□

Saber

- □□□□□Scaling Zero-Shot Reference-to-Video Generation
- □□□□□https://arxiv.org/abs/2512.06905
- arxiv 2025

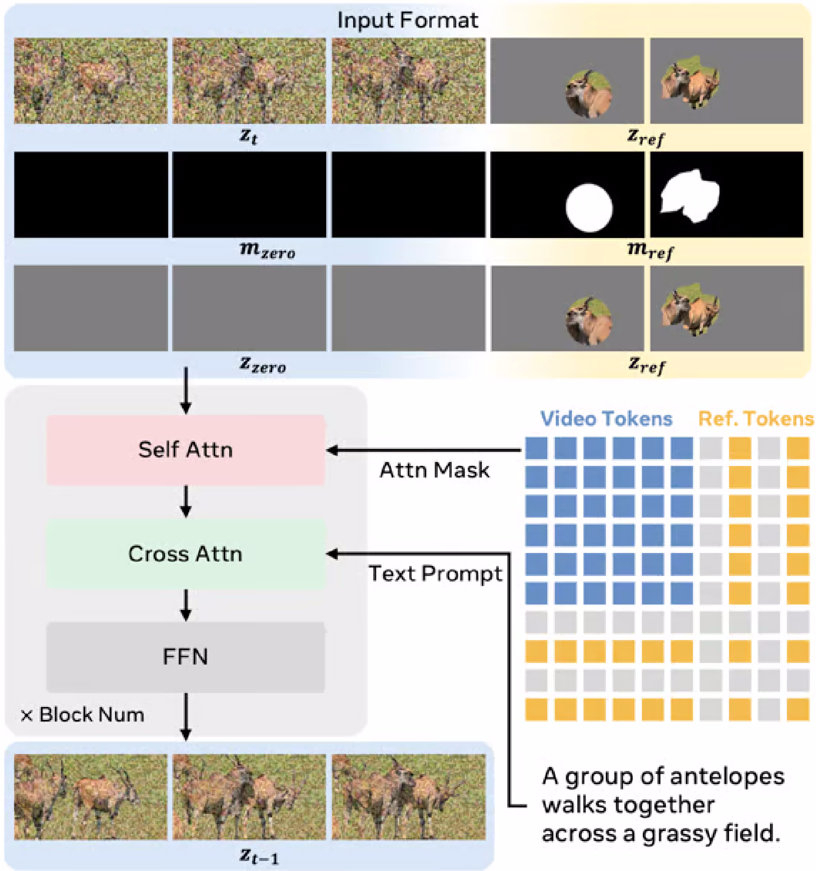

Figure 3 Model design overview. Masked frames serve as reference images and are concatenated to the video tokens in latent space. Self-attention enables interaction between video and reference tokens under the attention mask, while cross-attention incorporates text guidance for semantic alignment. The VAE, text encoder, and timestep components are omitted for clarity.

□□□□□□□□□□□□□□□□□□□□□□□□S2V□□□□□□□□□□R2V□□□□□□□□□□□R2V□□□□□□□□□□□R2V□□□□□□-□□-□□□□□□□□□□□□□□□□□□□□□□□□□scaling□□□□□□□□□□saber□□□□□□T2V□□□-□□□□□□□□□□□R2V□□□□

□□□□□□□□□□□□□mask□□□□□□□□□attention□□□□□□□□□□□□□□□□□□□□□□k□□□□□□□□□□□□□□mask□□□□□□□□□□□□□□□□□video token□□□□□□□□□□□□□□Wan2.1-I2V□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□mask□□□□□□□□□□□□□□

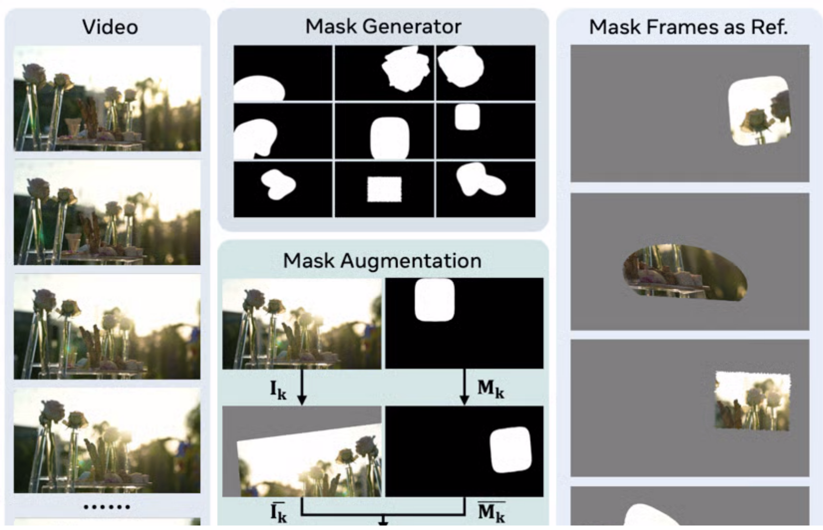

Figure 2 Masked reference generation. Given a video, the mask generator produces diverse random masks, which are then applied to each randomly sampled video frame with mask augmentation.

□□mask generator□□□□□□□□□□□□□□□□□□□□□□□□□□□□

□□□□self-attention□□□□□□ref image□□mask□□□□□□attention□□□□

Newer　　Older

2026-1-15
BindWeave

2026-1-14
MAGREF