Few-Shot Video Generation with Recurrent Motion Prediction using Pre-trained Image Diffusion Models

Ryan Li
Student ID: 06375418
lansong@stanford.edu

Churan He
Student ID: 06684508
lansong@stanford.edu

Yingying Chen
Student ID: 06287845
ych@stanford.edu

Summary

- **Objective:** Text-to-video generation with few-shot finetuning
- **Goal:** ***Guidance-free*** fewshot text-to-video generation with high video ***fidelity*** and ***temporal consistency***

Background

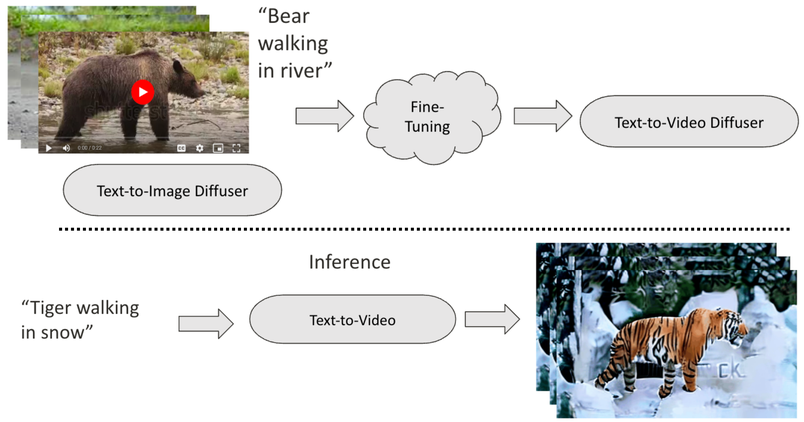
- Traditional Text-to-Video (T2V) training paradigms are ***prohibitively expensive.***
- To address these computational challenges, attempts have been made to finetune existing Text-to-Image (T2I) diffusers for T2V generation
- However, videos generated from existing T2I finetuning methods suffer from ***low temporal consistency*** and ***poor generalizability***, relying heavily on structural guidance or first-frame conditioning

Dataset

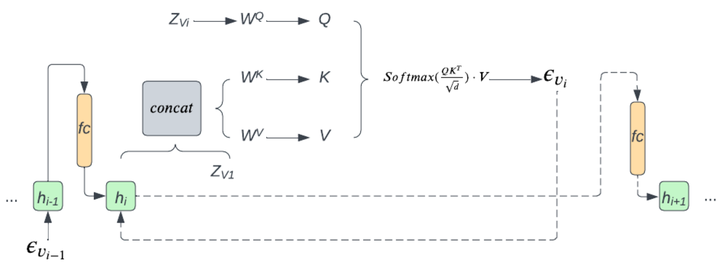
Selected 12 categories of few-shot training videos, each contains 8-16 videos and their corresponding text prompts.

- Videos: DAVIS dataset, Frozen in Time, and Shutterstock
- Text prompts: manually labeled based on video content and dataset labels

Technical Methods



- Given input video + text prompt describing the video, finetune a pretrained T2I model to generate a new, temporally consistent video given an edited text prompt



Recurrent Causal Attention Layer - a RNN-like spatio-temporal attention "memory" h_i that is updated with the features throughout all the frames.

Structured Sampling from Noise - the noise used to generate the following frame based upon the noise used to generate the previous frame

Evaluations & Results

- Evaluated temporal consistency, fidelity, and overall video quality via user studies.