FRESCO: Spatial-Temporal Correspondence for Zero-Shot Video Translation

Shuai Yang - Yifan Zhou - Ziwei Liu - Chen Change Loy
Peking University, S-Lab, Nanyang Technological University

Chen Change Loy

MMLab@NTU, S-Lab, Nanyang Technological University
Verified email at ntu.edu.sg - Homepage

Computer Vision Image Processing Machine Learning

FOLLOW

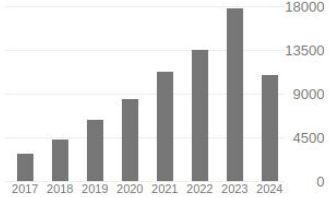
GET MY OWN PROFILE

TITLE	CITED BY	YEAR
Image Super-Resolution Using Deep Convolutional Networks C Dong, CC Loy, K He, X Tang IEEE Transactions on Pattern Analysis and Machine Intelligence 38 (2), 295 - 307	9642	2015
Learning a Deep Convolutional Network for Image Super-Resolution C Dong, CC Loy, K He, X Tang European Conference on Computer Vision, 184-199	6141	2014
ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks X Wang, K Yu, S Wu, J Gu, Y Liu, C Dong, Y Qiao, CC Loy Workshop and Challenge on Perceptual Image Restoration and Manipulation …	4279	2018
Accelerating the Super-Resolution Convolutional Neural Network C Dong, CC Loy, X Tang European Conference on Computer Vision, 391-407	3683	2016
MMDetection: OpenMMLab Detection Toolbox and Benchmark K Chen, J Wang, J Pang, Y Cao, Y Xiong, X Li, S Sun, W Feng, Z Liu, J Xu, ... arXiv preprint arXiv:1906.07155	2998	2019
WIDER FACE: A Face Detection Benchmark S Yang, P Luo, CC Loy, X Tang IEEE Conference on Computer Vision and Pattern Recognition, 5525-5533	2120	2016
Facial Landmark Detection by Deep Multi-task Learning Z Zhang, P Luo, CC Loy, X Tang European Conference on Computer Vision, 94-108	1778	2014
Learning to Prompt for Vision-Language Models K Zhou, J Yang, CC Loy, Z Liu International Journal of Computer Vision	1500	2022
Zero-Reference Deep Curve Estimation for Low-Light Image Enhancement C Guo, C Li, J Guo, CC Loy, J Hou, S Kwong, R Cong IEEE Conference on Computer Vision and Pattern Recognition	1423	2020
Hybrid Task Cascade for Instance Segmentation K Chen, J Pang, J Wang, Y Xiong, X Li, S Sun, W Feng, Z Liu, J Shi, ... IEEE Conference on Computer Vision and Pattern Recognition	1409	2019
PSANet: Point-wise Spatial Attention Network for Scene Parsing H Zhao, Y Zhang, S Liu, J Shi, CC Loy, D Lin, J Jia European Conference on Computer Vision	1189	2018

Cited by

VIEW ALL

	All	Since 2019
Citations	79246	68611
h-index	115	109
i10-index	249	246



Public access

VIEW ALL

0 articles — not available

77 articles — available

Based on funding mandates

Co-authors

VIEW ALL

- Xiaoou Tang — The Chinese University of Hong …
- Ziwei Liu — Assistant Professor, Nanyang Te…
- Dahua Lin — The Chinese University of Hong …
- Chao Dong — Shenzhen Institutes of Advanced…
- Ping Luo (羅平) — Associate Professor, The Univer…

Hi, I'm Shuai Yang (杨 帅)

This is my personal website. I'm an Assistant Professor with the Wangxuan Institute of Computer Technology, Peking University. I'm a member of the Spatial and Temporal Restoration, Understanding and Compression Team (STRUCT), and work with Prof. Zongming Guo and Prof. Jiaying Liu. I was a member of MMLab@NTU and worked with Prof. Chen Change Loy and Prof. Ziwei Liu.

I was a Research Assistant Professor with MMLab@NTU, Singapore, from Mar. 2023 to Feb. 2024. I was a Postdoctoral Research Fellow with MMLab@NTU, Singapore, from Oct. 2020 to Feb. 2023. I visited VITA group, Texas A&M University, from Sep. 2018 to Sep. 2019. I visited GSIP lab, National Institute of Informatics, Japan, from Mar. 2017 to Aug. 2017. My current research interests include image stylization and image editing.

FRESCO

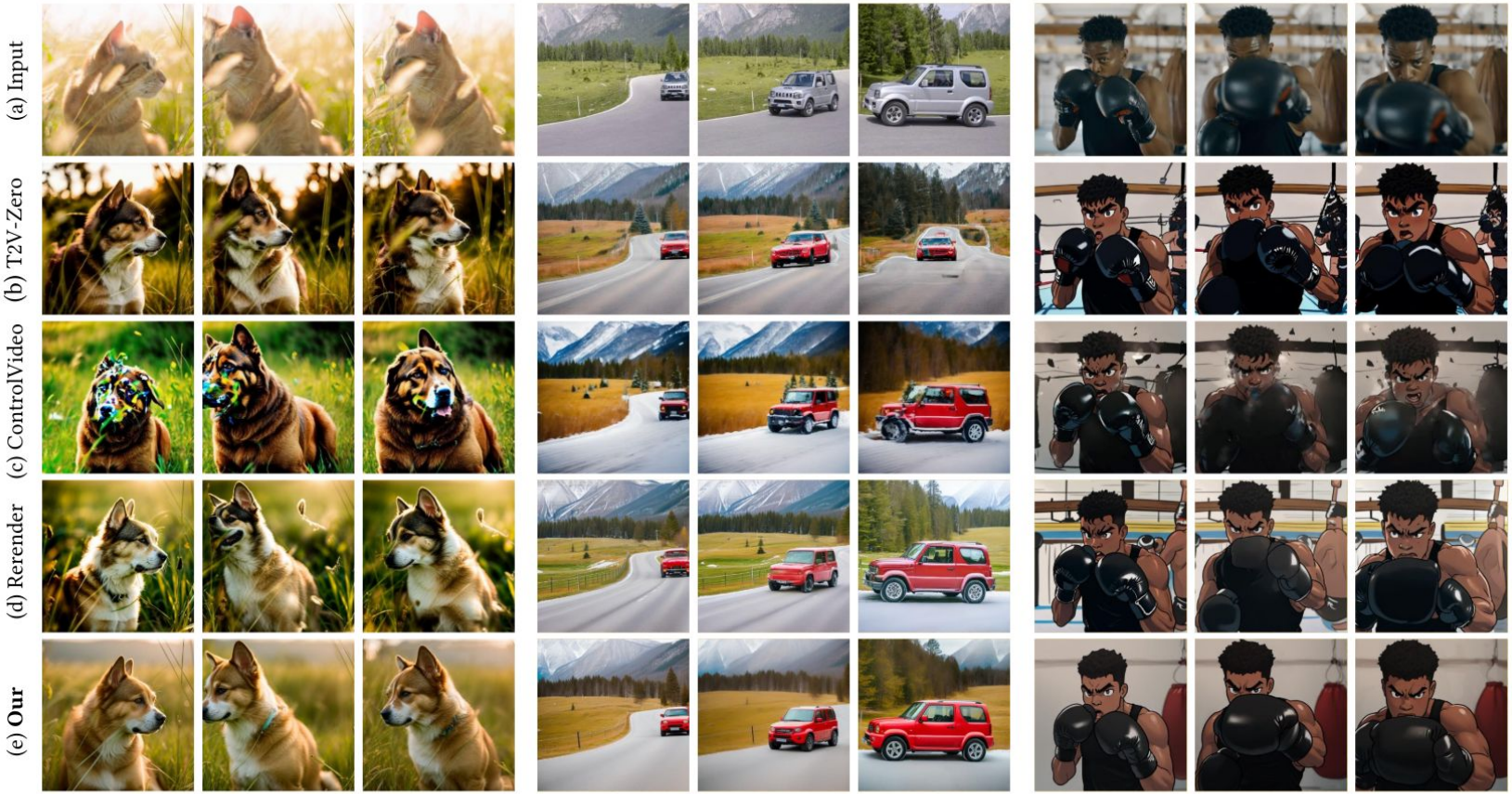
High-quality and coherent video translation based on pre-trained image diffusion model

Why we need zero shot editing?

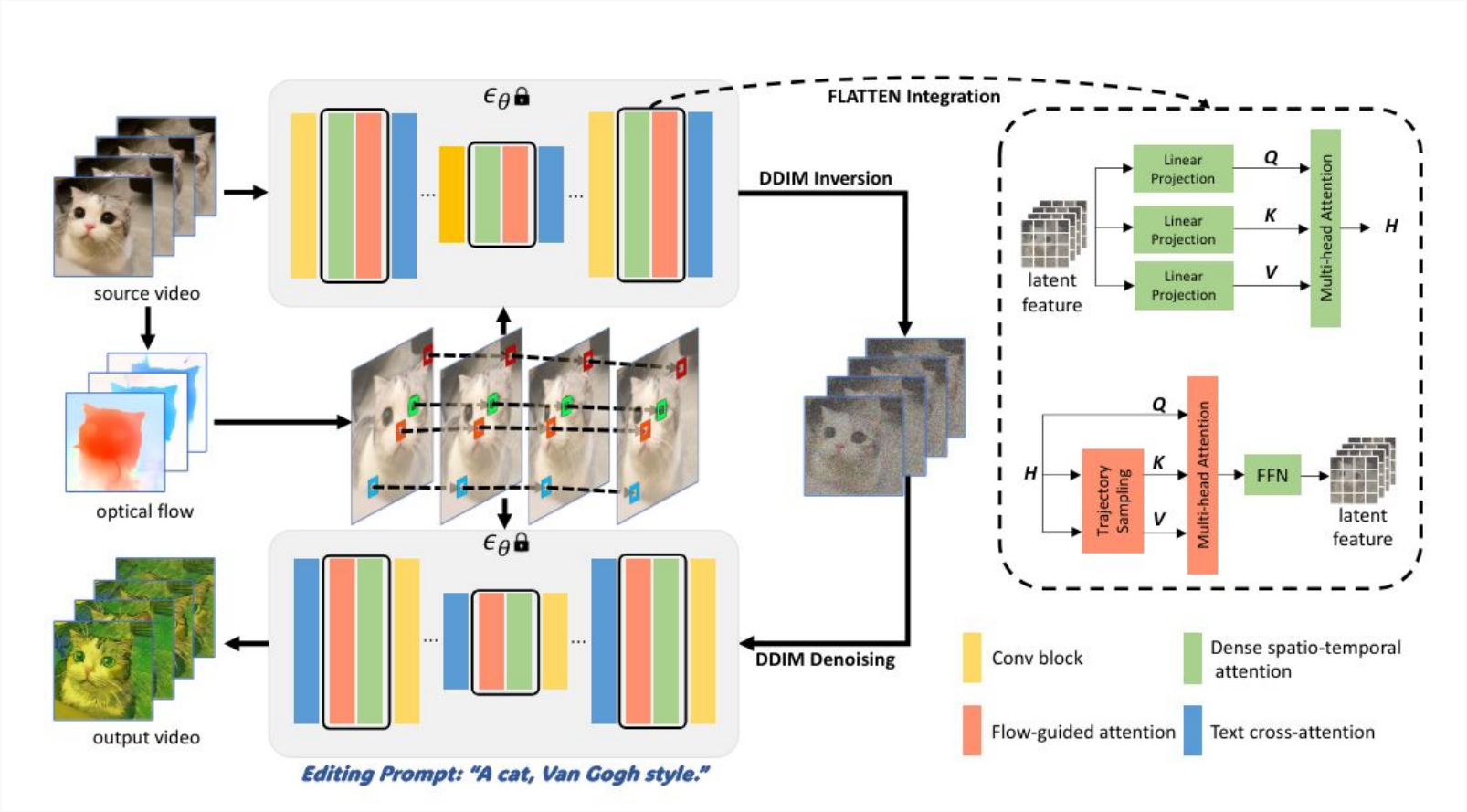
Learning Temporal-Coherent Motions: There are two main approaches:

- Extensive Video Datasets: Training video models on large video datasets is effective but can be expensive and computationally demanding.

- Fine-tuning Image Models on Single Videos: This approach is more accessible but suffers from limitations:
  - Not Cost-Effective: It might not be the most efficient use of resources for casual users.
  - Not Convenient: Fine-tuning models can be a complex process requiring technical expertise.

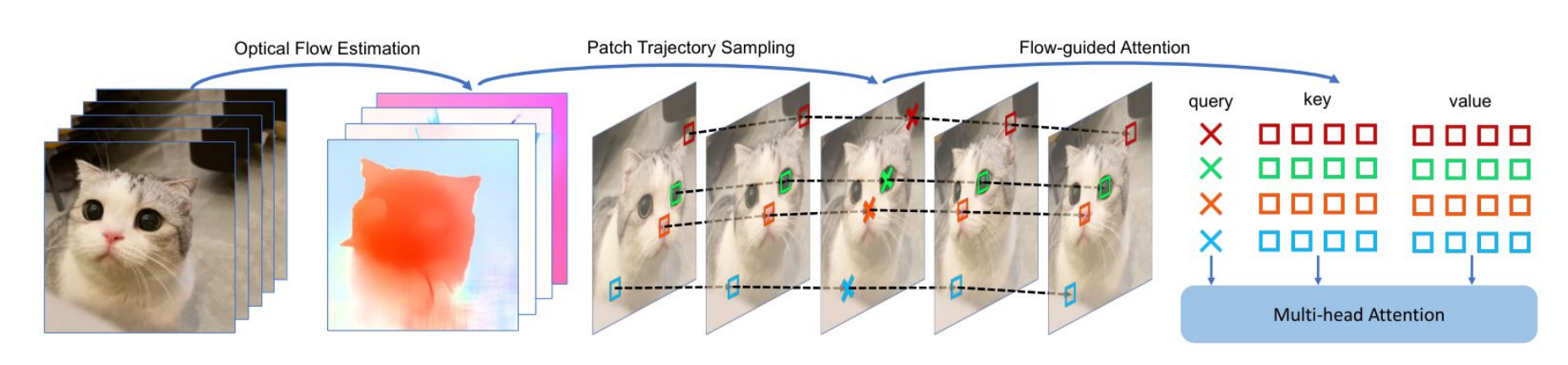

(a) Input
(b) T2V-Zero
(c) ControlVideo
(d) Rerender
(e) Our
Prompt: A dog in the grass in the sun
Prompt: A red car turns in the winter
Prompt: A black boxer wearing black boxing gloves punches towards the camera, cartoon style

Preliminaries: FLATTEN



Preliminaries: FLATTEN

Flow-guided Attention



Why Inversion-Free Frameworks?

Advantages of Inversion-Free Methods:

- Flexible Conditioning: Allows for more adaptable conditioning compared to inversion-based methods.
- Higher Compatibility: Better integration with customized models, enabling users to conveniently control the output appearance.

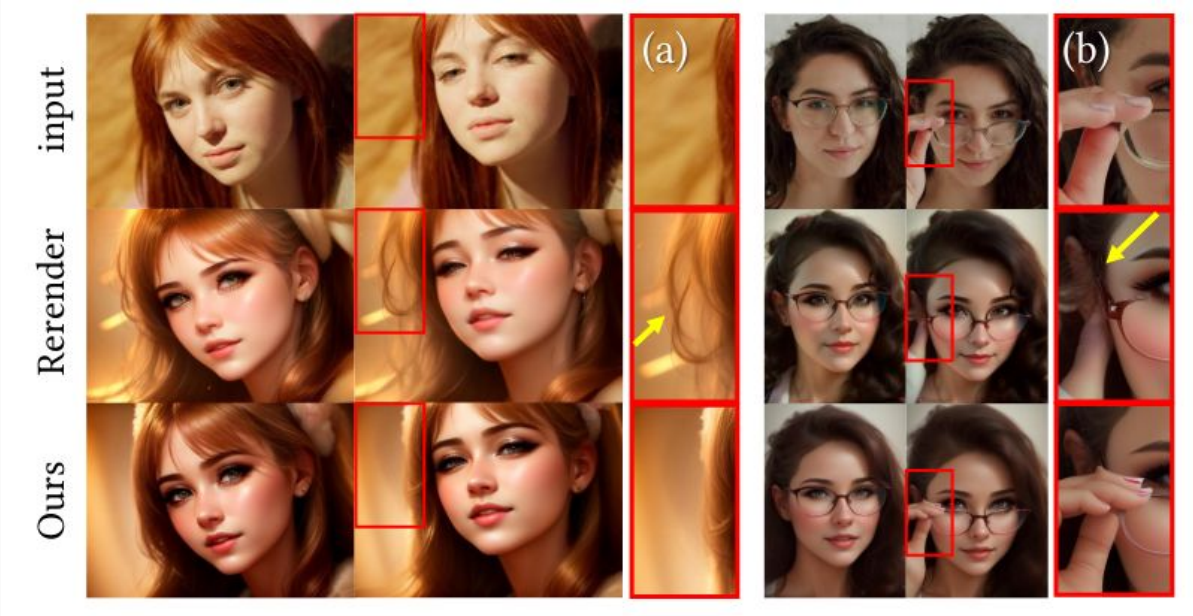
Challenges of Inversion-Free Methods:

- Prone to Flickering: Without the guidance of DDIM inversion features, the inversion-free framework can experience flickering issues.

Limitations of Existing Approaches for Temporal Consistency

Current methods (Rerender-A-Video, FLATTEN) rely on the original video's optical flow for guidance, but face challenges:

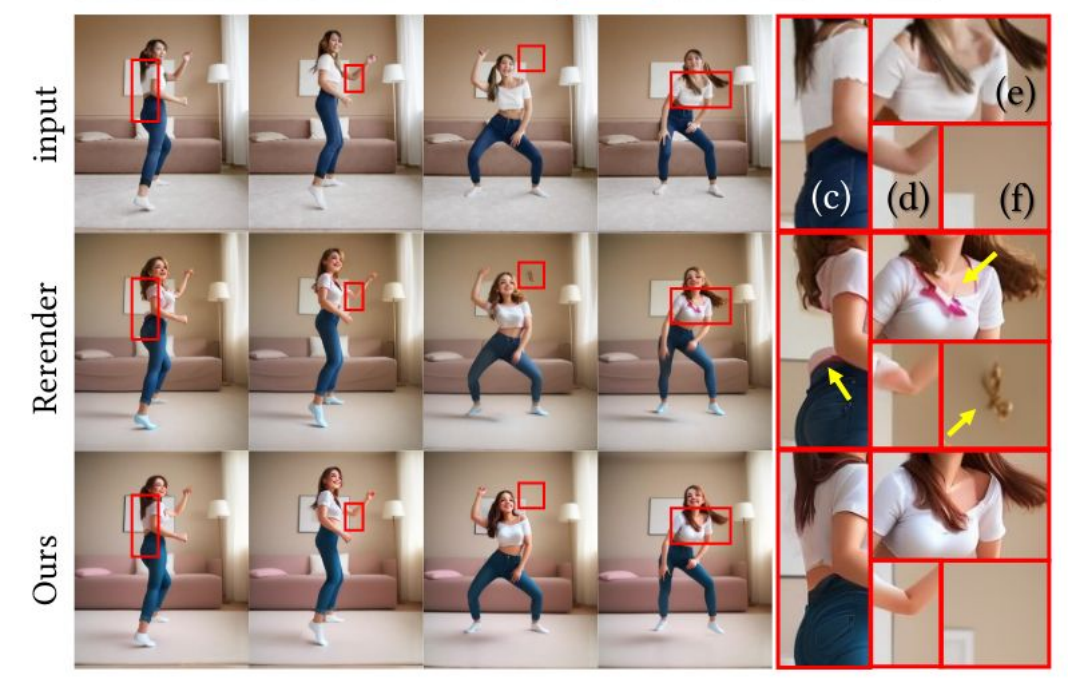
- 1- Inconsistency: Manipulations can alter optical flow, causing artifacts like foreground objects appearing in static backgrounds. (a)



Limitations of Existing Approaches for Temporal Consistency

Current methods (Rerender-A-Video, FLATTEN) rely on the original video's optical flow for guidance, but face challenges:

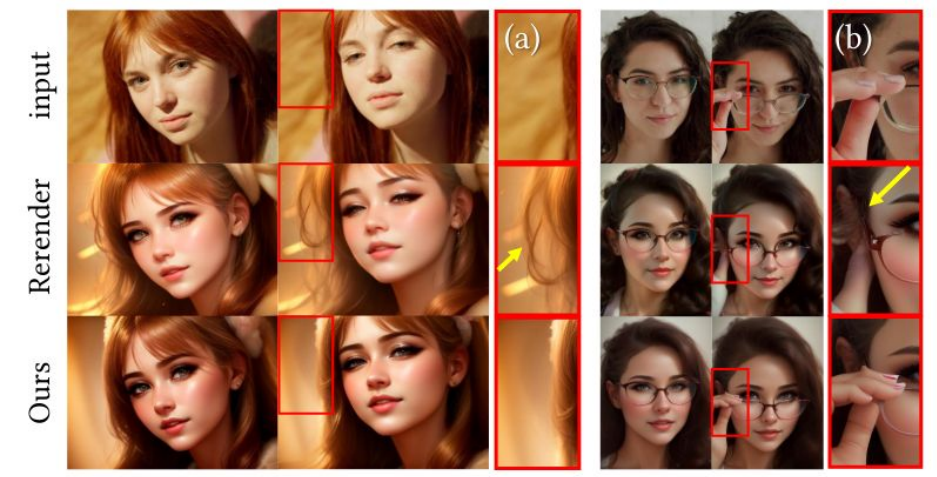
- 2- Undercoverage: Occlusions or rapid motion can lead to inaccurate optical flow. (c-d-e)



Limitations of Existing Approaches for Temporal Consistency

Current methods (Rerender-A-Video, FLATTEN) rely on the original video's optical flow for guidance, but face challenges:

- 3- Inaccuracy: Frame-by-frame generation suffers from error accumulation over time. (b)



(a)(f) inconsistent
(c)(d)(e) missing optical flow guidance
(b) error accumulation

Optimization during Inference

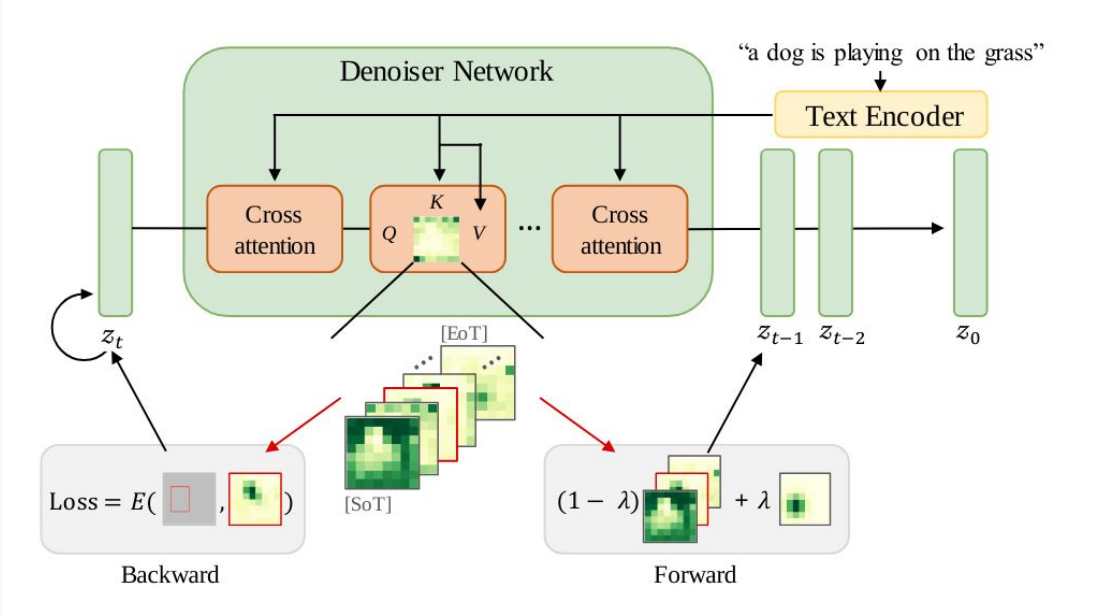
Training-Free Layout Control with Cross-Attention Guidance

Minghao Chen Iro Laina Andrea Vedaldi
Visual Geometry Group, University of Oxford
{minghao, iro, vedaldi}@robots.ox.ac.uk
silent-chen.github.io/layout-guidance

Attend-and-Excite: Attention-Based Semantic Guidance for Text-to-Image Diffusion Models

HILA CHEFER*, Tel Aviv University, Israel
YUVAL ALALUF*, Tel Aviv University, Israel
YAEL VINKER, Tel Aviv University, Israel
LIOR WOLF, Tel Aviv University, Israel
DANIEL COHEN-OR, Tel Aviv University, Israel

Optimization during Inference

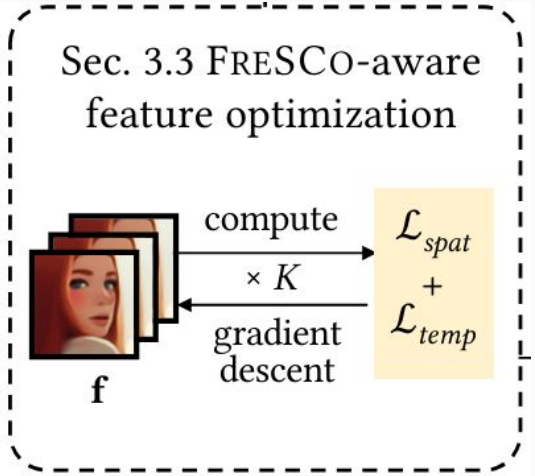


$$z_t' \leftarrow z_t - \alpha_t \cdot \nabla_{z_t} \mathcal{L}$$

FRESCO-Aware Feature Optimization

The input feature $\mathbf{f} = \{f_i\}_{i=1}^{N}$ of each decoder layer of U-Net is updated by gradient descent.

$$\hat{\mathbf{f}} = \arg\min_{\mathbf{f}} \mathcal{L}_{temp}(\mathbf{f}) + \mathcal{L}_{spat}(\mathbf{f})$$

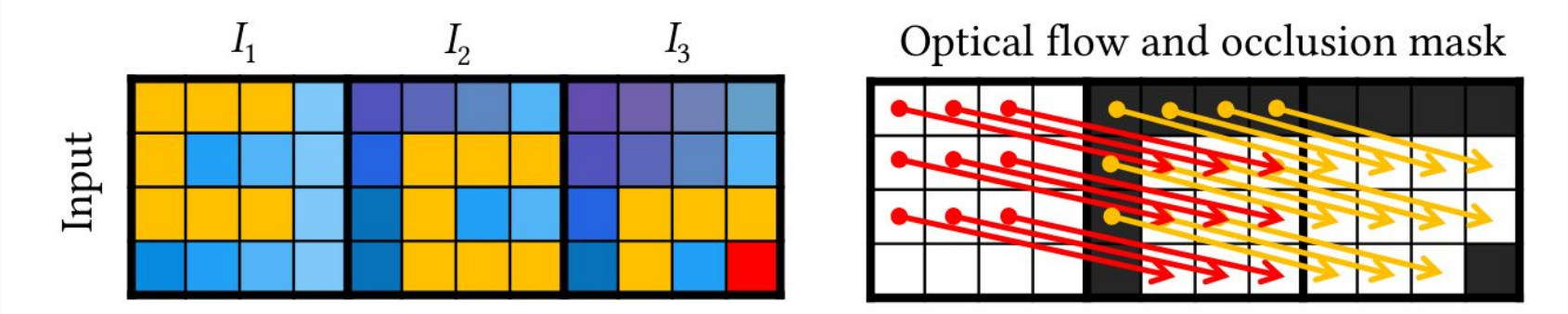


Temporal Consistency Loss

$$\mathcal{L}_{temp}(\mathbf{f}) = \sum_i \|M_i^{i+1}(f_{i+1} - w_i^{i+1}(f_i))\|_1$$

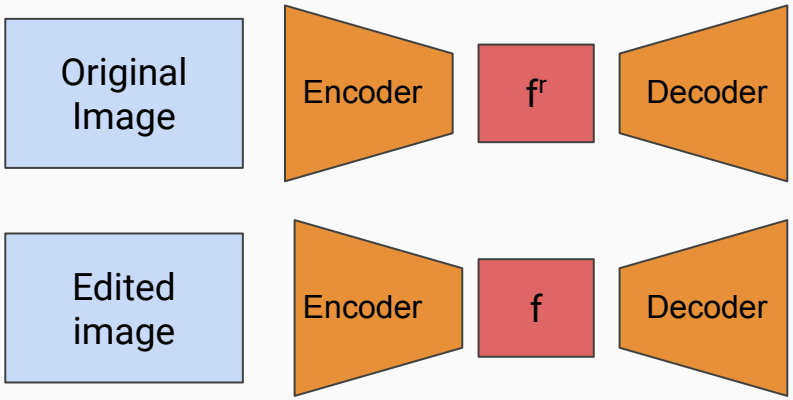
Where w_i^j and M_i^j are denoting the optical flow and occlusion mask from I_i to I_j.

The objective is to ensure that I'_i and I'_{i+1} share w_i^{i+1} in non-occluded regions.



Spatial Consistency Loss

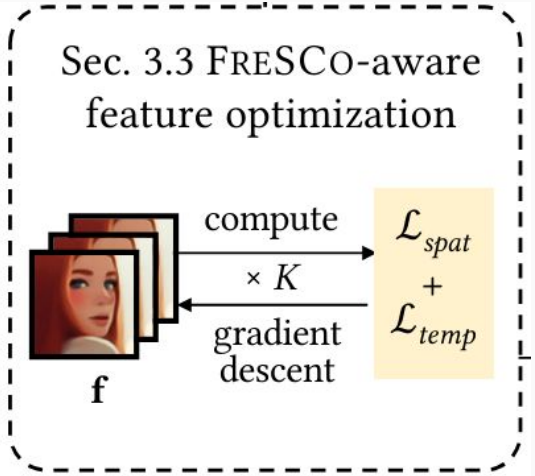
$$\mathcal{L}_{spat}(\mathbf{f}) = \lambda_{\text{spat}} \sum_i \| \tilde{f}_i \tilde{f}_i^{\top} - \tilde{f}_i^r \tilde{f}_i^{r\top} \|_2^2$$



FRESCO-Aware Feature Optimization

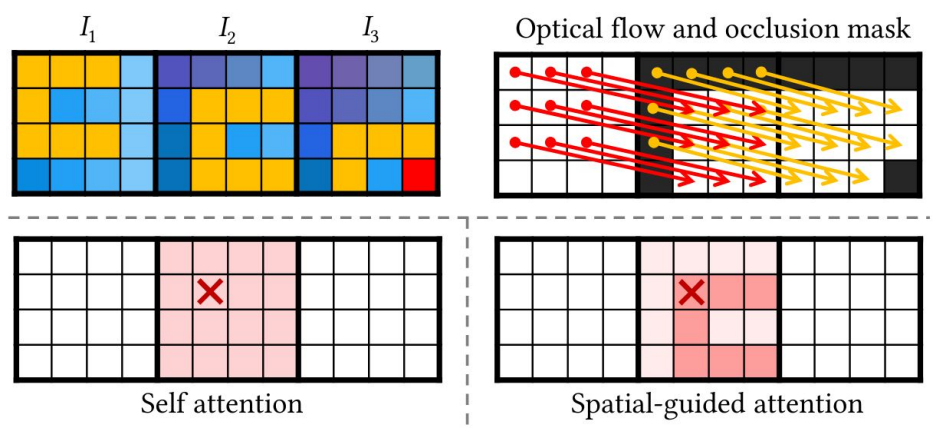
The input feature $\mathbf{f} = \{f_i\}_{i=1}^{N}$ of each decoder layer of U-Net is updated by gradient descent.

$$\hat{\mathbf{f}} = \arg\min_{\mathbf{f}} \mathcal{L}_{temp}(\mathbf{f}) + \mathcal{L}_{spat}(\mathbf{f})$$



Spatial-guided Attention

$$Q_i' = Softmax(\frac{Q_i^r K_i^{r\top}}{\lambda_s \sqrt{d}}) \cdot Q_i$$

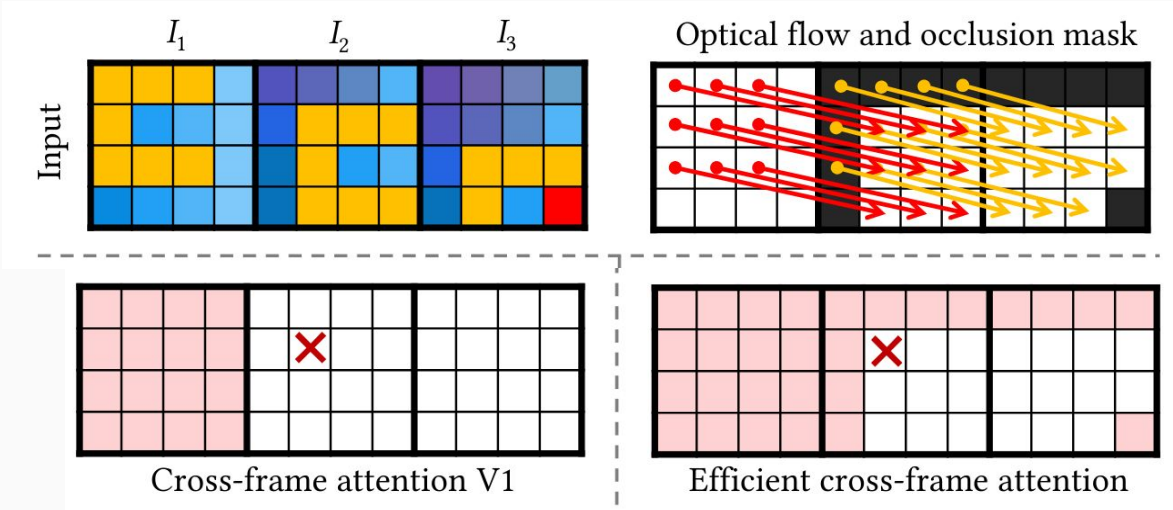


Efficient Cross-Frame Attention

$$V_i' = Softmax(\frac{Q_i'(K[p_u])^\top}{\sqrt{d}}) \cdot V[p_u]$$

Where P_u is the cross-frame index of all patches within the Unique Region.

Unique Region: Except for the first frame, we only reference to the areas of each frame that were not seen in its previous frame (i.e., the occlusion region).

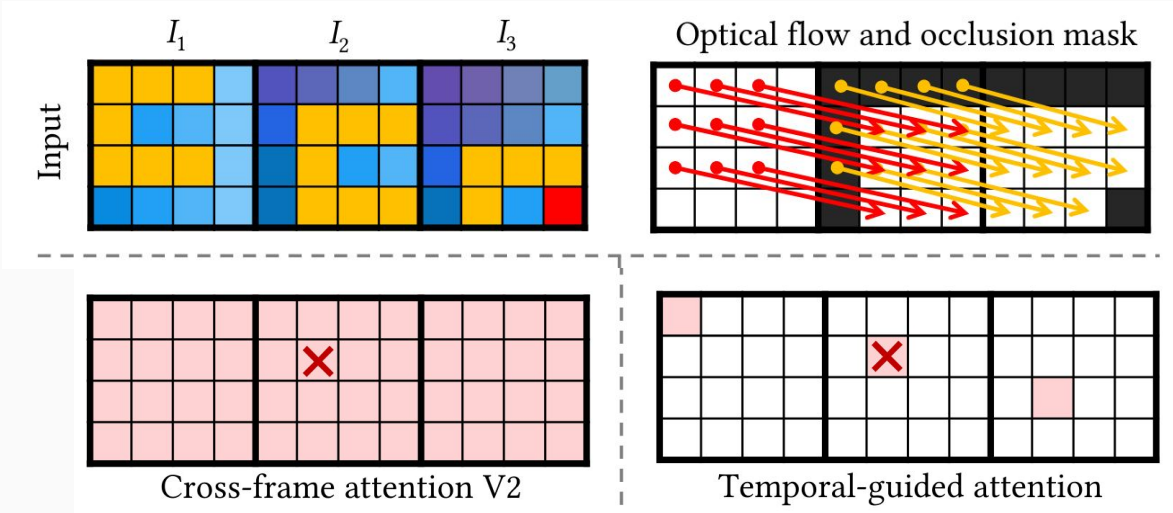


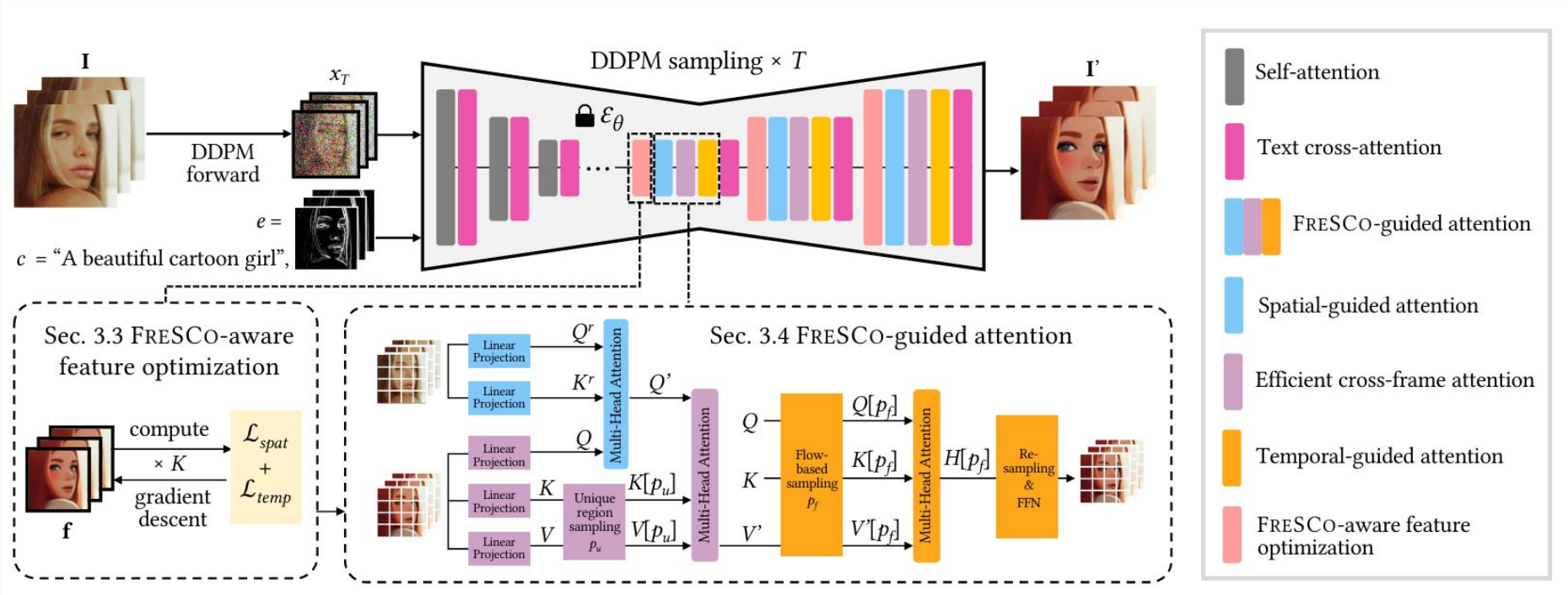
Temporal-Guided Attention

$$H[p_f] = Softmax(\frac{Q[p_f](K[p_f])^\top}{\lambda_t\sqrt{d}}) \cdot V'[p_f]$$

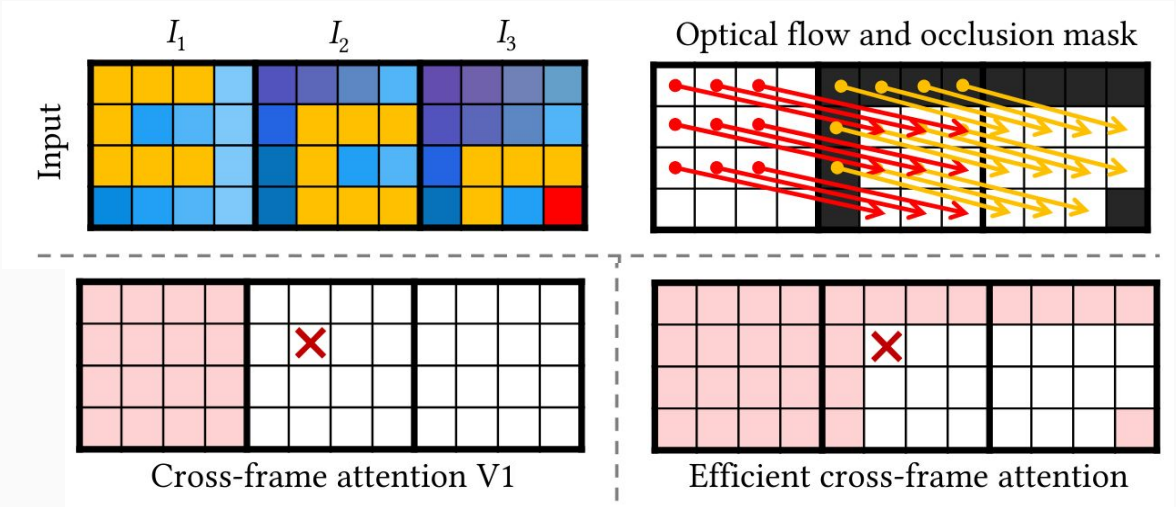
Where P_f is the cross-frame index of all patches on each optical flow.

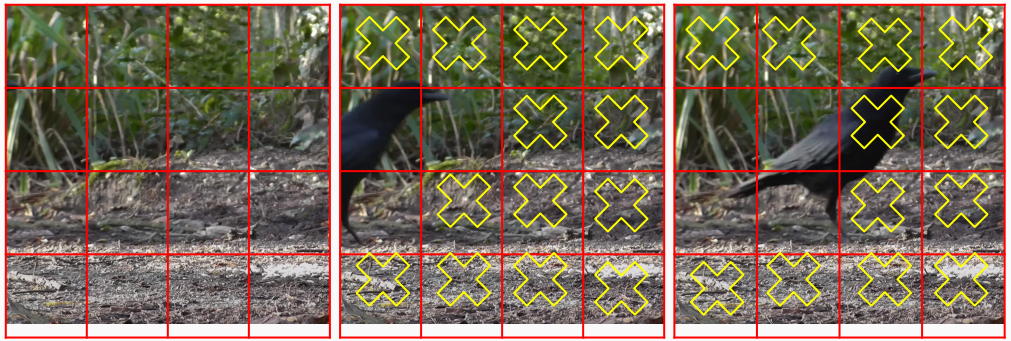
Flatten: Each patch can only attend to patches in other frames, which is unstable when a flow contains few patches.



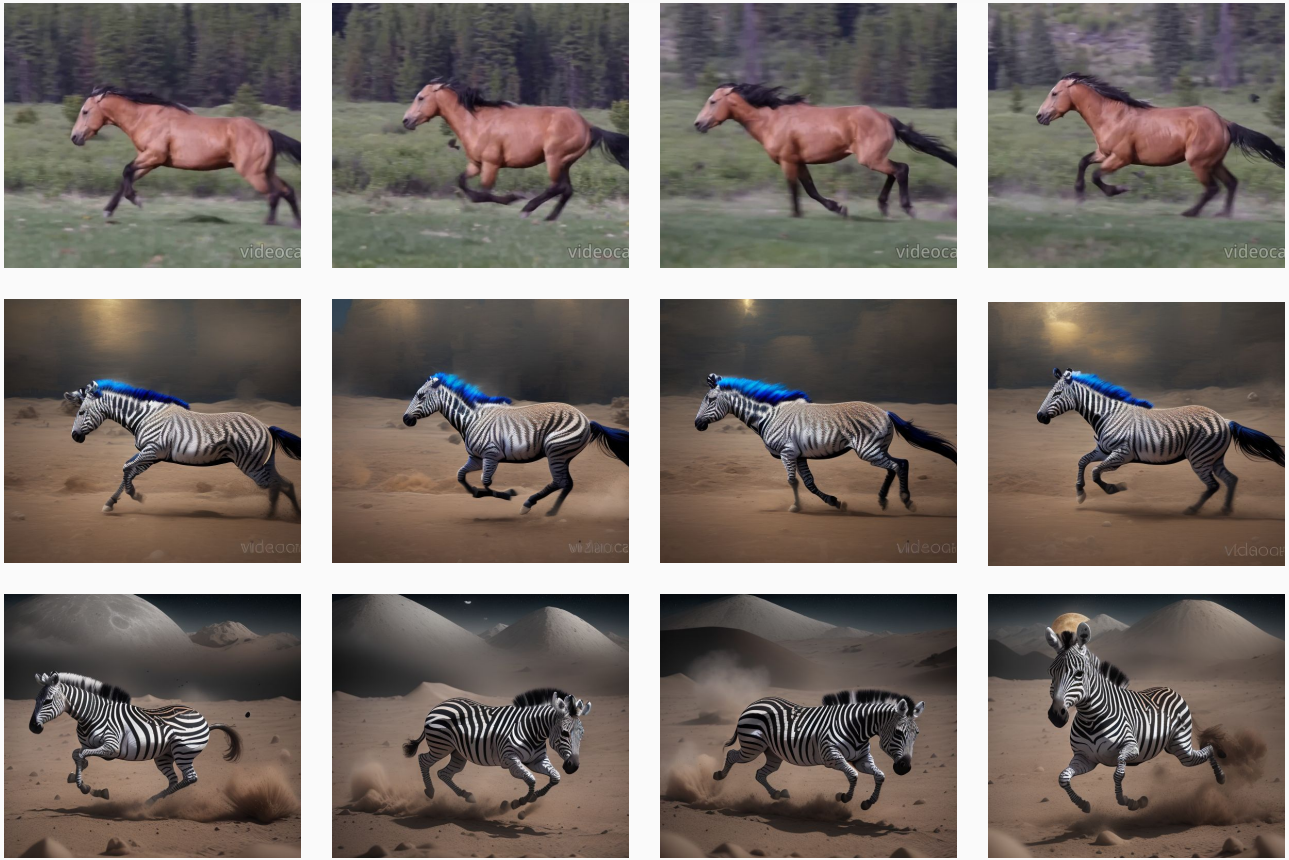

I
x_T
DDPM forward
DDPM sampling × T
$\mathcal{E}_\theta$
I'
e =
c = "A beautiful cartoon girl",
Sec. 3.3 FRESCO-aware feature optimization
compute
× K
gradient descent
$\mathcal{L}_{spat}$ + $\mathcal{L}_{temp}$
f
Sec. 3.4 FRESCO-guided attention
Linear Projection
Multi-Head Attention
Unique region sampling p_u
Flow-based sampling p_f
Re-sampling & FFN
Self-attention
Text cross-attention
FRESCO-guided attention
Spatial-guided attention
Efficient cross-frame attention
Temporal-guided attention
FRESCO-aware feature optimization

2 Limitations (Efficient Cross-Frame Attention)



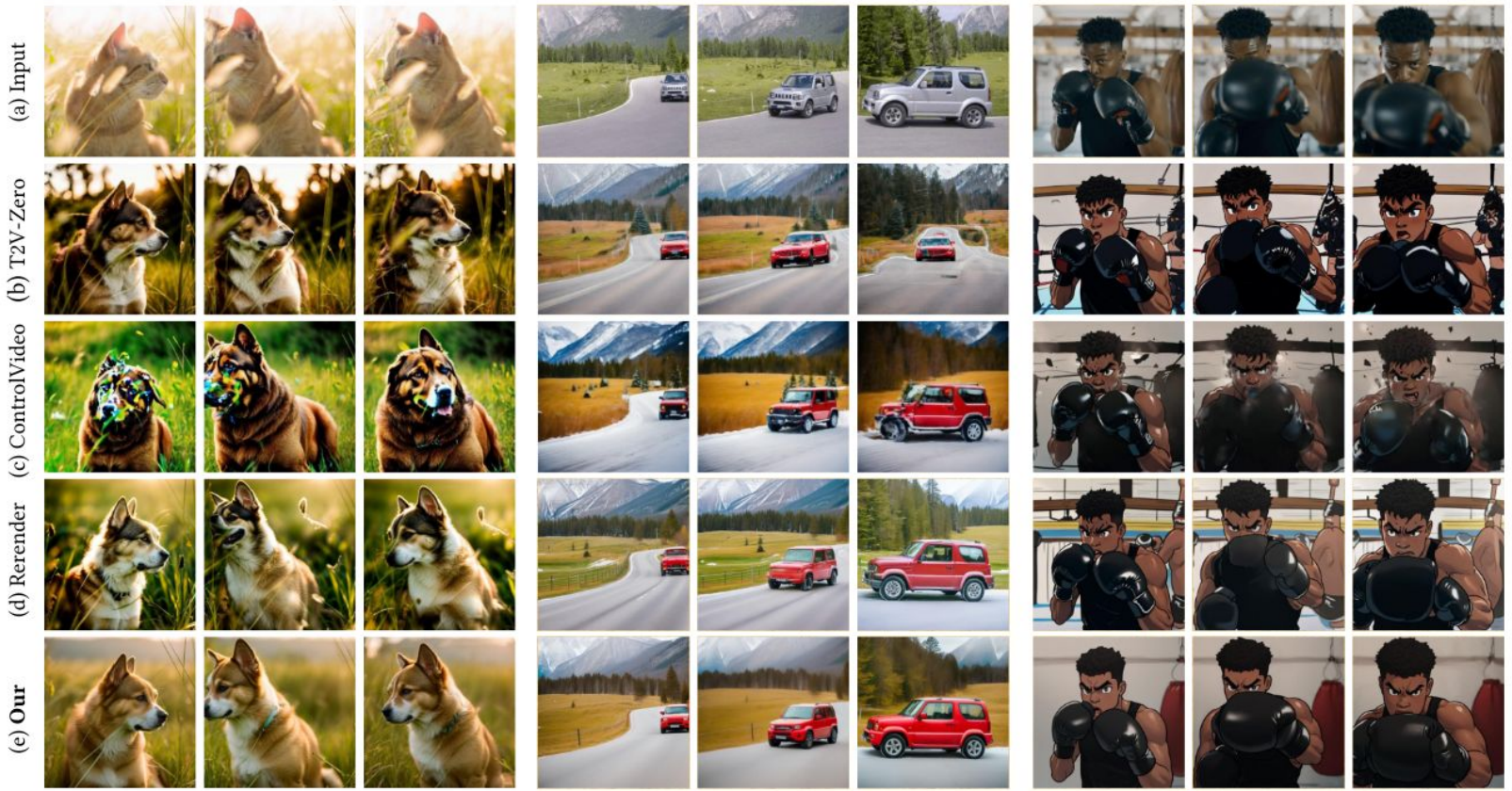


2 Limitations (ControlNet)

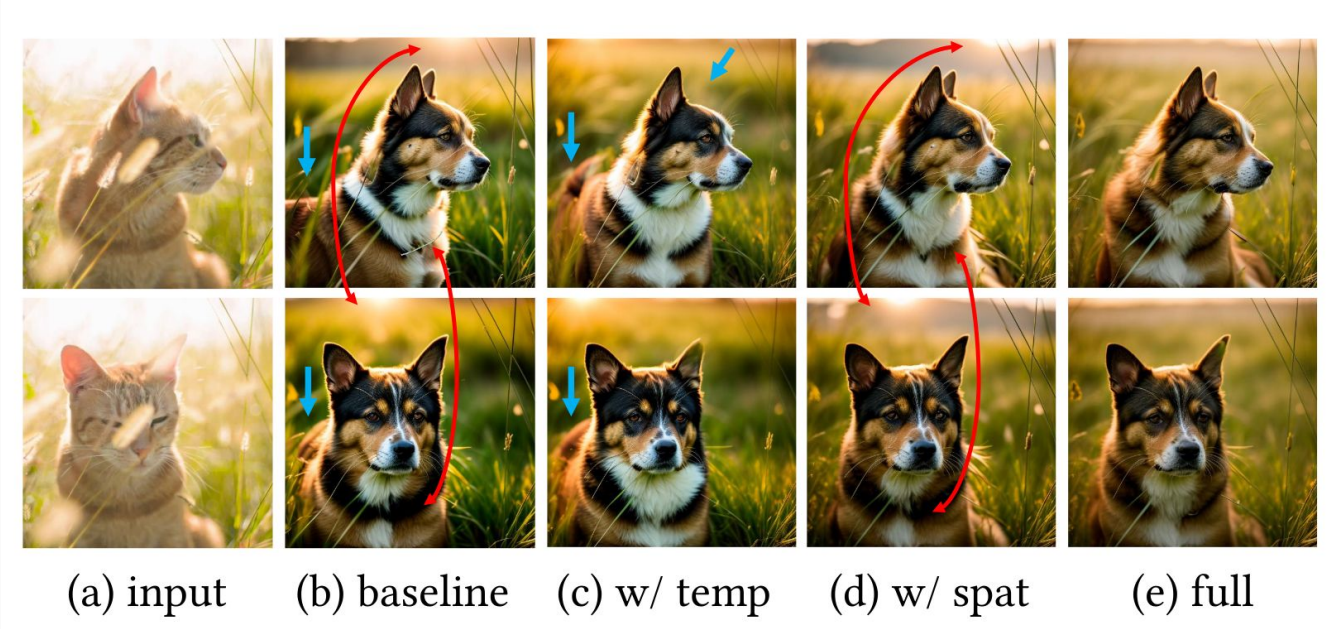


Prompt:
A zebra running on the moon, colorful

Comparison


(a) Input
(b) T2V-Zero
(c) ControlVideo
(d) Rerender
(e) Our
Prompt: A dog in the grass in the sun
Prompt: A red car turns in the winter
Prompt: A black boxer wearing black boxing gloves punches towards the camera, cartoon style

Ablation Study



Effect of incorporating spatial and temporal correspondences.
The blue arrows indicate the spatial inconsistency with the input frames.
The red arrows indicate the temporal inconsistency between two output frames.

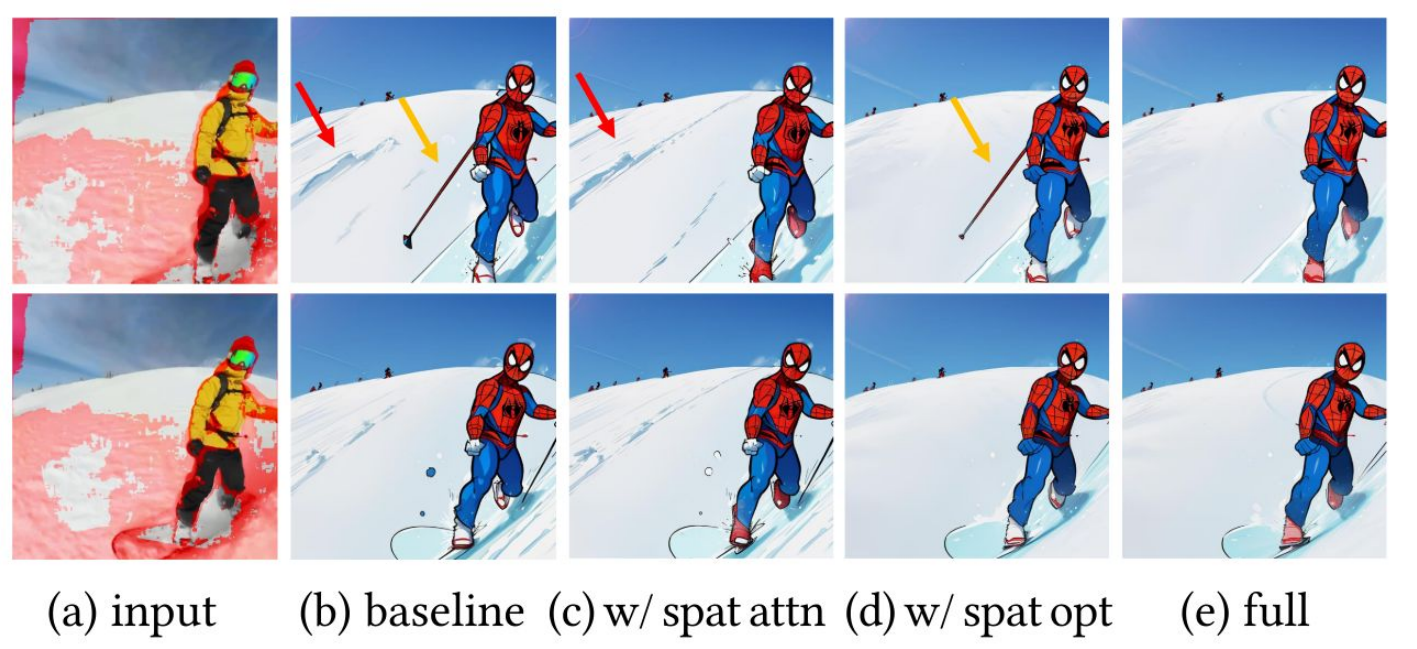
Ablation Study

Effect of attention adaptation and feature adaptation.

(a) Input
(b) only cross-frame attention
(c) attention adaptation
(d) feature adaptation
(e) both attention and feature adaptations

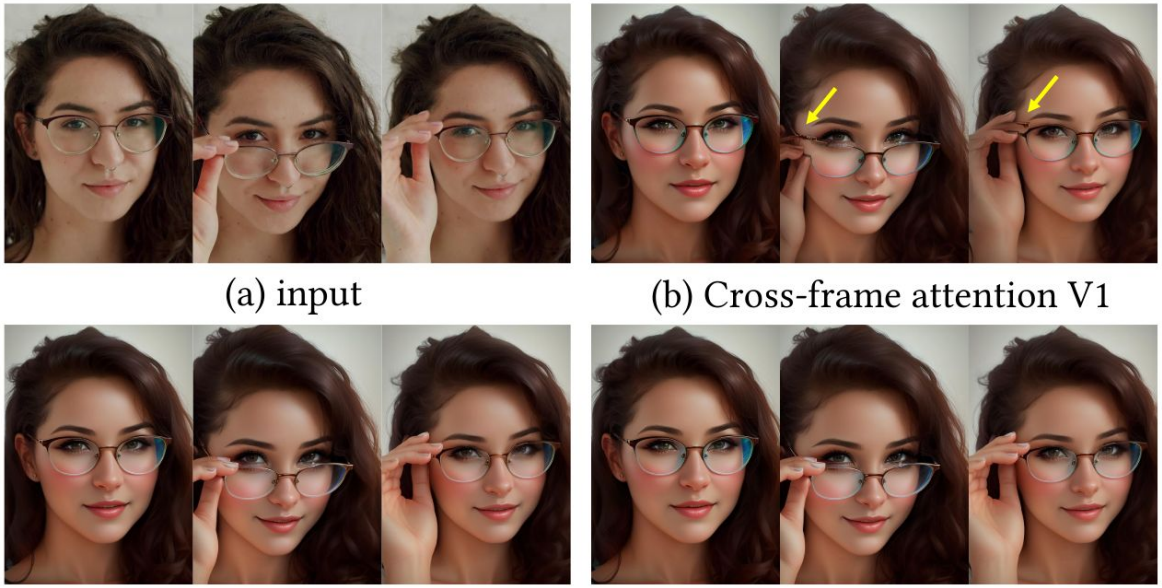
Prompt: A beautiful woman in CG style.

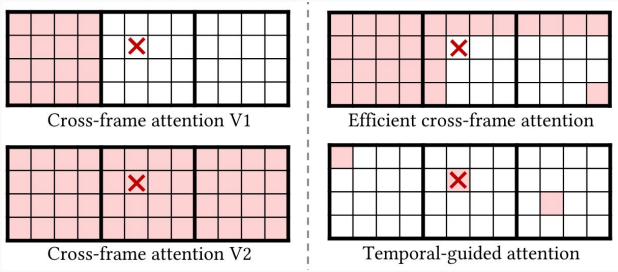
Ablation Study



Effect of incorporating spatial correspondence.
(a) Input covered with red occlusion mask.
(b)-(d) FRESCO spatial-guided attention and spatial consistency loss help reduce the inconsistency in ski poles (yellow arrows) and snow textures (red arrows), respectively.
Prompt: A cartoon Spiderman is skiing.

Ablation Study





Effect of efficient cross-frame attention. (a) Input. (b) Cross-frame attention V1 attends to the previous frame only, thus failing to synthesize the newly appearing fingers. (d) The efficient cross-frame attention achieves the same performance as (c) cross-frame attention V2, but reduces the region that needs to be attended to by 41.6% in this example. Prompt: A beautiful woman holding her glasses in CG style.

Limitations

Final Objective in Video Editing:

- Converting any object to any other object.
- Requires motion knowledge.

Limitations

- FateZero Examples:

Generated Results

Google Drive

Thank you!

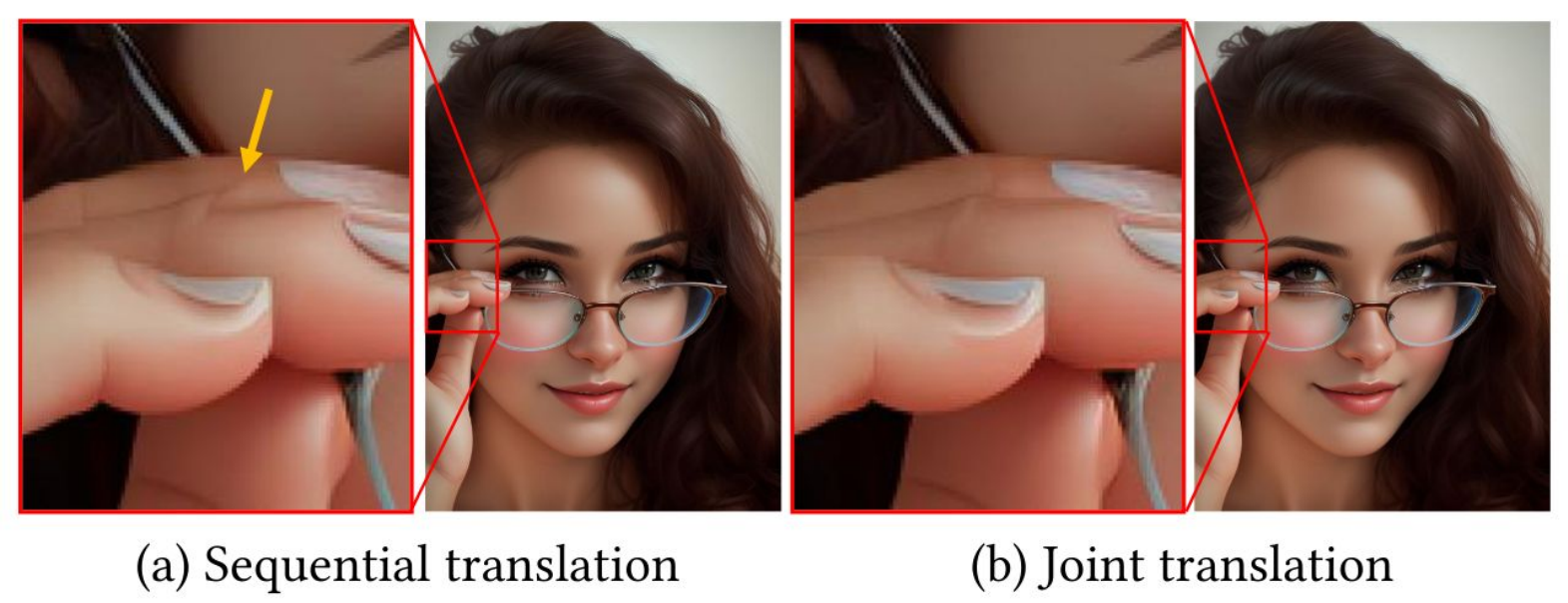
- Thank you for your attention!
- I appreciate your time and interest.
- If you have any questions, please feel free to ask.
- Contact information: alimohammadiamirhossein@gmail.com

Extra Slides

Gram Matrix

$$\left|G(v_1,\ldots,v_n)\right| = \begin{vmatrix} \langle v_1,v_1\rangle & \langle v_1,v_2\rangle & \cdots & \langle v_1,v_n\rangle \\ \langle v_2,v_1\rangle & \langle v_2,v_2\rangle & \cdots & \langle v_2,v_n\rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle v_n,v_1\rangle & \langle v_n,v_2\rangle & \cdots & \langle v_n,v_n\rangle \end{vmatrix}$$

Ablation Study



Effect of joint multi-frame translation.
Sequential translation relies on the previous frame alone.
Joint translation uses all frames in a batch to guide each other, thus achieving accurate finger structures by referencing to the third frame