KAIST AI
Graduate School of AI

DreamMotion: Space-Time Self-Similar Score Distillation for Zero-shot Video Editing

Hyeonho Jeong, Jinho Chang, Geon Yeong Park, Jong Chul Ye

ECCV

Key Ideas

- **TL;DR:** We present a zero-shot, text-driven, diffusion-based video editing framework.
- **DreamMotion:** The first framework that utilizes text-to-video score distillation for video editing.
- **Appearance Injection**: Video score distillation effectively introduces new content indicated by the target text.
- **Problem of Score Distillation**: Inaccurate gradients of the score distillation cause significant structure and motion deviation.
- **Structure Correction**: Self-similarity matching across spatial dimension using diffusion features ensures structural correspondence between the input video and the target video.
- **Temporal Smoothing**: Self-similarity matching across temporal dimension using diffusion features facilitates temporal smoothing.

Score Distillation Sampling

Score Distillation Sampling (SDS) & Delta Denoising Score (DDS)

$\epsilon_\emptyset$: T2V diffusion model, $x_0^{1:N}(\theta)$: Target video parameterized by θ, y : Target text

$\hat{x}_0^{1:N}$: Source video, $\hat{y}$: Source text

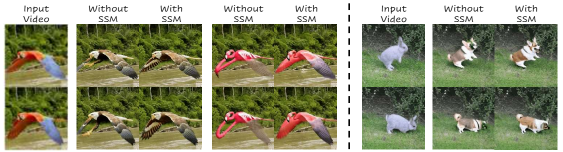
$$\mathcal{L}_{SDS}(\theta; y) = \left\| \epsilon_\emptyset(x_t^{1:N}(\theta), t, y) - \epsilon \right\|_2^2$$

$$\mathcal{L}_{DDS}(\theta; y) = \left\| \epsilon_\emptyset(x_t^{1:N}(\theta), t, y) - \epsilon_\emptyset(\hat{x}_t^{1:N}, t, \hat{y}) \right\|_2^2$$

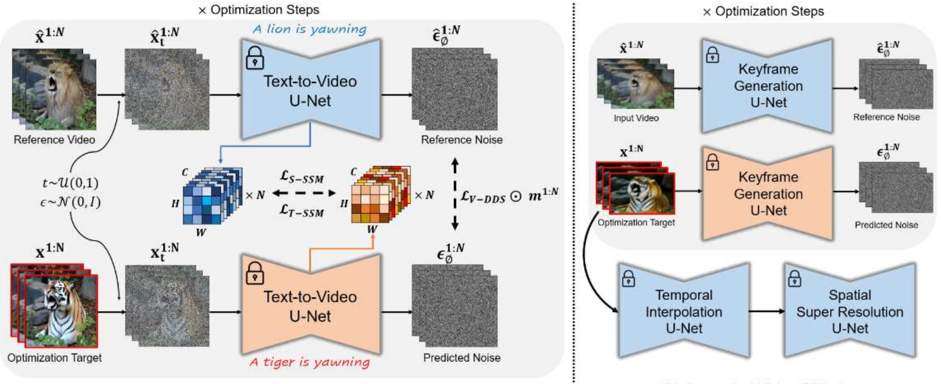
Why Score Distillation for Diffusion-based Video Editing?

- Conventional reverse diffusion process (ancestral sampling) struggles to reformulate real-world motion.
- Score Distillation-based gradients $\nabla_\theta \mathcal{L}_{SDS/DDS}$ enable optimizing a clean video variable $x_0^{1:N}(\theta)$ that already exhibits real-world, natural motion.

Necessity of Space-Time Self-Similarity



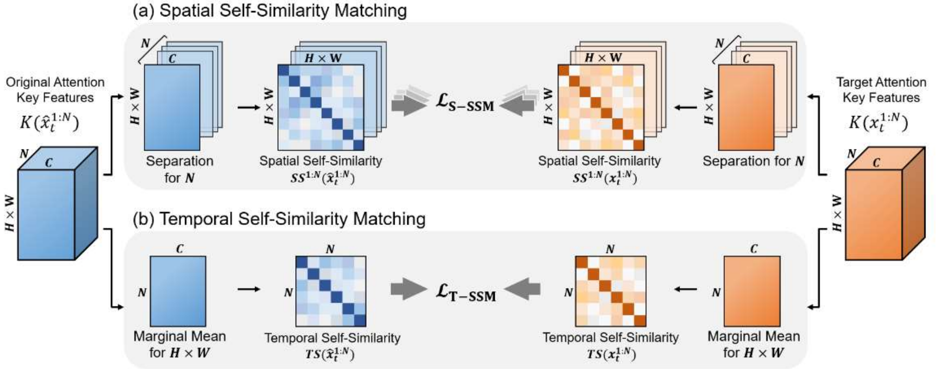
DreamMotion Overview



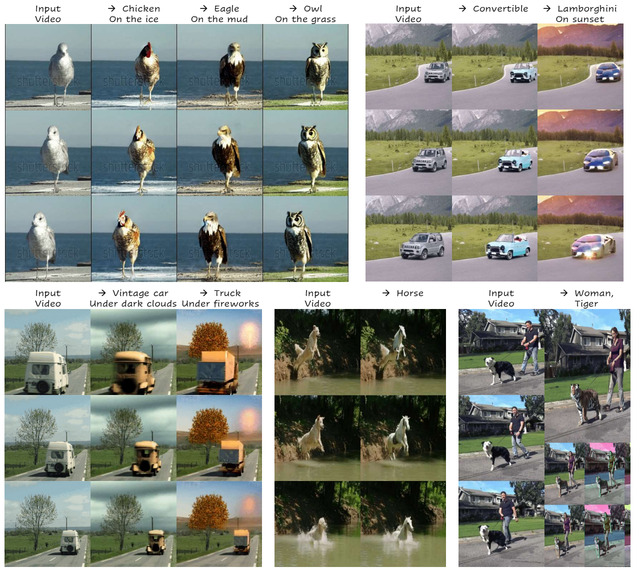
(a) Non-Cascaded Video Diffusion

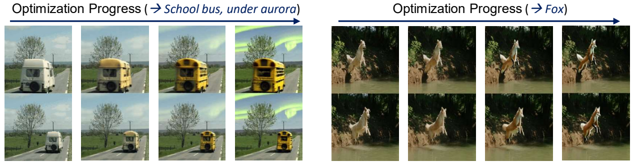
(b) Cascaded Video Diffusion

Space-Time Self-Similarity Matching



Experiment Results (Visit our page)




Optimization Progress (→ School bus, under aurora)
Optimization Progress (→ Fox)

Quantitative Comparison To Baselines

	Automatic Metrics				Human Evaluation		
Method	Text-Align	Frame-Con	Motion-Fidelity	Frame-LPIPS	Edit-Acc	Frame-Con	SM-Preserve
Tune-A-Video	0.8177	0.9218	0.6947	0.4172	3.52	2.82	2.89
ControlVideo	0.7850	0.9678	-	0.3763	2.74	2.68	2.03
Control-A-Video	0.7848	0.9297	0.8453	0.3829	2.17	2.16	2.18
Gen-1	<u>0.8192</u>	<u>0.9704</u>	-	-	3.31	<u>3.62</u>	2.95
Tokenflow	0.7813	0.9576	<u>0.9184</u>	<u>0.3427</u>	<u>3.63</u>	3.54	<u>3.92</u>
Ours (Zeroscope)	**0.8209**	**0.9726**	**0.9259**	**0.3042**	**4.14**	**4.21**	**4.33**