CVPR SEATTLE, WA JUNE 17-21, 2024

TRIP: Temporal Residual Learning with Image Noise Prior for Image-to-Video Diffusion Models

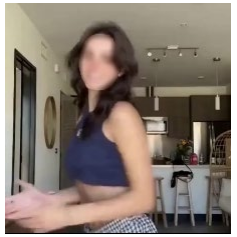
Zhongwei Zhang[1], Fuchen Long[2], Yingwei Pan[2], Zhaofan Qiu[2], Ting Yao[2], Yang Cao[1], Tao Mei[2]

[1]University of Science and Technology of China [2]HiDream.ai Inc.

Backgrounds

Ref Images:



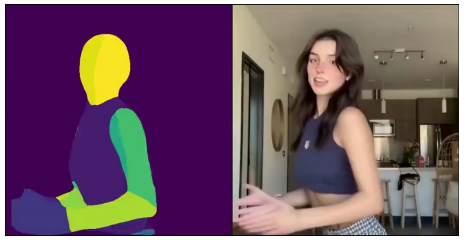


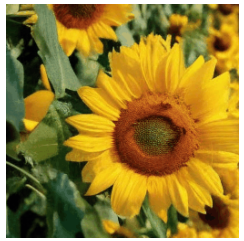

Unconditional、Motion driven(trajectory, motion sequence)、Text Prompt driven

Generated Videos:


CINN[1]


DragNUWA[2]


MagicAnimate[3]

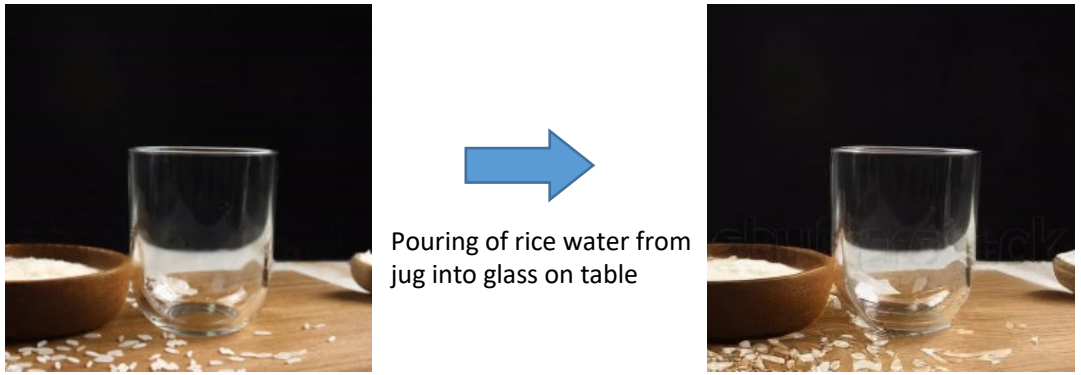


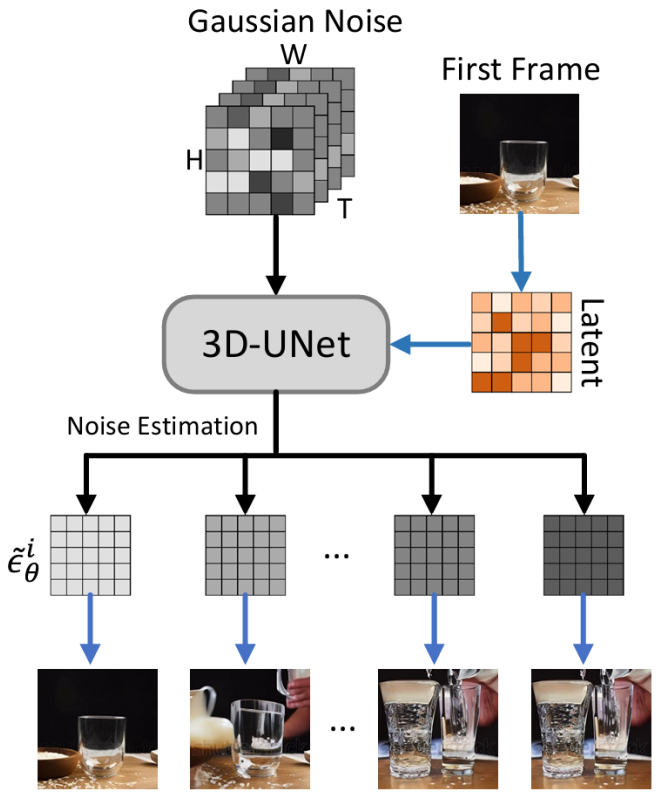
VideoComposer[4]

Our focus is: Text driven I2V task

[1] Dorkenwald, Michael, et al. Stochastic image-to-video synthesis using cinns. In CVPR, 2021.
[2] Yin, Shengming, et al. Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory. arXiv preprint arXiv:2308.08089, 2023.
[3] Xu, Zhongcong, et al. MagicAnimate: Temporally Consistent Human Image Animation using Diffusion Model. In CVPR, 2024.
[4] Alibaba Group. VideoComposer: Compositional Video Synthesis with Motion Controllability. In NeurIPS, 2023.

Limitation





(a) Independent noise prediction

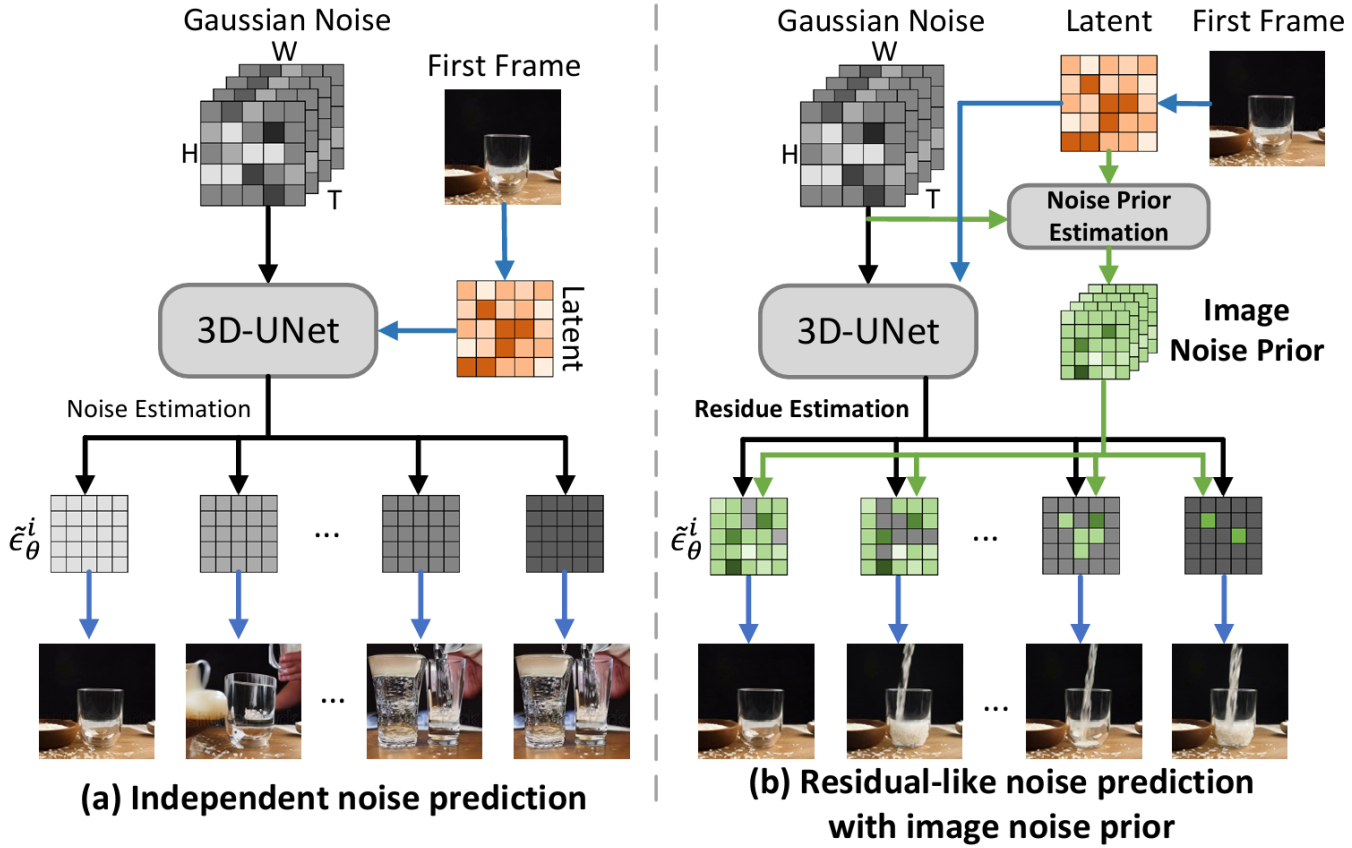
Challenge
- Align with
  - Text prompt
  - Given Image
- Temporal consistency

➢ Typical Image-to-Video Diffusion[1]:

- **Independent noise prediction** of each frame
- **Ignoring the inherent relation** between video frames
- Lack of **temporal coherence** modeling

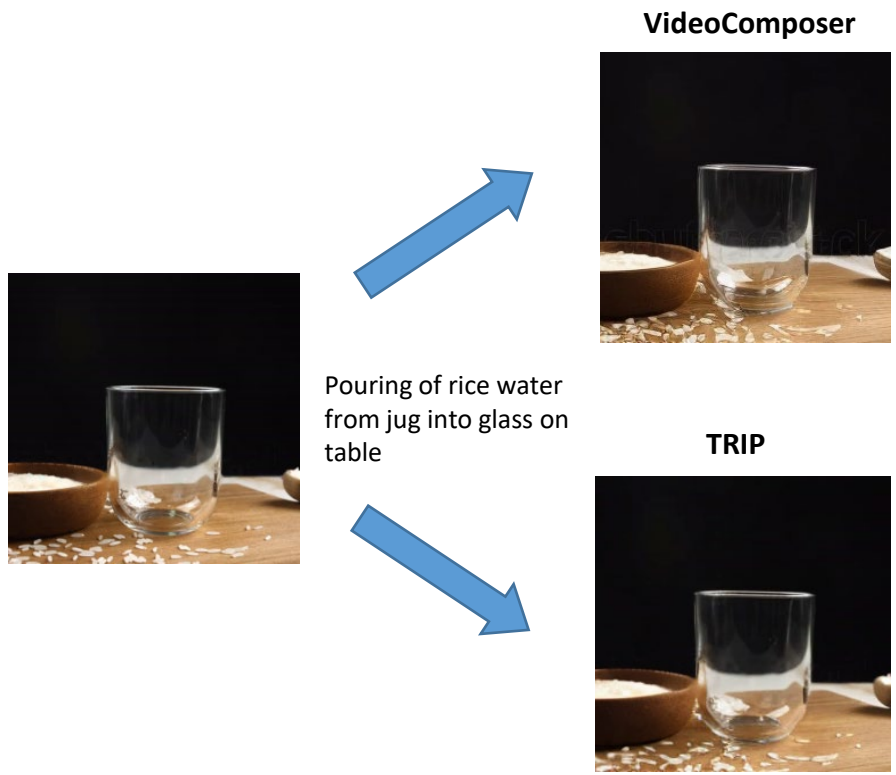
[1] Alibaba Group. VideoComposer: Compositional Video Synthesis with Motion Controllability. In NeurIPS, 2023.

Temporal Residual Learning with Image Noise Prior



(a) Independent noise prediction

(b) Residual-like noise prediction with image noise prior



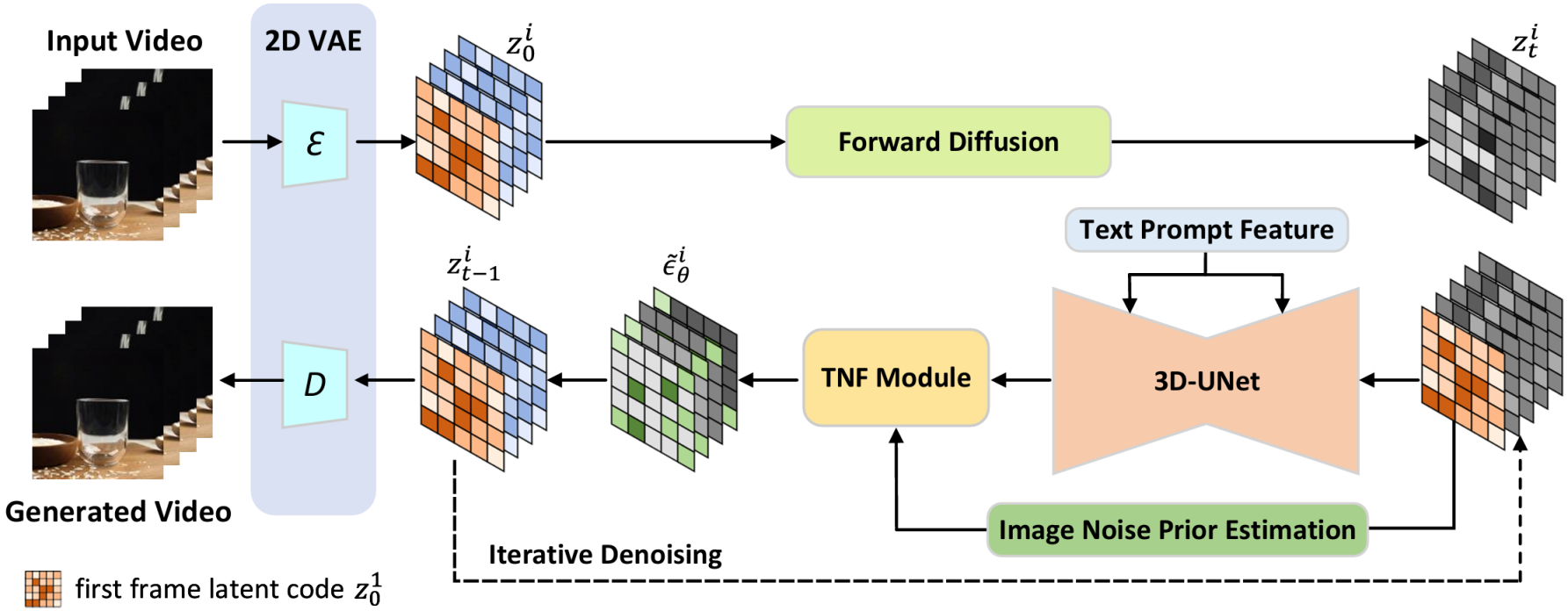
- Temporal Residual Learning with Image Noise Prior (TRIP): residual-like dual-path scheme
  - Take image noise prior as reference noise to amplify alignment across frames
  - Residual noise learning to capture motion dynamics
  - Attention mechanism for reference and residual noise merging

Better Temporal Consistency

Temporal Residual Learning with Image Noise Prior

CVPR SEATTLE, WA JUNE 17-21, 2024

➢ Framework:



Video Diffusion: $z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \epsilon \sim N(0, I)\ ,\ z_t = \{z_t^i\}_{i=1}^N$

Typical noise formulation:

$$z_0^i = \frac{z_t^i - \sqrt{1-\bar{\alpha}_t}\epsilon_t^i}{\sqrt{\bar{\alpha}_t}} \qquad \epsilon_t^i \leftarrow \epsilon_\theta^i(z_t, t, c)$$

TRIP noise formulation:

$$z_0^1 = \frac{z_t^i - \sqrt{1-\bar{\alpha}_t}\epsilon_t^{i\to 1}}{\sqrt{\bar{\alpha}_t}} \qquad \tilde{\epsilon}_\theta^i = \varphi(\epsilon_t^{i\to 1}, \Delta\tilde{\epsilon}_t^i)$$

Image noise prior — *Temporal residual noise*

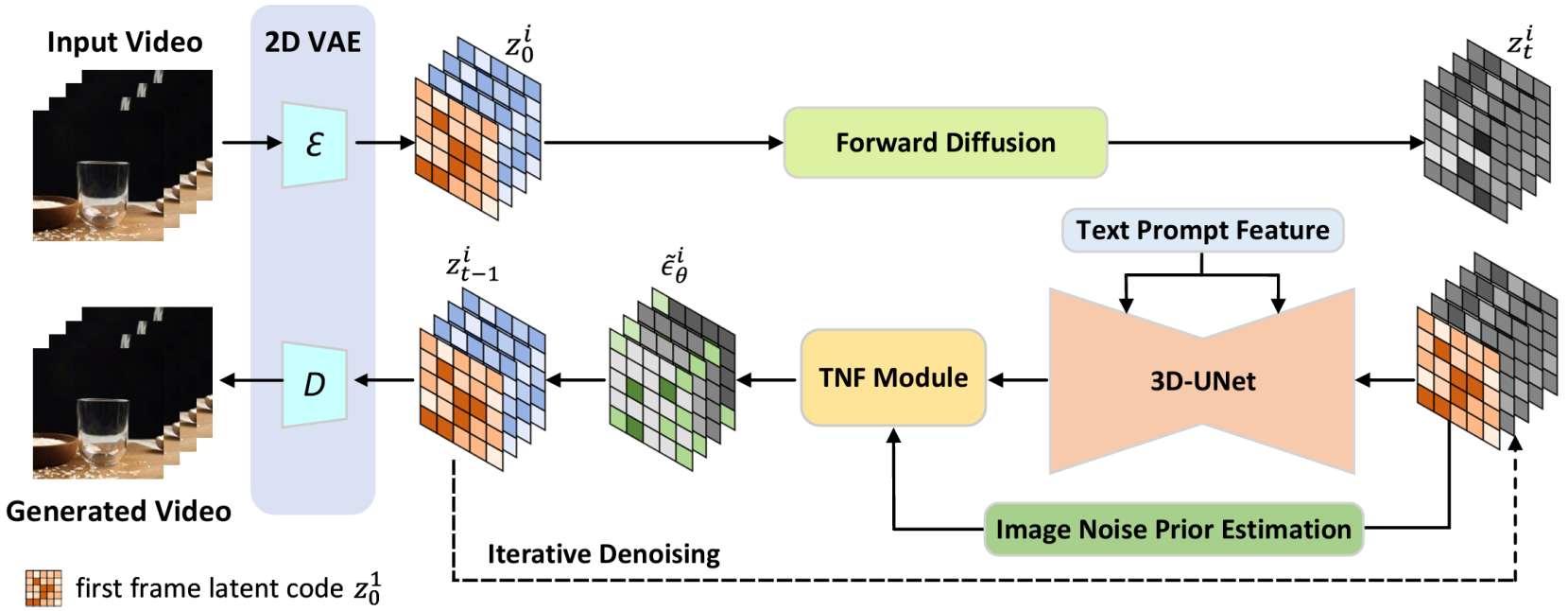
$$\epsilon_t^{i\to 1} = \frac{z_t^i - \sqrt{\bar{\alpha}_t} z_0^1}{\sqrt{1-\bar{\alpha}_t}} \qquad \epsilon_t^i \leftarrow \tilde{\epsilon}_\theta^i$$

➢ Key Steps:

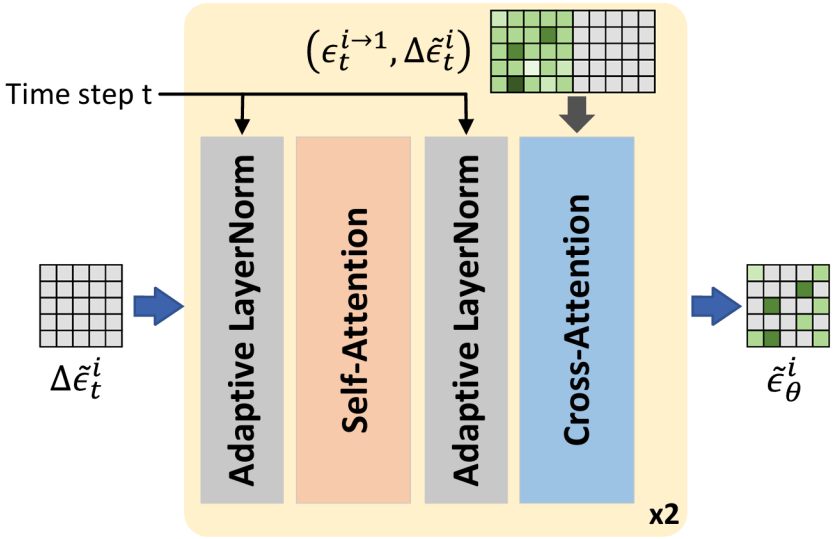
- Image Noise Prior Computation
- Temporal Residual Noise Prediction
- Attention mechanism for reference and residual noise merging(TNF module)

Temporal Residual Learning with Image Noise Prior

- Framework:



- TNF(Temporal Noise Fusion Module):



- Training loss:

$$\tilde{\mathcal{L}} = \mathbb{E}_{\epsilon\sim\mathcal{N}(0,I),t,c,i}[\|\epsilon_t^i - \tilde{\epsilon}_\theta^i\|^2]$$

$$\tilde{\epsilon}_\theta^i = \varphi(\epsilon_t^{i-1}, \Delta\tilde{\epsilon}_t^i)$$

Image noise prior

Temporal residual noise

Experimental Analysis

Table 1. Performances of F-Consistency (F-Consistency$_4$: consistency among the first four frames, F-Consistency$_{all}$: consistency among all frames) and FVD on WebVid-10M.

Approach	F-Consistency$_4$ (↑)	F-Consistency$_{all}$ (↑)	FVD (↓)
T2V-Zero [32]	91.59	92.15	279
VideoComposer [60]	88.78	92.52	231
TRIP	**95.36**	**96.41**	**38.9**

Table 2. Performances of averaged FID and FVD over four scene categories on DTDB dataset.

Approach	Zero-shot	FID (↓)	FVD (↓)
AL [11]	No	65.1	934.2
cINN [9]	No	31.9	451.6
TRIP	Yes	**24.8**	**433.9**

Table 3. Performances of FID and FVD on MSR-VTT dataset.

Approach	Model Type	FID (↓)	FVD (↓)
CogVideo [28]	T2V	23.59	1294
Make-A-Video [52]	T2V	13.17	-
VideoComposer [60]	T2V	-	580
ModelScopeT2V [59]	T2V	11.09	550
VideoComposer [60]	I2V	31.29	208
TRIP	I2V	**9.68**	**91.3**

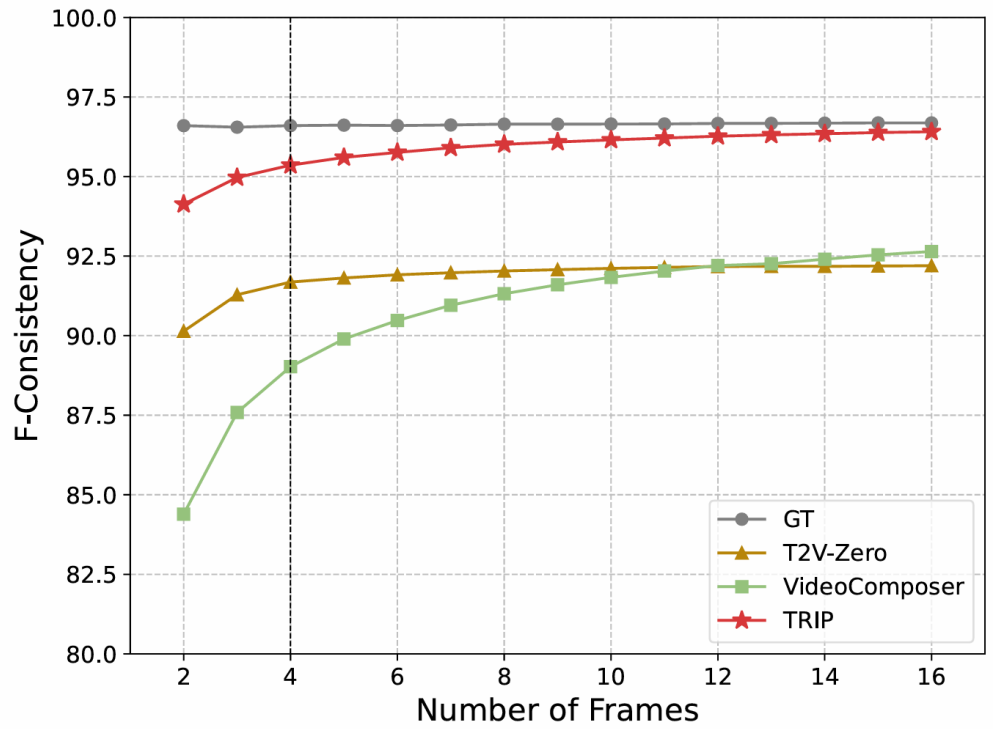


Figure 4. Performance comparisons of F-Consistency by using different number of frames on WebVid-10M.

Experimental Analysis

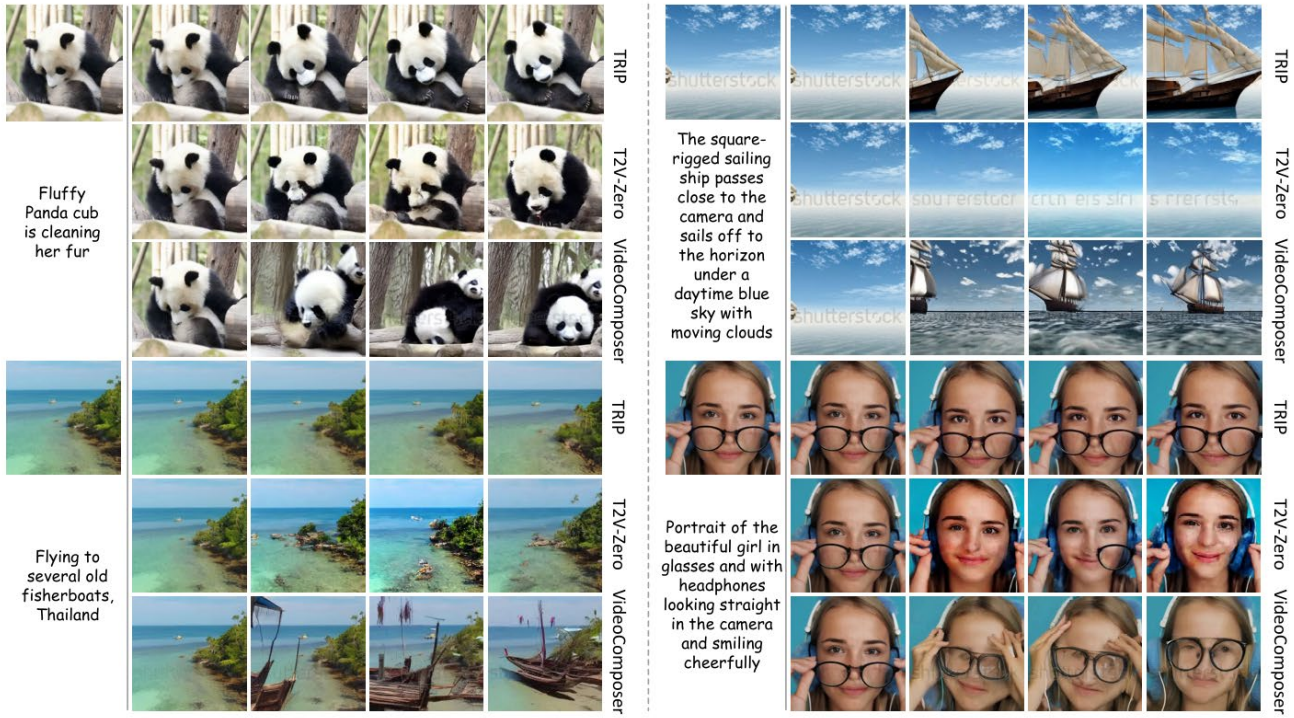


Figure 5. Examples of I2V generation results by three methods (VideoComposer, T2V-Zero, and our TRIP) on WebVid-10M dataset. We uniformly sample four frames of each generated video for visualization.

Better Temporal Consistency

Table 6. Evaluation of temporal residual learning in terms of F-Consistency and FVD on WebVid-10M dataset.

Model	F-Consistency$_4$ (↑)	F-Consistency$_{all}$ (↑)	FVD (↓)
TRIP$^-$	94.66	95.92	39.9
TRIP^W	95.22	95.96	43.0
TRIP	**95.36**	**96.41**	**38.9**

Figure 8. Visualization of I2V example with text prompt "Young male tourist taking a photograph of old city on sunset" by using variants of our TRIP.

More Results

Image T2V-Zero VideoComposer TRIP

A male fencer adjusts his epee mask and prepares to duel with his sparring partner in slow motion

Image T2V-Zero VideoComposer TRIP

Glass of red wine being poured

Image T2V-Zero VideoComposer TRIP

Giant spider crab

Image T2V-Zero VideoComposer TRIP

Few big purple plums rotating on the turntable water drops appear on the skin during rotation isolated on the white background, close-up macro

Image T2V-Zero VideoComposer TRIP

Asian elephants, thailand

Image T2V-Zero VideoComposer TRIP

Close-up view national flag of Iceland waving in the wind on a blue sky background without clouds, national symbol consists of a blue background bearing a red cross

More Results

Two parrots are sitting on a branch in the forest(SD-XL)

A castle in the middle of a forest with a river running through it(SD-XL)

WALL-E on the beach with a plant in his hand(SD-XL)

A fox is walking in the woods at night(SD-XL)

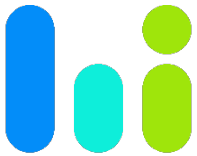
A ship sails in the sea(SD-XL)

Crabs with bubbles crawls on the beach with sunset in the background(SD-XL)

A bear is playing a guitar in front of a tank in the snow(SD-XL)

A giant robot with pumpkins on it(SD-XL)

Halloween pumpkin with glowing eyes and smoke coming out of it(SD-XL)



HiDream.ai

Thanks!

Contact: zhwzhang@mail.ustc.edu.cn

Project Page: https://trip-i2v.github.io/TRIP/