ICLR
International Conference On
Learning Representations

Looking Backward: Streaming Video-to-Video Translation with Feature Banks

ICLR 2025

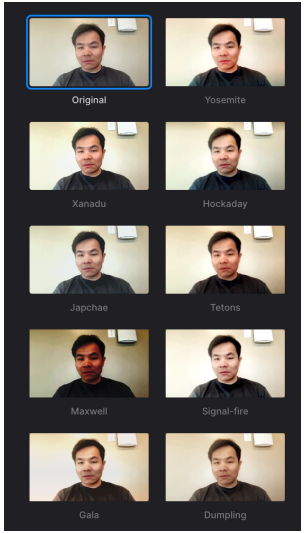
Feng (Jeff) Liang, Akio Kodaira, Chenfeng Xu, Masayoshi Tomizuka, Kurt Keutzer, **Diana Marculescu**

TEXAS
The University of Texas at Austin

Berkeley
UNIVERSITY OF CALIFORNIA

Challenges of traditional video-to-video

Input video



Pre-defined filters

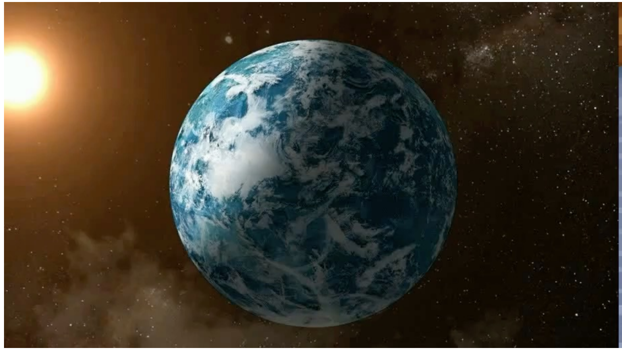
[Source: clideo]

Grayscale output video

Traditional filter-based video-to-video translation is **single-modal:**

- Video as the only input
- Limited filters with poor editing capabilities

Text-prompted video-to-video



Input video

Prompt : A pixel art of an artist's rendering of an earth in space.

Prompt : An artist's rendering of a Mars in space

Input video

Prompt : a woman wearing headphones in flat 2d anime.

Prompt : a Greek statue wearing headphones.

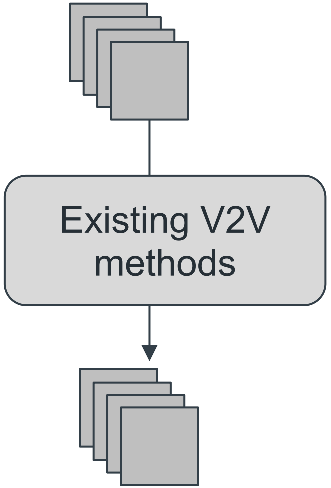
Video-to-video method comparison

	Traditional filters	Existing diffusion models [1-6]
Pros ✅	Real-time processing Unlimited length	Easy use with natural language Good edit capability
Cons ❌	Limited filters Bad edit capability	Only handle limited length, e.g., 4 sec Extremely slow, 1 min processing for one 4 sec edit

[1] Wu, Jay Zhangjie, et al. "Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.
[2] Qi, Chenyang, et al. "Fatezero: Fusing attentions for zero-shot text-based video editing." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.
[3] Zhang, Yabo, et al. "Controlvideo: Training-free controllable text-to-video generation." arXiv preprint arXiv:2305.13077 (2023).
[4] Geyer, Michal, et al. "Tokenflow: Consistent diffusion features for consistent video editing." arXiv preprint arXiv:2307.10373 (2023).
[5] Ouyang, Hao, et al. "Codef: Content deformation fields for temporally consistent video processing." arXiv preprint arXiv:2308.07926 (2023).
[6] Liang, Feng, et al. "FlowVid: Taming Imperfect Optical Flows for Consistent Video-to-Video Synthesis." arXiv preprint arXiv:2312.17681 (2023).

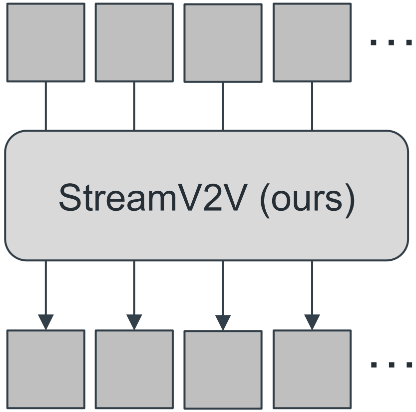
Batch and stream processing

Recorded video

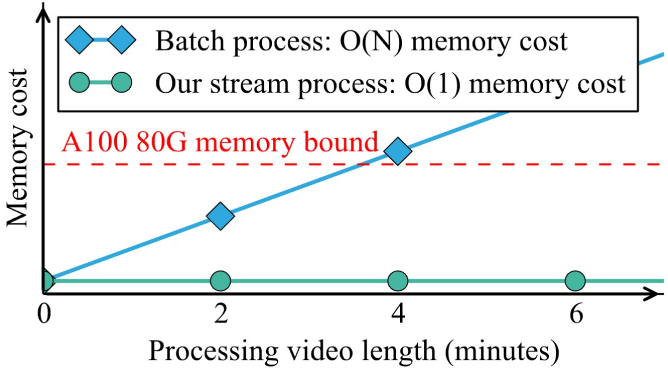


(a) Batch processing

Real-time video



(b) Stream processing

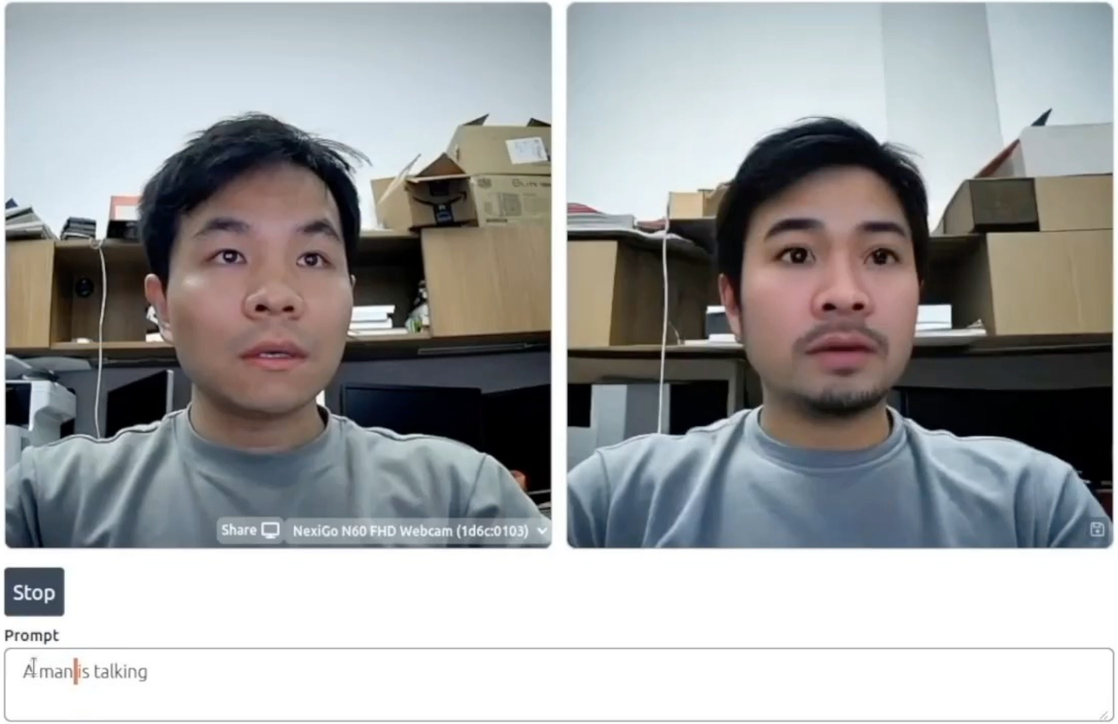


(c) Memory consumption comparsion

All frames loaded into GPU and processed in a batch

Process frame by frame so that we can handle unlimited frames in real-time

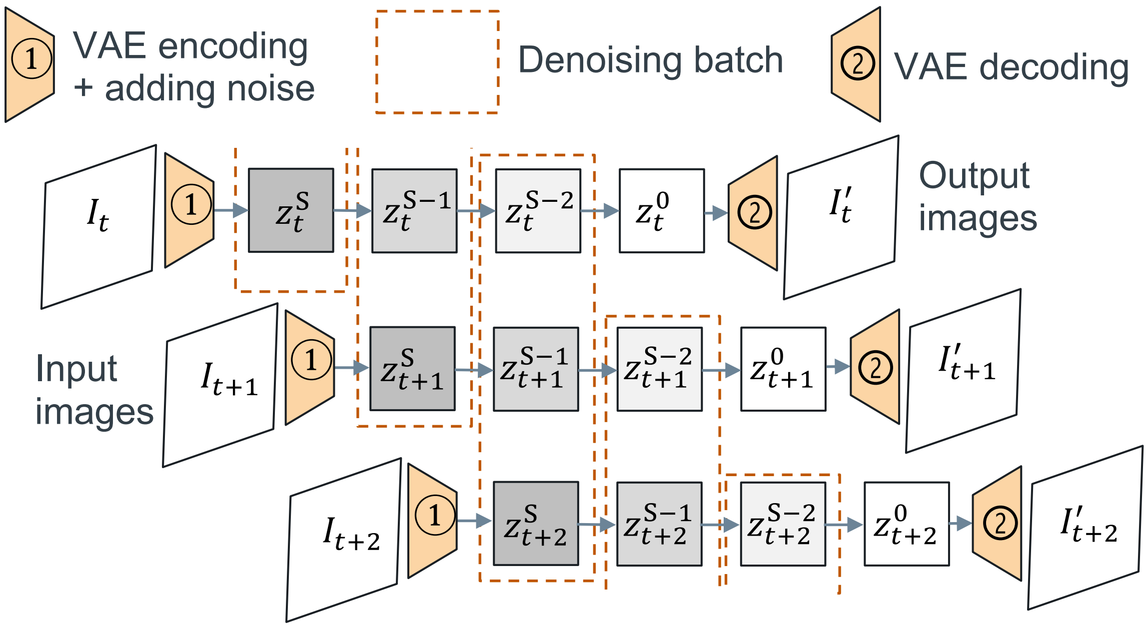
StreamV2V realtime demo on RTX 4090



Our StreamV2V supports face swap (e.g., to Elon Musk or Will Smith) and video stylization (e.g., to Claymation or doodle art)

Our starting point: StreamDiffusion + LCM

LCM [1] can generate images with 1-4 steps
StreamDiffusion [2] batchify the LCM for streaming images



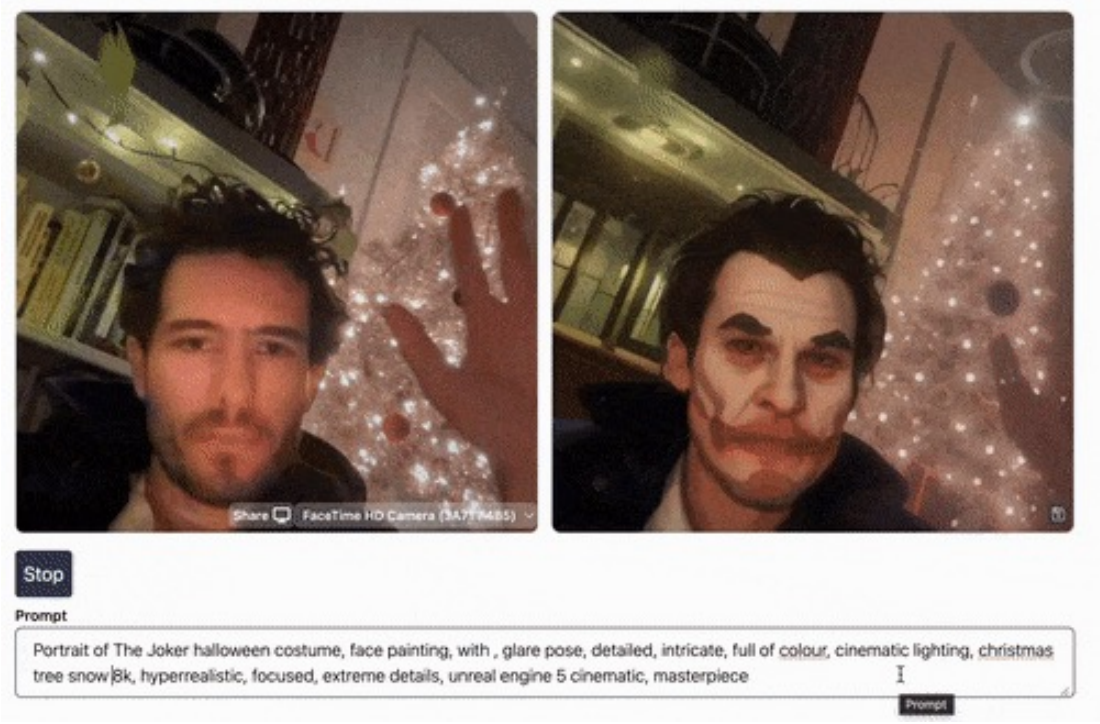
[1] Luo, Simian, et al. "Latent consistency models: Synthesizing high-resolution images with few-step inference." arXiv preprint arXiv:2310.04378 (2023).
[2] Kodaira, Akio, et al. "StreamDiffusion: A Pipeline-level Solution for Real-time Interactive Generation." arXiv preprint arXiv:2312.12491 (2023).

StreamDiffusion is an img2img model

[1] Luo, Simian, et al. "Latent consistency models: Synthesizing high-resolution images with few-step inference." arXiv preprint arXiv:2310.04378 (2023).
[2] Kodaira, Akio, et al. "StreamDiffusion: A Pipeline-level Solution for Real-time Interactive Generation." arXiv preprint arXiv:2312.12491 (2023).

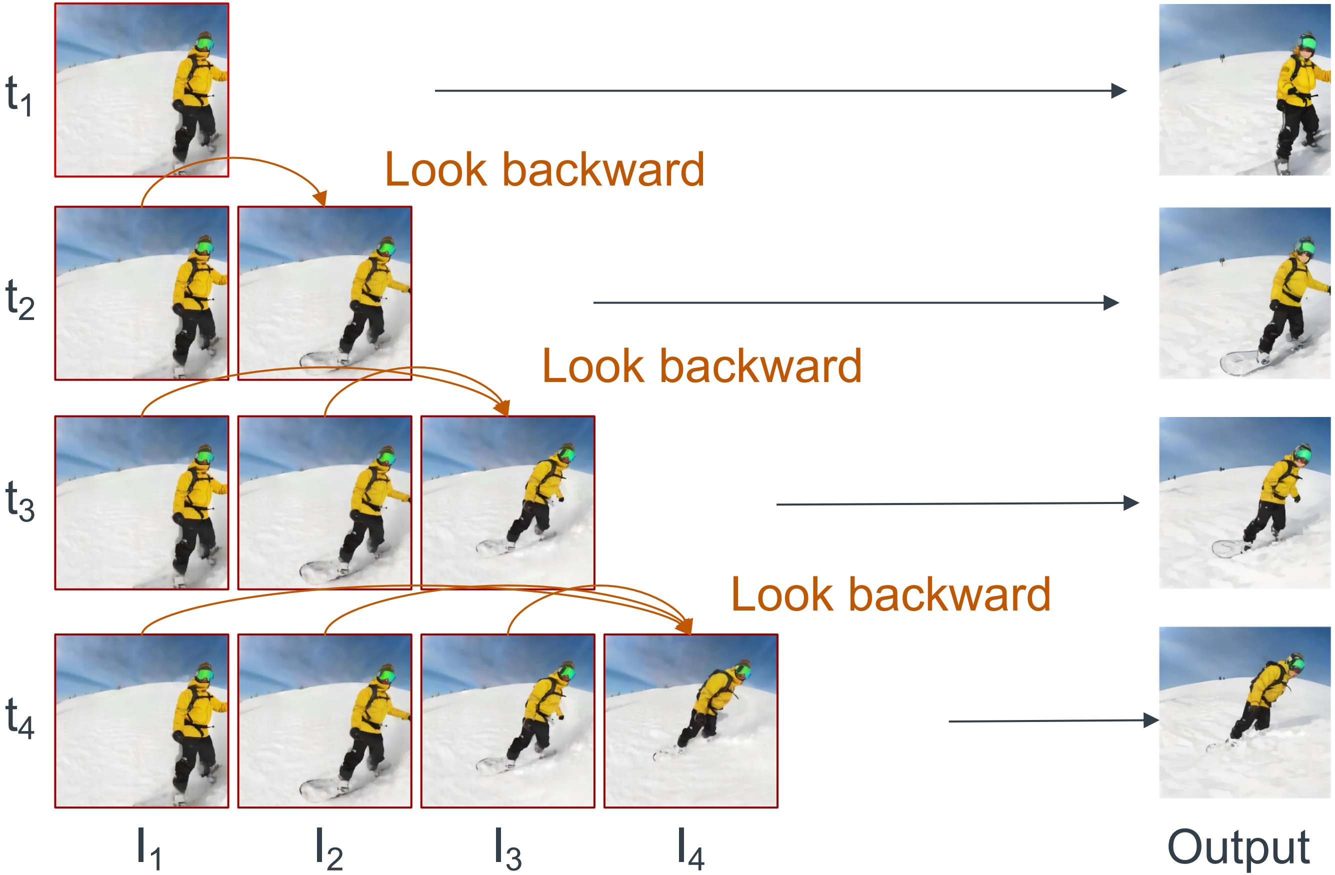
StreamDiffusion has poor consistency

However, StreamDiffusion is an image model, producing inconsistent outcome



[1] Luo, Simian, et al. "Latent consistency models: Synthesizing high-resolution images with few-step inference." arXiv preprint arXiv:2310.04378 (2023).
[2] Kodaira, Akio, et al. "StreamDiffusion: A Pipeline-level Solution for Real-time Interactive Generation." arXiv preprint arXiv:2312.12491 (2023).

Looking backward to improve consistency



“Looking backward” with diffusion features

Cached self-attention diffusion features

K_1 V_1 | K_2 V_2 | K_3 V_3 | K_4 V_4 | Q_5 K_5 V_5

I_1

I_2

I_3

I_4

I_5

[1] Tang, Luming, et al. "Emergent correspondence from image diffusion." Advances in Neural Information Processing Systems 36 (2023): 1363-1389.
[2] Luo, Grace, et al. "Diffusion hyperfeatures: Searching through time and space for semantic correspondence." Advances in Neural Information Processing Systems 36 (2024).

Diffusion features have semantic correspondace

Find the point $q_n = \arg\max Q_5(p) K_n^T$ where n = 1,2,3,4

$q_1 = \arg\max Q_5(p) K_1^T$

$Q_5(p)$

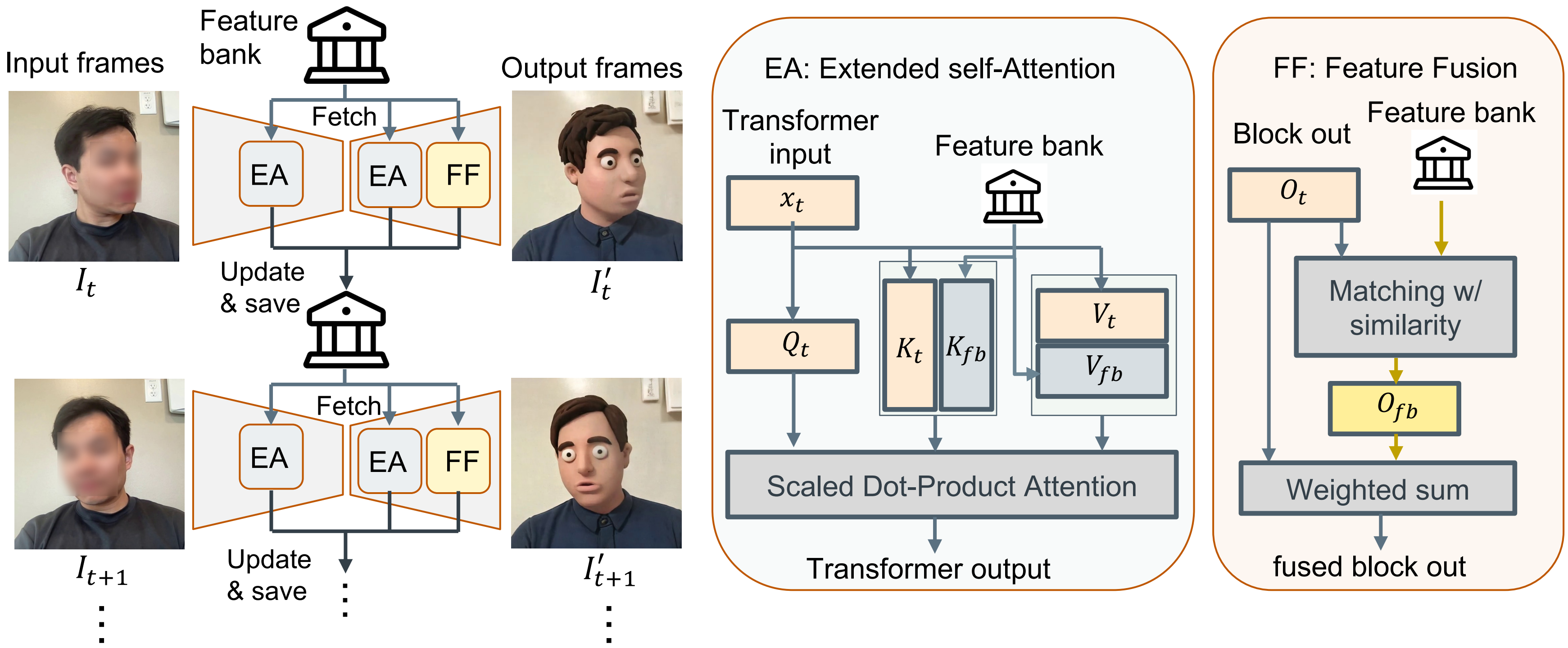
K_1 V_1 q_1 | K_2 V_2 | K_3 V_3 | K_4 V_4 | K_5 V_5 p

I_1 | I_2 | I_3 | I_4 | I_5

The cached self-attention diffusion features contain rich **semantic correspondence**

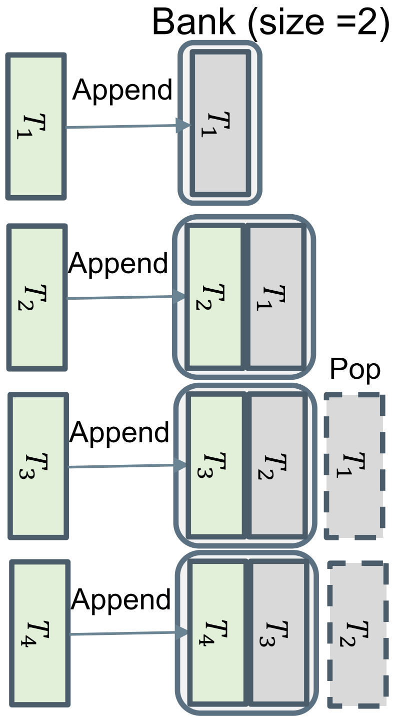
[1] Tang, Luming, et al. "Emergent correspondence from image diffusion." Advances in Neural Information Processing Systems 36 (2023): 1363-1389.
[2] Luo, Grace, et al. "Diffusion hyperfeatures: Searching through time and space for semantic correspondence." Advances in Neural Information Processing Systems 36 (2024).

Maintain a feature bank to track information



Dynamic merging for bank updates

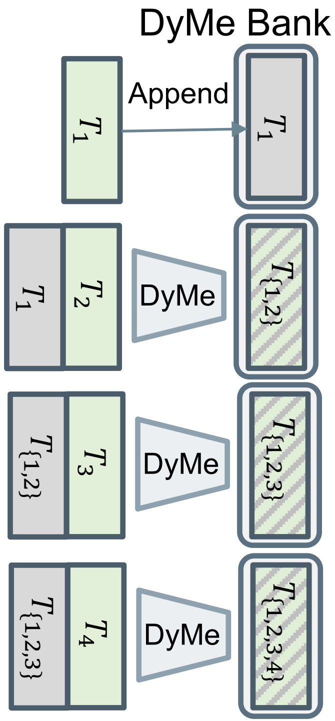
(a) Naïve queue



Problems of queue bank

- Limited span
- Redundant features

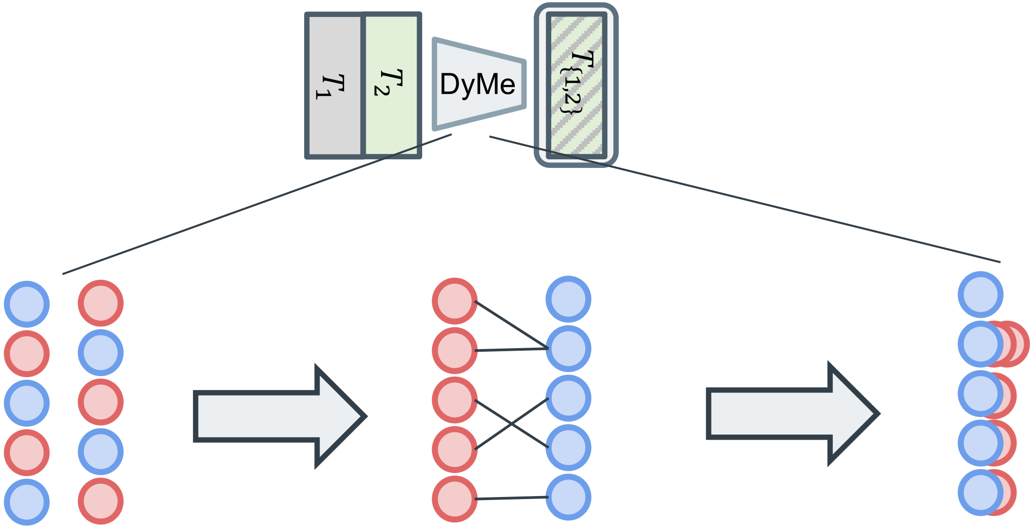
(b) Dynamic merging (ours)



Advantages of DyMe

- Unlimited span
- Compact size

Dynamic merging for bank updates



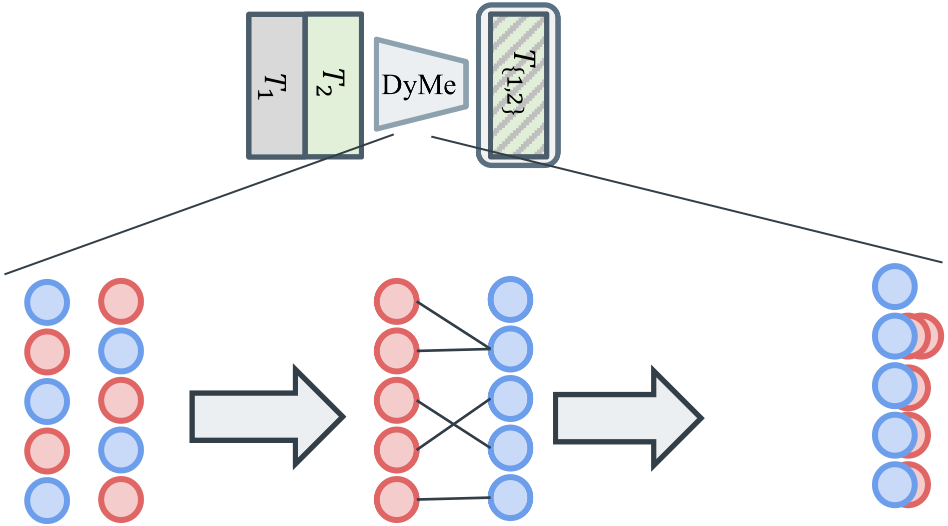
Step1: Randomly split tokens into two groups *Set A* and *Set B*

Step2: For every token in *Set A*, find its most similar token in *Set B*

Step3: Merge *Set A* to *Set B* by averaging

[1] Bolya, Daniel, et al. "Token merging: Your vit but faster." arXiv preprint arXiv:2210.09461 (2022).
[2] Bolya, Daniel, and Judy Hoffman. "Token merging for fast stable diffusion." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

Dynamic merging for bank updates



Step1: Randomly split tokens into two groups *Set A* and *Set B*

Step2: For every token in *Set A*, find its most similar token in *Set B*

Step3: Merge *Set A* to *Set B* by averaging

[1] Bolya, Daniel, et al. "Token merging: Your vit but faster." arXiv preprint arXiv:2210.09461 (2022).
[2] Bolya, Daniel, and Judy Hoffman. "Token merging for fast stable diffusion." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

Experiments: Quantitative results

Our evalution dataset contains 18 DAVIS videos and 67 video-prompt pairs

Table 1: **Quantitative metrics comparison.** We report the CLIP score and warp error to indicate the consistency of generated videos. We bold the **best** result and underline the second best.

	StreamDiffusion	CoDeF	Rerender	TokenFlow	FlowVid	StreamV2V (ours)
CLIP score ↑	95.24	96.33	96.20	**97.04**	96.68	96.58
Warp error ↓	117.01	116.17	107.00	114.25	111.09	**102.99**

CLIP score*: TokenFlow > FlowVid > StreamV2V > CoDeF > Rerender > StreamDiffusion

Warp error*: StreamV2V < Rerender < FlowVid < TokenFlow < CoDeF < StreamDiffusion

* Quantitative metrics of generative models cannot directly translate to the performance

Experiments: Qualitative comparison

Prompt : `A pixel art of a man doing a handstand on the street`

Input video

StreamV2V (ours)

StreamDiffusion [1]

CoDeF [2]

Rerender [3]

FlowVid [4]

[1] Kodaira, Akio, et al. "StreamDiffusion: A Pipeline-level Solution for Real-time Interactive Generation." arXiv preprint arXiv:2312.12491 (2023).
[2] Ouyang, Hao, et al. "Codef: Content deformation fields for temporally consistent video processing." arXiv preprint arXiv:2308.07926 (2023).
[3] Yang, Shuai, et al. "Rerender a video: Zero-shot text-guided video-to-video translation." SIGGRAPH Asia 2023 Conference Papers. 2023.
[4] Liang, Feng, et al. "FlowVid: Taming Imperfect Optical Flows for Consistent Video-to-Video Synthesis." arXiv preprint arXiv:2312.17681 (2023).

Experiments: User study results

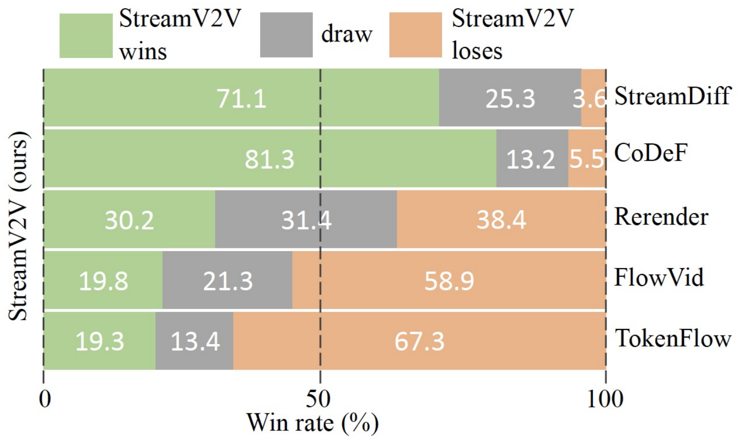


Figure 6: **User study comparison.** The win rate indicates the frequency our StreamV2V is preferred compared with certain counterpart.

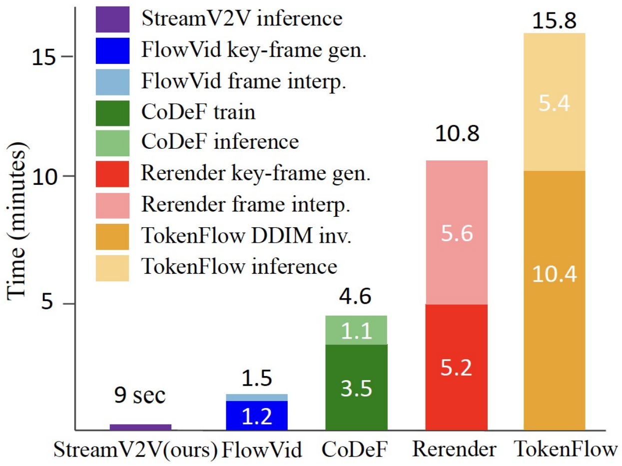


Figure 7: **Runtime breakdown** on one A100 GPU of generating a 4-second 512x512 resolution video with 30 FPS.

Regarding performance, StreamV2V is
- Better than StreamDiffusion, CoDeF
- Comparable with Rerender
- Worse than FlowVid, TokenFloe

Regarding speed, StreamV2V is
- 10X faster than FlowVid
- 72X faster than Rerender
- 100X faster than TokenFlow

Ablation : EA and FF

Extended self-Attention (EA) and Feature Fusion (FF)

Warp Error: 85.2 Warp Error: 74.0 Warp Error: 80.4 Warp Error: 73.4

Summary of StreamV2V

- StreamV2V is the one of the first approaches to **tackle real-time video-to-video translation** for streaming videos
- StreamV2V employs a simple yet effective looking-backward principle by **maintaining a feature bank to improve consistency**
- StreamV2V develop **a dynamic feature bank updating strategy** that merges redundant features, ensuring the feature bank remains both compact and descriptive