□

home
papers
study
life

About me

VideoLCT

- □□□□□Long Context Tuning for Video Generation
- □□□□□https://arxiv.org/abs/2503.10589
- ICCV2025

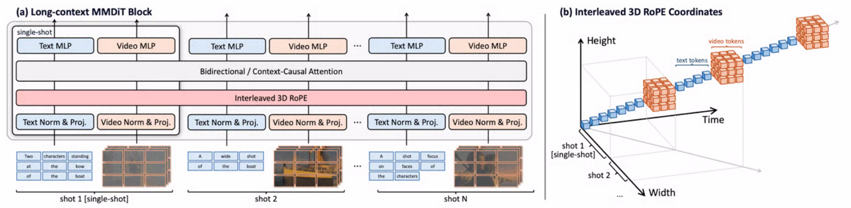

Figure 3. **Architecture Designs.** (a) *Long-context MMDiT block.* We expand the attention operation to all text and video tokens within a scene, and apply independent noise levels to individual shots. The interleaved 3D RoPE assigns distinct coordinates for each shot. (b) *Interleaved 3D RoPE coordinates.* At shot-level, text tokens precede video tokens along the space diagonal. At scene-level, tokens are arranged shot by shot, forming an interleaved "[text]-[video]-[text]-..." pattern along the space diagonal.

□□

□□□□□□□□□□□□□□□□□□□□□□□□□□token□□□□□□□prompt□□□□□□□□□□□□□□□□□□□□□□□3D□□diffusion step□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□

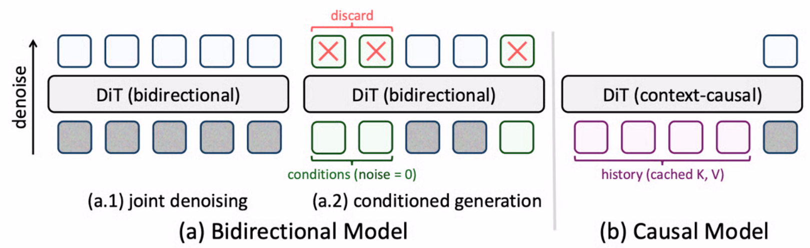

Figure 4. **Inference Modes.** (a) *Bidirectional* model enables (a.1) joint or (a.2) visual-conditioned generation, while (b) *context-causal* model supports auto-regressive generation.

□□□KV-cache□□□□□□□

□□□□□□□□□□□□□□□□□□□□□□□□□diffusion step□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□

- □□□□□□□500K□□□□□□
- □□□128 H800
- □□□

Newer — Older

2026-1-6
StoryMem

2026-1-4
MotionInversion

Archive RSS feed GitHub Email QR Code