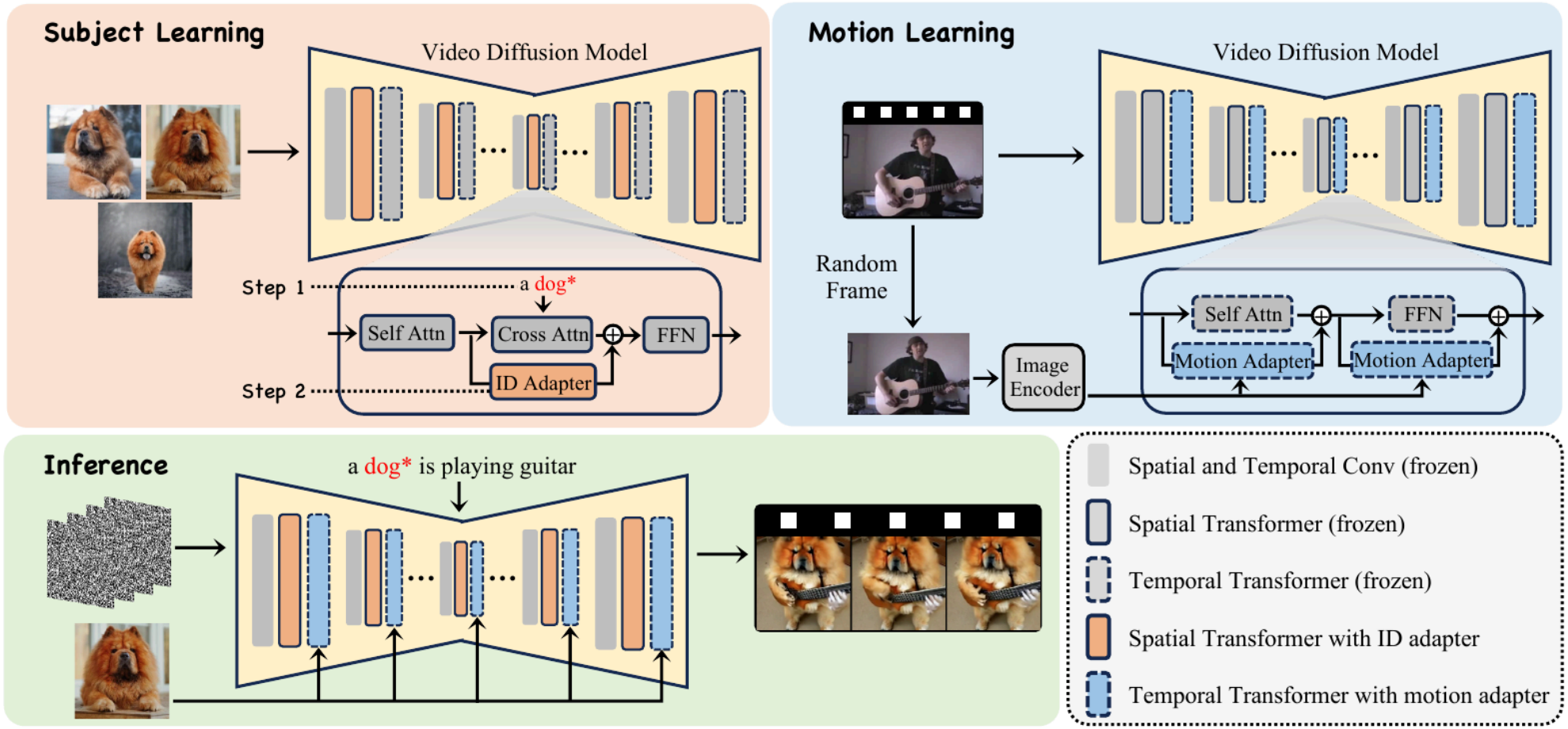
DreamVideo: Composing Your Dream Videos with Customized Subject and Motion

Wei et al, CVPR 2024

Overall Architecture: DreamVideo



where the baseline model is ModelScopeT2V (Wang et al, 2023).

Overall Architecture: ModelScopeT2V

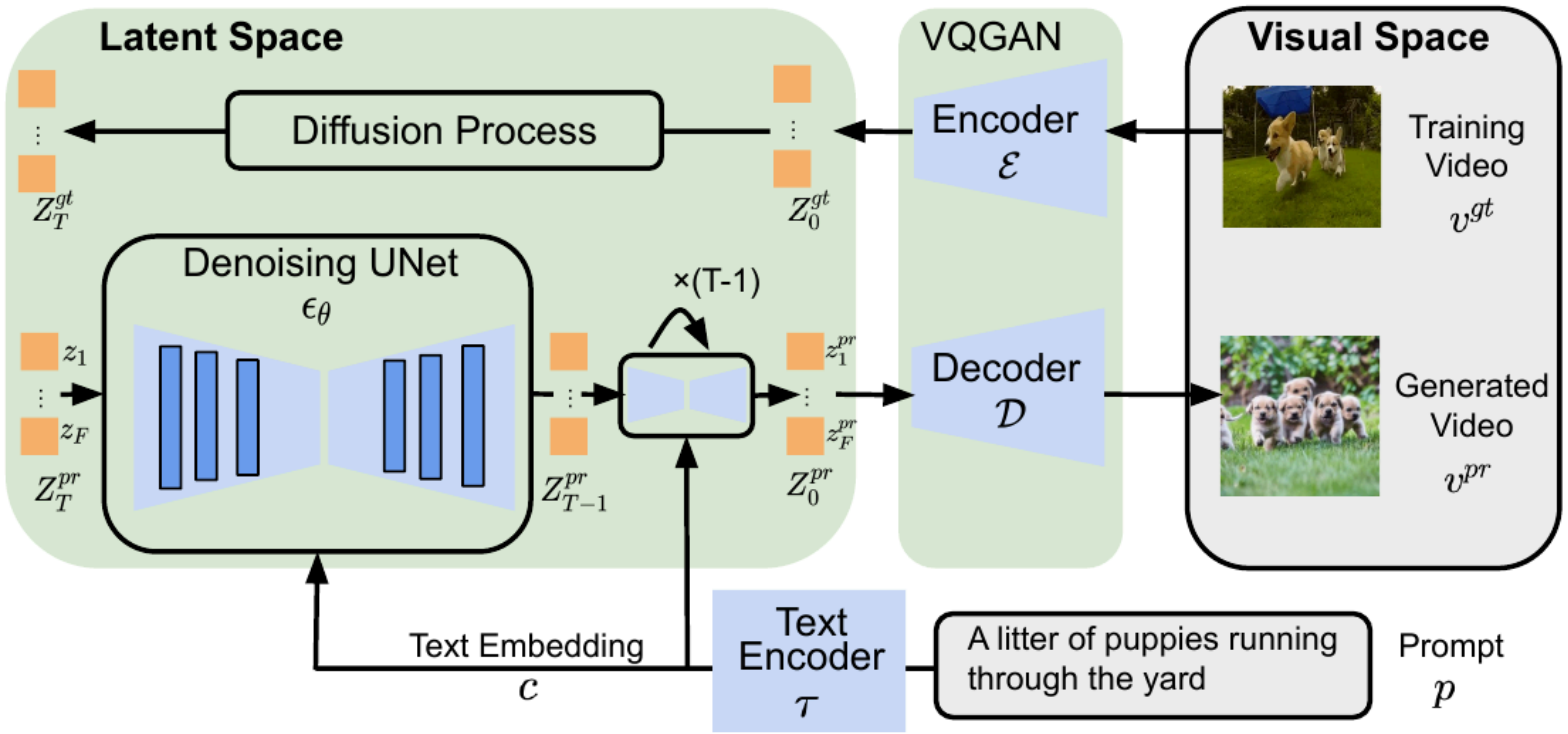


Figure 2: **The overall architecture of ModelScopeT2V.** Here, a text encoder τ encodes the prompt p into text embedding c. Then, input the embedding c into the UNetϵ_θ, directing the denoising process. During training, a diffusion process is performed, transitioning from Z_0^{gt} to Z_T^{gt}; so the denoising UNet could be trained on these latent variables. Conversely, during inference, random noise Z_0^{pr} is sampled and utilized for the denoising procedure.

Overall Architecture: ModelScopeT2V

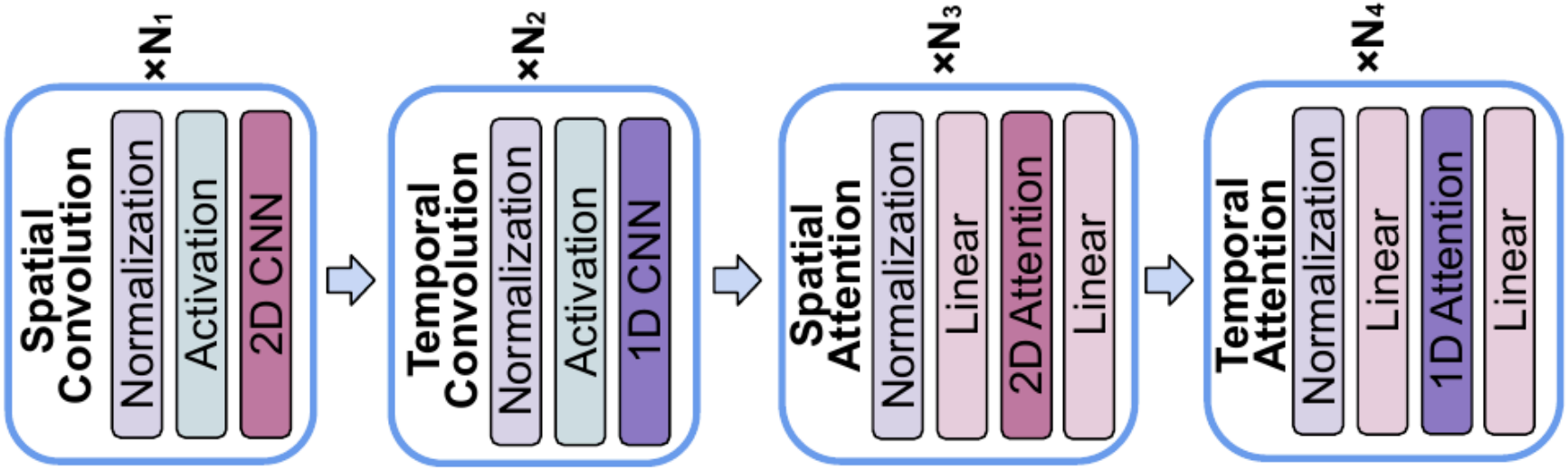
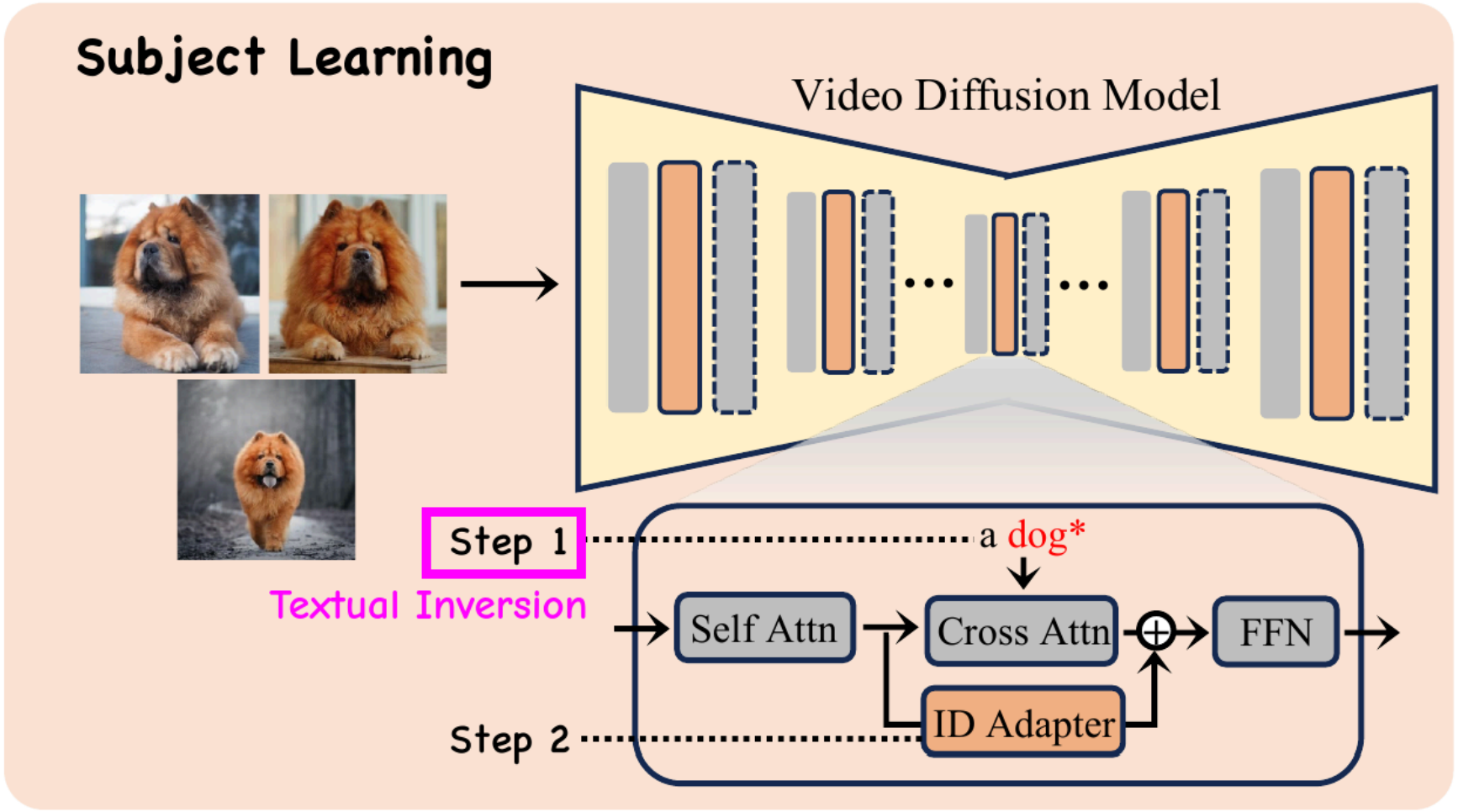


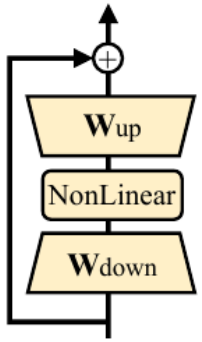
Figure 3: **The structure of the spatio-temporal block.** It includes four modules, *i.e.*, spatial convolution, temporal convolution, spatial attention, and temporal attention. The main layers for these modules are marked in different colors.

where

1. the **spatial attention blocks** are used for the **subject learning**
2. and the **temporal attention blcoks** are used for the **motion learning**.

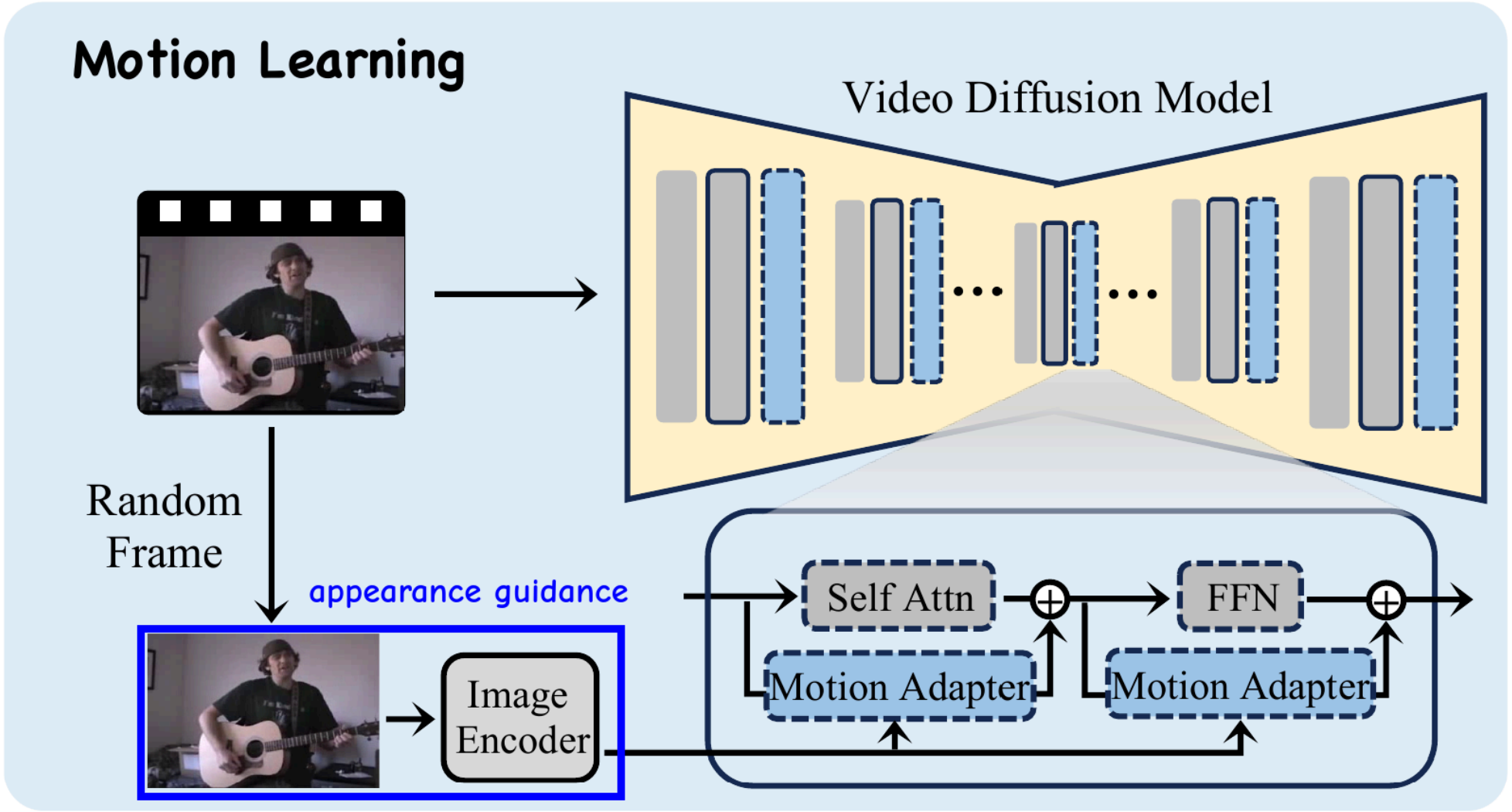
Subject Learning

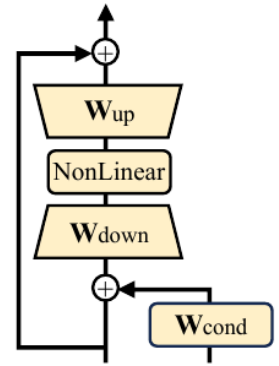




(a) Identity Adapter

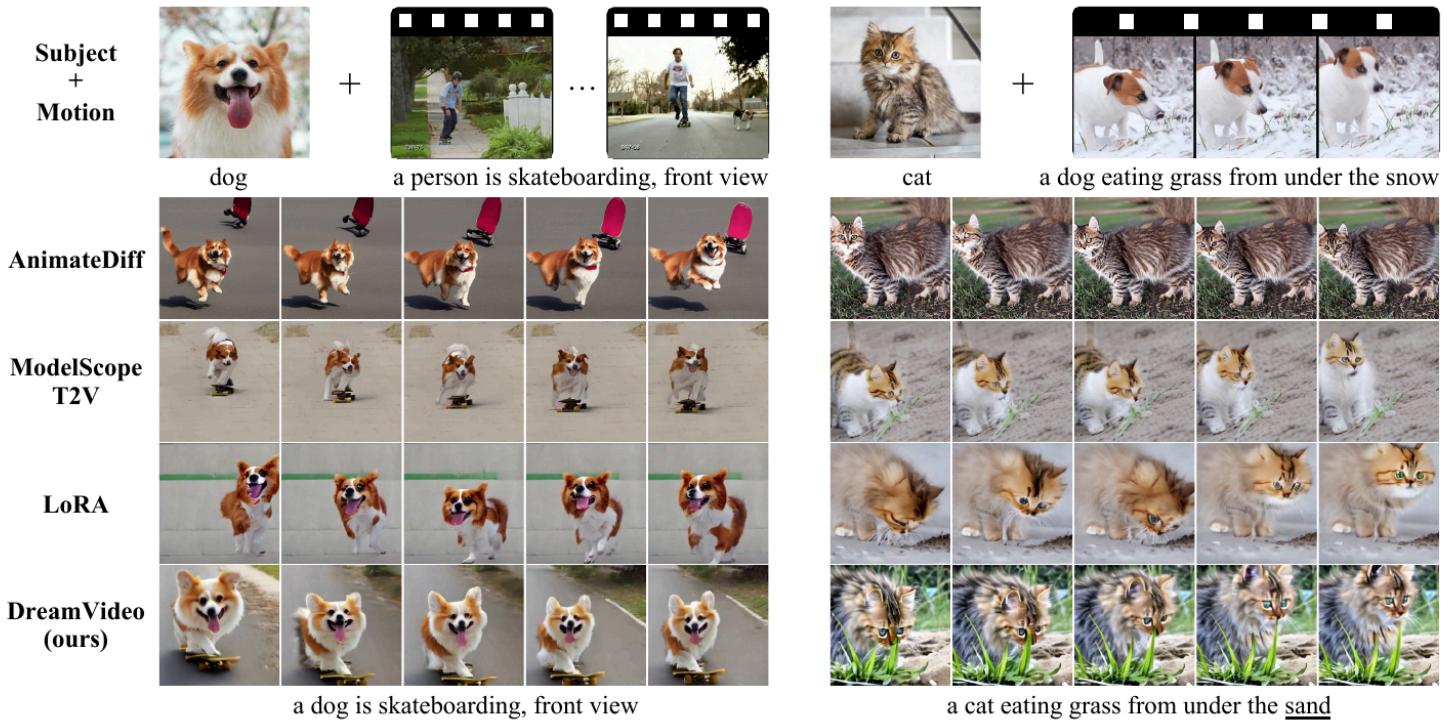
Motion Learning





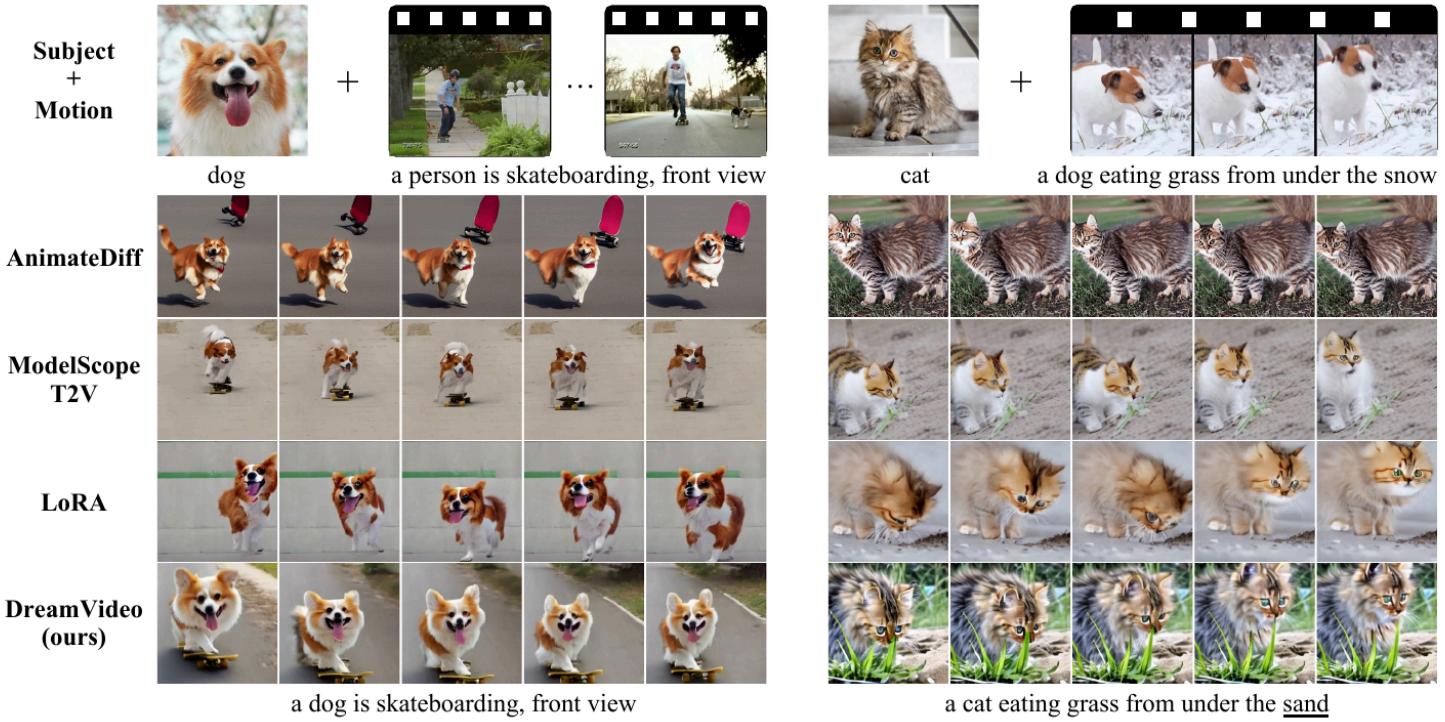
(b) Motion Adapter

Qualitative Comparison: Subjet and Motion Customization



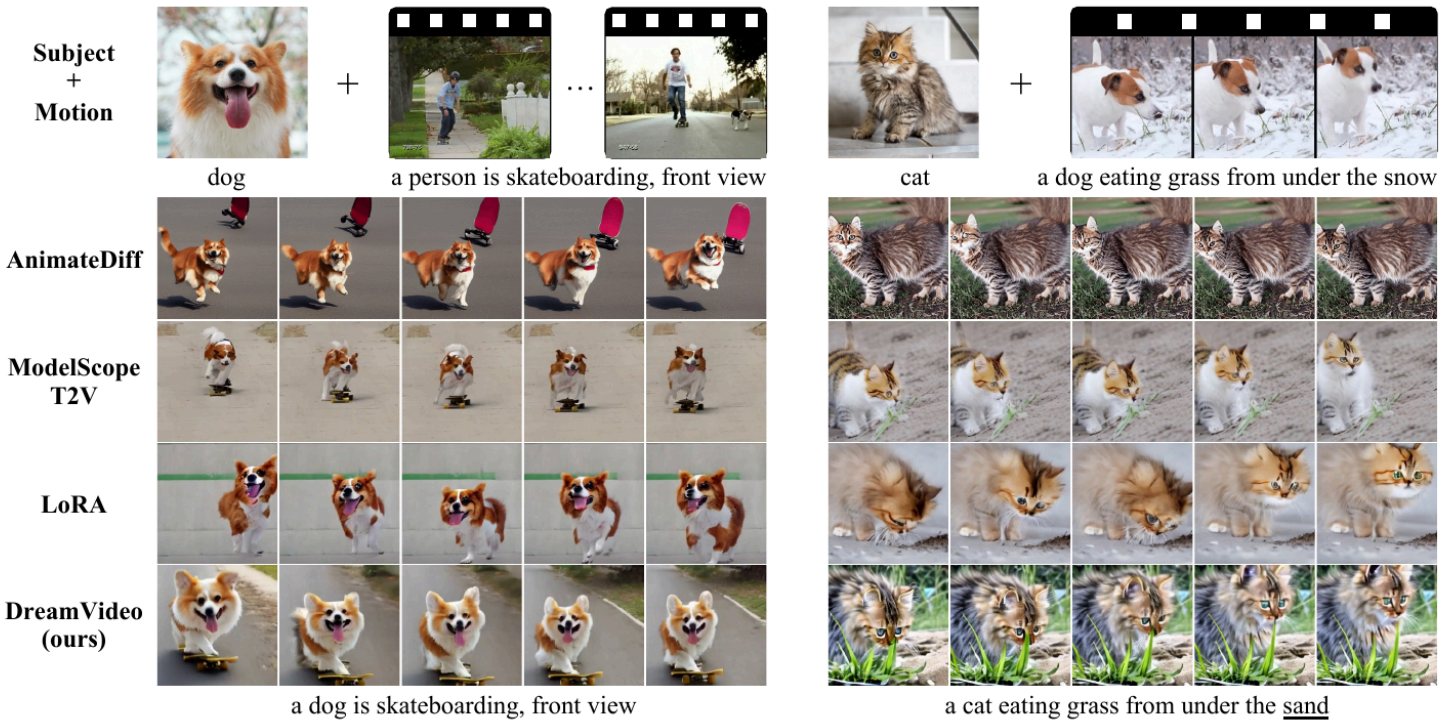
AnimateDiff (Guo et al, ICLR2024) — A newly-initialized **motion module** is trained from scratch. And it will be **appended to the DreamBooth-fine-tuned IMAGE diffusion model** to generate videos. That is, the DreamBooth learns the subject and the motion module learns the motion.

Qualitative Comparison: Subjet and Motion Customization



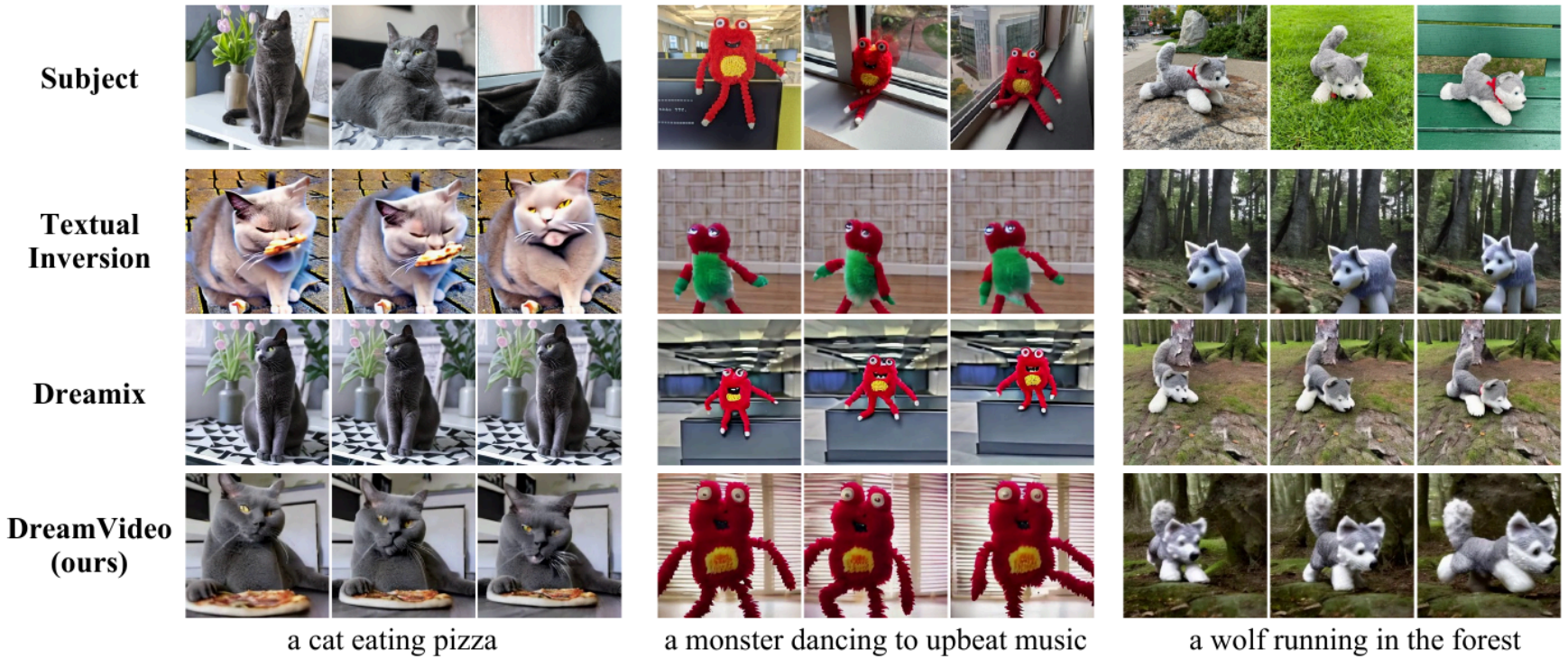
ModelScopeT2V (Wang et al, 2023) — the spatial/temporal parameters of the UNet are only fine-tuned, while other parameters are frozen.

Qualitative Comparison: Subjet and Motion Customization



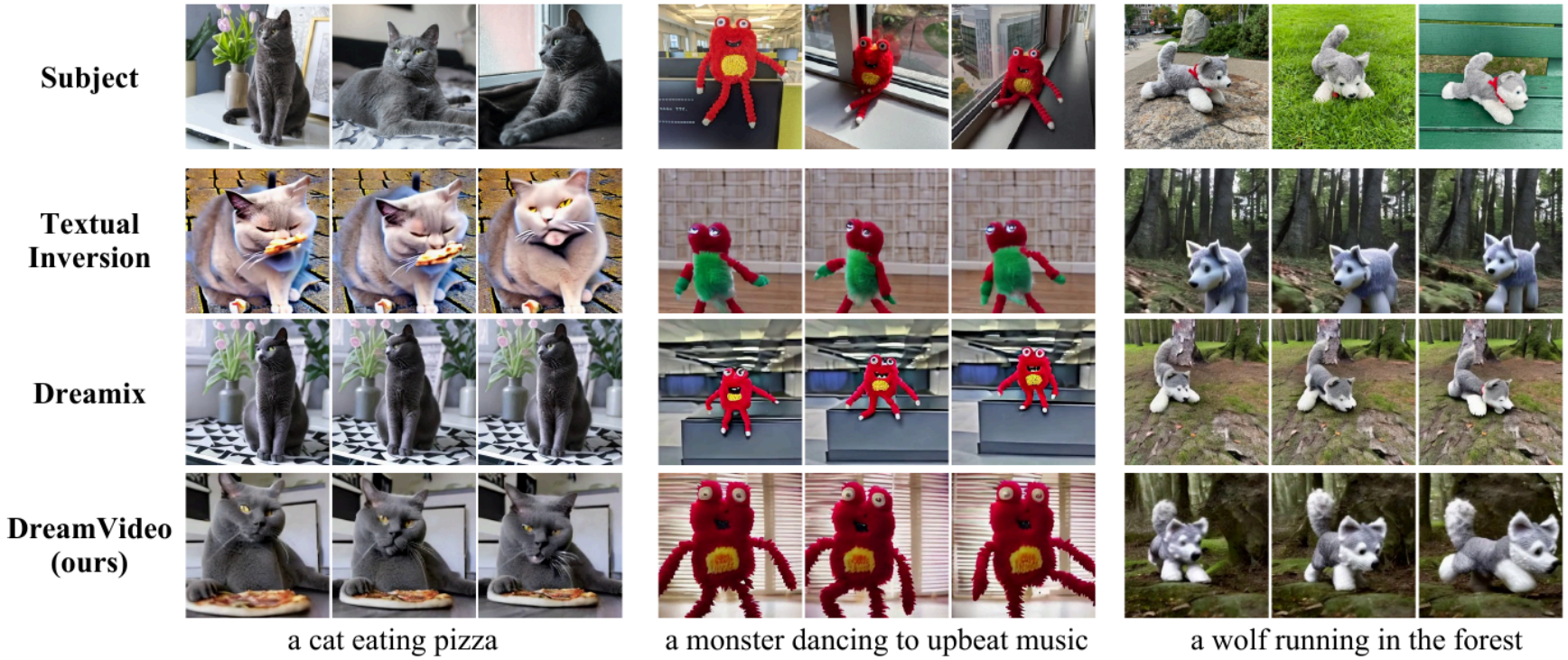
ModelScopeT2V (LoRA fine-tuned) — LoRA parameters are added to the key and value matrices in the cross-attention layers to learn a subject. For motion learning, LoRA parameters are added to the key/value matrices in all self-attention layers.

Qualitative Comparison: Subjet Customization



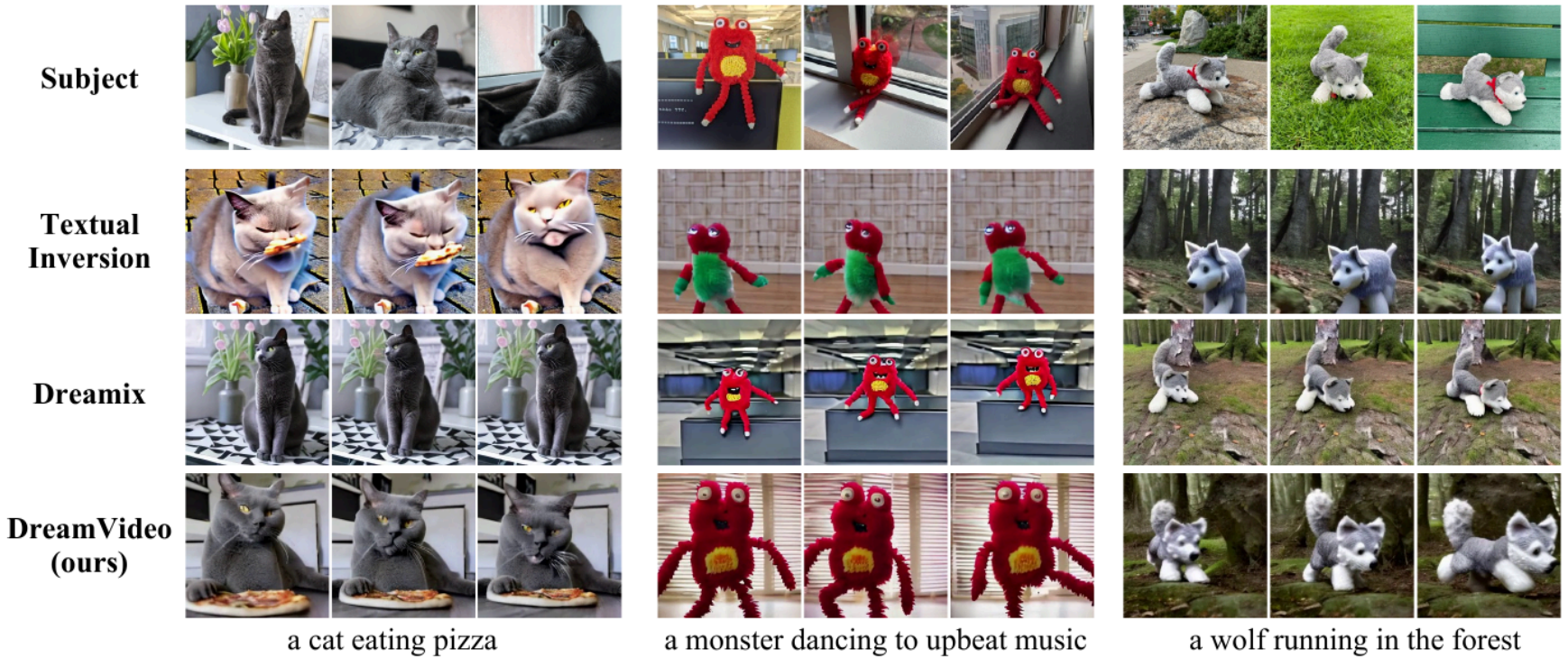
The baseline model is ModelScopeT2V.

Qualitative Comparison: Subjet Customization



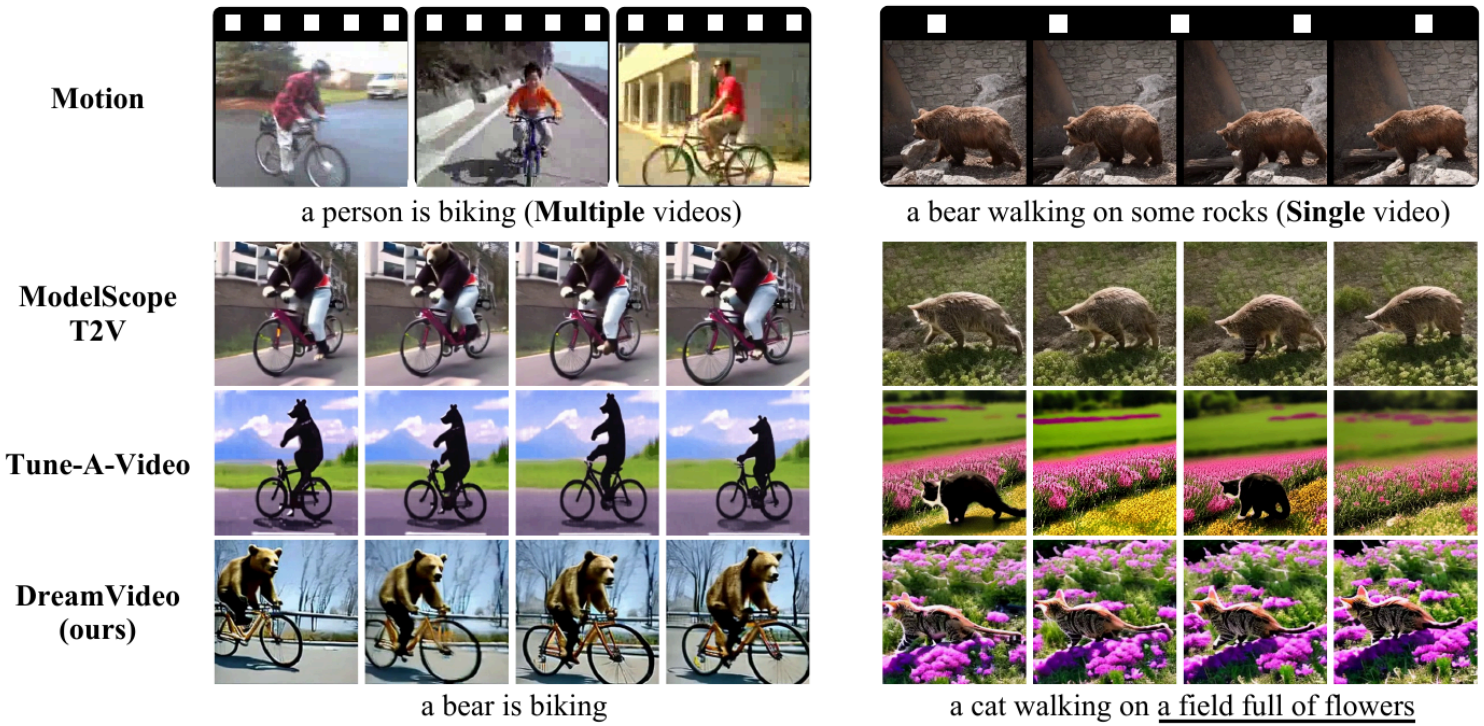
Dreamix (Molad et al, 2023) — The spatial parameters in the UNet are only trained, while the temporal parameters are frozen. The subject token is `"sks"`. It seems to be just the Dreambooth for T2V.

Qualitative Comparison: Subjet 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