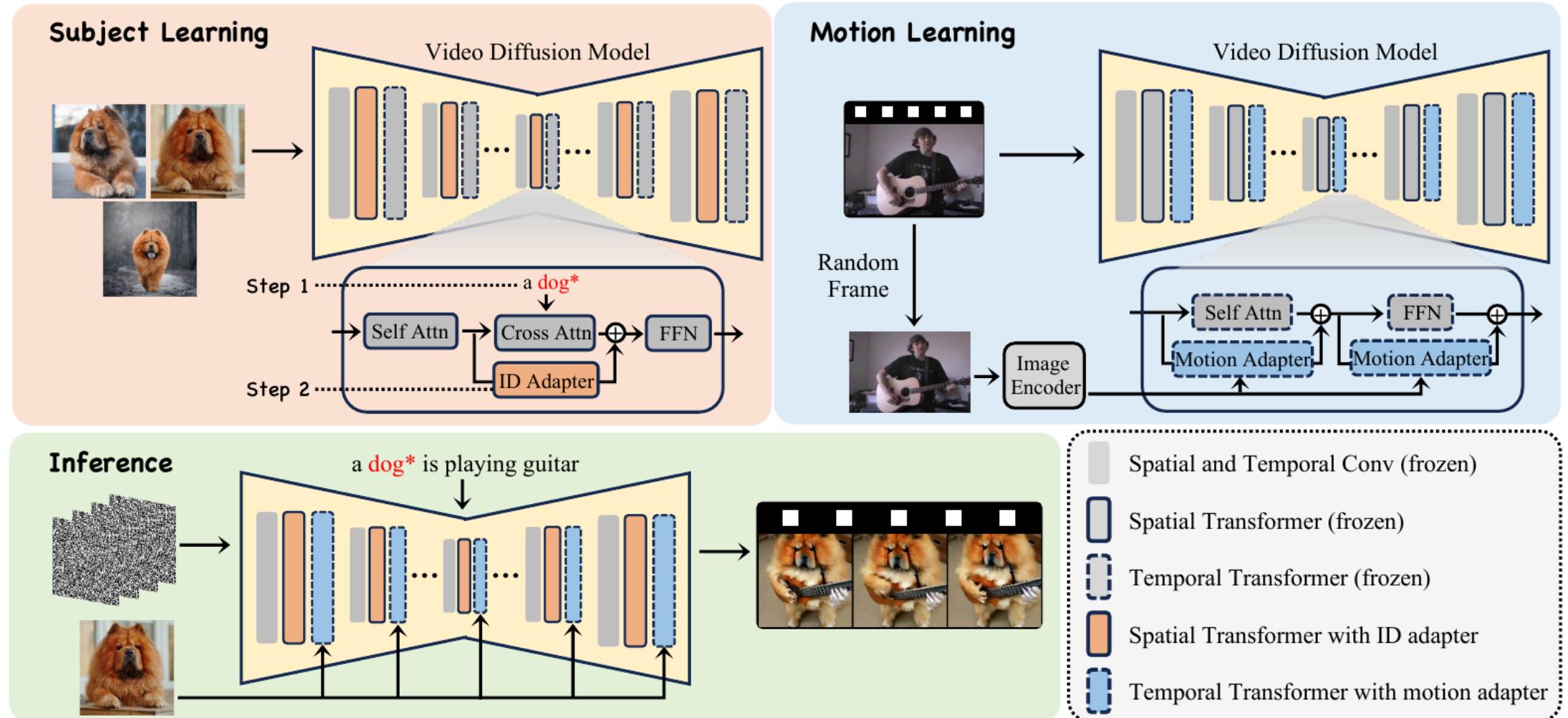


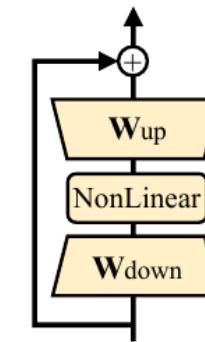
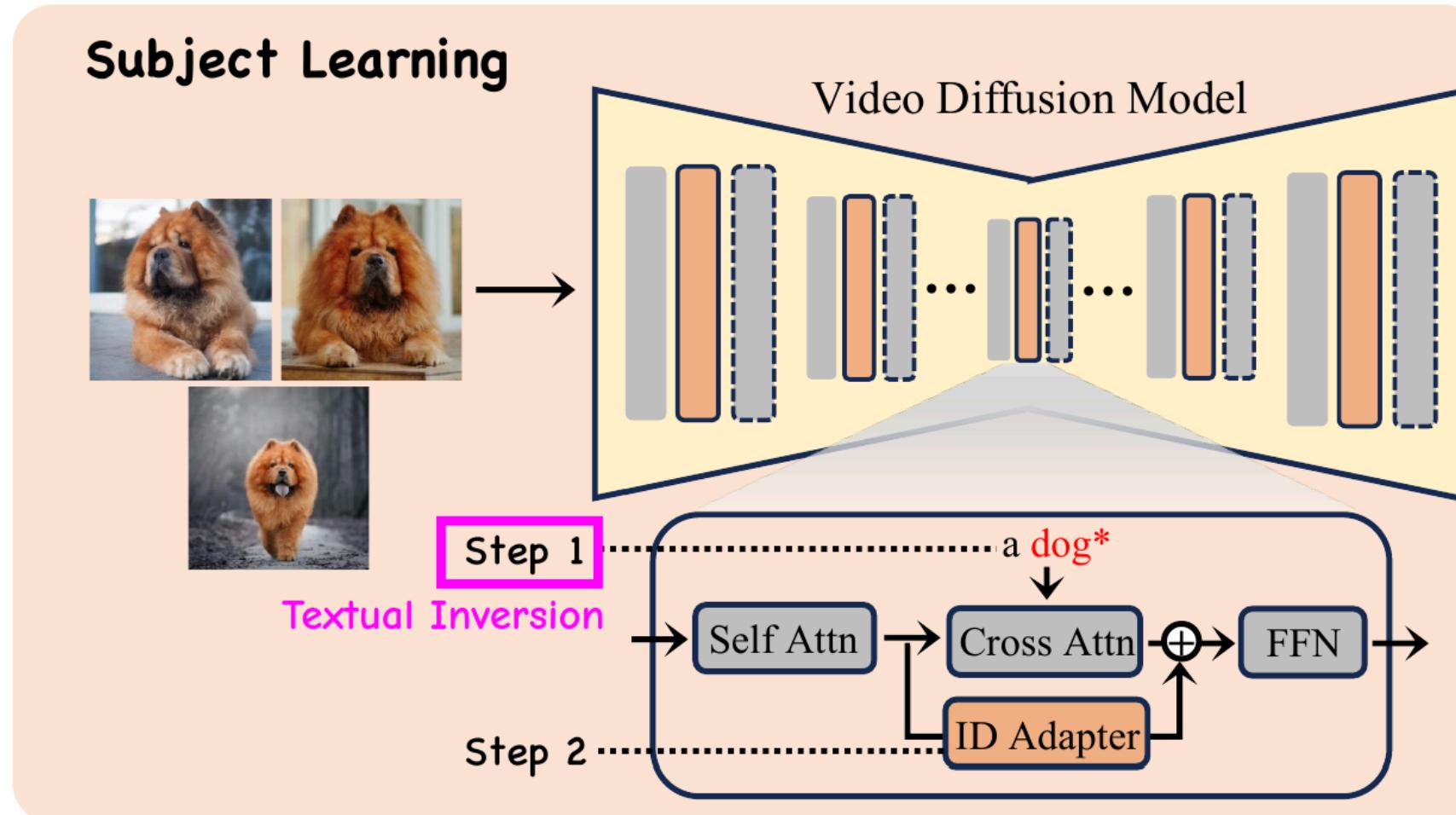
DreamVideo: Composing Your Dream Videos with Customized Subject and Motion

Wei et al, CVPR 2024

Overall Architecture



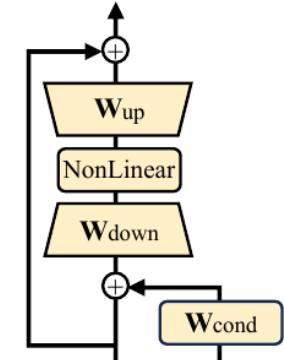
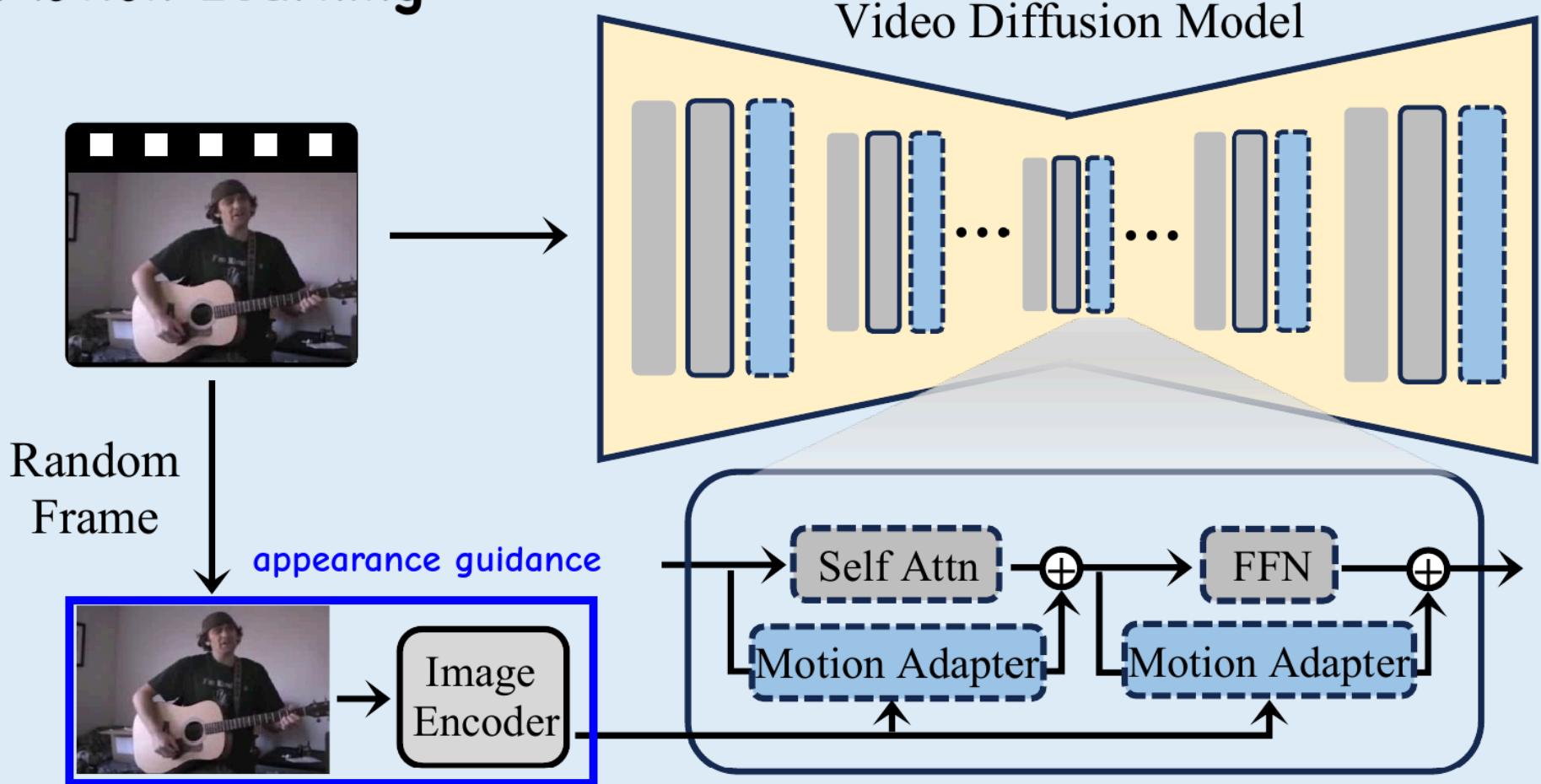
Subject Learning



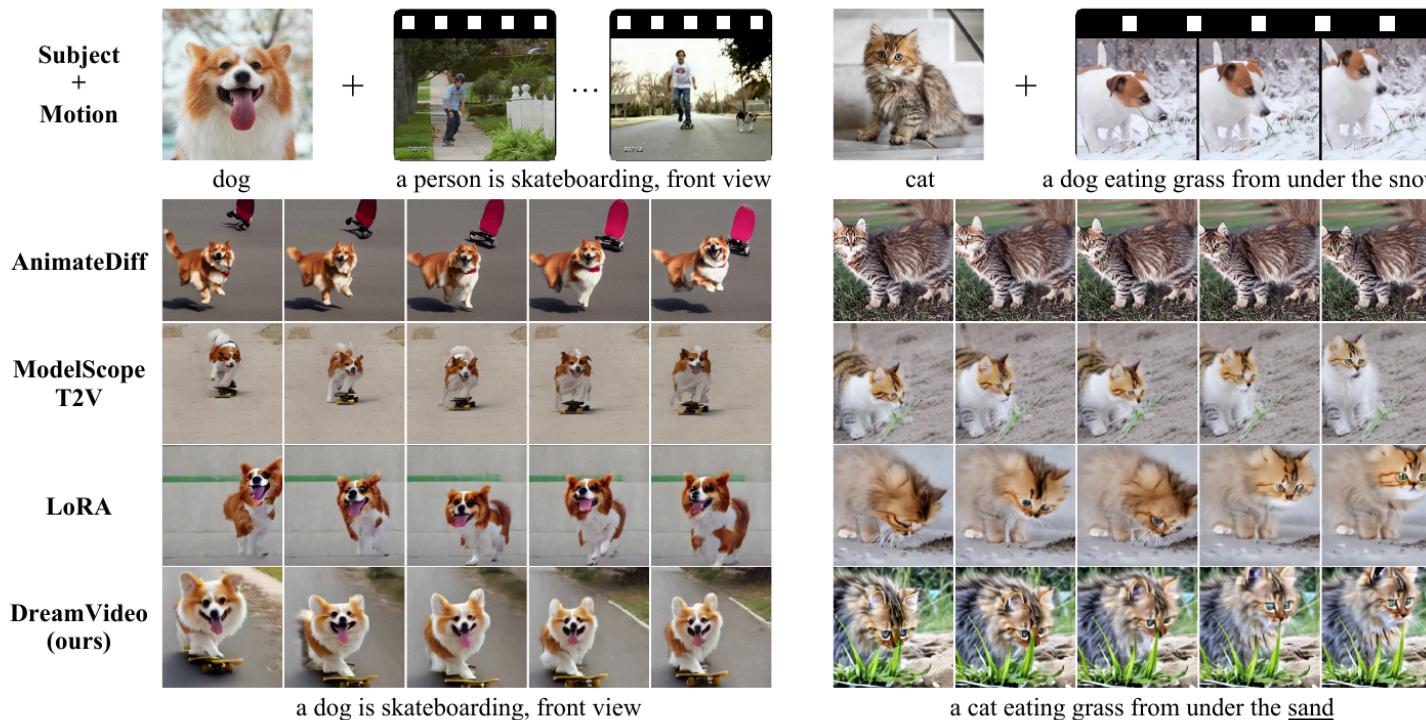
(a) Identity Adapter

Motion Learning

Motion Learning

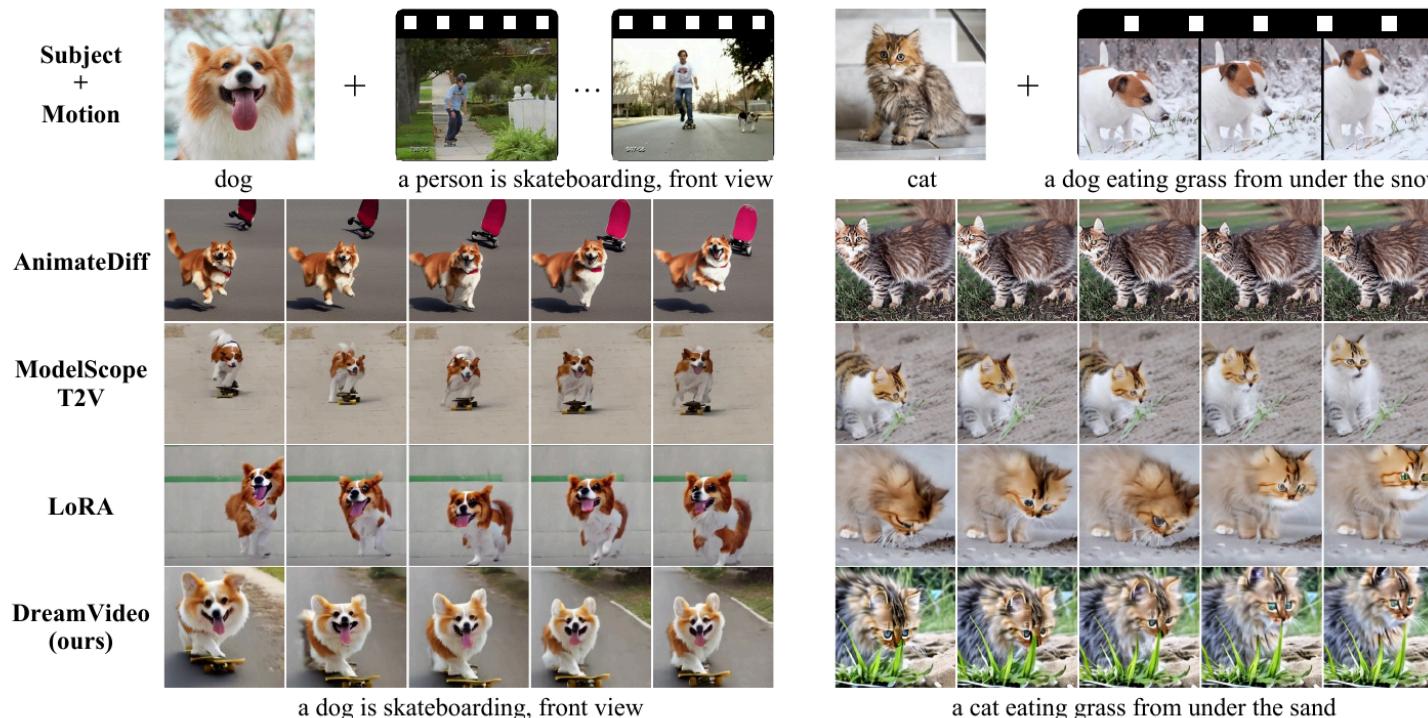


Qualitative Comparison: Subjet and Motion Customization



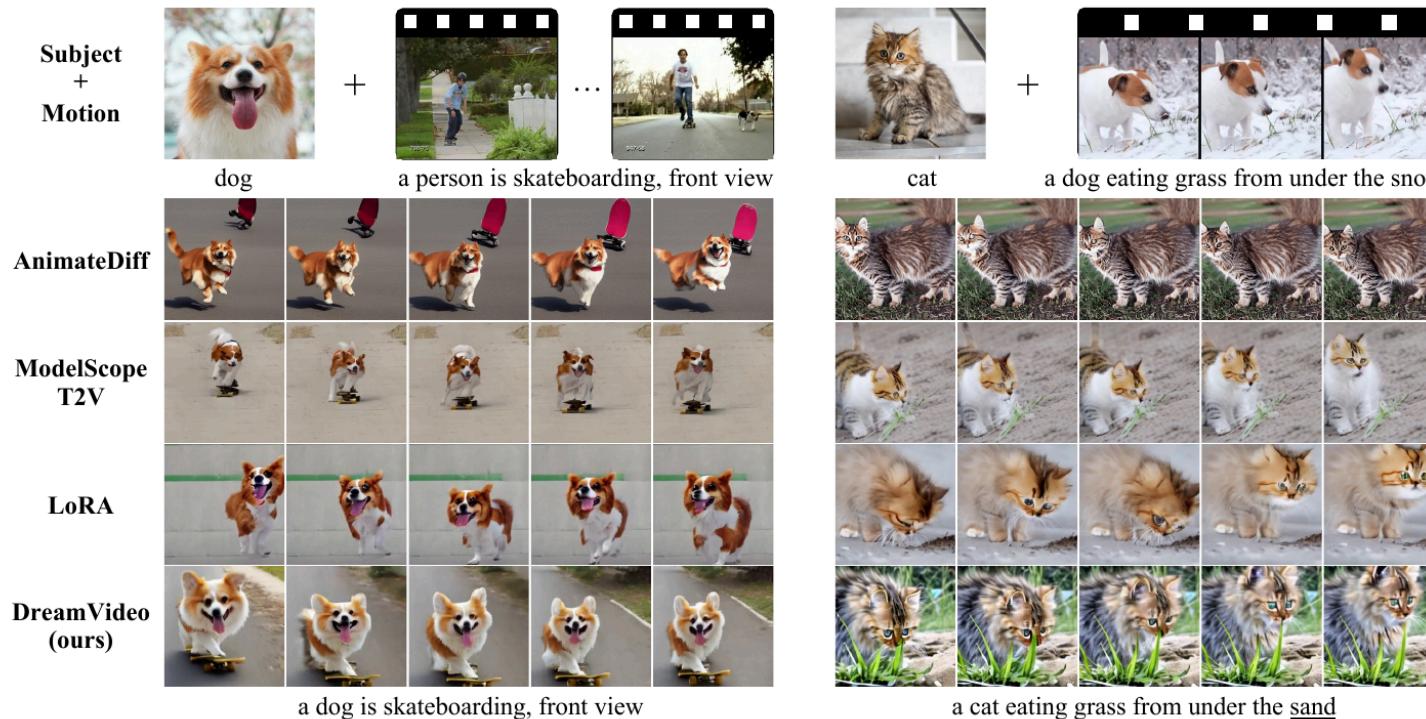
AnimateDiff (Guo et al, ICLR2024) — A newly-initialized **motion module** is trained from scratch. And it will be **appended to the DreamBooth-fine-tuned IMAGE diffusion model** to generate videos. That is, the DreamBooth learns the subject and the motion module learns the motion.

Qualitative Comparison: Subject and Motion Customization



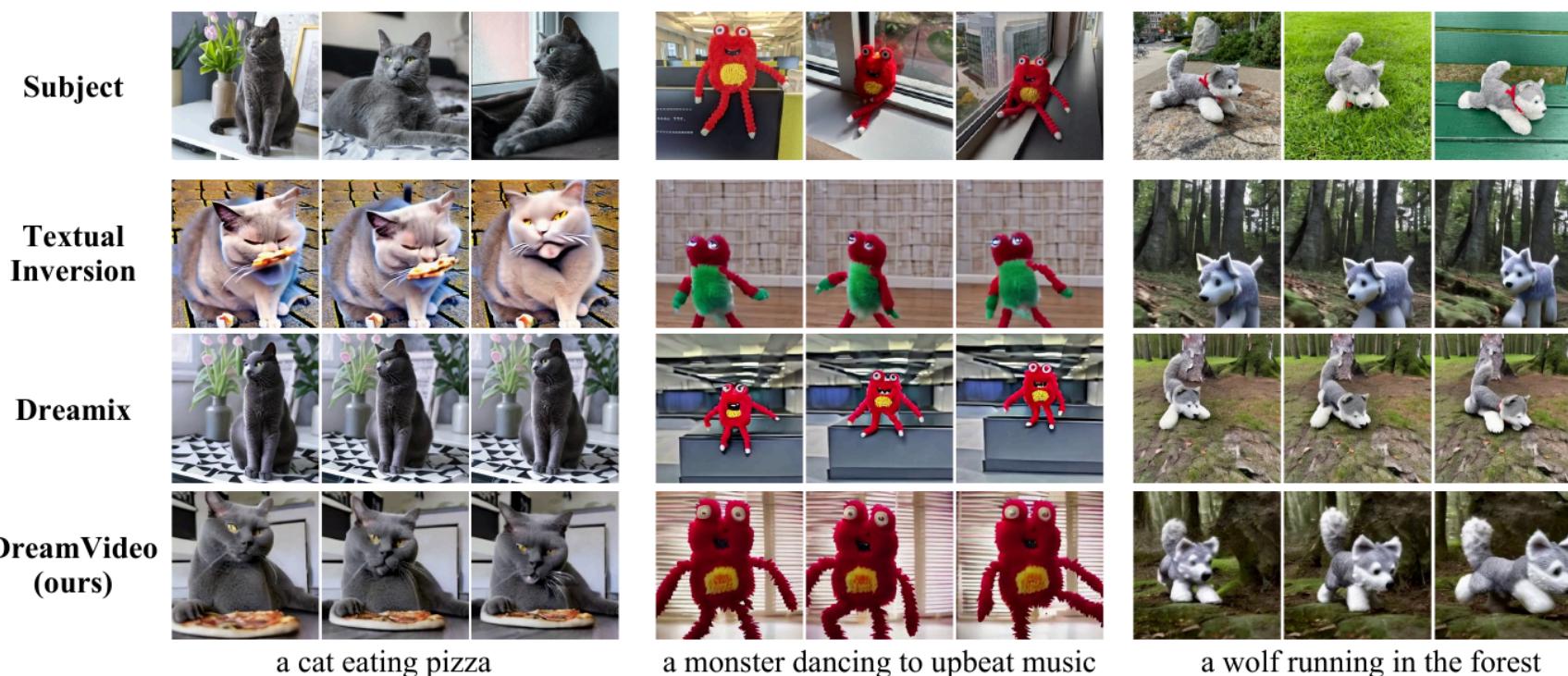
ModelScopeT2V (Wang et al, 2023) — the spatial/temporal parameters of the UNet are only fine-tuned, while other parameters are frozen.

Qualitative Comparison: Subject and Motion Customization



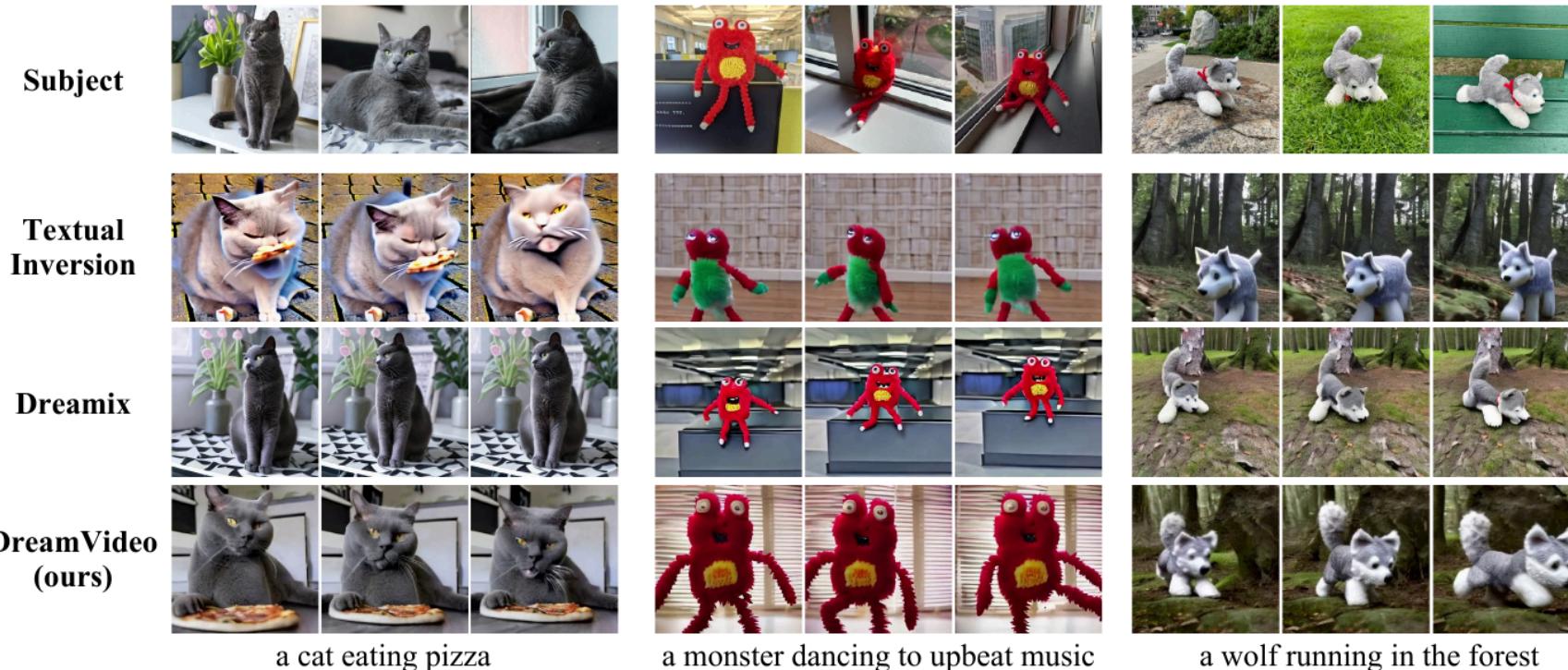
ModelScopeT2V (LoRA fine-tuned) — LoRA parameters are added to the key and value matrices in the cross-attention layers to learn a subject. For motion learning, LoRA parameters are added to the key/value matrices in all self-attention layers.

Qualitative Comparison: Subject Customization



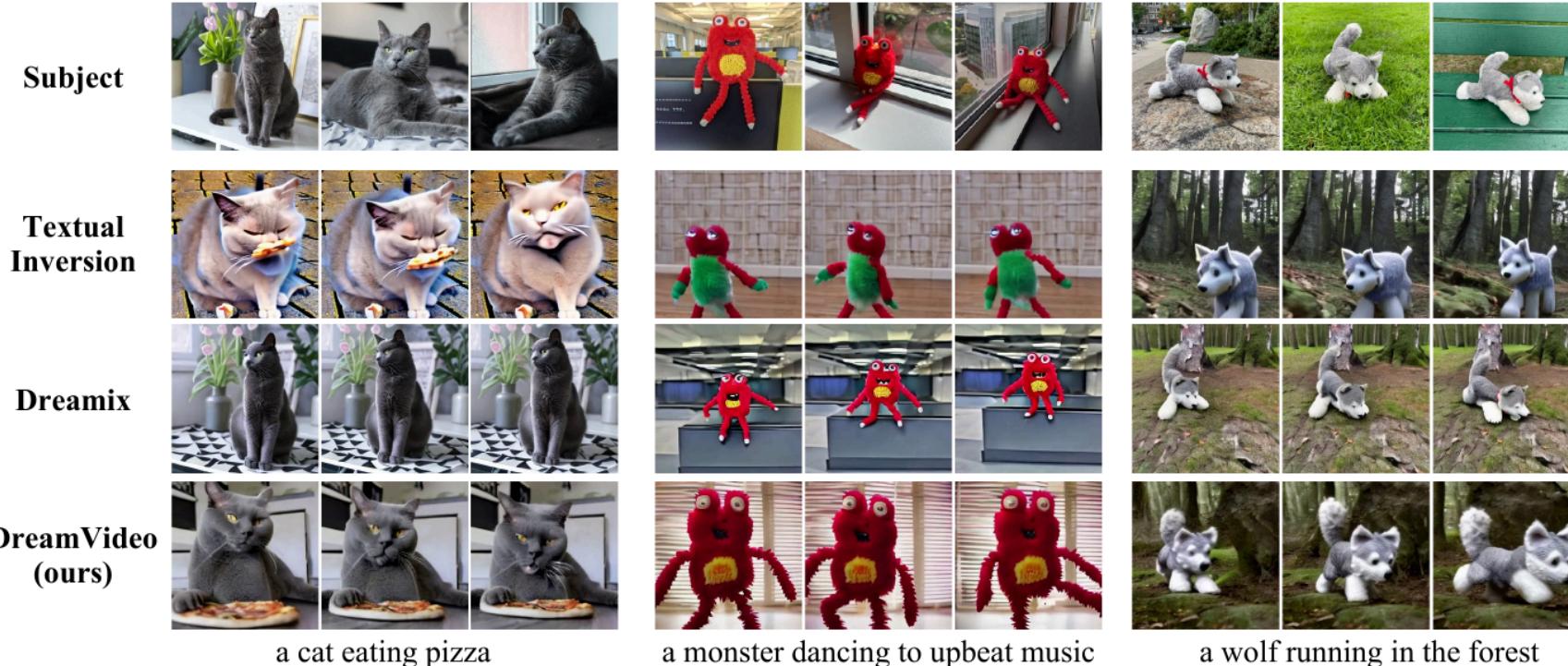
The baseline model is ModelScopeT2V.

Qualitative Comparison: Subject Customization



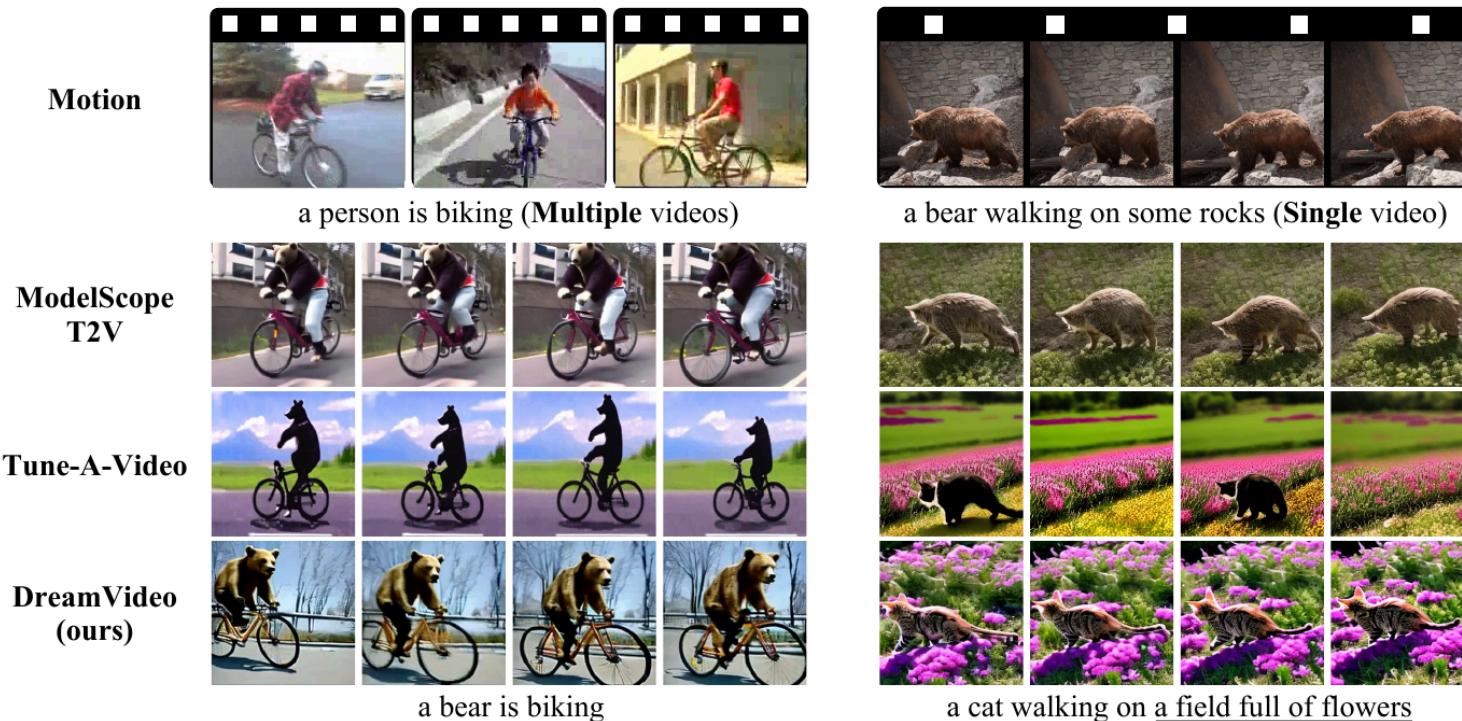
Dreamix (Molad et al, 2023) — The spatial parameters in the UNet are only trained, while the temporal parameters are frozen. The subject token is "skz". It seems to be just the Dreambooth for T2V.

Qualitative Comparison: Subject Customization



Textual Inversion (Gal et al, 2022) — Just textual inversion for the video diffusion model.

Qualitative Comparison: Motion Customization



- ModelScopeT2V — the temporal parameters of the UNet are trained, while the spatial parameters are frozen.
- [Tune-A-Video \(Wu et al 2022\)](#) — It seems to be based on an image diffusion model (Stable Diffusion v1.5) like the AnimateDiff.

Ablation Study

