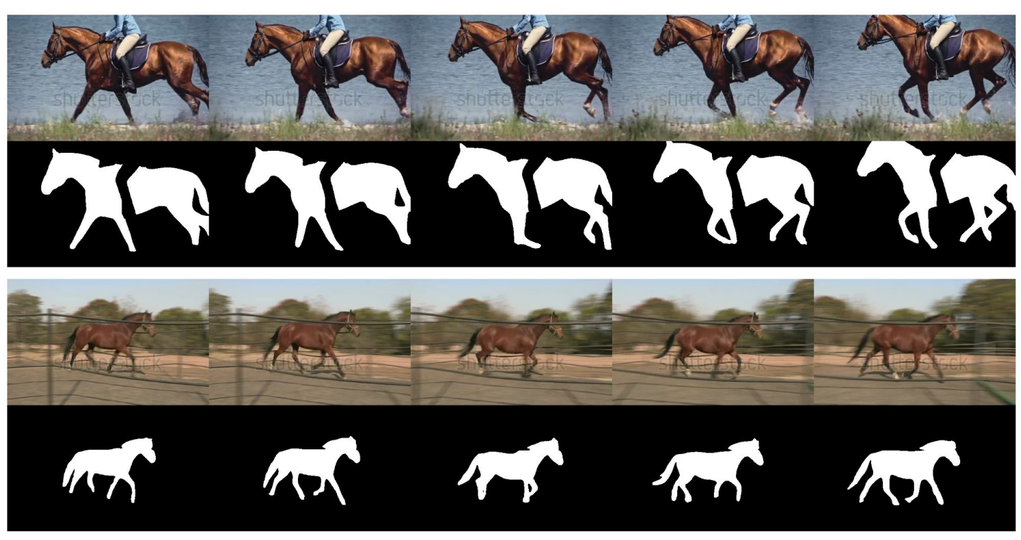

- *Upcoming*: Finetuning T2I models on limited video data [3]



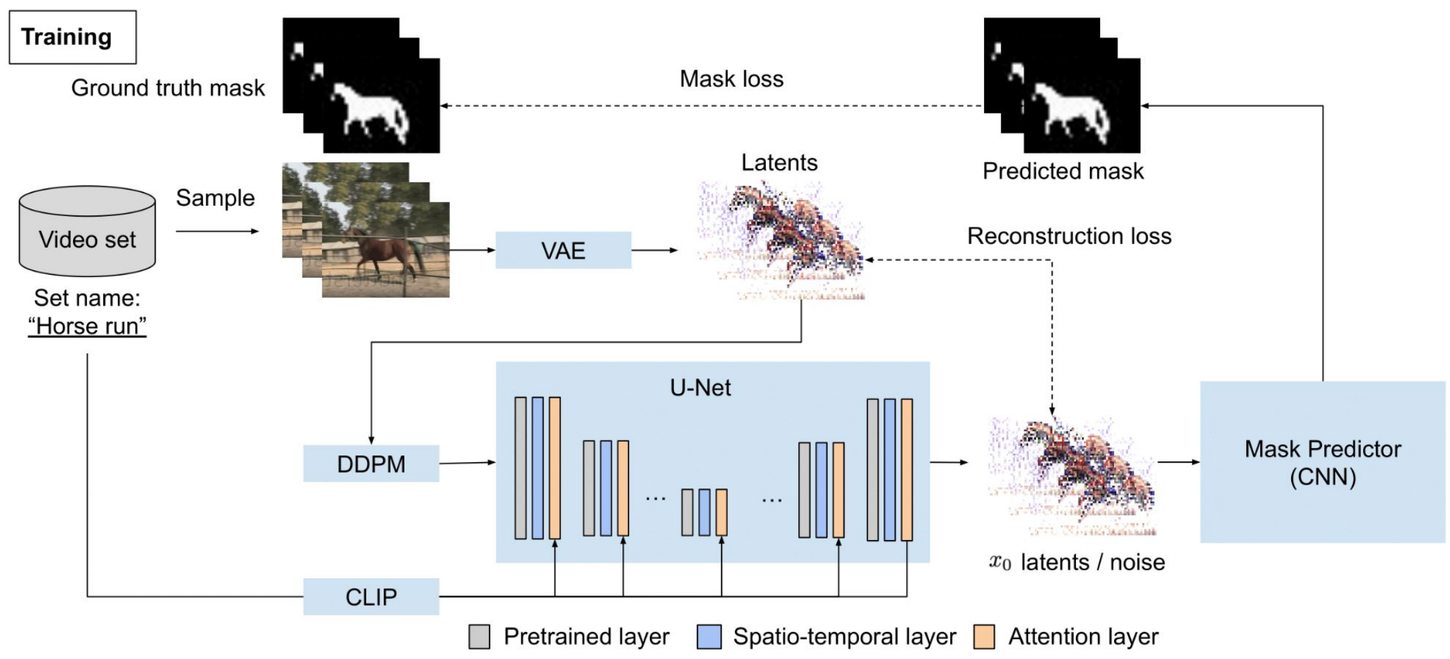

(a) The proposed 1D conv (b) Temporal-spatial motion learning layer

## Dataset

- Few-shot dataset from LAMP
- Off-the-shelf foreground masks

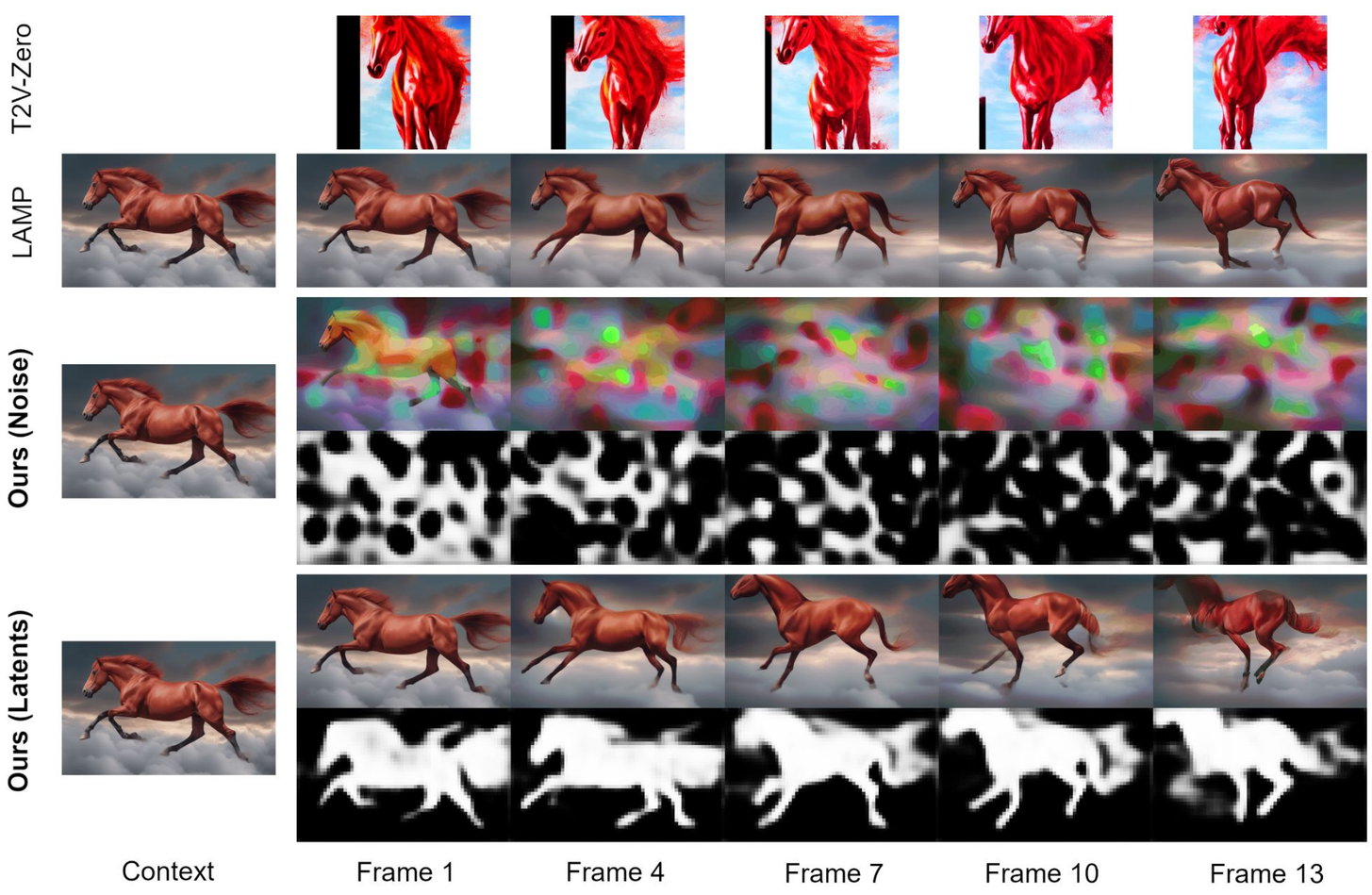


## Method




## Qualitative Results

*"A red horse runs in the sky"*




## Quantitative Results

- Metrics use CLIP embeddings and measure cosine similarity

| Model | Alignment ↑ | Consistency ↑ | Diversity ↓ |
|---|---|---|---|
| Text2Video-Zero [2] | 27.84 | 94.10 | **82.26** |
| LAMP [3] | 29.01 | **98.12** | 86.51 |
| Ours (Noise) | 23.37 | 94.91 | 96.66 |
| Ours (Latents) | **29.28** | 97.79 | 86.64 |

## References

[1] Icons taken from Noun Project.
[2] Khachatryan et al. Text2Video-Zero. ICCV 2023.
[3] Wu et al. LAMP: Learn A Motion Pattern for Few-Shot-Based Video Generation. arXiv 2023.