VMC: Video Motion Customization using Temporal Attention Adaption for Text-to-Video Diffusion Models

Hyeonho Jeong*, Geon Yeong Park*, Jong Chul Ye

KAIST

Project | Paper

CVPR JUNE 17-21, 2024 SEATTLE, WA

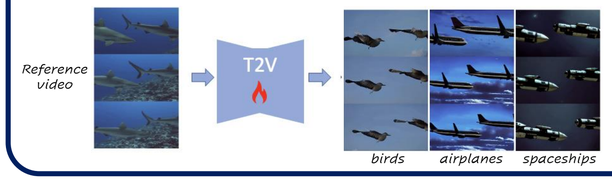
What is Motion Customization?

Appearance Customization (a.k.a. 'personalization')

: Given an image of a particular subject, adapt T2I model to generate images of that subject in various contexts.

Motion Customization

: Given a video of a particular motion, adapt T2V model to create videos that *(i) accurately reproduce the motion* but in *(ii) entirely distinct visual contexts*.

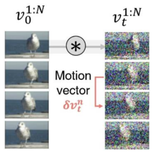


Key Idea

Frame residuals contains motion patterns.

→ Motion Distillation Objective

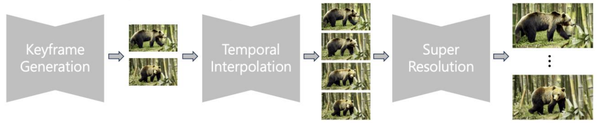
→ Appearance-invariant prompts



Solutions

Keyframes determine motion.

- Optimize keyframe generator
- Freeze temporal interpolation & super resolution models

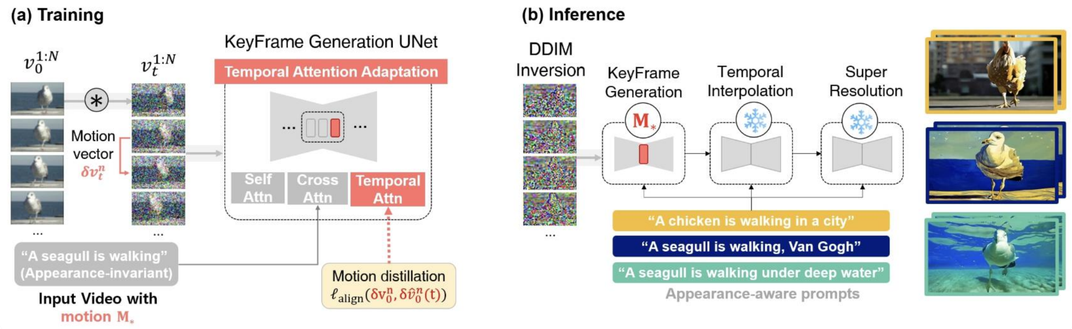


Temporal attentions have robust temporal prior.

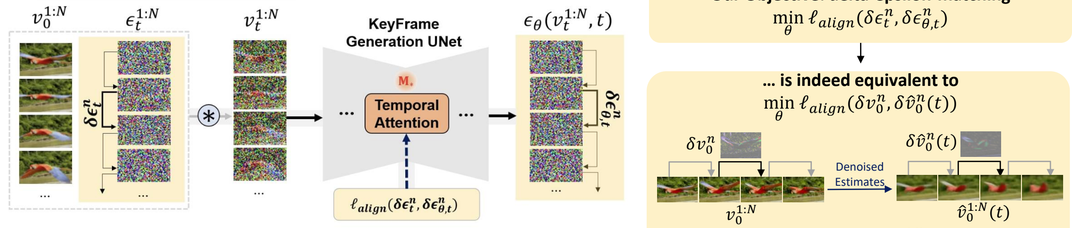
- Optimize temporal attentions only
- Freeze self & cross attentions

→ *Lightweight Training (15GB VRAM & < 5 minutes)*

VMC Overview



Motion Distillation



$$\min_\theta \mathbb{E}_{x_t,t,\epsilon}\left[\| \epsilon_\theta(x_t,t)-\epsilon \|\right] \quad \text{VS} \quad \min_\theta \mathbb{E}_{x_t,n,t,\epsilon}\left[\| \delta\epsilon^n_\theta(x_t^{1:N},t)-\delta\epsilon^n \|\right]$$

Appearance-Invariant Prompts

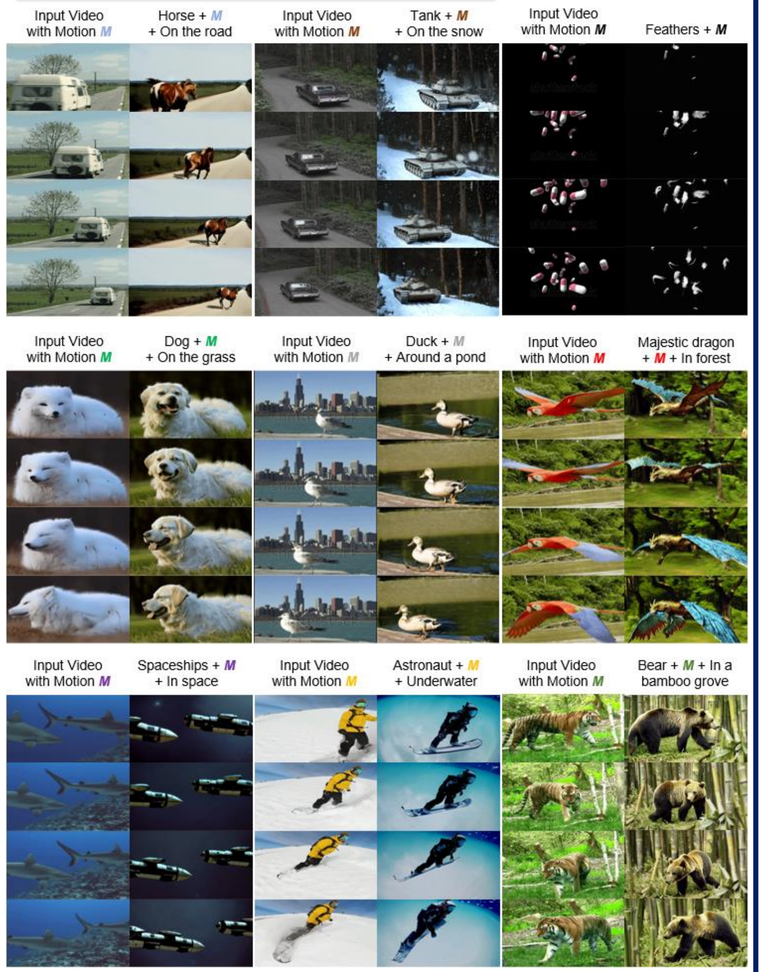
- ~~"A cat is roaring on the grass under the tree."~~ → "A cat is roaring."
- "Sharks are swimming ~~in the ocean on a coral reef~~." → "Sharks are moving."

Quantitative Results

	Automatic Evaluation (CLIP)		Human Evaluation			
	Text Alignment	Temporal Consistency	Motion Preservation	Appearance Diversity	Text Alignment	Temporal Consistency
Video Composer	0.798	0.958	3.45	3.43	2.96	3.03
Gen-1	0.780	0.957	3.46	3.17	2.87	2.73
Tune-A-Video	0.758	0.947	3.50	2.88	2.67	2.80
Control-A-Video	0.764	0.952	2.75	2.45	2.07	2.00
VMC (Ours)	**0.801**	**0.959**	**4.42**	**4.54**	**4.56**	**4.57**

- **Text-Alignment**: Average cosine similarity between target prompt and output frames
- **Temporal Consistency**: Average cosine similarity between all pairs of output frames

Qualitative Results (Visit our page)



VMC uses *Show-1* as the cascaded video diffusion backbone.

- Zhang, David Junhao, et al. "Show-1: Marrying pixel and latent diffusion models for text-to-video generation." *arXiv preprint arXiv:2309.15818* (2023).