RUTGERS
THE STATE UNIVERSITY OF NEW JERSEY

Meta

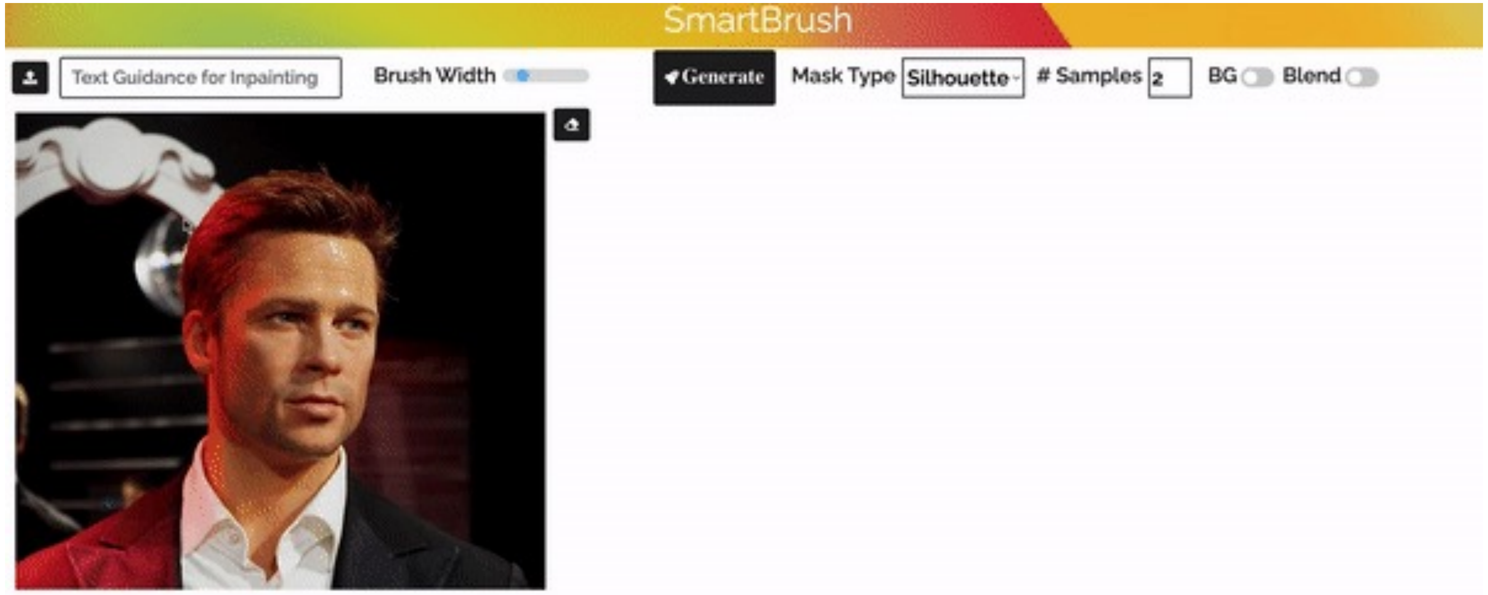
AVID: Any-Length Video Inpainting with Diffusion Model

Zhixing Zhang, Bichen Wu, Xiaoyan Wang, Yaqiao Luo, Luxin Zhang,
Yinan Zhao, Peter Vajda, Dimitris N. Metaxas, Licheng Yu

Text-to-image Diffusion Models

Diffusion Models have been shown to generate high-quality images according to input text prompts.

Diffusion Models for Image Inpainting



Mask Type Silhouette

Samples 2

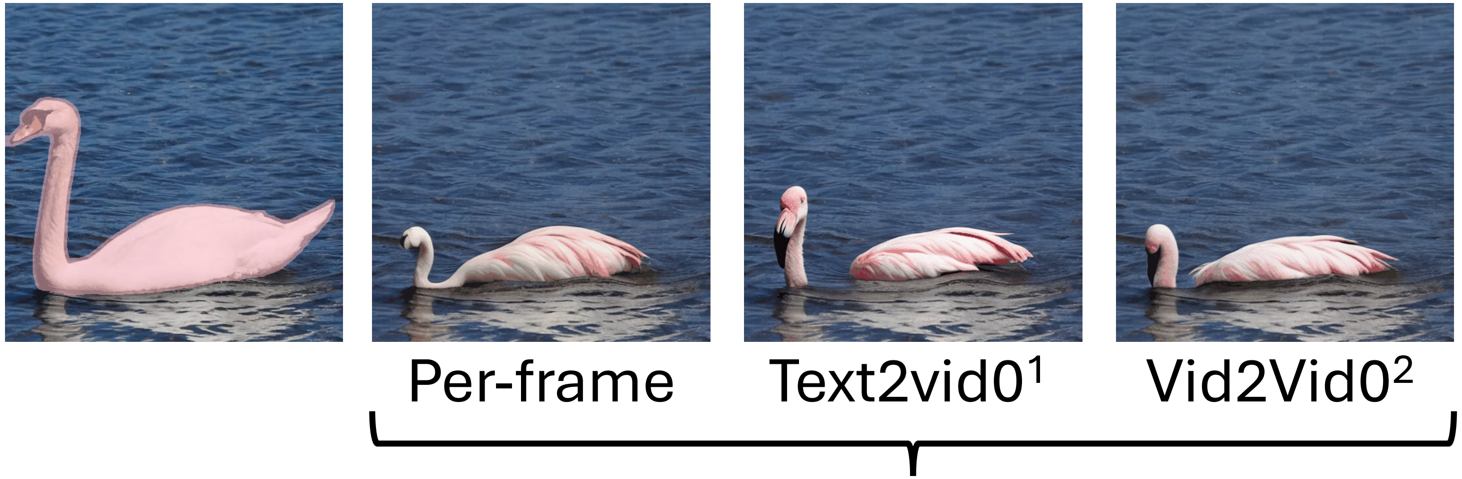
- Text prompt + mask region -> image inpainting
  - Object replacement
  - Re-texturing
  - …

[1] Xie, Shaoan, et al. "Smartbrush: Text and shape guided object inpainting with diffusion model." In CVPR 2023.

Can we do the same on videos?

Existing Methods

Object swap: *"A flamingo swimming in a lake."*



Per-frame

Text2vid0[1]

Vid2Vid0[2]

Training-free

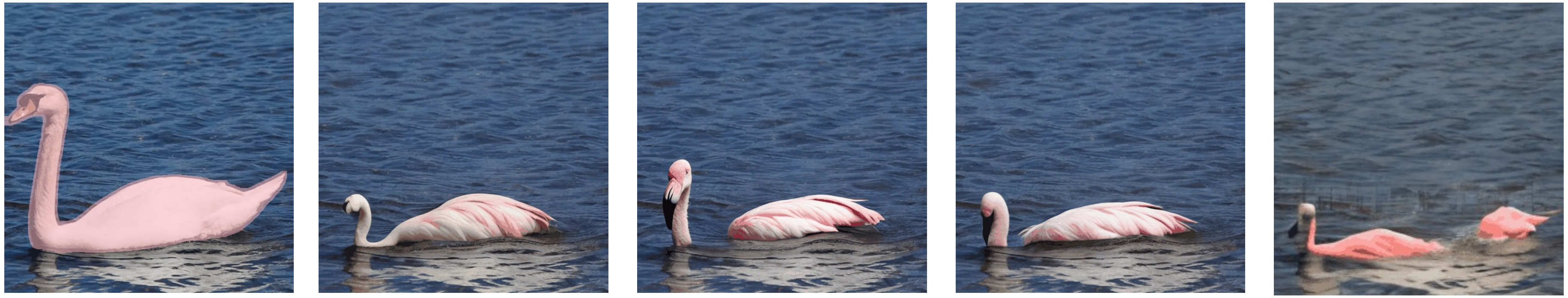
- Poor temporal consistency.

[1] Khachatryan, Levon, et al. "Text2video-zero: Text-to-image diffusion models are zero-shot video generators." In CVPR 2023.

[2] Wang, Wen, et al. "Zero-shot video editing using off-the-shelf image diffusion models."

Existing Methods

Object swap: *"A flamingo swimming in a lake."*



Per-frame Text2vid0[1] Vid2Vid0[2] VC[3]

Training-free

- Poor per-frame quality.
- Fixed video duration.

[1] Khachatryan, Levon, et al. "Text2video-zero: Text-to-image diffusion models are zero-shot video generators." In CVPR 2023.

[2] Wang, Wen, et al. "Zero-shot video editing using off-the-shelf image diffusion models."

[3] Wang, Xiang, et al. "Videocomposer: Compositional video synthesis with motion controllability." In NeurIPS 2023.

Challenges

- Temporal consistency

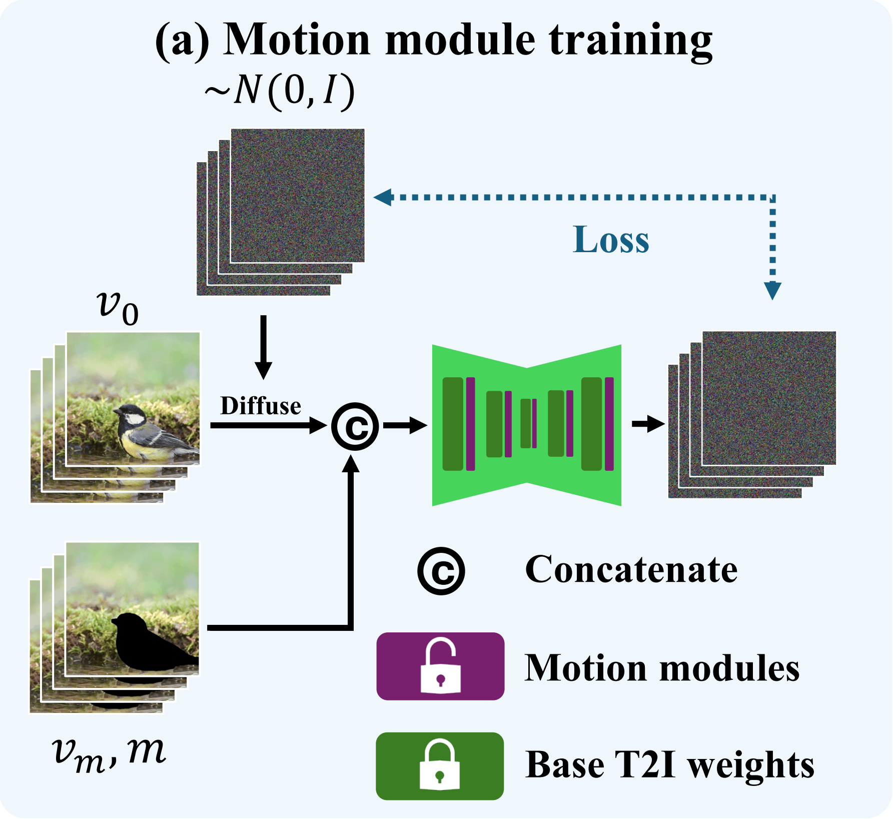
- Various editing types -> different levels of structural fidelity
  - Object swap (e.g. sedan->sport cat) ❌
  - Retexturing (e.g. white coat-> red one) ✅
  - Uncropping (e.g. 256x512->512x512) ❌

- Arbitrary duration

Method Overview

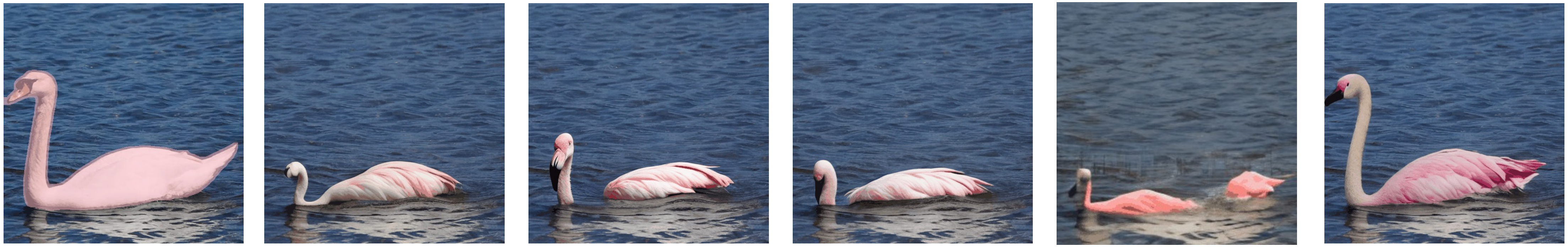
- Temporal consistency -> motion modules
- Various fidelity requirements -> adjustable structure guidance
- Arbitrary duration -> zero-shot any-length video inference
  - Temporal MultiDiffusion
  - Middle-frame Attention Guidance

Motion Modules



Motion Modules

Object swap: *“A flamingo swimming in a lake.”*



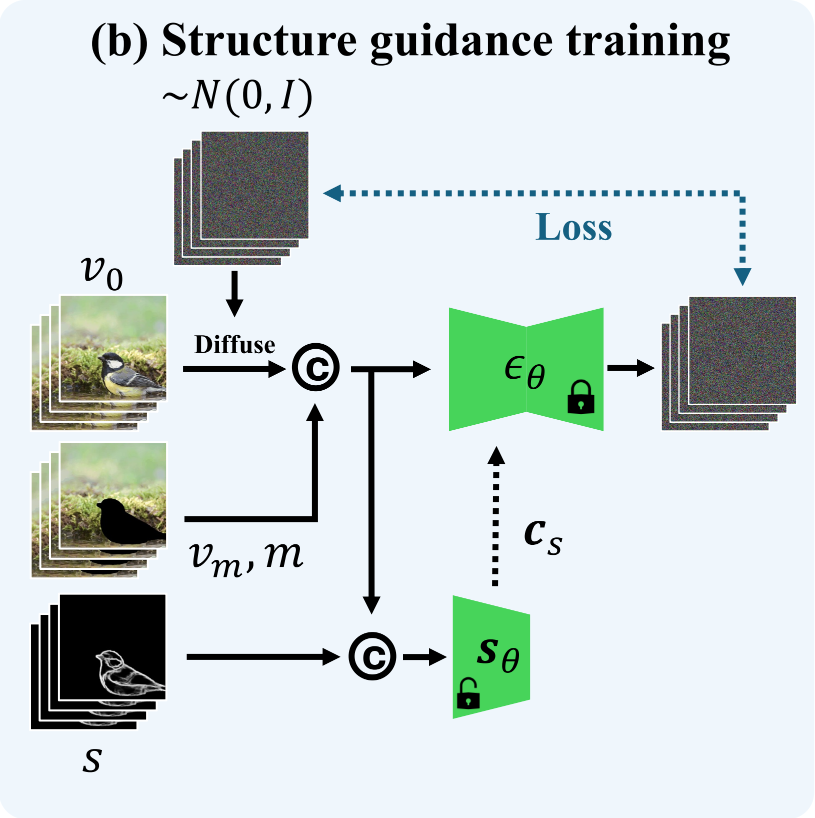
Per-frame Text2vid0[1] Vid2Vid0[2] VC[3] Ours

[1] Khachatryan, Levon, et al. "Text2video-zero: Text-to-image diffusion models are zero-shot video generators." In CVPR 2023.

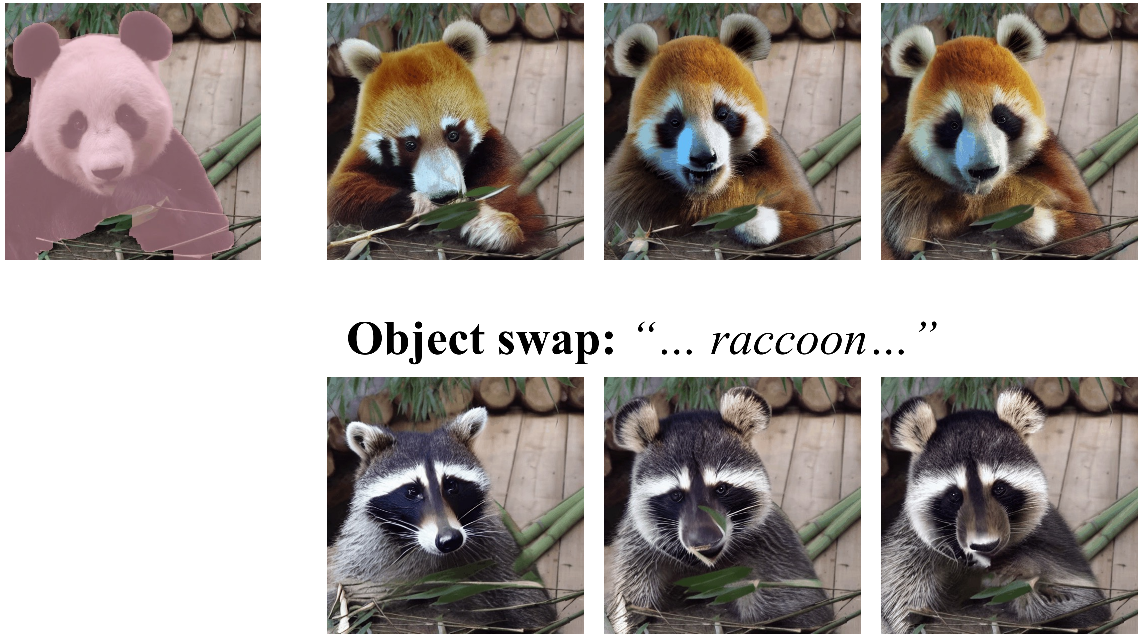
[2] Wang, Wen, et al. "Zero-shot video editing using off-the-shelf image diffusion models.”

[3] Wang, Xiang, et al. "Videocomposer: Compositional video synthesis with motion controllability." In NeurIPS 2023.

Structural Guidance



Structure Guidance



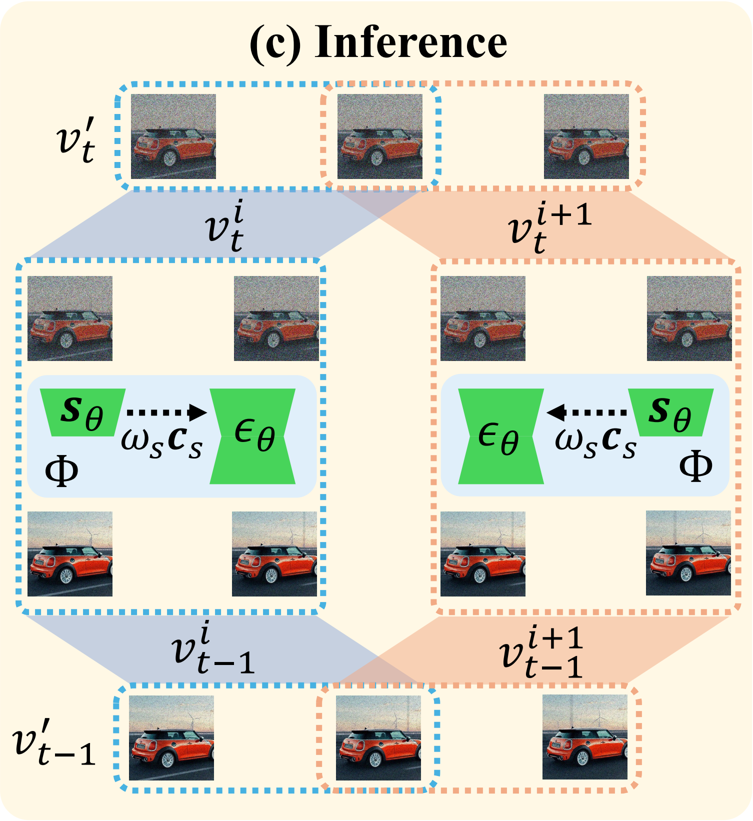
Source

Re-texturing: *"... golden furred..."*

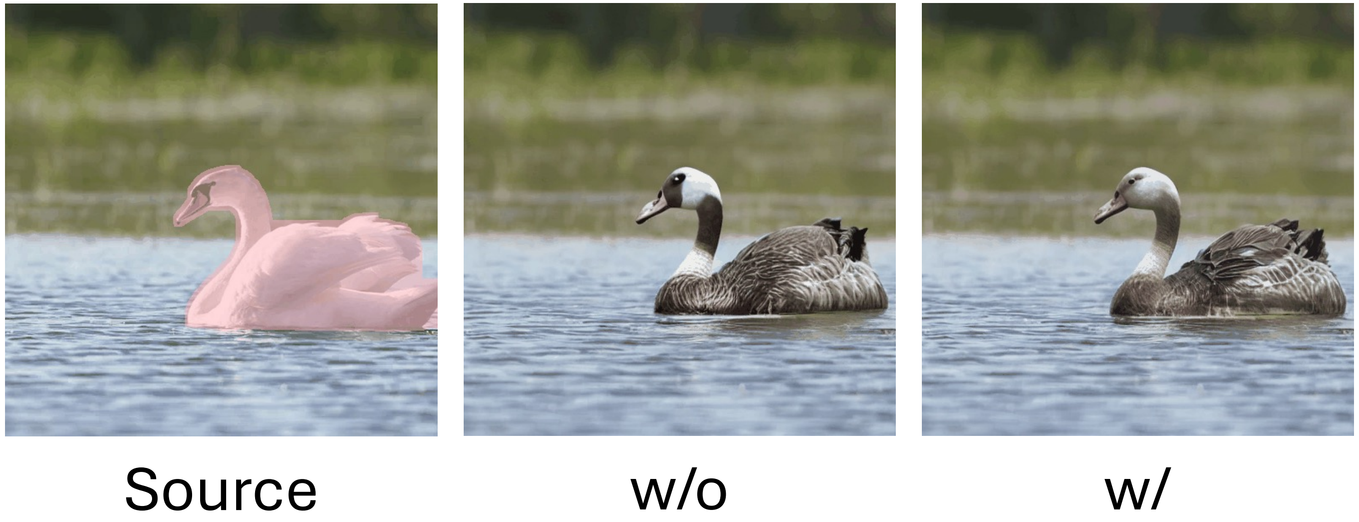
Object swap: *"... raccoon..."*

$\omega_s = 0.0$ $\omega_s = 0.5$ $\omega_s = 1.0$

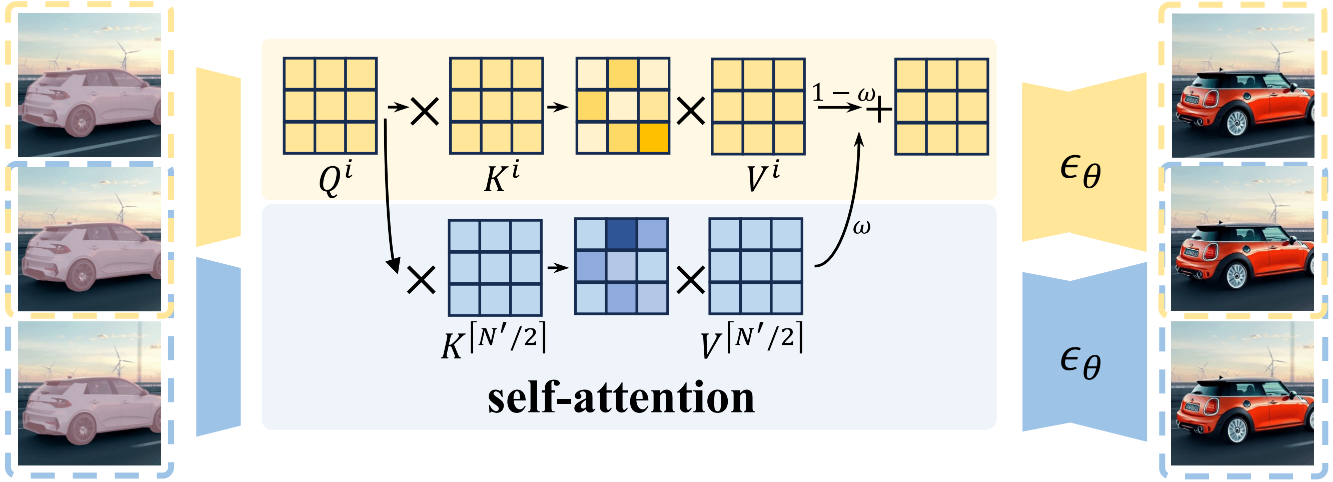
Temporal MultiDiffusion



"… a goose…"



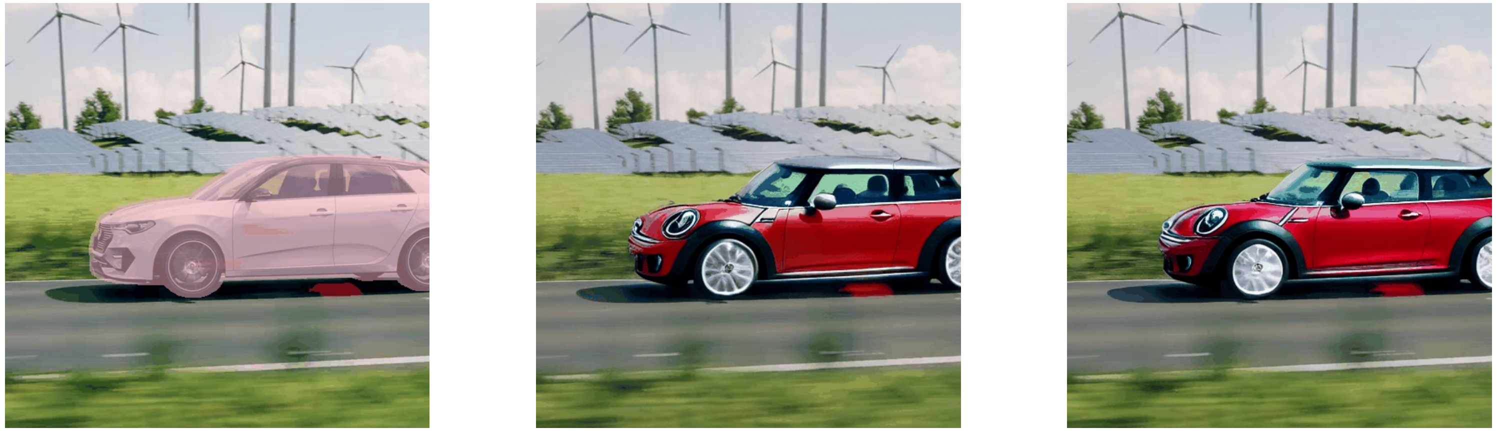
Middle-frame Attention Guidance



$$\text{Attention}(\psi^i) = \text{softmax}\left(\frac{Q^i {K^i}^T}{\sqrt{d}}\right) V^i \cdot (1-\omega) + \text{softmax}\left(\frac{Q^i {K^{\lceil N'/2 \rceil}}^T}{\sqrt{d}}\right) V^{\lceil N'/2 \rceil} \cdot \omega$$

Middle-frame Attention Guidance

“A MINI Cooper driving down the road.”

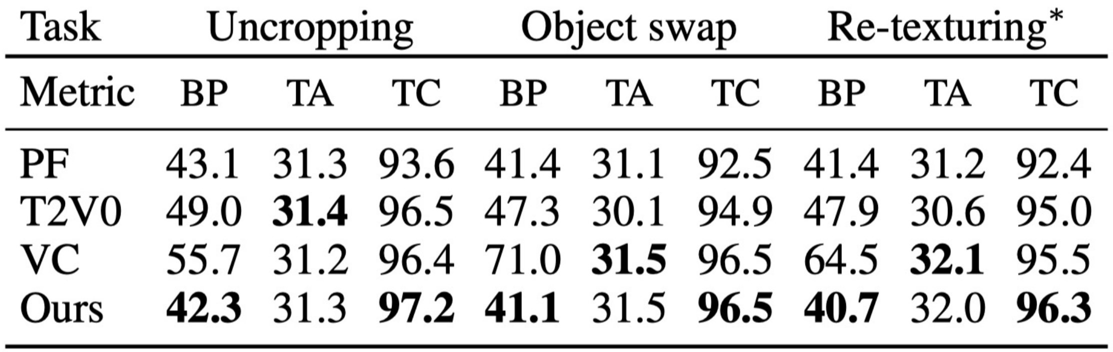


Source

w/o

w/

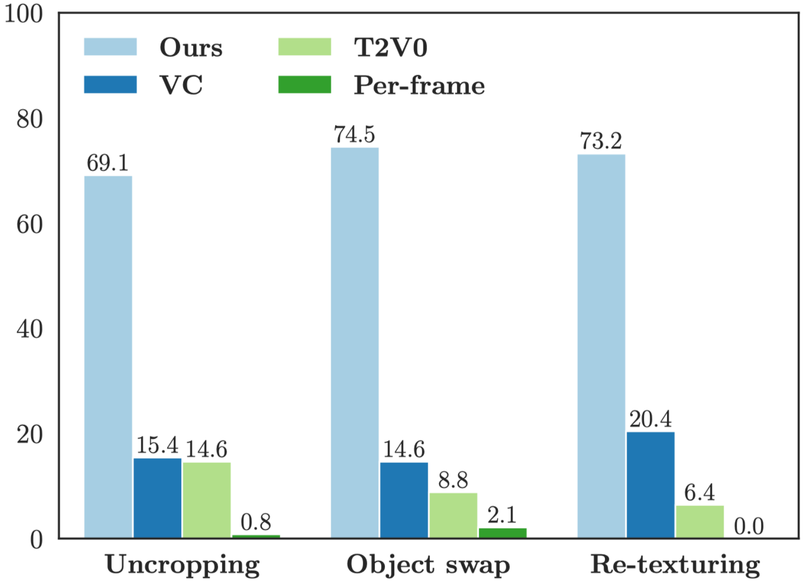
Comparisons



Task	Uncropping			Object swap			Re-texturing*		
Metric	BP	TA	TC	BP	TA	TC	BP	TA	TC
PF	43.1	31.3	93.6	41.4	31.1	92.5	41.4	31.2	92.4
T2V0	49.0	**31.4**	96.5	47.3	30.1	94.9	47.9	30.6	95.0
VC	55.7	31.2	96.4	71.0	**31.5**	96.5	64.5	**32.1**	95.5
Ours	**42.3**	31.3	**97.2**	**41.1**	31.5	**96.5**	**40.7**	32.0	**96.3**

- Background preservation (BP) ↓
- Text alignment (TA) ↑
- Temporal consistency (TC) ↑

Comparisons



User Study

Results

Future Work

- Better video foundation model
    - Geometrical understanding
    - Text-alignment
    - …

- More efficient generation

- Discontinuity handling

Thanks