Phantom

- □□□□□Phantom: Subject-Consistent Video Generation via Cross-Modal Alignment
- □□□□□https://arxiv.org/abs/2502.11079
- ICCV2025

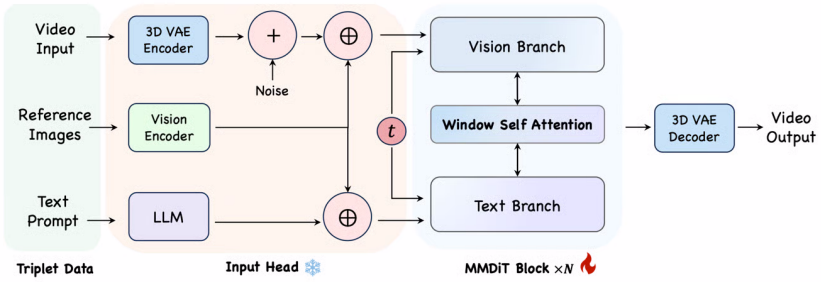


Figure 4. Overview of the ***Phantom*** architecture. Triplet data is encoded into latent space at the input head, and after combination, it is processed through modified MMDiT blocks to learn the alignment of different modalities.

□□□□□subject-to-video□□□□□□□□□□pipeline□□□□□□□□ □□□□□□□□pipeline□

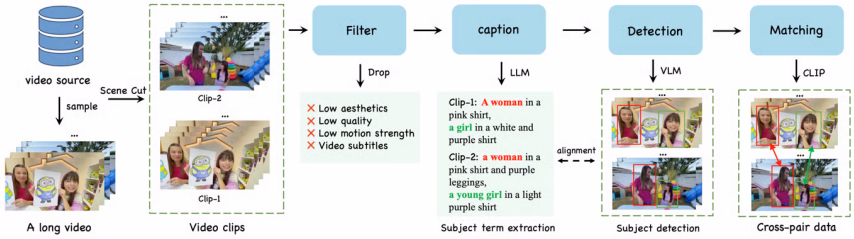
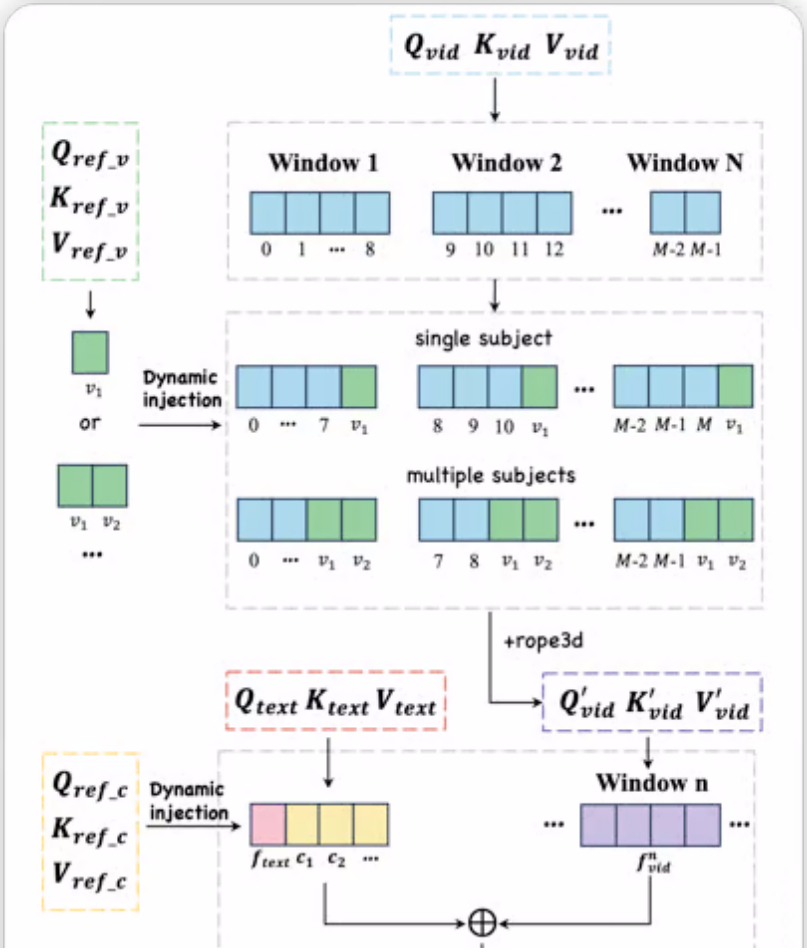


Figure 3. Data processing pipeline for cross-modal video generation. The process involves filtering, adding captions, detection, and matching stages to extract subjects from video clips and align them with the text prompts, ensuring consistent video generation.

□□□□□□□□□□□□□□□□□VAE□CLIP□□□□□□□□□□□VAE□□□toekn□□□token□□□□□CLIP□□□ □□□text token□□□□□□□□□□□□□□□□□□□□□□□□□□□□□window self attention□□□□□□□□□□□□□□ □□□□□□□□□□attention□□□□



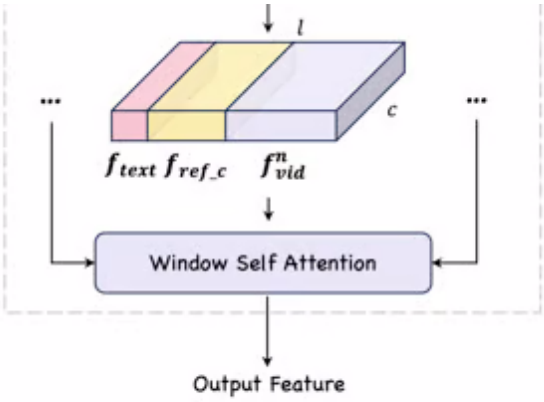


Figure 5. **Dynamic injection strategy** and attention calculation for single or multiple reference subjects in each MMDiT block.

Newer Older

2026-1-13
VACE

2025-12-10
LoRAinLoRA