Georgia Tech

Auto-Regressive Video Diffusion for Scalable Control

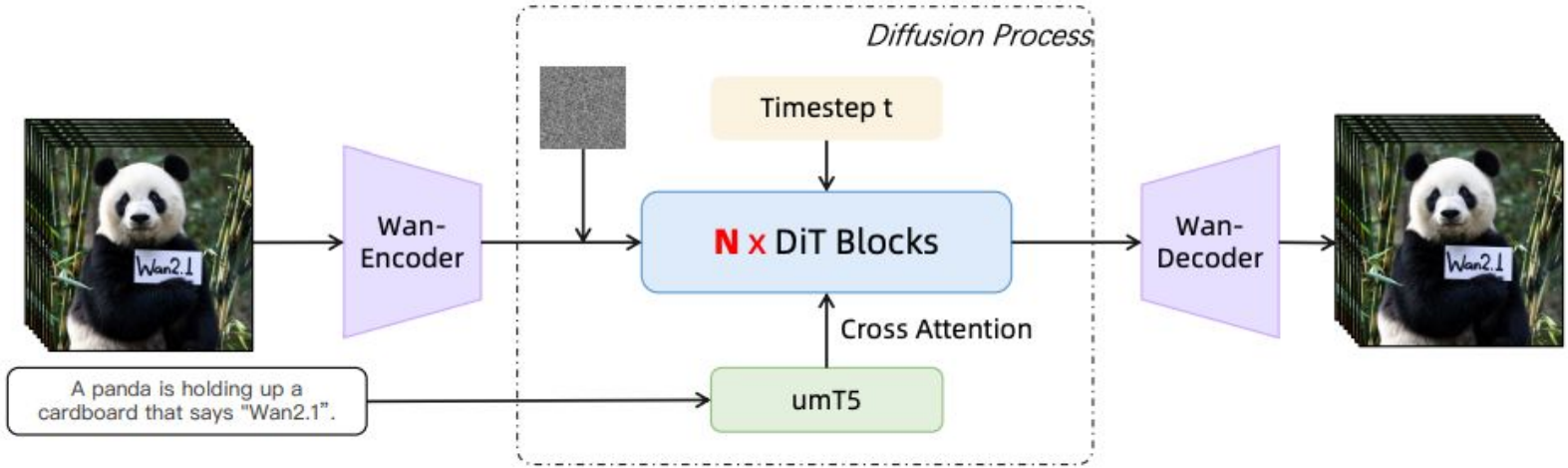
AE8803 Proposal Presentation

Chenxiao Gao, Yixin Zhang

Video Diffusion Models (VDM)

Leverages the flexibility of Diffusion Models to create high-fidelity, temporally consistent video clips

Causal video auto-encoders → Diffusion Transformers → Decoder

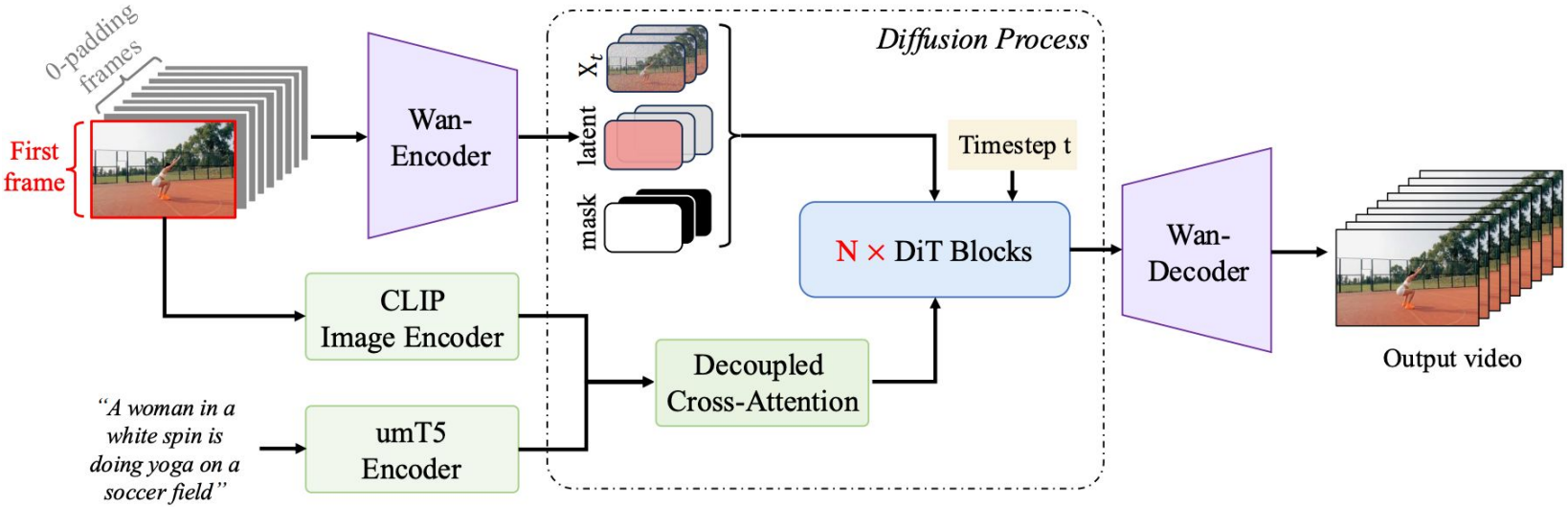


Wang, Ang, et al. "Wan: Open and advanced large-scale video generative models." *arXiv preprint arXiv:2503.20314* 3.4 (2025): 6.

Video Diffusion Models (VDM)

Flexible conditioning: Text-Image-to-Video pipeline

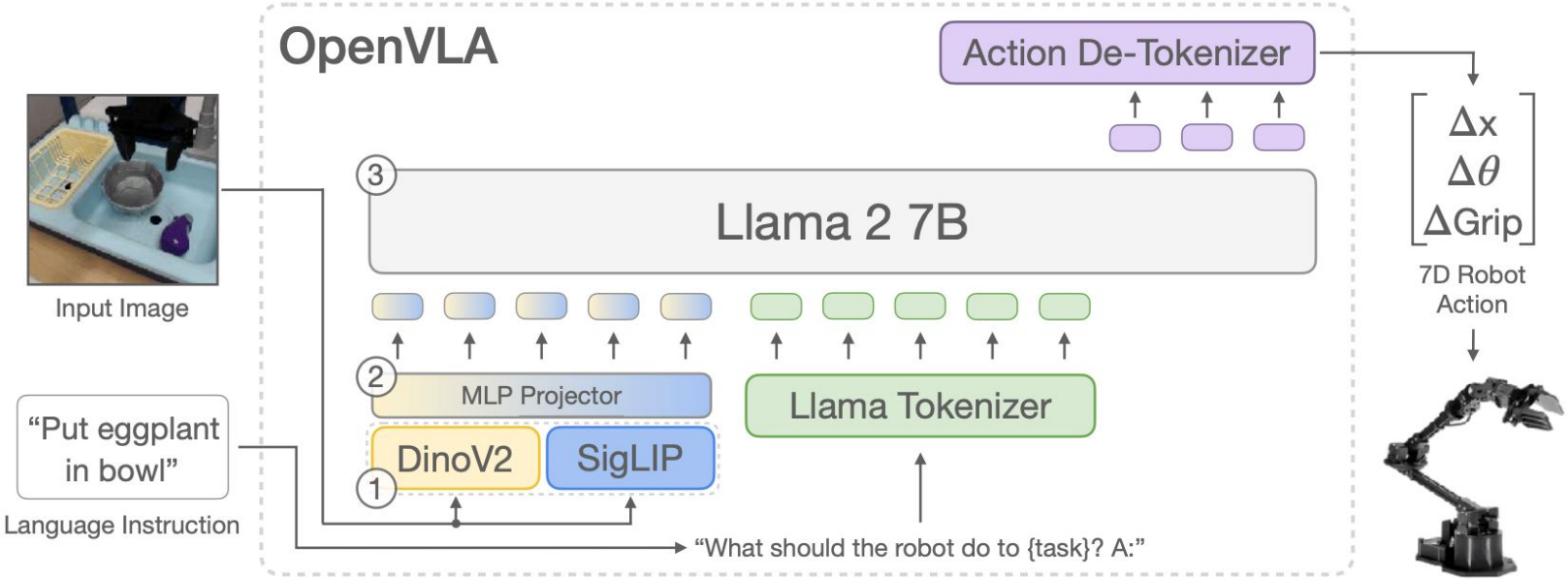
A lively, brownish-yellow puppy running fast across a lush green meadow …



Wang, Ang, et al. "Wan: Open and advanced large-scale video generative models." *arXiv preprint arXiv:2503.20314* 3.4 (2025): 6.

Vision-Language-Action Models (VLAs)

Leverages LLM backbone to predict the desired position of the robot end-effector



Kim, Moo Jin, et al. "Openvla: An open-source vision-language-action model." *arXiv preprint arXiv:2406.09246* (2024).

Motivation

Scalability Gap

- Action-labeled data is scarce; we need policies that scale using unlabeled data.

Generalization Gap

- Backbones must generalize across diverse instructions and scenarios

Video diffusion model generalizes, but it is not decision-aware

Vision-Language-Action model is powerful, but we cannot train it at scale

Problem Formulation

The decision process from **UniPi** and recent **Video-Action-Models**

- Language instruction ℓ
- Image observations x_t

Du, Yilun, et al. "Learning universal policies via text-guided video generation." *Advances in neural information processing systems* 36 (2023): 9156-9172.

Pai, Jonas, et al. "mimic-video: Video-action models for generalizable robot control beyond vlas." *arXiv preprint arXiv:2512.15692* (2025).

Proposed Pipeline

By decomposing

$$\pi(a_{0:H}|x_0,\ell) = \int \left(\prod_{t=0}^{H} \mathcal{T}(x_{t+1}|x_{0:t},\ell)\pi_{\text{IDM}}(a_t|x_{0:t},x_{t+1},\ell) \right) dx_{1:H+1}$$

A two-stage plan emerges

- Autoregressive video diffusion

$$\mathcal{T}(x_{t+1}|x_{0:t},\ell)$$

- Action prediction with inverse dynamics models (IDMs)

$$\pi_{\text{IDM}}(a_t|x_{0:t},x_{t+1},\ell)$$

Proposed Pipeline

Why this two-stage approach?

- Better internal representation from video pretraining

Challenges

- Inference latency
- Collaboration between the two models

Proposed Pipeline

Why autoregressive?

- Better reactivity to environment stochasticity
- Flexible decision horizon

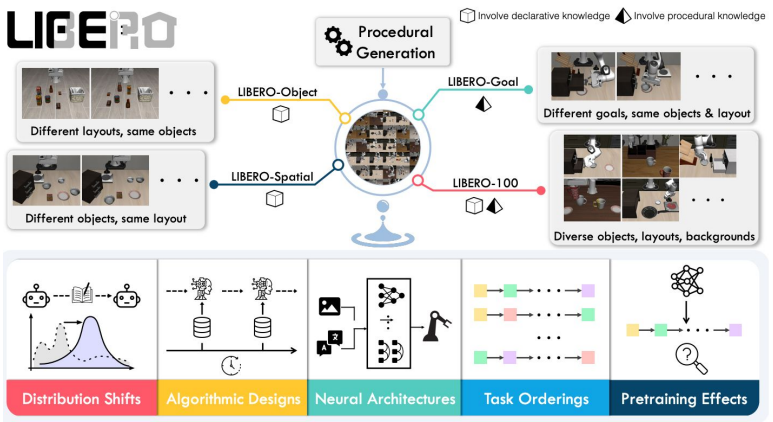
Challenges

- Inference cost
- Exposure bias

Datasets and Models

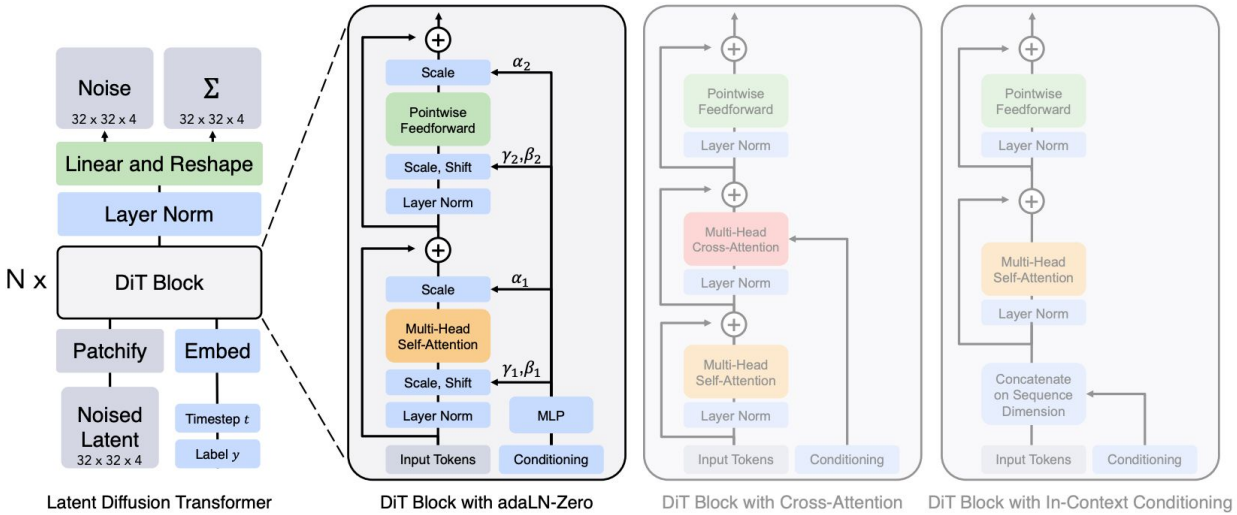
- Autoregressive video diffusion model training w/o action label
- IDM Training w/ action label
- Fine-tuning the models with simulation environments (Libero)

Ego4d



Libero

Datasets and Models



DiT with causal masking

[1] Peebles, Xie, et al. "Scalable Diffusion Models with Transformers"

Proposed Pipeline

Post-training

- A chain-of-thought (CoT) interpretation provided by [1]
- Fine-tuning the models with reinforcement learning (video/action prediction)
  - reweighted score matching / flow matching [2]

$$\pi^*(\boldsymbol{a}|\boldsymbol{s}) = \pi_{\text{old}}(\boldsymbol{a}|\boldsymbol{s})\bar{g}_f\left(\frac{Q(\boldsymbol{s},\boldsymbol{a})-\nu(\boldsymbol{s})}{\lambda}\right), \quad \bar{g}_f(x) = \begin{cases} (f')^{-1}(x) & \text{if } x > f'(0) \\ 0 & \text{if } x \leqslant f'(0) \end{cases}$$

$$\mathcal{L}(\theta) = \mathbb{E}_{\boldsymbol{s},\boldsymbol{a}_0,\epsilon}\left[\bar{g}_f\left(\frac{Q(\boldsymbol{s},\boldsymbol{a}_t)-\nu(\boldsymbol{s})}{\lambda}\right)\|\boldsymbol{v}_\theta(\boldsymbol{s},\boldsymbol{a}_t,t)-\boldsymbol{v}_{t|0}\|^2\right]$$

[1] Zhao, Qingqing, et al. "Cot-vla: Visual chain-of-thought reasoning for vision-language-action models." Proceedings of the Computer Vision and Pattern Recognition Conference. 2025.

[2] Ma, Haitong, et al. "Efficient online reinforcement learning for diffusion policy." arXiv preprint arXiv:2502.00361 (2025).

Expected Results

- IDM Training achieves a certain level of success rate
- Post-training further improves performance

[1] Ma, Haitong, et al. "Efficient online reinforcement learning for diffusion policy." arXiv preprint arXiv:2502.00361 (2025).

Summary

- Motivation: improve Scalability and Generalization
- Proposed pipeline
  - train autoregressive video backbone on actionless dataset
  - train IDM on target embodiment dataset
  - (optional)post-training w/ weighted RL
- IDM achieve a certain level of SR w/o post-trianing
- Post-training further improves performance