

Presenter

**Pushpen Bikash
Goala**

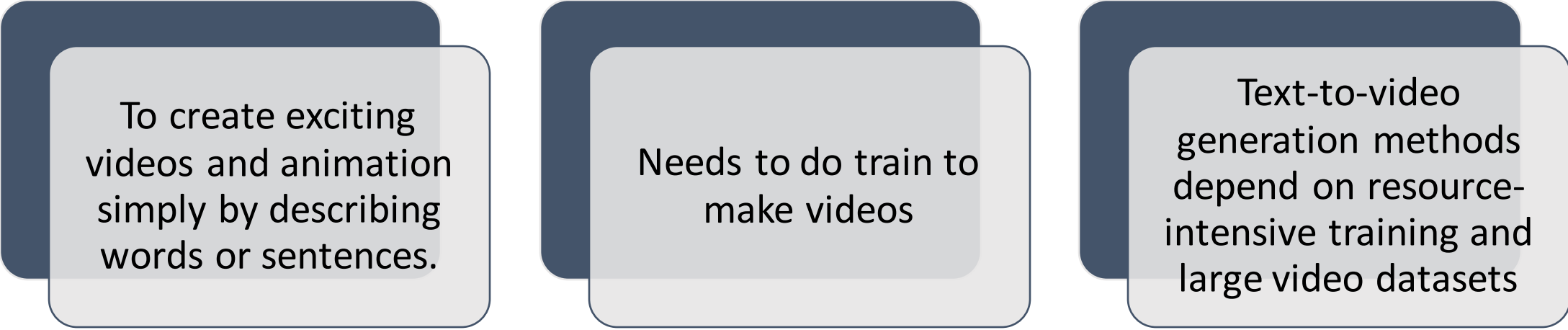
2nd year
Computer Science
Ph.D. Student at
GC CUNY

Text2Video-Zero: Text-to-Image Diffusion Models are Zero-Shot Video Generators

Levon Khachatryan^{1*} Andranik
Movsisyan^{1*} Vahram Tadevosyan^{1*} Roberto
Henschel^{1*} Zhangyang Wang^{1,2} Shant
Navasardyan¹ Humphrey Shi^{1,3,4} ¹Picsart AI
Research (PAIR) ²UT Austin ³U of Oregon ⁴UIUC

<https://github.com/picsart-ai-research/text2video-zero>

Motivation



To create exciting videos and animation simply by describing words or sentences.

Needs to do train to make videos

Text-to-video generation methods depend on resource-intensive training and large video datasets

Motivation

A novel zero-shot text-to-video generation task and a cost-effective solution.

Training Free

Does not need extensive computational power or multiple GPUs

Makes video generation more accessible

Contribution

- A text-guided video generation and editing
- Only use a pre-trained text-to-image diffusion model without any further fine-tuning or optimization
- Consistent generation
 - Via encoding motion dynamics in the latent codes
 - Reprogramming each frame's self-attention using a new cross-frame attention
- The approach can be extended to
 - Conditional and content-specialized video generation
 - Instruction-guided video editing



Related Work

Initial text-to-image synthesis methods were relying on techniques like template-based generation and feature matching

The advent of GANs led to more advanced deep learning-based approaches

- StackGAN, AttnGAN, and MirrorGAN
- Generate image with improved quality and diversity

Transformers further advanced text-to-image synthesis

- Dall-E, Parti, and Make-a-Scene

Diffusion models, significantly enhance text-to-image synthesis quality

- GLIDE, Dall-E 2, LDM/SD, Imagen, and Versatile Diffusion

Adapting diffusion models for video generation is challenging due to their probabilistic generation procedure, making temporal consistency difficult to maintain

Related Work

In Text-to-video synthesis current approaches leveraging autoregressive transformers and diffusion processes for the generation

- NUWA, Phenaki, CogVideo, Video Diffusion Models (VDM), Imagen Video, Make-A-Video, Gen-1, and Tune-A-Video
- Require substantial computing resources and training

The authors approach, in contrast, is training-free and does not need extensive computational power or multiple GPUs, making video generation more accessible

- Relies on an optimization process and is heavily dependent on the reference video

Problem Formulation

- Current text-to-video synthesis methods necessitate either expensive training on large-scale text-video paired data or tuning on a reference video
- zero-shot text-to-video
 - Given a text description τ and a positive integer $m \in \mathbb{N}$, m = number of frames
 - The objective is to design a function F
 - F produces video frames $V \in \mathbb{R}^{m \times H \times W \times 3}$ (for a predefined resolution $H \times W$) with temporal consistency
 - F should not require any training or fine-tuning on a video dataset.

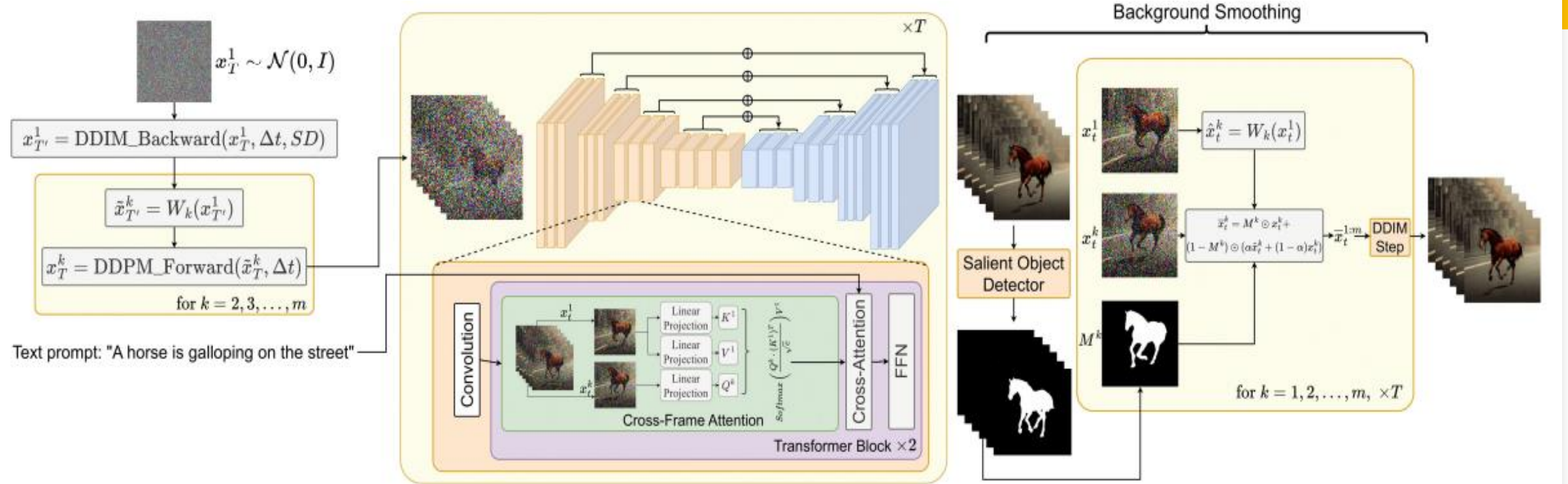
Method

- The authors use the text-to-image synthesis capabilities of Stable Diffusion (SD)
- To generate videos rather than images, SD should operate on sequences of latent codes
- The naive approach involves independently sampling m latent codes from a standard Gaussian distribution and applying DDIM (Denoising diffusion implicit models) sampling to obtain the corresponding tensors
 - Leads to random generation of images that share only the semantics described by τ
 - Lacks object appearance or motion coherence

Method

- To address this issue, the authors propose
 - Motion dynamics between the latent codes
 - To maintain global scene time consistency
 - Cross-frame attention mechanism
 - To preserve the appearance and identity of the foreground object

Method



Method

- Steps of Motion Dynamics
 - Randomly sample the latent code of the first frame
 - Perform $\Delta t \geq 0$ DDIM backward steps on the latent code using the SD model and get the corresponding latent.
 - Define a direction δ for the global scene and camera motion, with default values being the main diagonal direction.
 - For each frame to be generated, compute the global translation vector δk and to control the global motion amount use the hyperparameter λ .
 - Apply the constructed motion flow to the first latent and denote the resulting sequence.
 - Perform Δt DDPM (Denoising diffusion probabilistic models) forward steps on each of the latent and get the corresponding latent codes.

The latent codes lead to improved temporal consistency of the global scene and background

Method

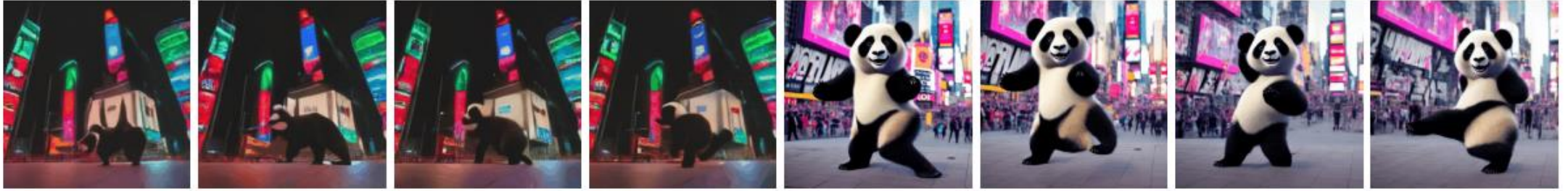
- Initial latent codes do not fully constrain colors, identities, or shapes
 - Resulting in temporal inconsistencies, especially for the foreground object.
- Reprogramming *Cross-Frame Attention*
 - Modify the pre-trained SD by replacing each self-attention layer with a cross-frame attention with attention for each frame being on the first frame
- *In the original SD architecture, each self-attention layer takes a feature map and linearly projects it into query, key, and value features, computing the layer output using a specific formula.*
- Here each attention layer receives m inputs, producing m queries, keys, and values.
- By using cross-frame attention
 - The appearance and structure of objects and background, as well as identities, are carried over from the first frame to subsequent frames
 - Significantly increasing the temporal consistency of the generated frames.

Implementation

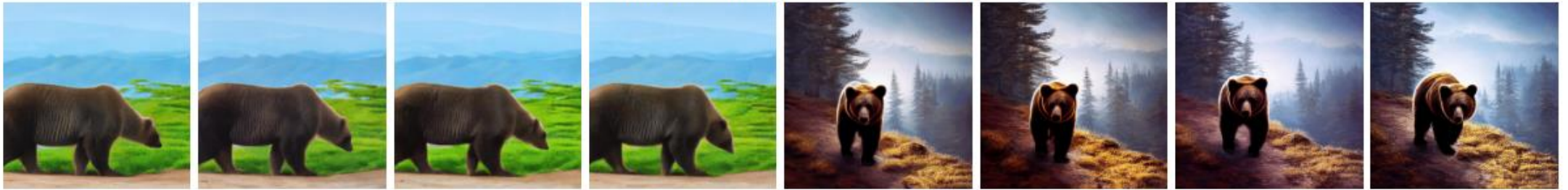
- Take the Stable Diffusion code with its pre-trained weights from version 1.5 as the basis and implement their modifications.
- Generate $m = 8$ frames with 512×512 resolution for each video.
- For a conditional generation, they use the codebase3 of ControlNet and apply their methods to the basic diffusion process.
- For Video Instruct-Pix2Pix, they use the codebase4 of Instruct Pix2Pix.

Results

- Text2Video-Zero successfully generates time-consistent videos with great temporal consistency and identity preservation of the objects
- The authors compare their method with two publicly available baselines CogVideo and Tune-A-Video
- CogVideo
 - Compare in pure text-guided video synthesis settings
 - CLIP score: Indicates video-text alignment
 - Randomly take 25 videos generated by CogVideo and synthesize corresponding videos using the same prompts according to their method.
 - The CLIP score for CogVideo is 29.63 and for their method is 31.19, respectively
 - Better text to video alignment



(a) A panda dancing on times square.



(b) A bear walking on a mountain.



(c) A cat running on the lawn.

Qualitative comparison between CogVideo and proposed method

Results

- Compare with Tune-A-Video and Instruct-Pix2Pix
 - Instruct-Pix2Pix, Tune-A-Video shows
 - A good editing performance per frame, it lacks temporal consistency
 - Proposed model solve this issue

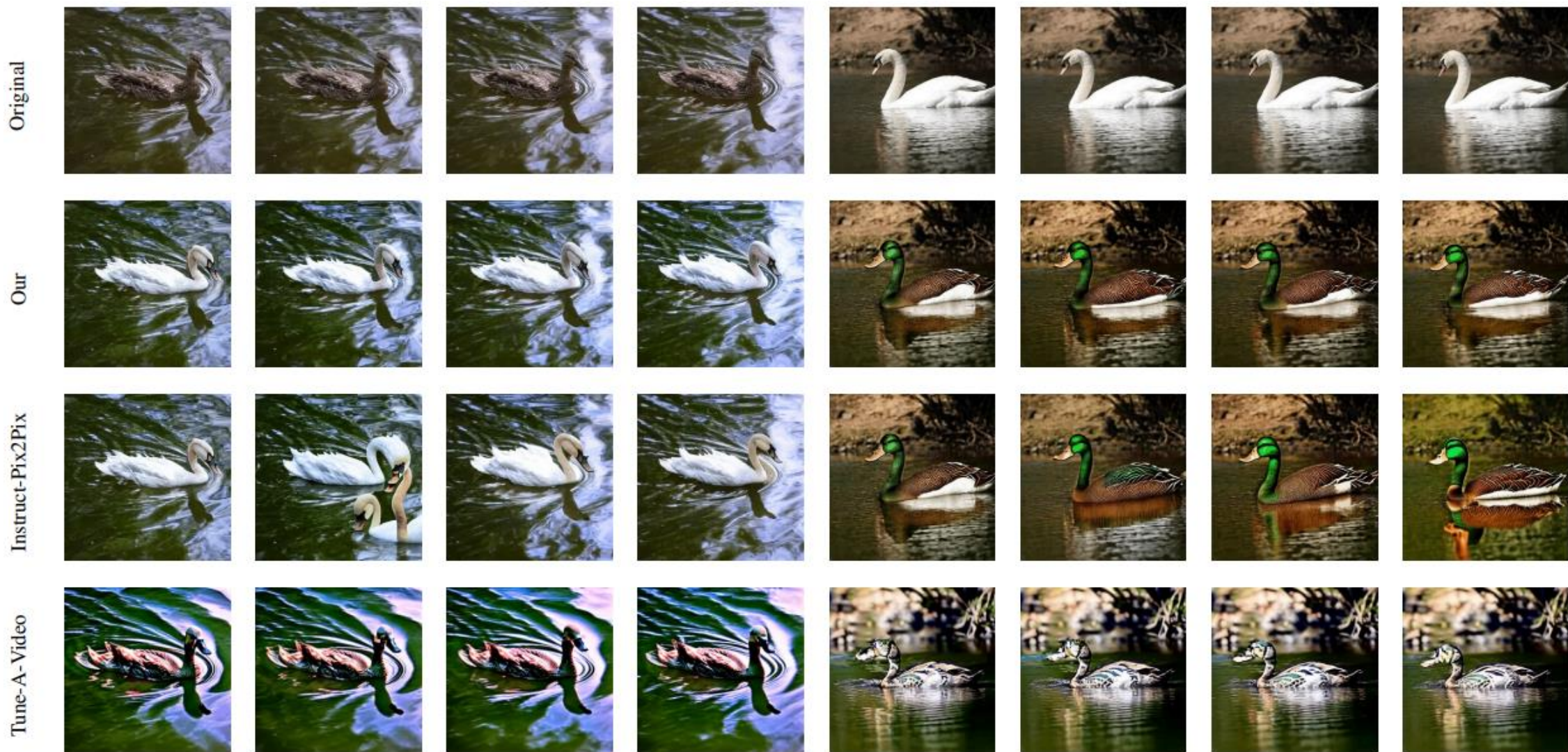


Figure 24: Text guided video editing using our method compared to Instruct-Pix2Pix [2] frame-by-frame and Tune-A-Video [41]. The left half of each row is generated with the instruction “replace mallard with swan” for our method and Instruct-Pix2Pix and “a **swan** swimming on water” for Tune-A-Video. The right half of each row is generated with the text prompt “replace swan with mallard” and “a **mallard** swimming on water” respectively.

Conclusion

- Proposed a time-consistent video generation method by utilizing the text-to-image diffusion models
- Does not require any optimization or fine-tuning making text-video generation affordable
- Demonstrated the effectiveness of their method for various applications, including
 - Conditional and specialized video generation and
 - Video Instruct-Pix2Pix, which is instruction-guided video editing