□

PolyVivid

- □□□□□PolyVivid: Vivid Multi-Subject Video Generation with Cross-Modal Interaction and Enhancement
- □□□□□https://arxiv.org/abs/2506.07848
- NIPS 2025

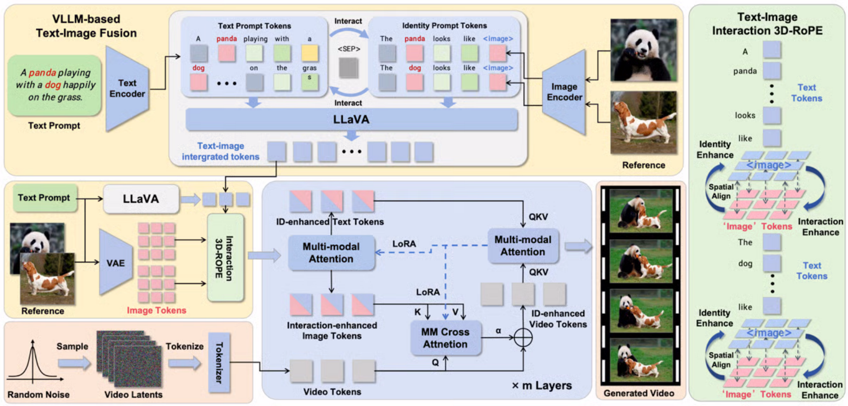

Figure 2: Framework of our PolyVivid: the text prompt and reference image are fused by the VLLM-based text-image fusion module. Then, a 3D RoPE-based identity-interaction enhancement module is employed to enhance the text-image interaction. The enhanced image tokens are injected by an MM cross-attention module, which helps preserve the identities while ensuring good subject interaction.

□□□□□□□□□S2V□□□□□□□□□□□

□□□□□LLaVA□□prompt□□□□□□□□□□□□□□□□□□□□□□□□□□□□-□□VAE□□□□□□□□□□□token□□□□□□□token□□□□□□□□□□3D-RoPE□□□□□□□□□□□□□MM-Atention□□□□□□□□□□□□□□□□□□□□token□□□token□□□□□□□token□□□□□□□□□□□□□□□□□□□□□□□□□MM-Attention□□□□□□block□□□□□□□□□□□□□□□□□□□□□□□token□□MM-DiT□□□□□□□□□□□token□□□adapter□

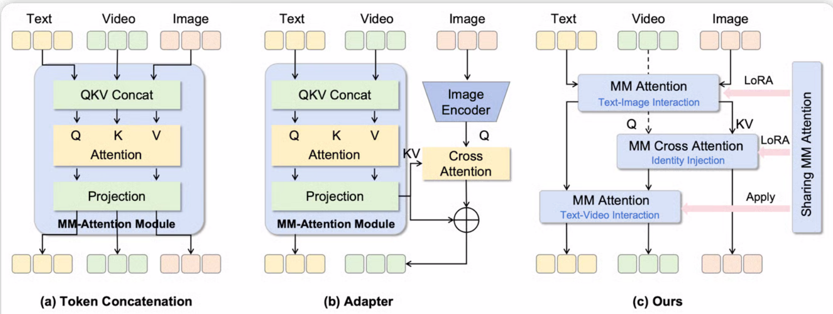

Figure 3: Comparison of the condition injection strategies for MM-DiT.

□□□□□token□□□□□□□□□□□□□□□□□□□□□□□token□□□video token□□□□□□□□attention□□□□□□□□□□□□□token□□□□□□□□□adapter□□□□□□□□□adapter□□□attention□□□□□□□□□□□□□□□□□token□□□□□□□□□□-□□□attention□□□□□□□□□□□□-□□□□□attention□□□□□□□□□□□□□□□Attention□□□□□□□□lora□□□□□□□-□□□□□attention□□□□id□□□□□□□□□lora□□□□□□

Newer
Older
2026-1-14
MAGREF
2026-1-13
VACE