BindWeave

- □□□□□BINDWEAVE: SUBJECT-CONSISTENT VIDEO GENERATION VIA CROSS-MODAL INTEGRATION
- □□□□□https://arxiv.org/abs/2510.00438
- arxiv 2025□ICLR2026□□4466□

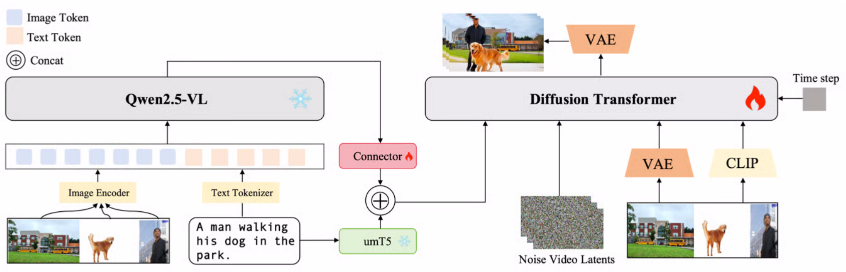

Figure 2: Framework of our method. A multimodal large language model performs cross-modal reasoning to ground entities and disentangle roles, attributes, and interactions from the prompt and optional reference images. The resulting subject-aware signals condition a Diffusion Transformer through cross-attention and lightweight adapters, guiding identity-faithful, relation-consistent, and temporally coherent video generation.

□□MLLM□□□□□□□□□□□hidden state□□□□□□□□□MLP□□□□□□□□□□□□□□□□□□□□□□□□concate□□□□□□□□□□□□□□□□□□□□cross attention□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□vae□□□□□□□□□□□□Wan2.1-I2V□□□□□□□□□□□□□□□□

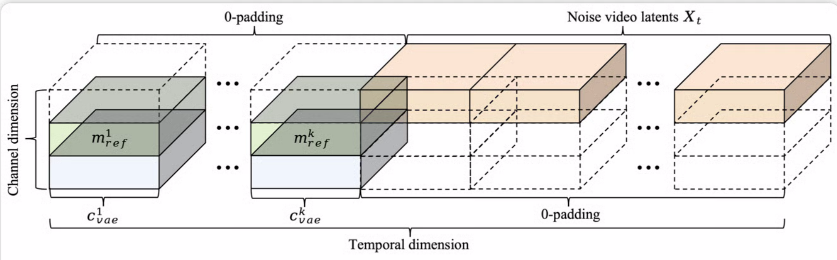

Figure 3: Illustration of our adaptive multi-reference conditioning strategy.

□□□□Wan2.1-I2V□□□□□□clip□□□cross attention□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□

Newer Older

2026-1-4
MotionInversion

2026-1-15
Saber