TokenFlow: Consistent Diffusion Features for Consistent Video Editing

ICLR 2024

20240501

Jaihoon Kim

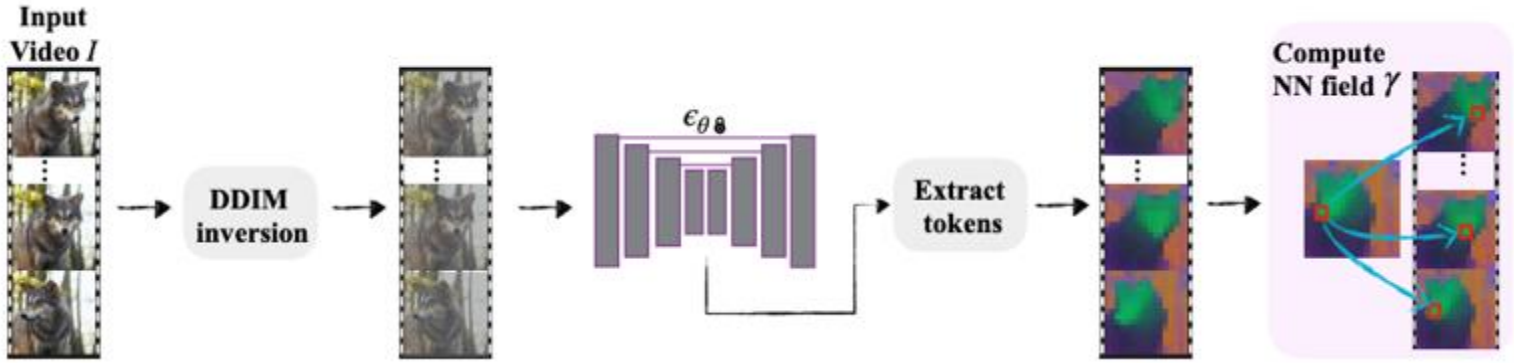
KAIST Visual AI Group

Key Ideas

Stage 1:

- Obtain $\{X_1^T, \dots, X_n^T\}$ using DDIM inversion and extract feature maps from the SA of the U-Net
- Compute the correspondences using the nearest-neighbor search

$$\gamma^{i\pm}[p] = \arg\min_q \mathcal{D}\left(\phi(\boldsymbol{x}^i)[p], \phi(\boldsymbol{x}^{i\pm})[q]\right)$$



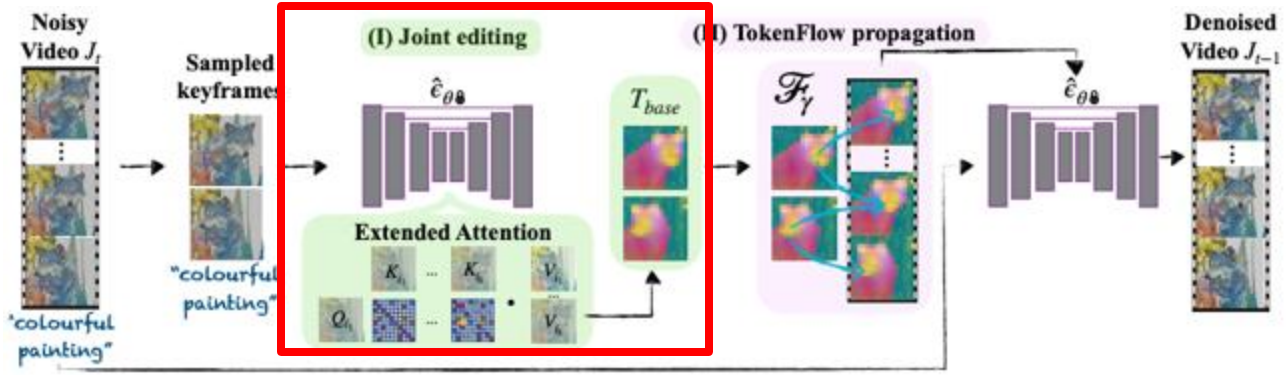
Key Ideas

Stage 2:

- Randomly sample keyframes and compute the extend-attention

$$\texttt{ExtAttn}\left(\boldsymbol{Q}^i;[\boldsymbol{K}^{i_1},\dots\boldsymbol{K}^{i_k}]\right)=\texttt{Softmax}\left(\frac{\boldsymbol{Q}^i\left[\boldsymbol{K}^{i_1},\dots\boldsymbol{K}^{i_k}\right]^T}{\sqrt{d}}\right)$$

$$\phi(\boldsymbol{J}^i)=\hat{\boldsymbol{A}}\cdot[\boldsymbol{V}^{i_1},\dots\boldsymbol{V}^{i_k}]\quad\text{where}\quad\hat{\boldsymbol{A}}=\texttt{ExtAttn}\left(\boldsymbol{Q}^i;[\boldsymbol{K}^{i_1},\dots\boldsymbol{K}^{i_k}]\right)$$



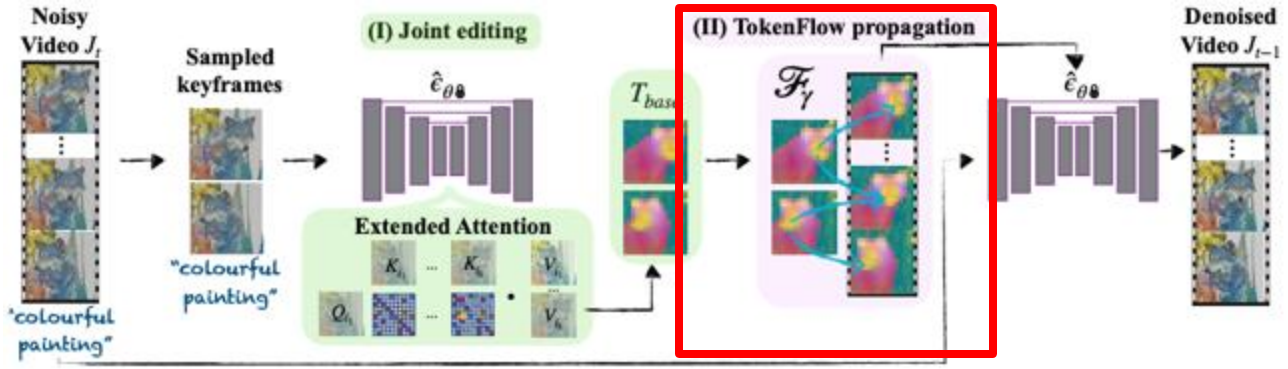
Key Ideas

Stage 2:

- Propagate outputs of extended-attention module to the rest of the frames using the correspondences

$$\mathcal{F}_\gamma(\mathbf{T}_{base}, i, p) = w_i \cdot \phi(\boldsymbol{J}^{i+})[\gamma^{i+}[p]] \; + \; (1 - w_i) \cdot \phi(\boldsymbol{J}^{i-})[\gamma^{i-}[p]]$$

- One-step denoise the entire frames



Results

PnP-Diffusion: Framewise editing
Fate-Zero, Text2Video-Zero, Re-render a video: Self-attention inflation
Tune-a-Video, Gen-1: Trains on video dataset

	Warp-err ↓ $(\times 10^{-3})$	User preference of our method	CLIP score ↑
LDM recon.	2.0	–	0.23
PnP-Diffusion	11.3	94%	0.33
Text2Video-Zero	12.5	78%	0.33
Tune-a-Video	30.0	82%	0.31
Fate-Zero	6.9	71%	0.32
Gen1	–	70%	0.32
Rerender-a-Video	1.8	71%	0.32
Ours *w joint attention*	5.9	90%	0.33
Ours *w/o rand keyframes*	3.7	–	0.33
Ours	3.0	–	0.33

Results

Input | **TokenFlow** | PnP | FateZero

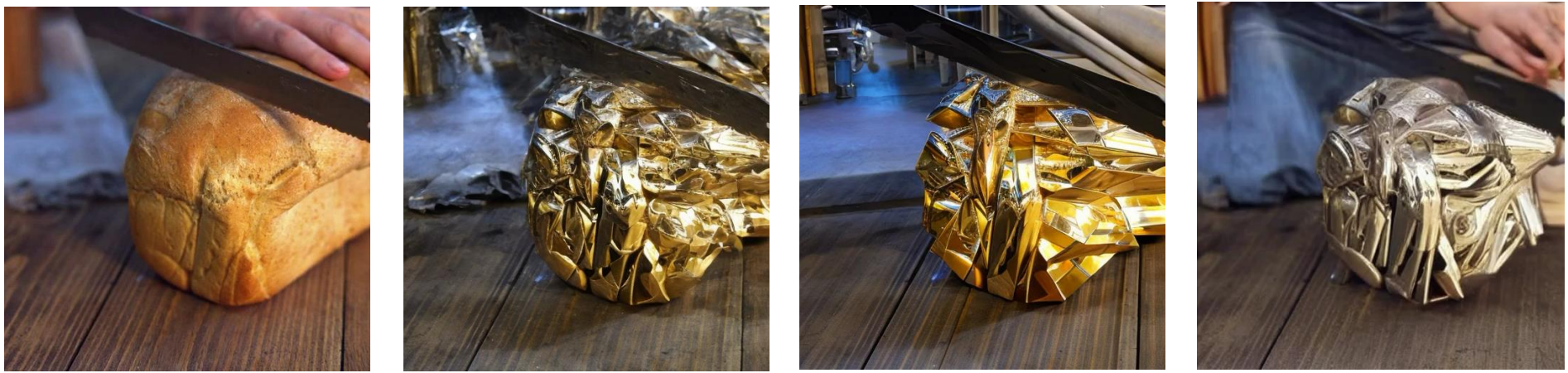


Figure 7: **Limitations.** Our method edits the video according to the feature correspondences of the original video, hence it cannot handle edits that requires structure deviations.