RAVE: Randomized Noise Shuffling for Fast and Consistent Video Editing with Diffusion Models

Ozgur Kara[1*] Bariscan Kurtkaya[2*] Hidir Yesiltepe[4] James M. Rehg[1,3] Pinar Yanardag[4]

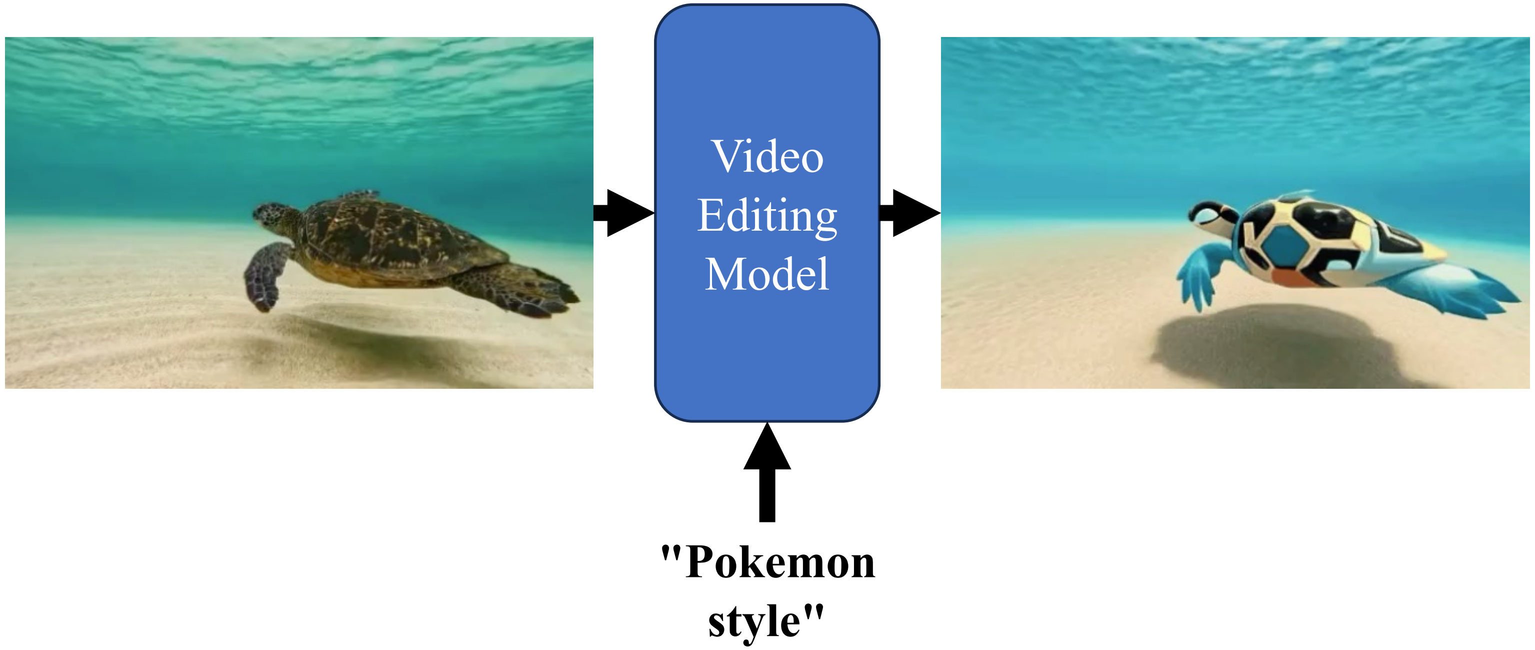
[1]Georgia Institute of Technology [2]KUIS AI Center [3]University of Illinois Urbana-Champaign [4]Virginia Tech

CVPR 2024 (Highlight)

https://rave-video.github.io/

Introduction

Problem Definition: Text guided video editing



Introduction

Problem Definition: Text guided video editing

TL; DR: RAVE is a fast

Introduction

Problem Definition: Text guided video editing

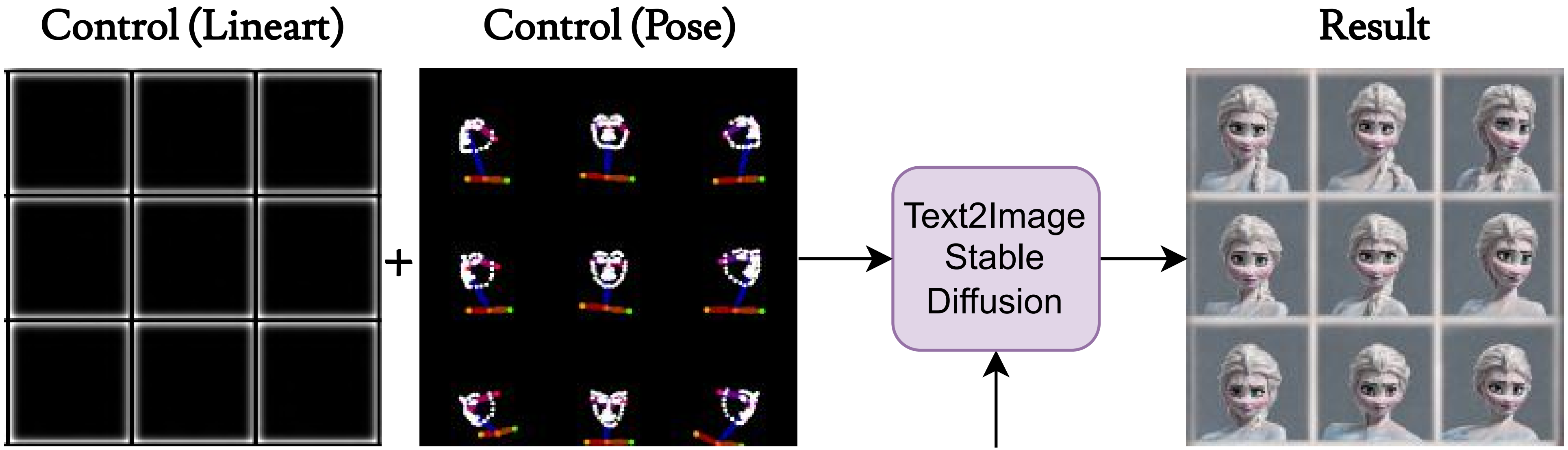
TL; DR: RAVE is a fast, zero-shot framework for text-guided video editing

Introduction

Problem Definition: Text guided video editing

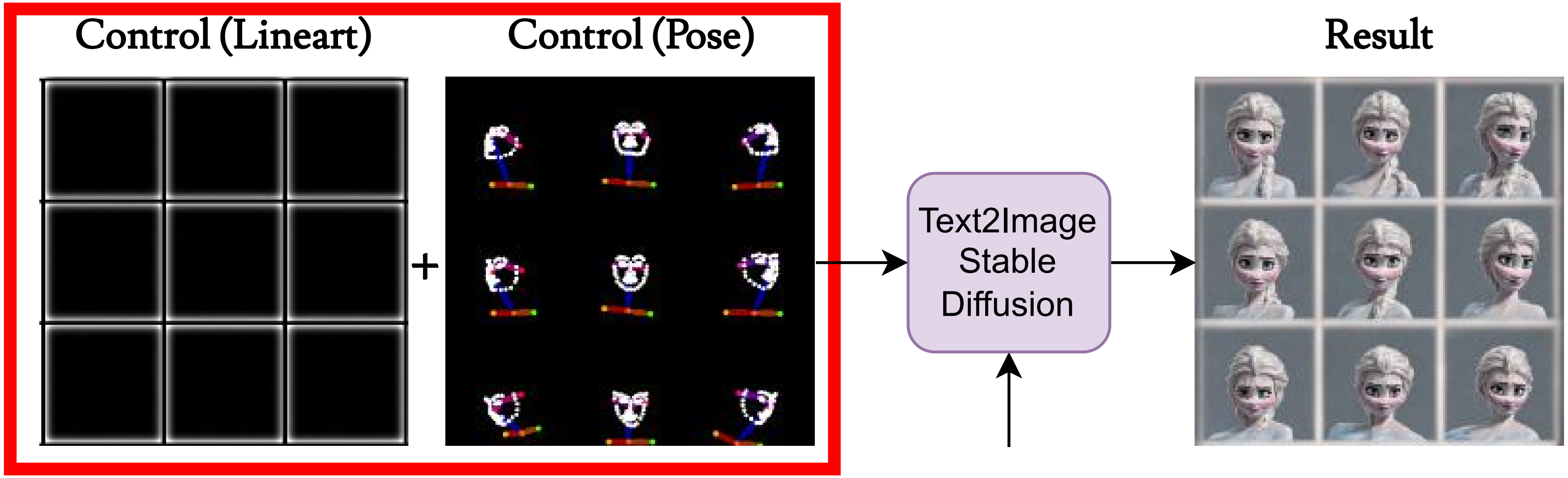
TL; DR: RAVE is a fast, zero-shot framework for text-guided video editing, compatible with off-the-shelf pretrained text-to-image (T2I) diffusion models.

Motivation – Grid Trick



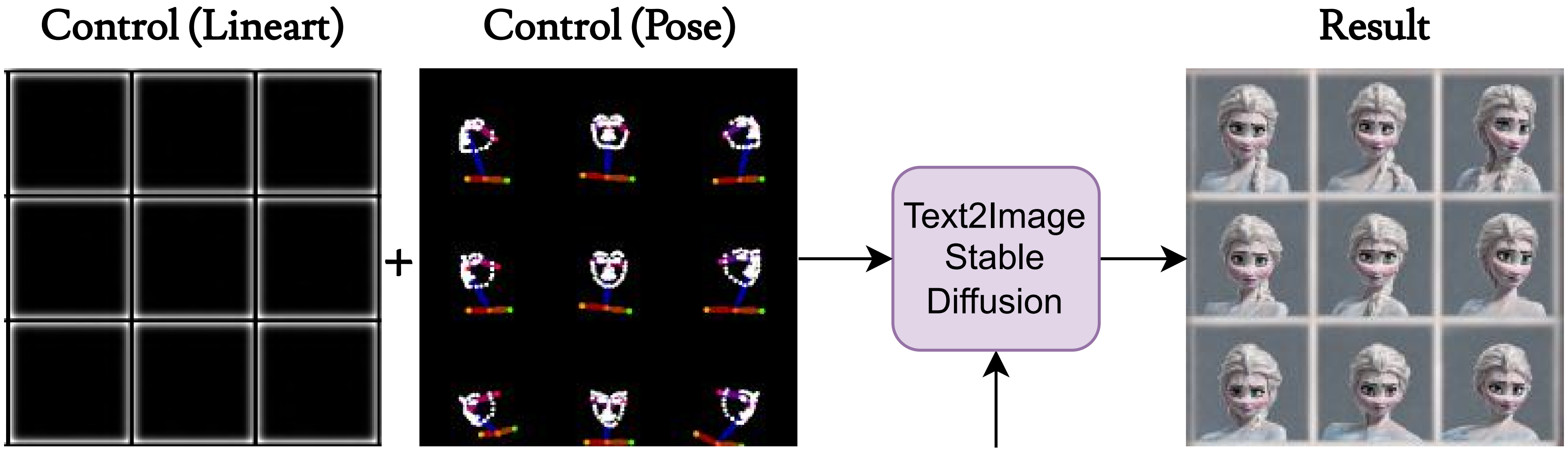
"(a character sheet of Elsa from different angles with a gray background:1.4), white hair,blue eyes open, cinematic lighting, Hyperrealism, depth of field, photography..."

Motivation – Grid Trick



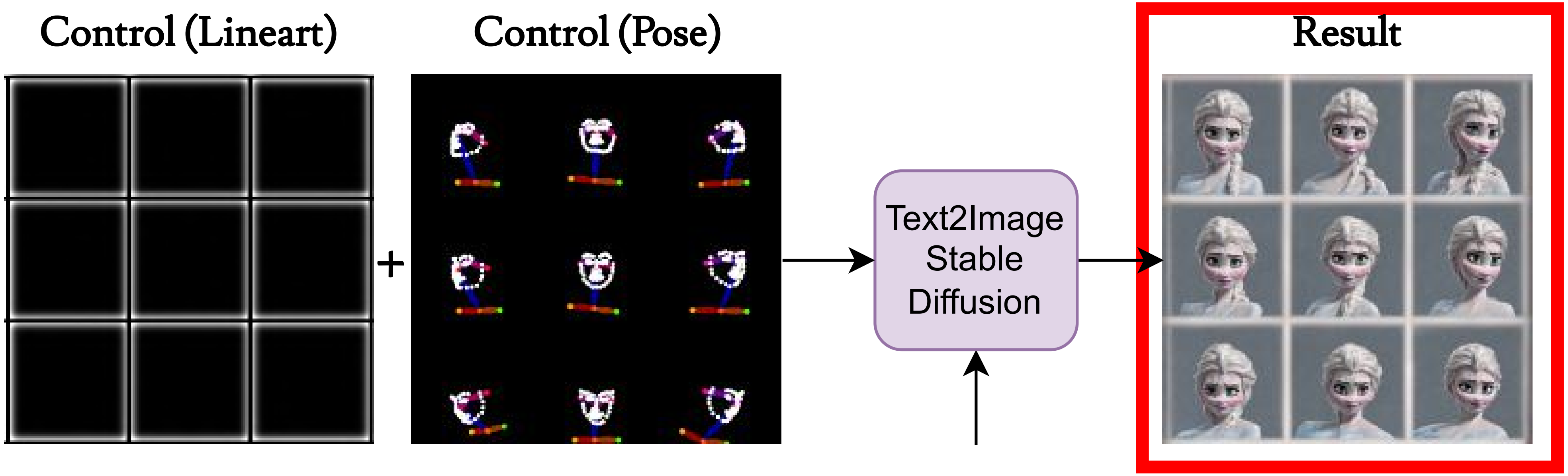
"(a character sheet of Elsa from different angles with a gray background:1.4), white hair,blue eyes open, cinematic lighting, Hyperrealism, depth of field, photography..."

Motivation – Grid Trick



"(a character sheet of Elsa from different angles with a gray background:1.4), white hair,blue eyes open, cinematic lighting, Hyperrealism, depth of field, photography..."

Motivation – Grid Trick



"(a character sheet of Elsa from different angles with a gray background:1.4), white hair,blue eyes open, cinematic lighting, Hyperrealism, depth of field, photography..."

Motivation - Extension to Video Domain

Question: How to extend the ‘grid trick’ for zero-shot video editing?

Motivation - Extension to Video Domain

Question: How to extend the 'grid trick' for zero-shot video editing?

Answer 1: Processing grids independently?

"a pink car in a snowy landscape, sunset lighting"

Answer 1: Grid

Motivation - Extension to Video Domain

Question: How to extend the ‘grid trick’ for zero-shot video editing?

~~**Answer 1:** Processing grids independently?~~

Answer 2: Adapting sparse-causal (SC) attention using grids?

"a pink car in a snowy landscape, sunset lighting"

~~**Answer 1:** Grid~~

Answer 2: Grid + SC

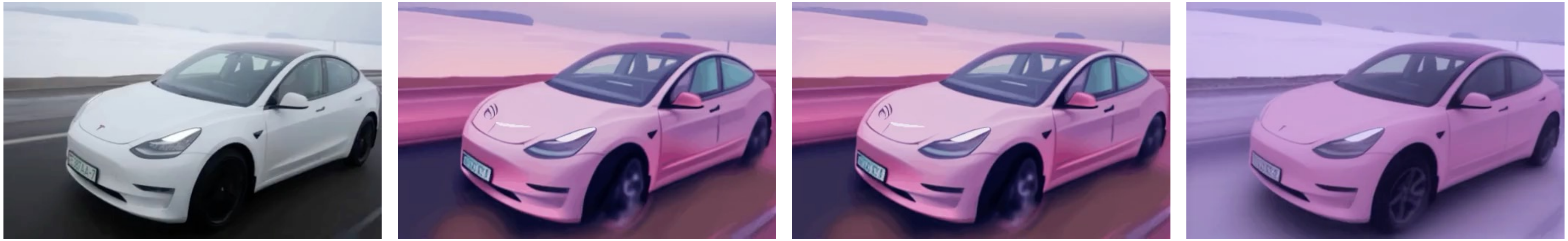
Motivation - Extension to Video Domain

Question: How to extend the 'grid trick' for zero-shot video editing?

~~**Answer 1:** Processing grids independently?~~

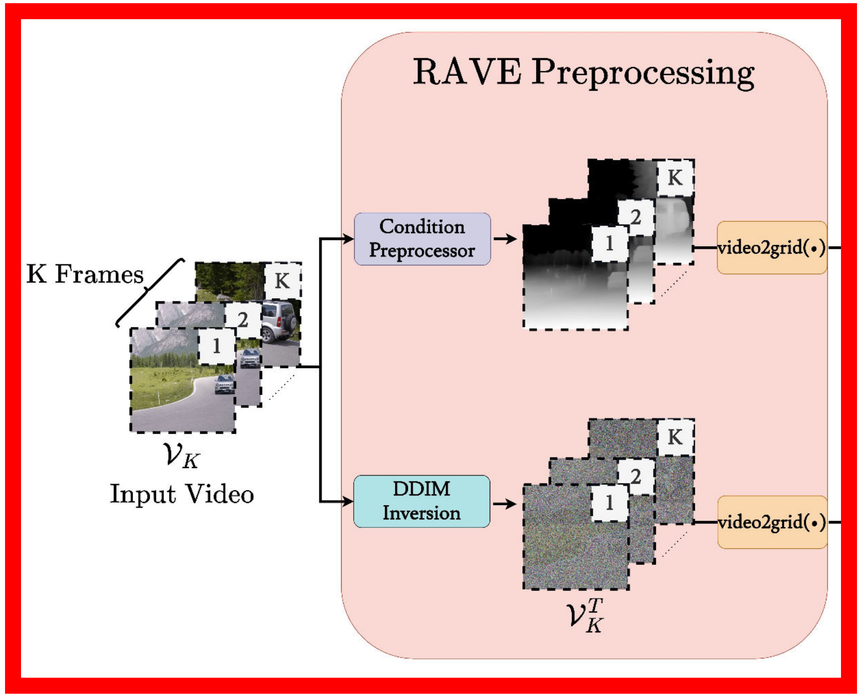
~~**Answer 2:** Adapting sparse-causal (SC) attention using grids?~~

Answer: **RAVE**

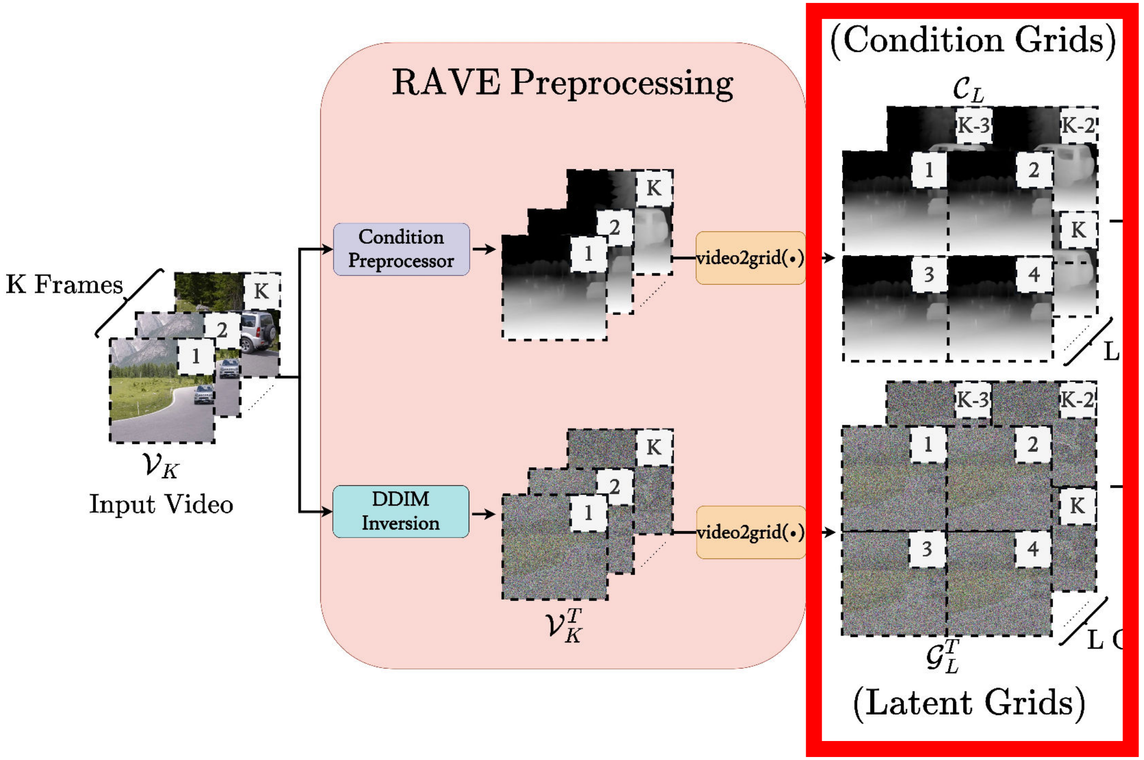


"a pink car in a snowy landscape, sunset lighting" | ~~**Answer 1:** Grid~~ | ~~**Answer 2:** Grid + SC~~ | **Answer:** RAVE

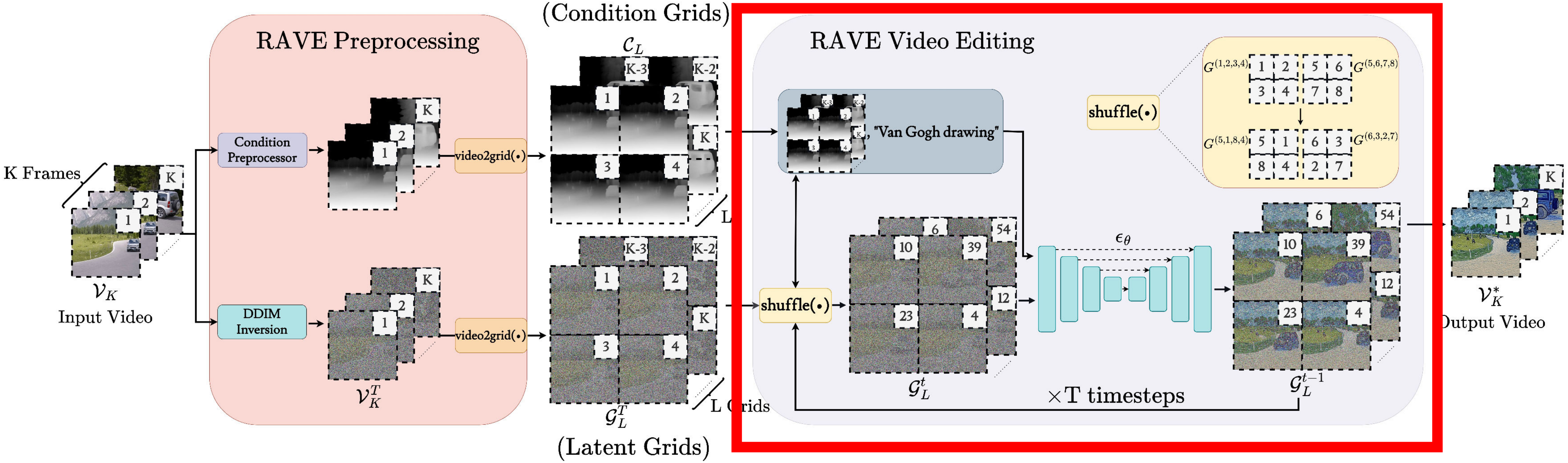
Methodology



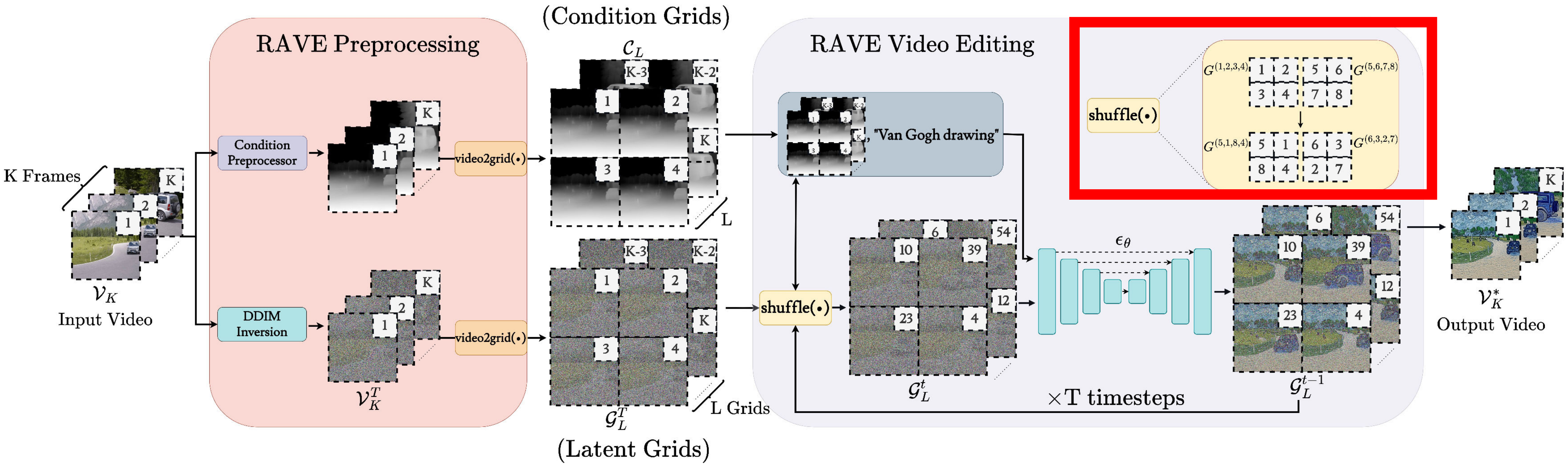
Methodology



Methodology



Methodology



Quantitative Results

Method	CLIP-F ($\times 10^{-2}$) ↑			WarpSSIM ($\times 10^{-2}$) ↑			CLIP-T ($\times 10^{-2}$) ↑			Q_{edit} ($\times 10^{-5}$) ↑			User Study ↑			Runtime ↓
	8-frames	36-frames	90-frames	8-frames	36-frames	90-frames	8-frames	36-frames	90-frames	8-frames	36-frames	90-frames	Q1 (GE)	Q2 (TC)	Q3 (TA)	90-frames
Text2Video-Zero	95.49	92.89	94.35	67.97	36.65	71.57	*29.46*	29.42	*29.73*	20.02	10.78	21.27	*47.95%*	24.87%	*52.56%*	5:33
Rerender	92.87	89.71	90.63	68.57	44.54	74.56	25.65	27.42	27.55	17.66	12.24	20.51	17.44%	23.33%	17.18%	5:24
TokenFlow	*95.80*	*93.17*	*95.92*	**74.03**	**50.97**	*80.40*	28.27	28.29	29.53	*20.92*	*14.41*	*23.74*	44.10%	*68.97%*	43.59%	5:24 (4.14)
Pix2Video	89.96	-	-	24.78	-	-	28.01	-	-	5.61	-	-	N/A	N/A	N/A	-
RAVE - w/o shuffle	93.98	89.90	92.49	*71.78*	47.26	76.58	28.78	*29.49*	29.71	20.66	13.94	22.76	N/A	N/A	N/A	N/A
RAVE	**95.95**	**93.18**	**95.99**	71.44	*48.81*	**80.51**	**29.51**	**29.93**	**29.76**	**21.08**	**14.60**	**23.95**	**90.51%**	**82.82%**	**86.67%**	**4:28 (3:13)**

Table 1. ***Quantitative comparison.*** CLIP-F, WarpSSIM, CLIP-T, and Q_{edit} metrics are reported individually on videos of 8, 36, and 90 frames. The user study section reports the frequency of each method chosen among the top two edits for General Editing (Q1 (GE)), Temporal Consistency (Q2 (TC)), and Textual Alignment (Q3 (TA)). The last column presents video-editing runtime in 'minutes:seconds' format for 90 frames for the entire pipeline, including preprocessing and editing stages (parentheses indicate runtime w/o preprocessing). '-' denotes methods that cannot be measured due to excessive memory requirements, while 'N/A' indicates that the value is not available.

Quantitative Results

Method	CLIP-F ($\times 10^{-2}$) $\uparrow$			WarpSSIM ($\times 10^{-2}$) $\uparrow$			CLIP-T ($\times 10^{-2}$) $\uparrow$			Q_{edit} ($\times 10^{-5}$) $\uparrow$			User Study $\uparrow$			Runtime $\downarrow$
	8-frames	36-frames	90-frames	8-frames	36-frames	90-frames	8-frames	36-frames	90-frames	8-frames	36-frames	90-frames	Q1 (GE)	Q2 (TC)	Q3 (TA)	90-frames
Text2Video-Zero	95.49	92.89	94.35	67.97	36.65	71.57	*29.46*	29.42	*29.73*	20.02	10.78	21.27	*47.95%*	24.87%	*52.56%*	5:33
Rerender	92.87	89.71	90.63	68.57	44.54	74.56	25.65	27.42	27.55	17.66	12.24	20.51	17.44%	23.33%	17.18%	5:24
TokenFlow	*95.80*	*93.17*	*95.92*	**74.03**	**50.97**	*80.40*	28.27	28.29	29.53	*20.92*	*14.41*	*23.74*	44.10%	*68.97%*	43.59%	5:24 (4.14)
Pix2Video	89.96	-	-	24.78	-	-	28.01	-	-	5.61	-	-	N/A	N/A	N/A	-
RAVE - w/o shuffle	93.98	89.90	92.49	*71.78*	47.26	76.58	28.78	*29.49*	29.71	20.66	13.94	22.76	N/A	N/A	N/A	N/A
RAVE	**95.95**	**93.18**	**95.99**	71.44	*48.81*	**80.51**	**29.51**	**29.93**	**29.76**	**21.08**	**14.60**	**23.95**	**90.51%**	**82.82%**	**86.67%**	**4:28 (3:13)**

Table 1. ***Quantitative comparison.*** CLIP-F, WarpSSIM, CLIP-T, and Q_{edit} metrics are reported individually on videos of 8, 36, and 90 frames. The user study section reports the frequency of each method chosen among the top two edits for General Editing (Q1 (GE)), Temporal Consistency (Q2 (TC)), and Textual Alignment (Q3 (TA)). The last column presents video-editing runtime in 'minutes:seconds' format for 90 frames for the entire pipeline, including preprocessing and editing stages (parentheses indicate runtime w/o preprocessing). '-' denotes methods that cannot be measured due to excessive memory requirements, while 'N/A' indicates that the value is not available.

Quantitative Results

Method	CLIP-F ($\times 10^{-2}$) ↑			WarpSSIM ($\times 10^{-2}$) ↑			CLIP-T ($\times 10^{-2}$) ↑			Q_{edit} ($\times 10^{-5}$) ↑			User Study ↑			Runtime ↓
	8-frames	36-frames	90-frames	8-frames	36-frames	90-frames	8-frames	36-frames	90-frames	8-frames	36-frames	90-frames	Q1 (GE)	Q2 (TC)	Q3 (TA)	90-frames
Text2Video-Zero	95.49	92.89	94.35	67.97	36.65	71.57	*29.46*	29.42	*29.73*	20.02	10.78	21.27	*47.95%*	24.87%	*52.56%*	5:33
Rerender	92.87	89.71	90.63	68.57	44.54	74.56	25.65	27.42	27.55	17.66	12.24	20.51	17.44%	23.33%	17.18%	5:24
TokenFlow	*95.80*	*93.17*	*95.92*	**74.03**	**50.97**	*80.40*	28.27	28.29	29.53	*20.92*	*14.41*	*23.74*	44.10%	*68.97%*	43.59%	5:24 (4.14)
Pix2Video	89.96	-	-	24.78	-	-	28.01	-	-	5.61	-	-	N/A	N/A	N/A	-
RAVE - w/o shuffle	93.98	89.90	92.49	*71.78*	47.26	76.58	28.78	*29.49*	29.71	20.66	13.94	22.76	N/A	N/A	N/A	N/A
RAVE	**95.95**	**93.18**	**95.99**	71.44	*48.81*	**80.51**	**29.51**	**29.93**	**29.76**	**21.08**	**14.60**	**23.95**	**90.51%**	**82.82%**	**86.67%**	**4:28 (3:13)**

Table 1. ***Quantitative comparison.*** CLIP-F, WarpSSIM, CLIP-T, and Q_{edit} metrics are reported individually on videos of 8, 36, and 90 frames. The user study section reports the frequency of each method chosen among the top two edits for General Editing (Q1 (GE)), Temporal Consistency (Q2 (TC)), and Textual Alignment (Q3 (TA)). The last column presents video-editing runtime in 'minutes:seconds' format for 90 frames for the entire pipeline, including preprocessing and editing stages (parentheses indicate runtime w/o preprocessing). '-' denotes methods that cannot be measured due to excessive memory requirements, while 'N/A' indicates that the value is not available.

Qualitative Results

Input Video

"A white cat"

"A dinosaur"

"A shiny silver robotic wolf, futuristic"

Qualitative Results

Input Video

"Swarovski blue crystal swan"

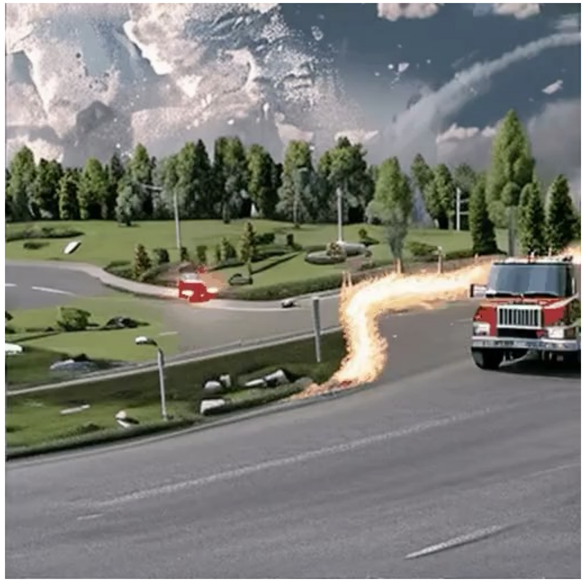
"crochet swan"

Qualitative Results – Extreme Shape Editing

Input Video

"Switzerland SBB CFF FFS train"

"a tractor"



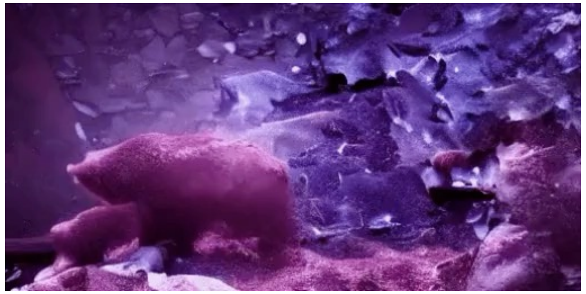
"a firetruck"

Comparisons to Baselines

"Mysterious purple and blue hues dominate, with twinkling stars and a glowing moon in the backdrop"

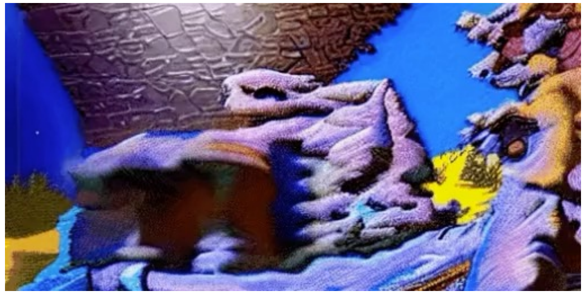
RAVE

RAVE w/o Shuffle



Tokenflow [1]

FateZero [2]



Rerender-A-Video [3]

Text2Video-Zero [4]

Ablation Study

"dark chocolate cake"

"RAVE"

"w/o Shuffling"

"w/o DDIM Inversion"

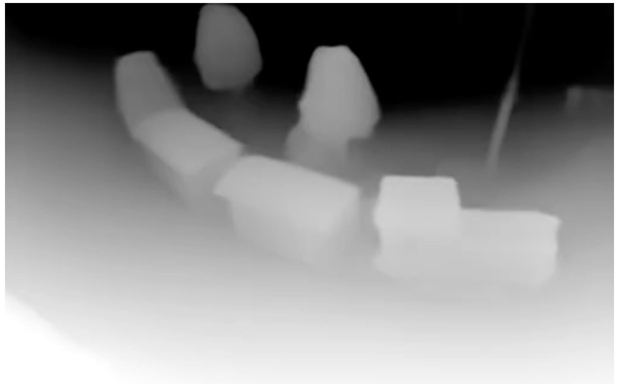
Ablation Study - Conditions

"dark chocolate cake"

"RAVE (Depth)"

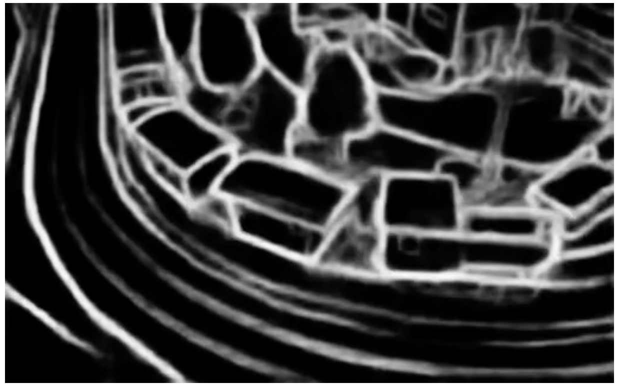
"w/ Lineart"

"w/ Softedge"



"Depth Control"

"Lineart Control"



"Softedge Control"

Project Webpage & Demo

Project Webpage

https://rave-video.github.io/

Huggingface Demo

https://huggingface.co/spaces/ozgurkara/RAVE

References

[1] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. arXiv preprint arXiv:2307.10373, 2023.

[2] Chenyang QI, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 15932–15942, 2023.

[3] Shuai Yang, Yifan Zhou, Ziwei Liu, , and Chen Change Loy. Rerender a video: Zero-shot text-guided video-to-video translation. In ACM SIGGRAPH Asia Conference Proceedings, 2023.

[4] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 15954–15964, 2023.