VACE

- □□□□□VACE: All-in-One Video Creation and Editing
- □□□□□https://arxiv.org/abs/2503.07598
- ICCV2025

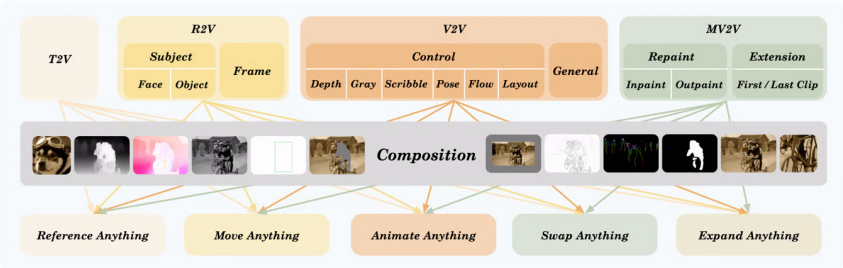

Figure 2. **Task categories covered by VACE.** Four basic tasks can be combined to create a vast number of possibilities.

□□□□□□□□□□video DiT□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□T2V, R2V, V2V, MV2V□□□Video Condition Unit □VCU□□□□□□[T, F, M]□T□□□□□□□□F□□M□□□□□□□□□□□□□□□□□□□□□□□□□□□VCU□□□□□□□□□□□□

Table 1. The formal representation of frames (Fs) and masks (Ms) under the four basic tasks. Frames and masks are aligned spatially and temporally.

Tasks	Frames (Fs) & Masks (Ms)
T2V	$F = \{0_{h\times w}\} \times n$ $M = \{1_{h\times w}\} \times n$
R2V	$F = \{r_1, r_2, ..., r_l\} + \{0_{h\times w}\} \times n$ $M = \{0_{h\times w}\} \times l + \{1_{h\times w}\} \times n$
V2V	$F = \{u_1, u_2, ..., u_n\}$ $M = \{1_{h\times w}\} \times n$
MV2V	$F = \{u_1, u_2, ..., u_n\}$ $M = \{m_1, m_2, ..., m_n\}$

□□□□□□□□□□□□□□□□□□□□□□□□□□□□

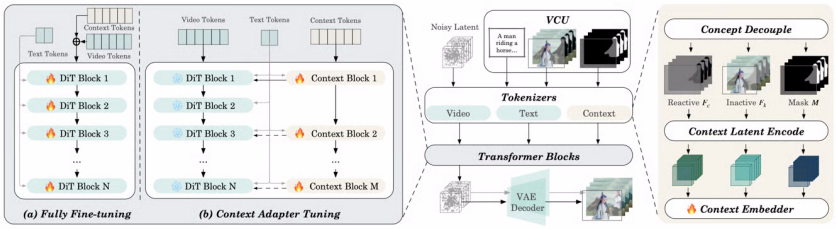

Figure 3. **Overview of VACE Framework.** Frames and masks are tokenized through Concept Decoupling, Context Latent Encode and Context Embedder. To achieve training with VCU as input, we employ two strategies, (a) Fully Fine-tuning and (b) Context Adapter Tuning. The latter converges faster and supports pluggable features.

□VCU□□tokenizer□□embedder□□□□□Context token□(a)□□□□□□□□context token□video token□□□□□□□□DiT□□□□□□□(b)□□ControlNet□□□□□block□□Context token□□□□□□block□□□□□□□□□□□□□□□□□R2V□□□□□□□□Video Token□□□□□□□□ref image□□□□□□□□□token□□□□□□□□□□

Newer Older

2026-1-14
PolyVivid

2026-1-13
Phantom