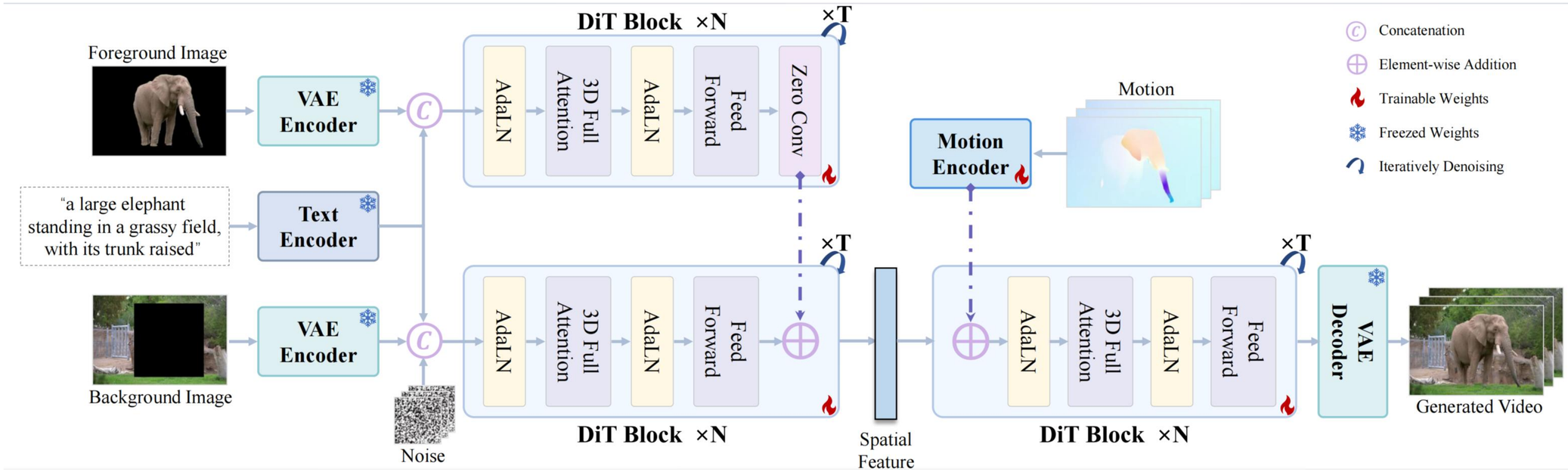
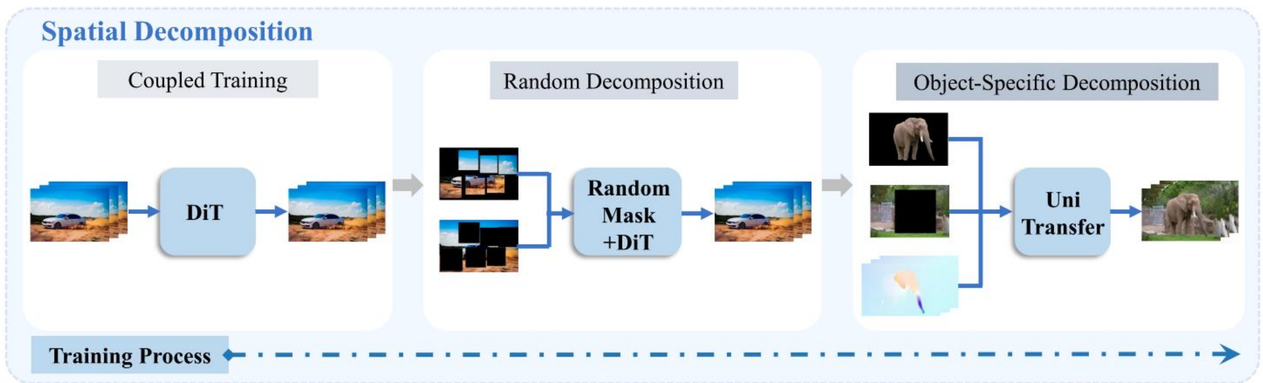
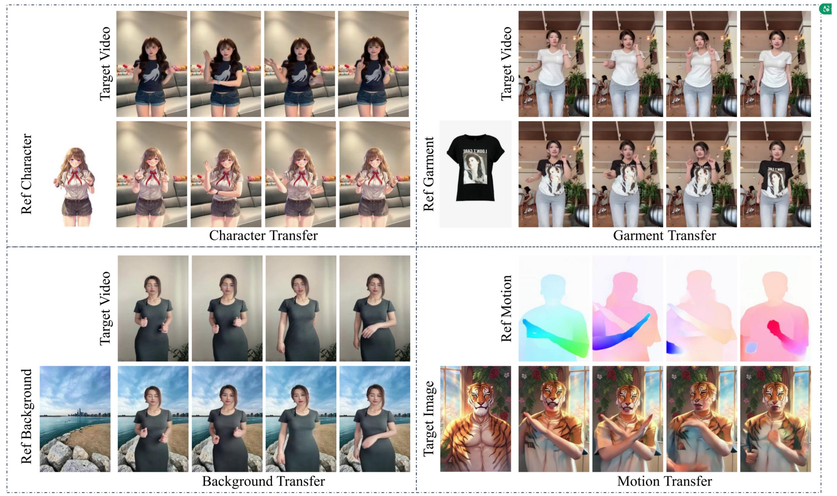
UniTransfer: Video Concept Transfer via Progressive Spatial and Timestep Decomposition

Guojun Lei, Rong Zhang, Tianhang Liu,
Chi Wang, Hong Li, Zhiyuan Ma, Weiwei Xu
Zhejiang University Zhejiang Gongshang University

https://yu-shaonian.github.io/UniTransfer-Web/







$$\mathcal{L}(\theta) = \begin{cases} \|\epsilon - \hat{\epsilon}_\theta(x_t, \tau_{crs}, \mathcal{U}, t)\|_2^2, & t \in [t_c, T-1] \\ \|\epsilon - \hat{\epsilon}_\theta(x_t, \tau_{mid}, \mathcal{U}, t)\|_2^2, & t \in [t_f, t_c) \\ \|\epsilon - \hat{\epsilon}_\theta(x_t, \tau_{fine}, \mathcal{U}, t)\|_2^2, & t \in [0, t_f) \end{cases}$$

- We propose a DiT-based image-guided video concept transfer framework UniTransfer, which incorporates progressive spatial and timestep decomposition.
- We introduce a self-supervised pretraining strategy based on randomized masking to enhance the disentangled representation learning
- We further introduce an LLMs-guided chain-of-prompt mechanism to achieve the timestep decomposition. This progressive prompting strategy guides the generation process with stage-specific instructions, improving the VCT generation quality.