DataCrunch.io

Generative World Modelling for Humanoid Robots

Revontuli Team

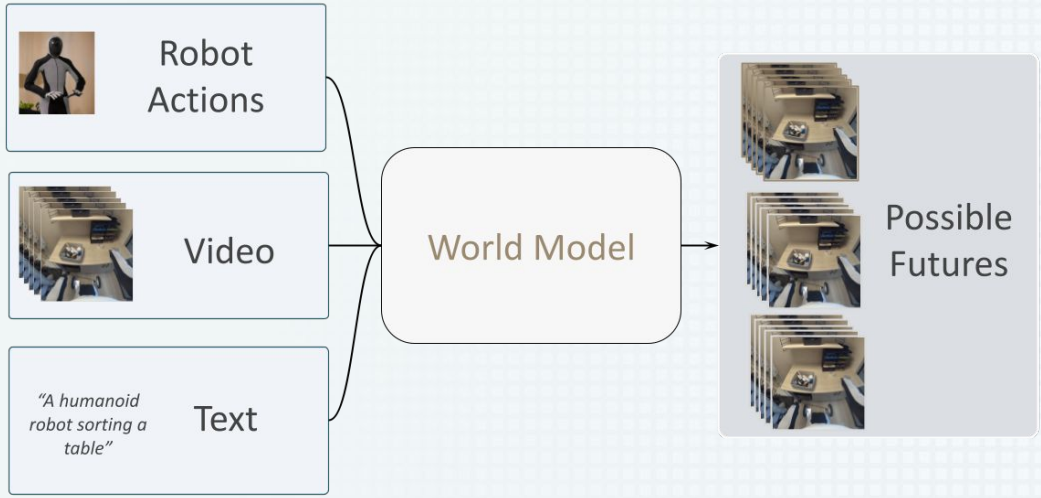
Riccardo Mereu*, Aidan Scannell*, Yuxin Hou, Yi Zhao,
Aditya Jitta, Antonio Dominguez, Luigi Acerbi, Amos Storkey, Paul Chang

Data Crunch · Aalto University · The University of Edinburgh · University of Helsinki · DR

*Equal contribution

World Models

*"In machine learning, a **world model** is a computer program that can imagine how the world evolves in response to an agent's behavior"*

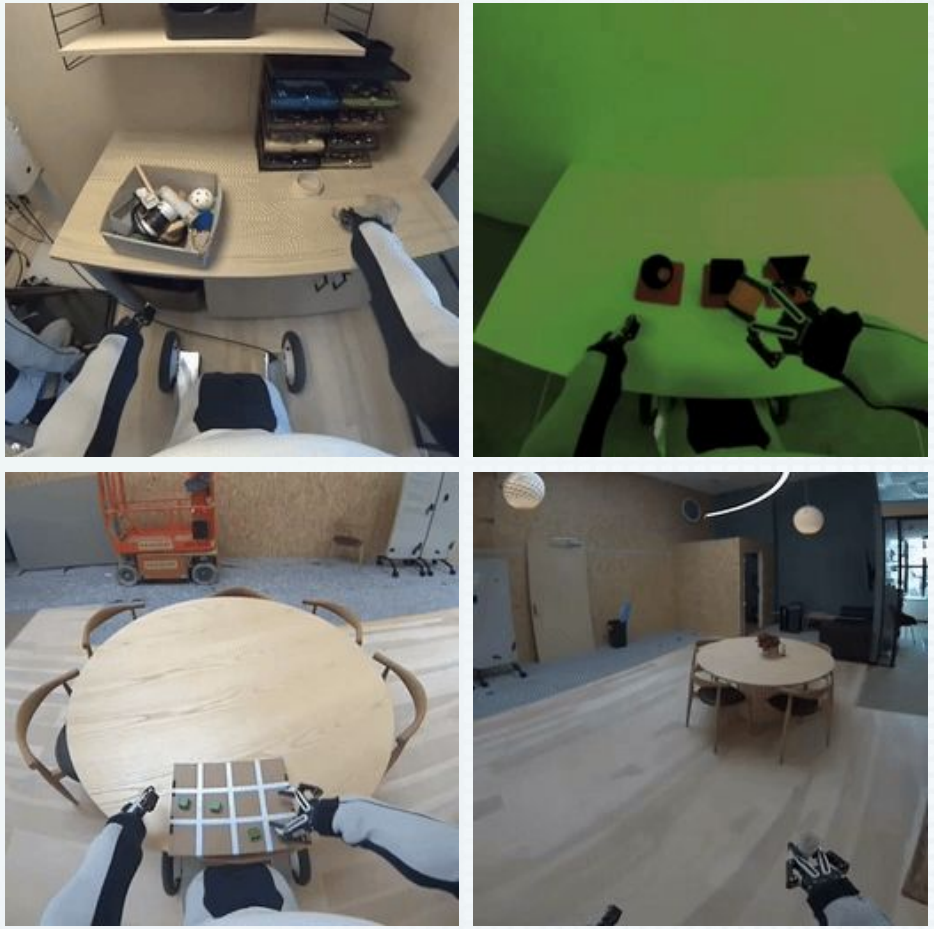


1X World Model Challenge

Challenges Track:

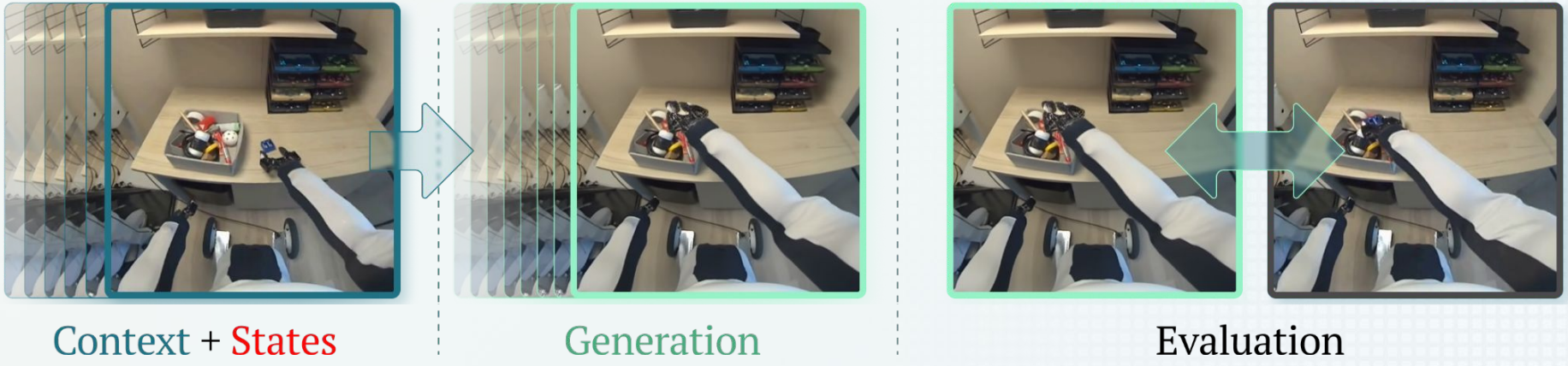
1. Sampling Track
2. Compression Track

100h videos collected with 1X EVE Humanoid Robots



Sampling Task

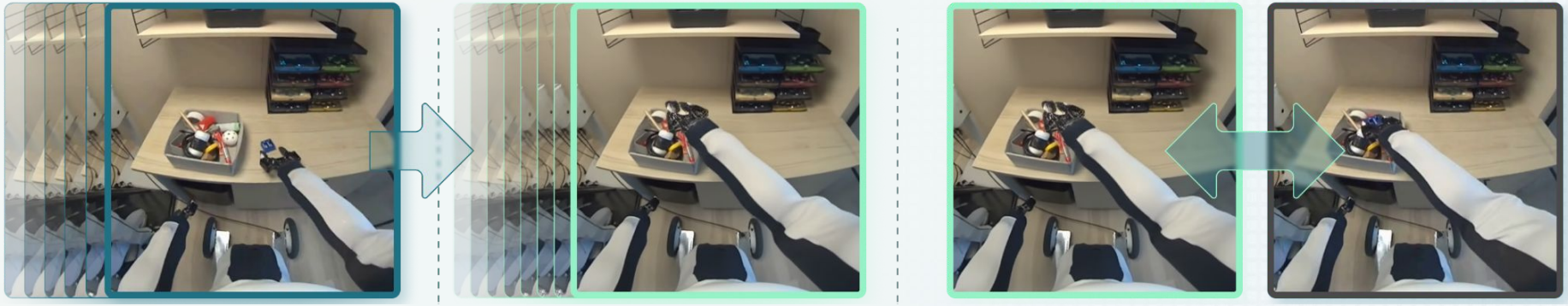
Sampling Task



- **Input:** **17 frames** and **77 robot states**
- **Evaluation:** PSNR between **77th predicted frame** and **ground truth**

$$\mathrm{PSNR}(\mathrm{target}, \mathrm{pred}) = 10 \cdot \log_{10}\left(\frac{\mathrm{MAX}^2}{\mathrm{MSE}(\mathrm{target}, \mathrm{pred})}\right)$$

Sampling Task

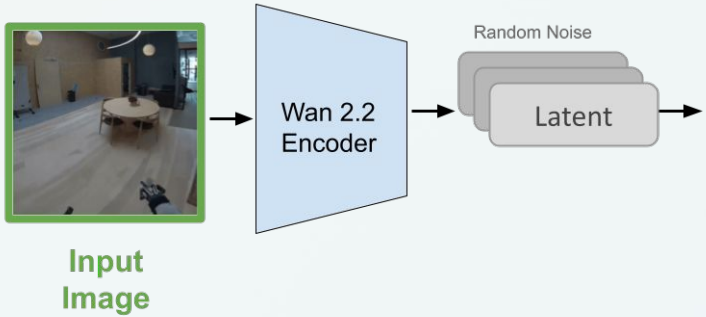


- Leverage SoTA video generative models
  → **Wan 2.2 TI2V-5B**
- Post-train the model on 1X WM Dataset

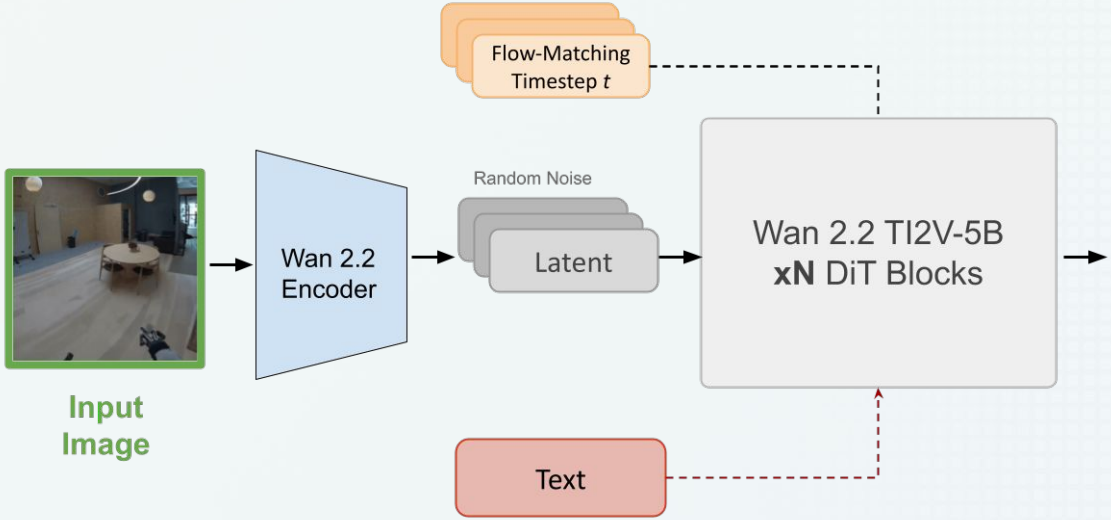
From Wan 2.2 TI2V-5B

Input
Image

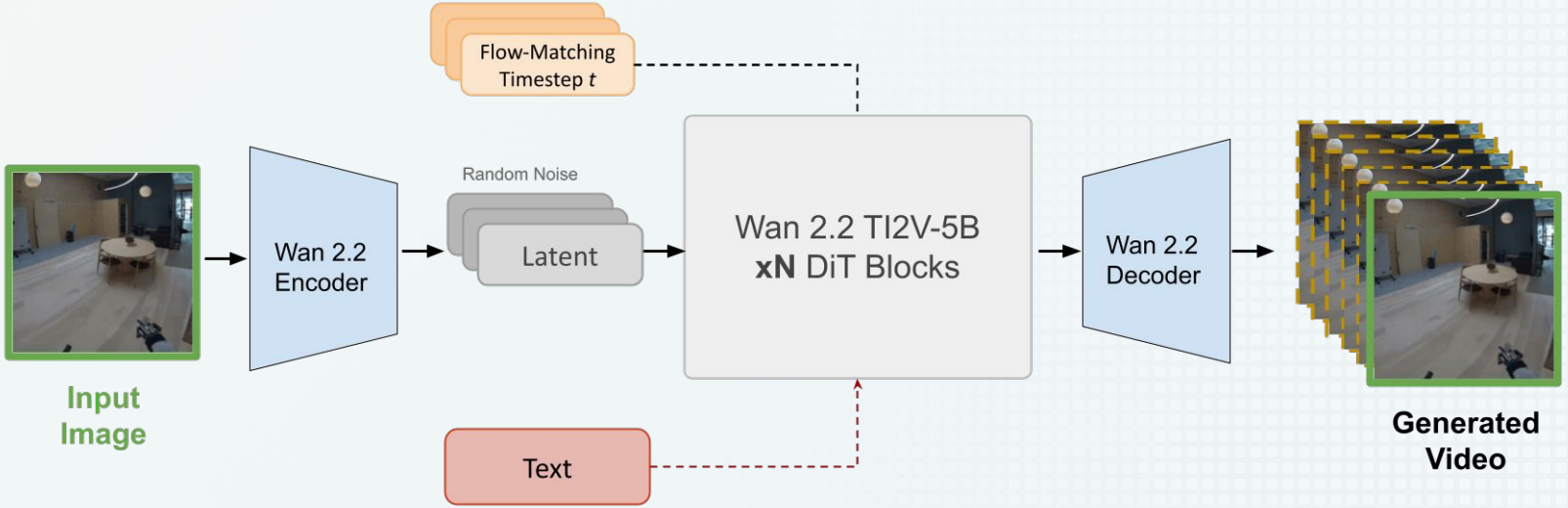
From Wan 2.2 TI2V-5B



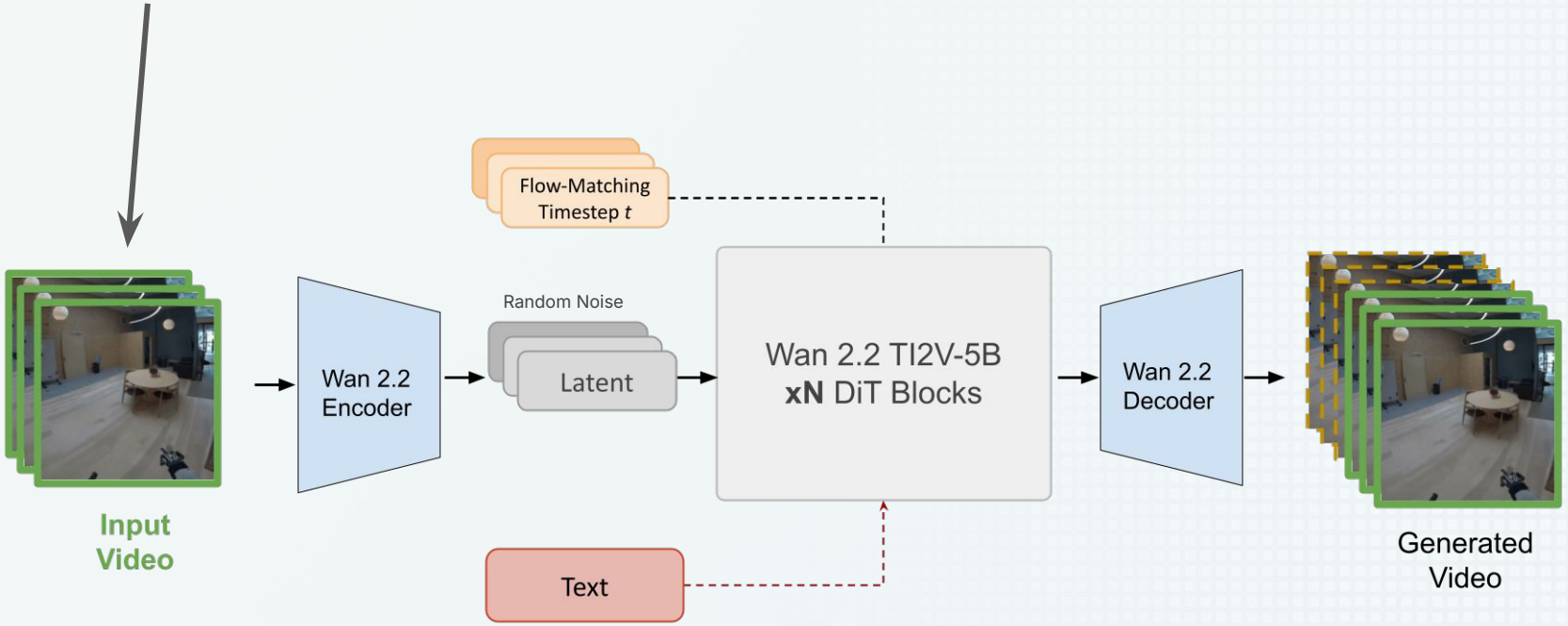
From Wan 2.2 TI2V-5B



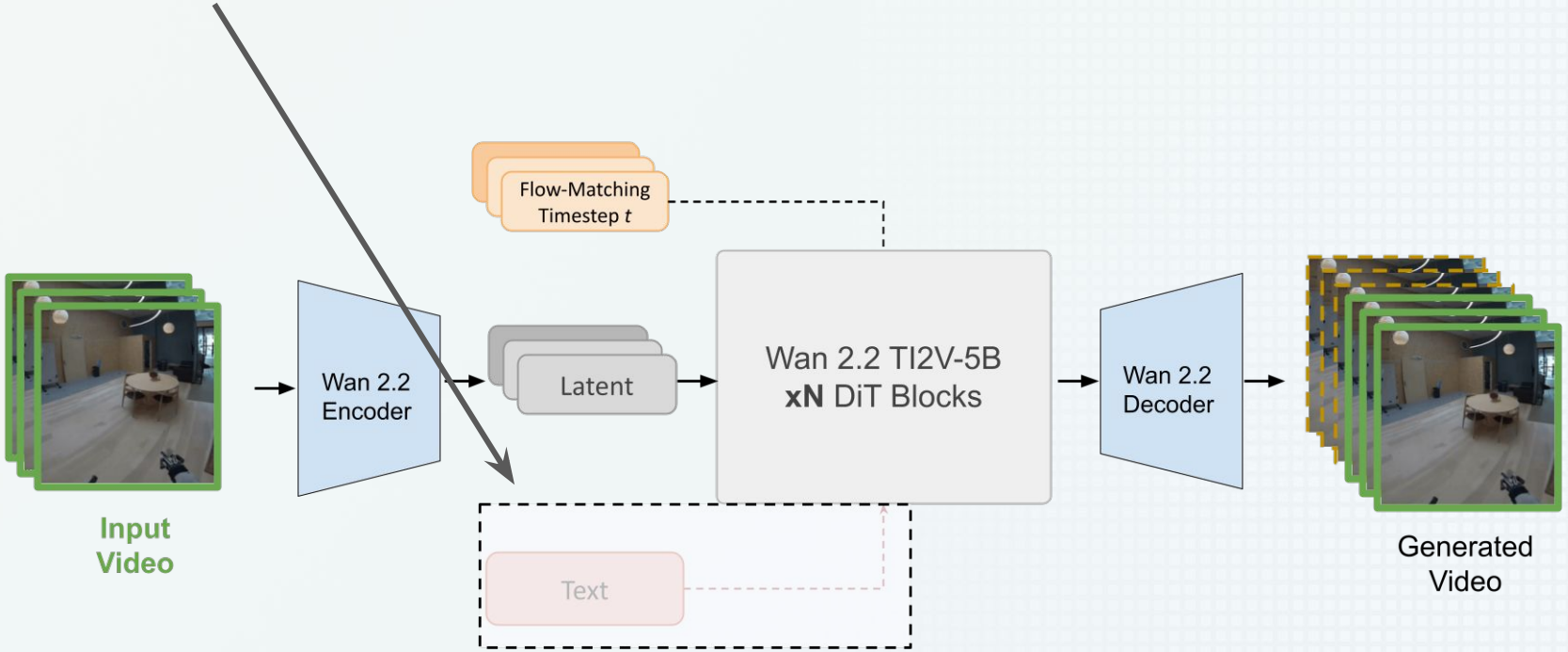
From Wan 2.2 TI2V-5B



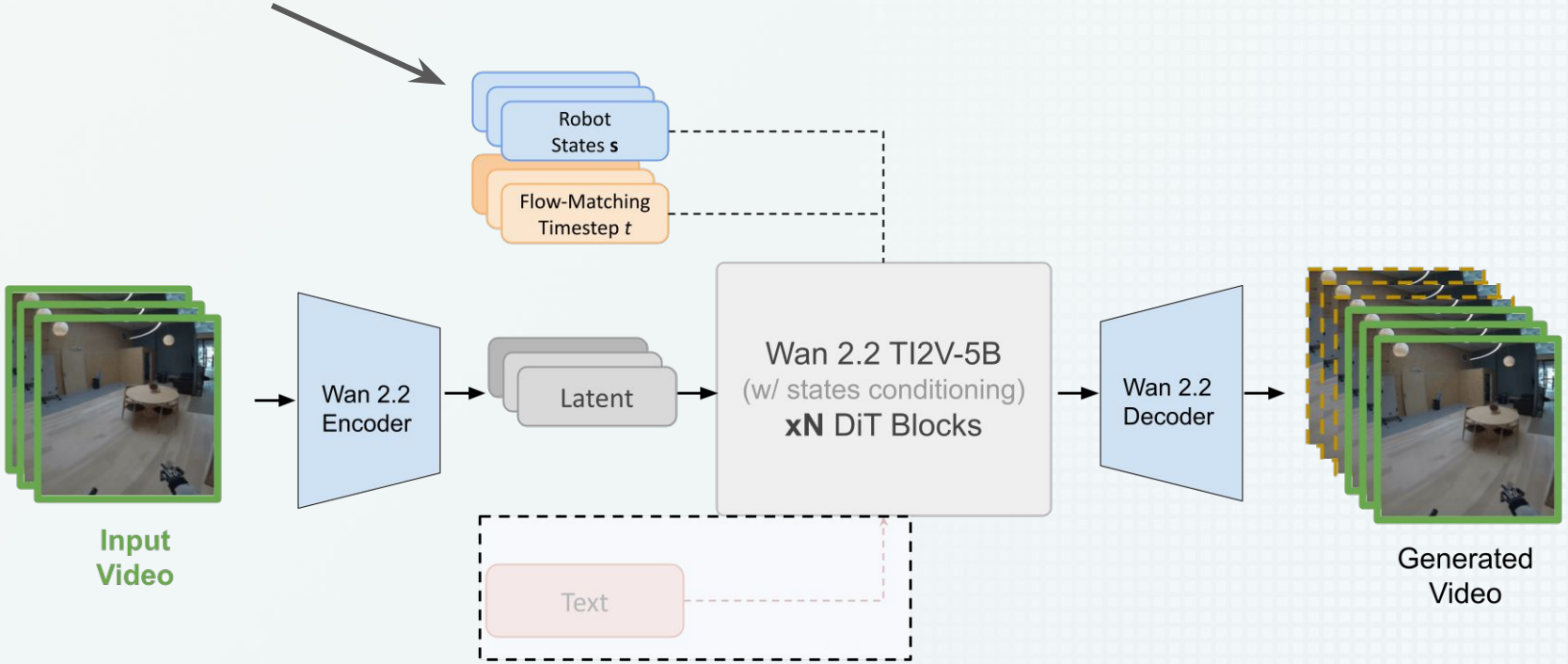
1. Add Video Conditioning



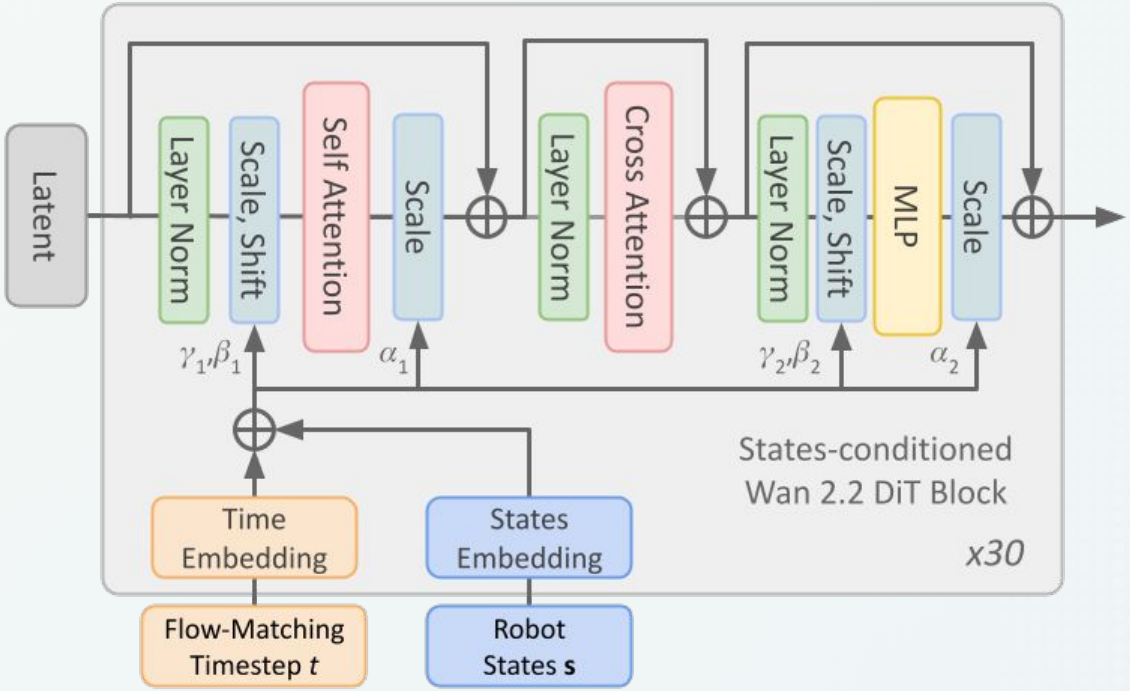
2. text_prompt = “”



3. Add Robot Conditioning



DiT Block: State Conditioning



- Flow Matching timestep *t* → AdaLN
- Robot's States **s** → AdaLN-Zero

Real vs. Generated Videos

GT Video

Generated Video

Real vs. Generated Videos

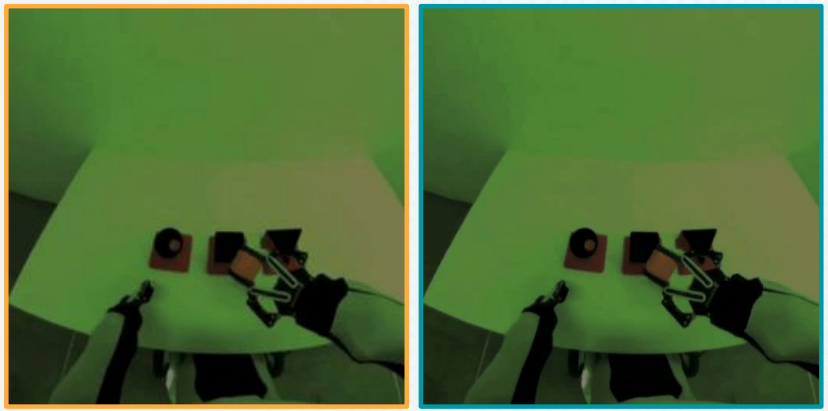
GT Video

Generated Video

Failure Modes

GT Video Generated Video

Object Coherence



GT Video Generated Video

Color Consistency

Inference Ensemble Approach

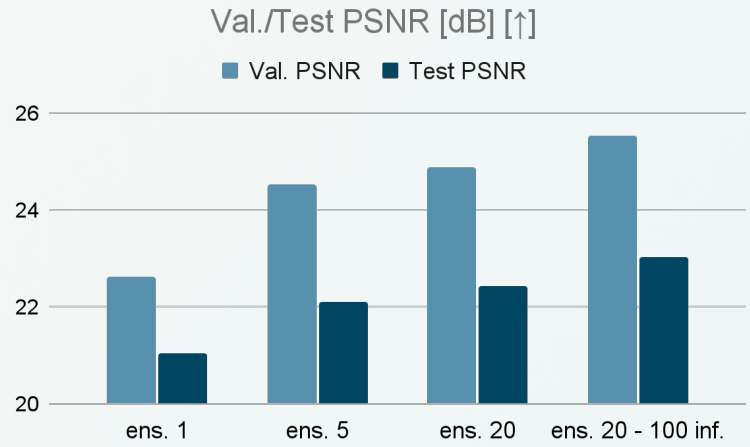
- Use the model to make multiple predictions
- Captures the uncertainty
- Averages out complicated areas in the generated frames

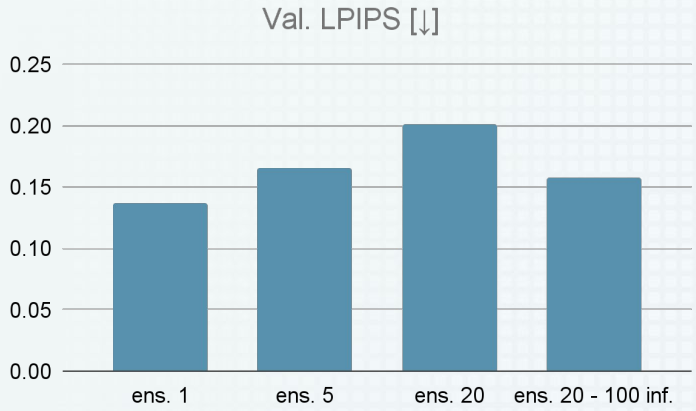
Generate N samples

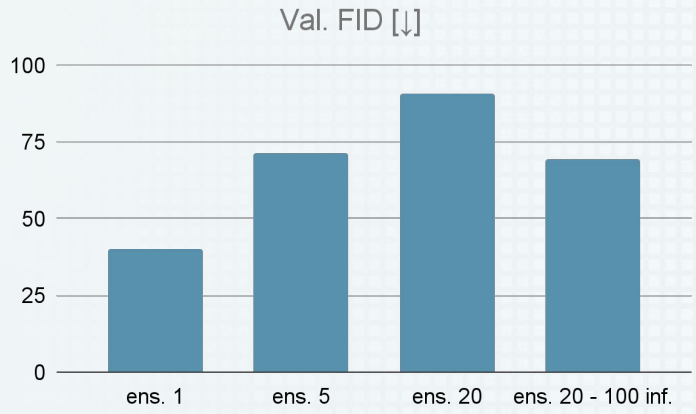
Ensemble Prediction

GT Video

Quantitative Comparison



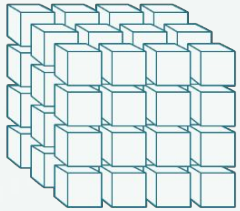




TL;DR: increasing the number of ensemble samples leads go less realistic videos, but achieves better PSNR.

Compression Task

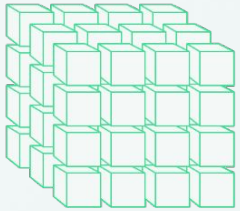
Compression Task



Context + States

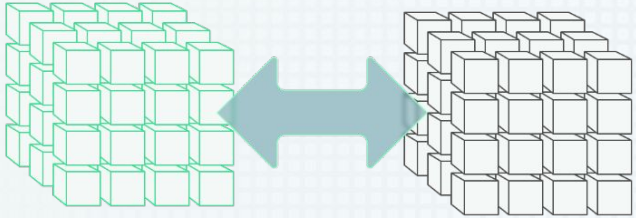
Past tokens (3×32×32)

Robot states



Generation

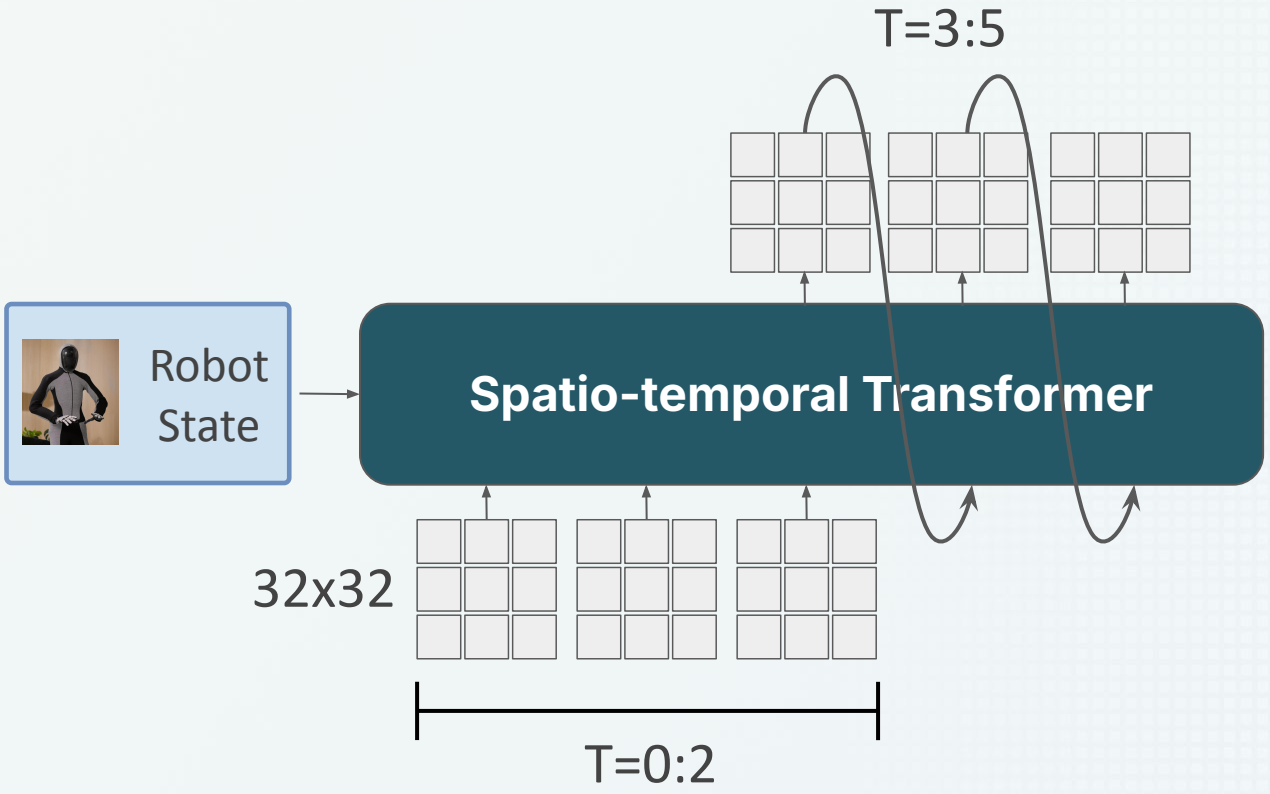
Future tokens (3×32×32)



Evaluation

$$\text{Top-500-CE}(\text{target}, \text{pred})$$

Architecture: ST-Transformer



Memory Complexity

- Tokens per sample: 5×32×32=5,120
  - T=5: number of time steps
  - S=1024: number of spatial locations
- Full attention memory complexity scales with

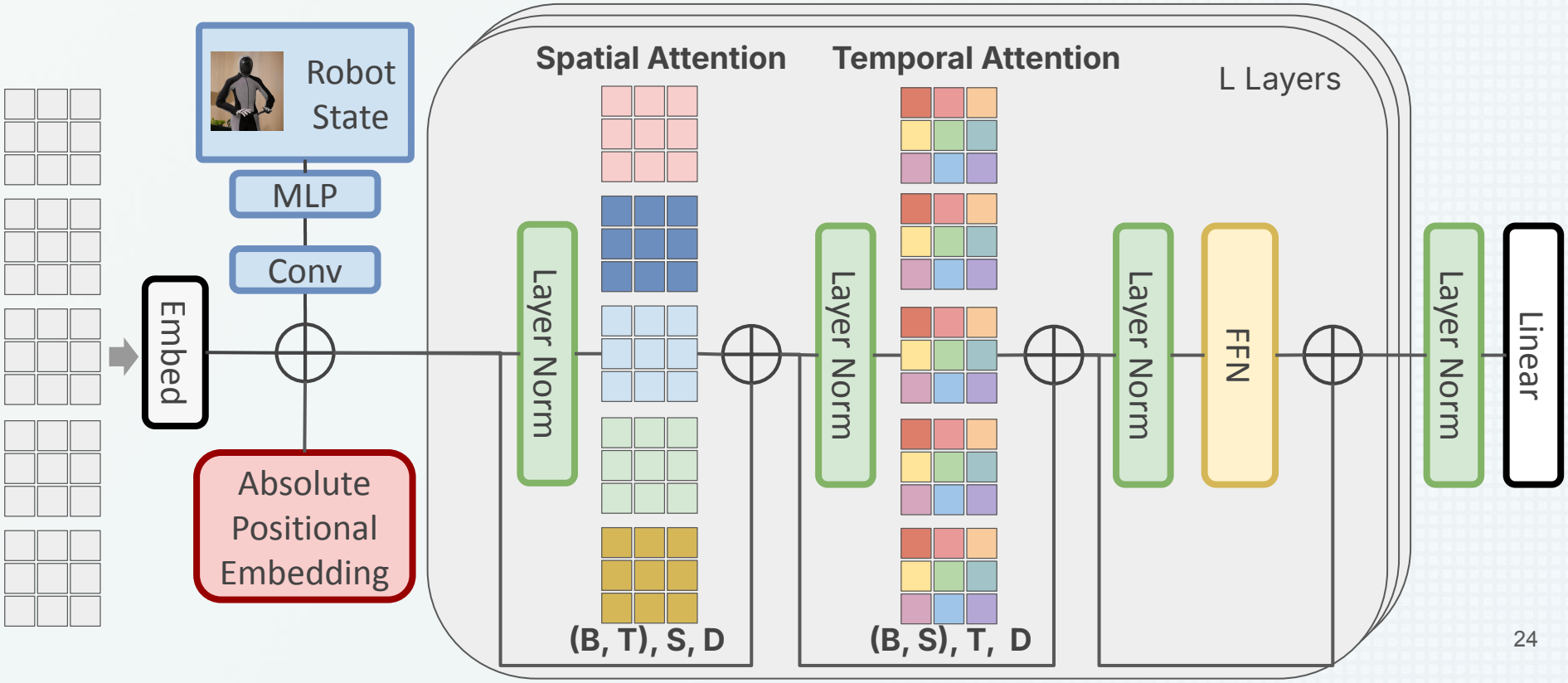
$$\mathcal{O}\left((TS)^2\right)$$

- Spatio-temporal attention scales with

$$\mathcal{O}\left(TS^2 + T^2S\right)$$

Method	Memory Complexity	Relative Cost
Full Attention	$(5 \times 1024)^2 = 26.2 \times 10^6$	100%
Spatio-temporal Attention	$5 \times 1024^2 + 5^2 \times 1024 = 5.26 \times 10^6$	20%

Architecture: ST-Transformer



Inference

Autoregressive inference

- **Autoregressive** over **time**
- **Parallel** over **space**

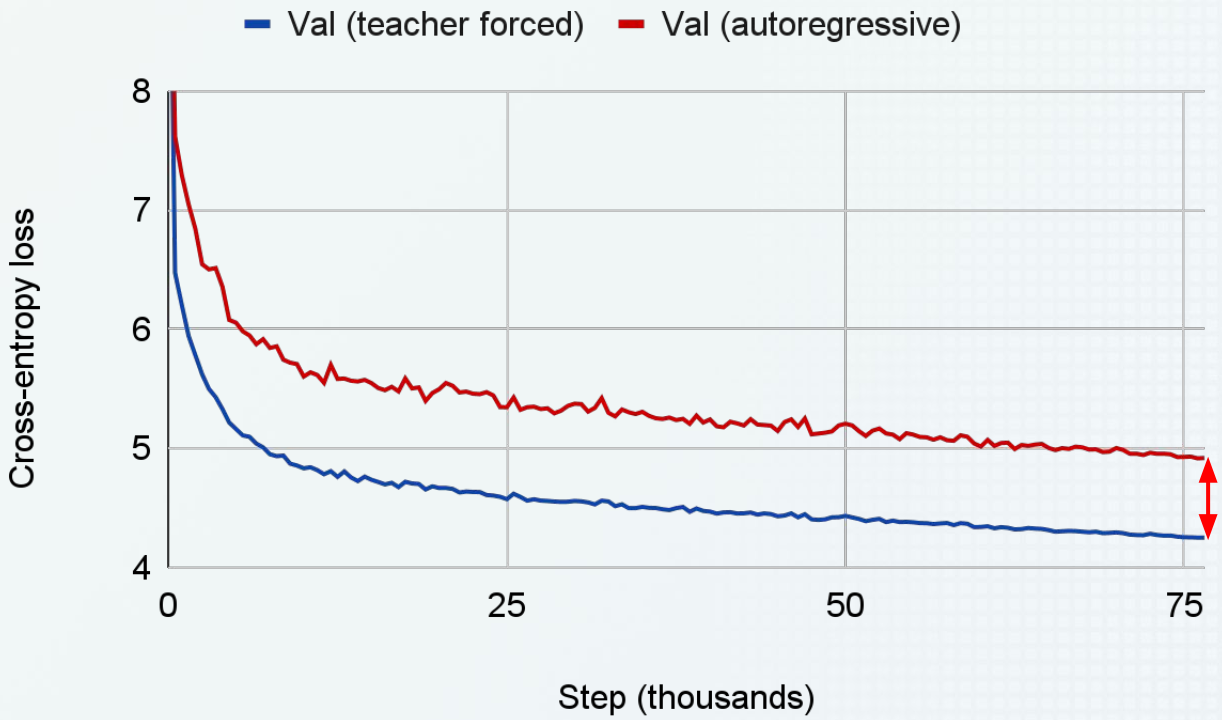
$$p(\mathbf{z}_{3:5} \mid \mathbf{z}_{<3}, \text{robot states}) = \prod_{t=3}^{5} f_\theta(\mathbf{z}_t \mid \mathbf{z}_{<t}, \text{robot states})$$

Greedy strategy

- At each step, the model picks the **most likely next token** instead of sampling randomly.

$$\mathbf{z}_t = \arg\max_{\mathbf{z}} f_\theta(\mathbf{z} \mid \mathbf{z}_{<t}, \text{robot states})$$

Results


Val (teacher forced)
Val (autoregressive)
Cross-entropy loss
8
7
6
5
4
0
25
50
75
Step (thousands)

Thanks for listening!

Contacts:
riccardo.mereu@aalto.fi
aidan.scannell@ed.ac.uk

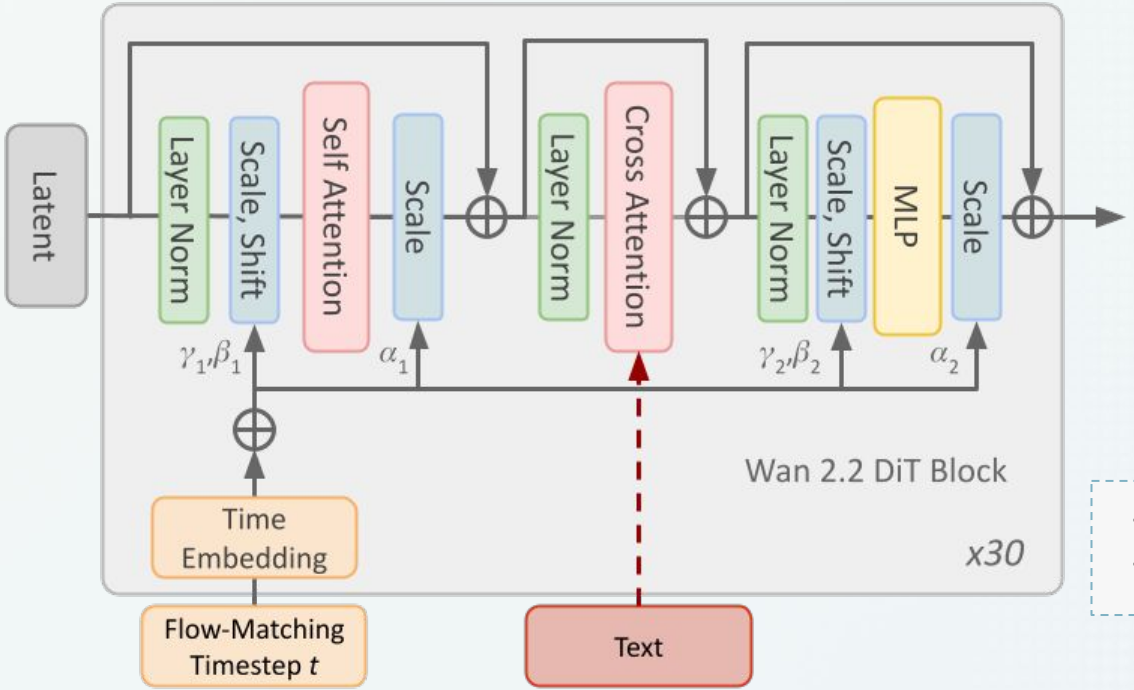
Sampling Training Details

- **LoRA config:**
  - lora_rank=32
  - Applied to all DiT layers
  - Wan2.2-VAE frozen
- **Optimizer:** AdamW w/ const. LR (4e-4) for
- **Batch size:** 128 (x8 GPUs = 1024 effective BS)
- Trained for 36h to achieve Val. PSNR = 23.04 dB

Compression Training Details

- **Model Config:**
  - **Num Layers**: 24
  - **Num heads**: 8
  - **Embedding dim**: 512
- **Parameters:** ~136 M
- **Train Config:**
  - **Optimizer:** AdamW w/ linear decay LR, 8e-4 peak, 2000 warm-up steps
  - **Weight Decay:** 0.05 (matrices only)
  - **Batch Size:** 20 (x8 B200 GPUs = 160 effective batch size)
  - **Epochs:** 80
- Trained for 17h to achieve Val. CE = 4.92

DiT Block: Wan 2.2 TI2V-5B

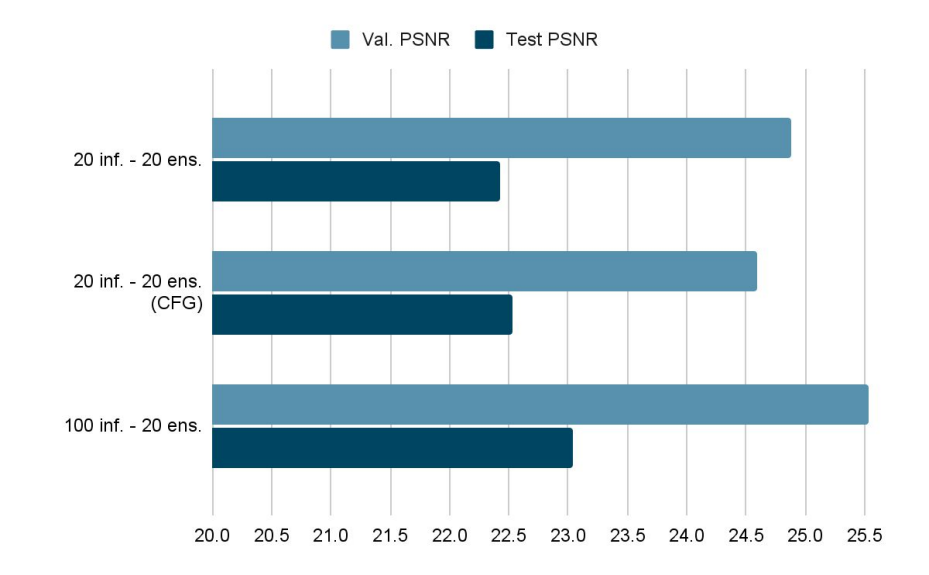


- Flow Matching timestep t → AdaLN
- Text Prompt → Cross-Attention

Team	PSNR [dB] [↑] Test	Val	Rank
Revontuli	**23.04**	**25.53**	1st
Duke	21.56	25.30	2nd
Micheal	18.51	–	3rd

Team	CE Loss [↓] Test	Val	Rank
Revontuli	**6.64**	**4.92**	1st
Duke	7.50	5.60	2nd
a27sridh	7.99	–	3rd

Number of Inference Steps



- Increasing the number of inference steps Val. PSNR increases and we get +0.5dB on the Public Test PSNR