SkyReels-A2

- □□□□□SkyReels-A2: Compose Anything in Video Diffusion Transformers
- □□□□□https://arxiv.org/abs/2504.02436
- arxiv 2025

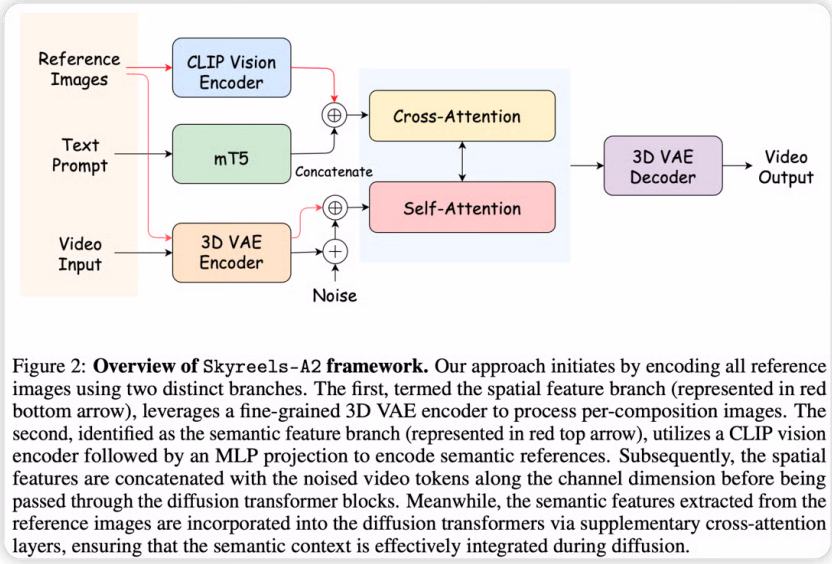

Figure 2: **Overview of** `Skyreels-A2` **framework.** Our approach initiates by encoding all reference images using two distinct branches. The first, termed the spatial feature branch (represented in red bottom arrow), leverages a fine-grained 3D VAE encoder to process per-composition images. The second, identified as the semantic feature branch (represented in red top arrow), utilizes a CLIP vision encoder followed by an MLP projection to encode semantic references. Subsequently, the spatial features are concatenated with the noised video tokens along the channel dimension before being passed through the diffusion transformer blocks. Meanwhile, the semantic features extracted from the reference images are incorporated into the diffusion transformers via supplementary cross-attention layers, ensuring that the semantic context is effectively integrated during diffusion.

Subject-to-Video□□□□□□□□□□□□□□□□□□□□□□□□□□□CLIP□□□□□□□□concatenate□□cross-attention□□□□□VAE□□□□padding□□□□□□□□3D VAE□□□□□□□□□□□□□□□□□□□□concatenate□□□Wan-I2V□□

Newer Older

2026-1-4
MotionInversion

2026-1-15
BindWeave