NEURAL INFORMATION PROCESSING SYSTEMS

Video-Mirai: Autoregressive Video Diffusion Models Need Foresight

Yonghao Yu[1] Lang Huang[2] Runyi Li[1] Zerun Wang[1] Toshihiko Yamasaki[1]

[1] The University of Tokyo [2] National Institute of Informatics



Video-Mirai rollouts · 5 s · 832 × 480

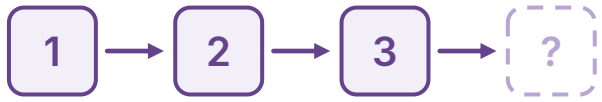
UTokyo

NII Inter-University Research Institute Corporation Research Organization of Information and Systems National Institute of Informatics

Project page
y0uroy.github.io/Video-Mirai

PROBLEM

Every emitted segment is a commitment



Segments are emitted one at a time, each from the past only.
Once shown, a segment cannot be revised.

MOTION

“A bear climbing a tree.”

× **Causal-Forcing:**
the climb resets.

IDENTITY & SCENE

“Gwen Stacy reading a book, tilt up.”

× **Causal-Forcing:**
the woman and the room change.

Causal-Forcing





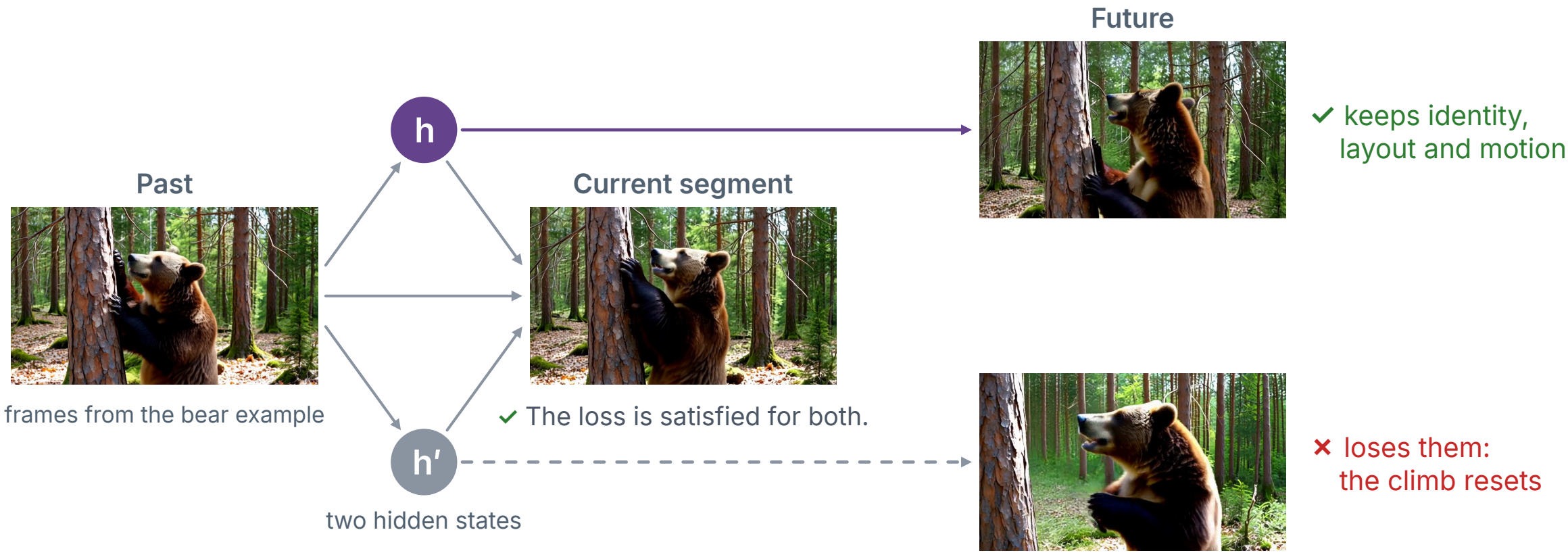
Video-Mirai (ours)





WHY IT HAPPENS

Output-level supervision leaves a planning gap



Representation-level planning gap: the loss does not ask the state to keep what the future needs.

IDEA

Future frames supervise representations, never generator inputs

Training the future is a teacher



past current future

Causal generator
hidden state h

Foresight target

supervises h during training

Inference strictly causal, unchanged



past current future

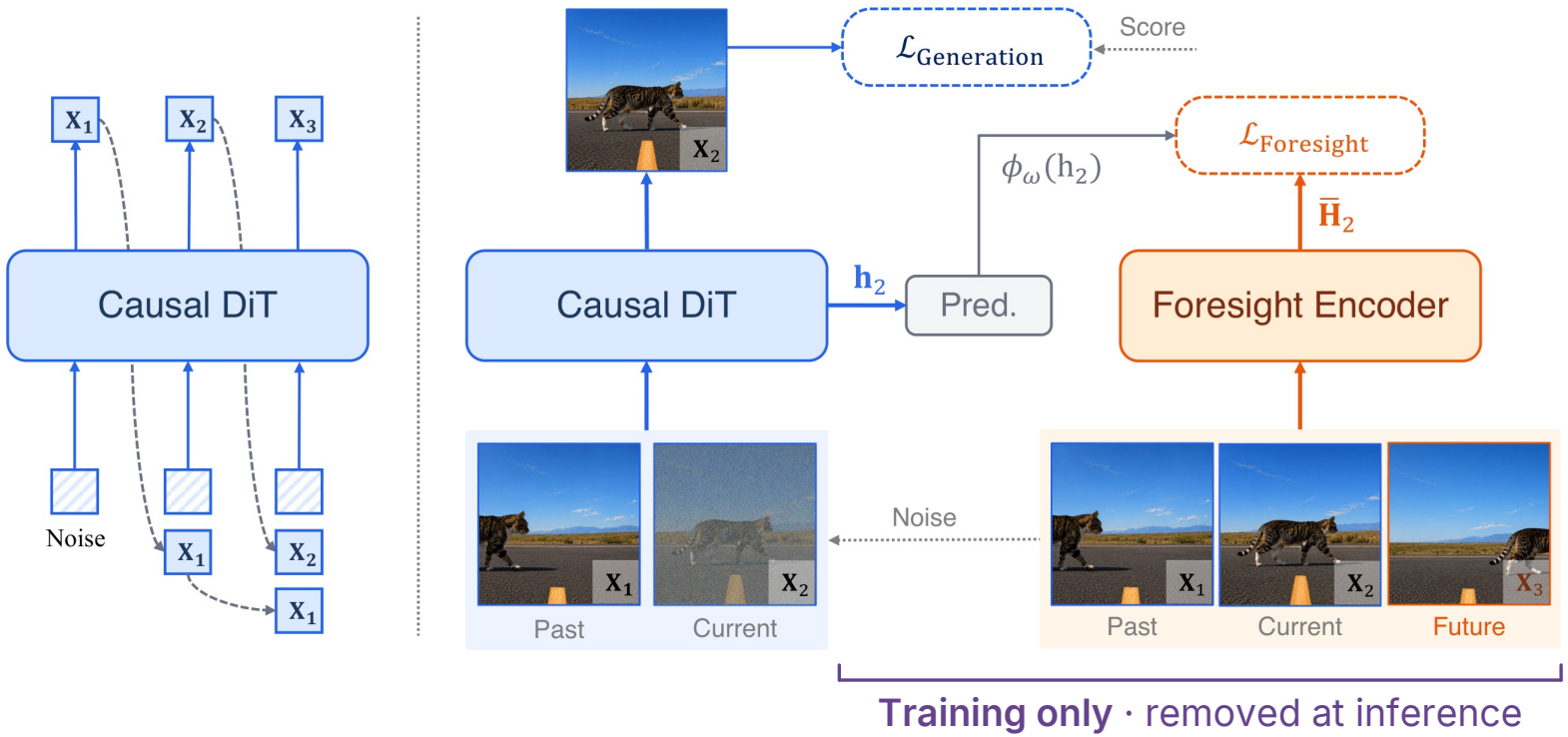
not yet generated

Same causal generator

No change in architecture, compute or KV cache.

METHOD

Align causal states with a future-aware target



Training only · removed at inference

1 Causal rollout
exactly as at inference

2 Frozen foresight encoder
reads the whole rollout;
target $\bar{H}$ averages the current and the next segment

3 Predictor φ + cosine loss
mid-layer state h → $\bar{H}$;
the target is stop-gradient

4 Inference
encoder and predictor removed:
0 extra cost

$$\bar{\mathbf{H}}_i = \frac{1}{|\Delta|} \sum_{\delta \in \Delta} \mathbf{H}^{L'}_{i+\delta}$$

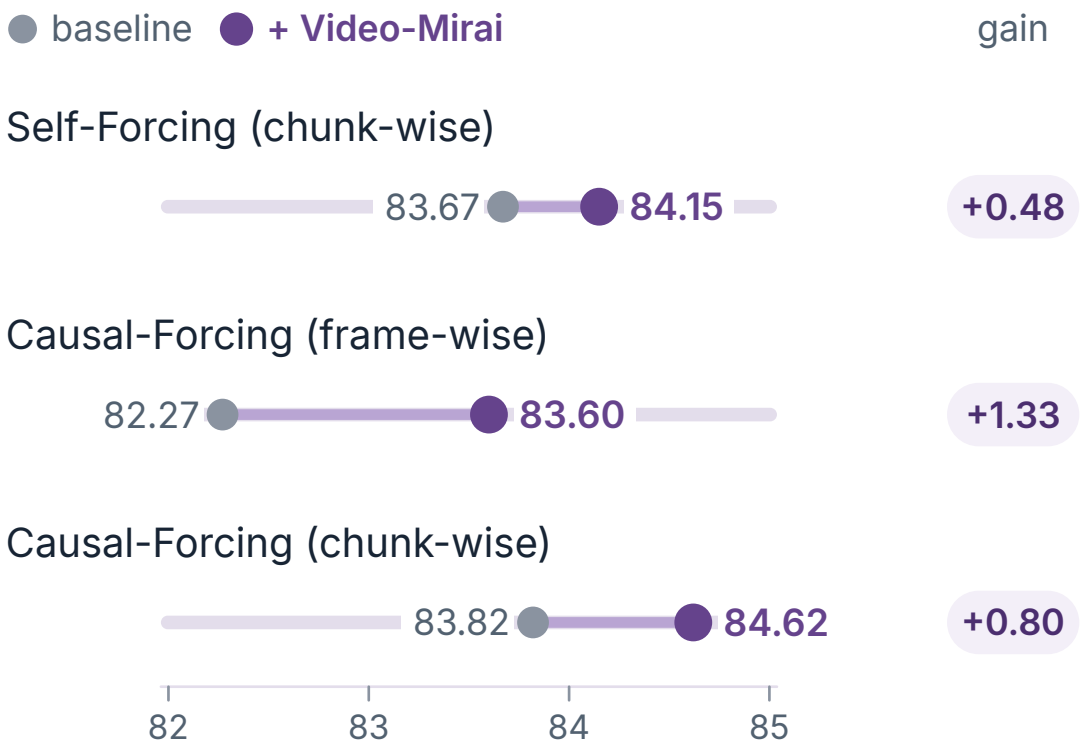
$$\ell^{\mathrm{F}}_i = 1 - \cos\big(\phi_\omega(\mathbf{h}^L_i),\ \mathrm{sg}[\bar{\mathbf{H}}_i]\big)$$

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{Generation}} + \lambda \underbrace{\frac{1}{N-K} \sum_{i=1}^{N-K} \ell^{\mathrm{F}}_i}_{\mathcal{L}_{\text{Foresight}}}$$

RESULTS

Drop-in gains on three autoregressive generators

VBench Total, 5-second videos



Throughput and latency unchanged: 17.0 FPS and 0.69 s chunk-wise, 8.9 FPS and 0.45 s frame-wise.

Causal-Forcing	Video-Mirai (ours)
	

"A person is climbing a rope."

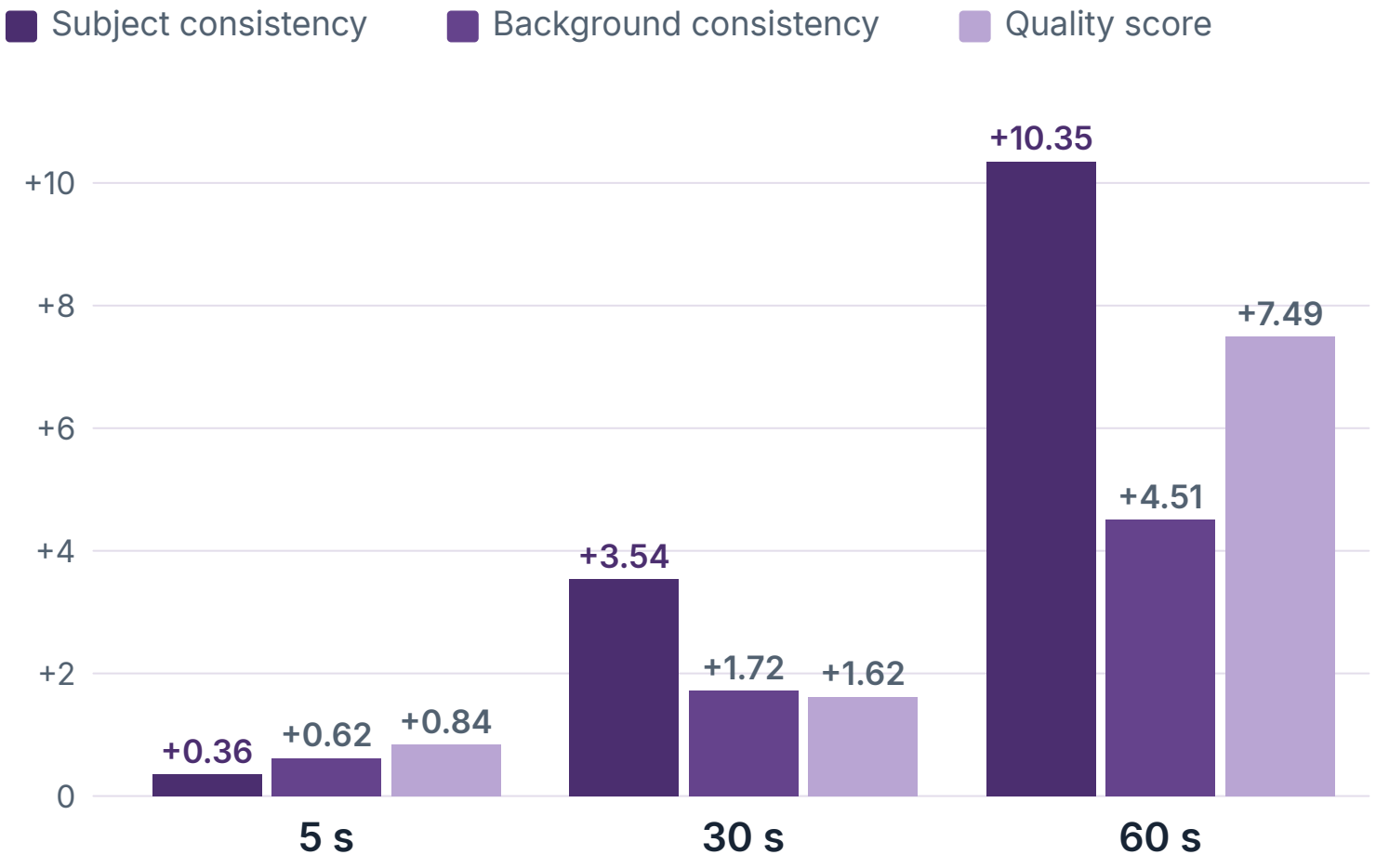
	

"The bund Shanghai, zoom in."

LONG ROLLOUTS

Gains grow with rollout length

Gain over Causal-Forcing (chunk-wise)



+10.35

Subject consistency at 60 s

Causal-Forcing 73.74 → 84.09

+7.49

Quality score at 60 s

Causal-Forcing 65.81 → 73.30

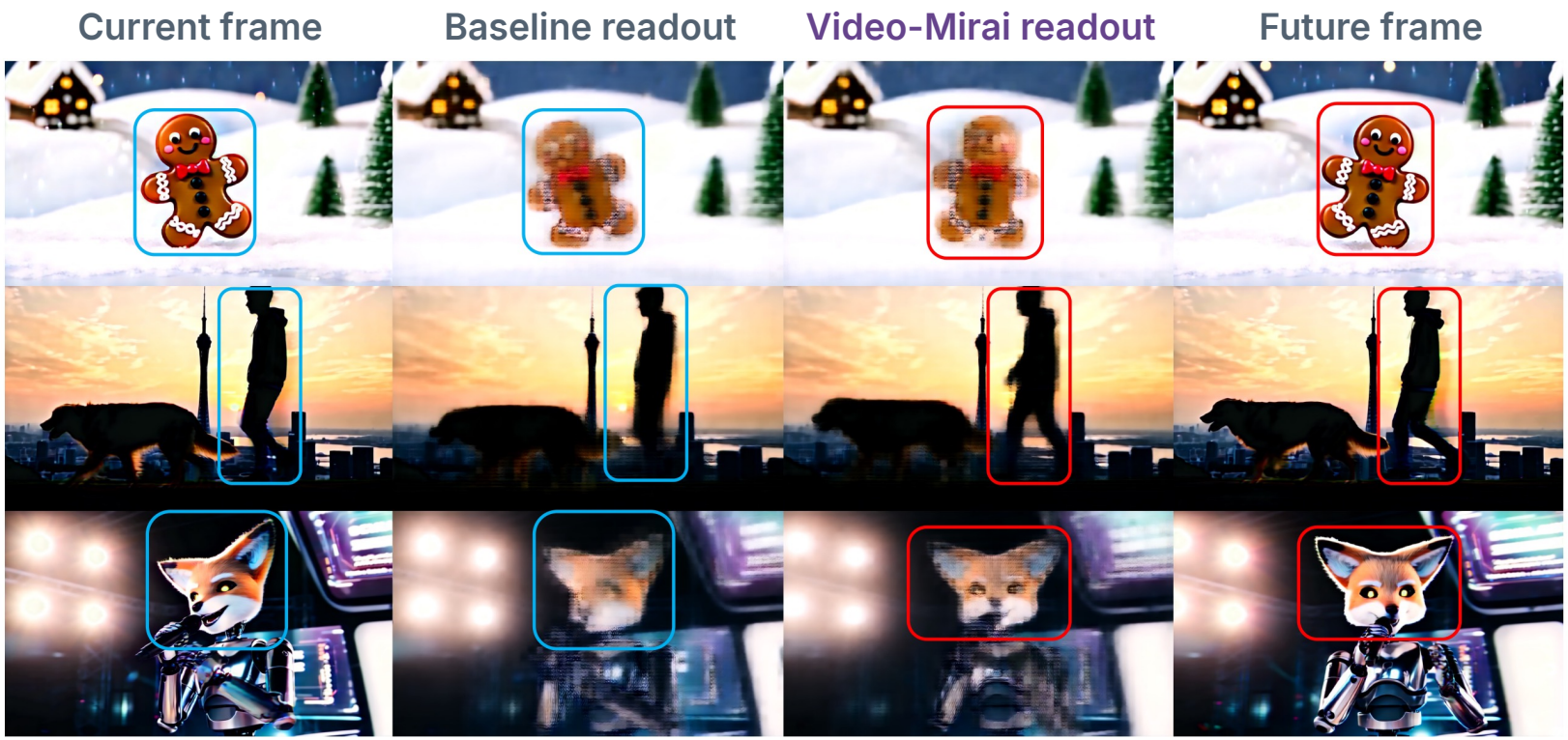
Trained on 5-second clips.

5 s: VBench prompts.

30 s and 60 s: 200 MovieGen prompts, VBench prompt-agnostic protocol.

WHAT THE GENERATOR LEARNS

The future becomes decodable from current features



Layer-15 readout from the current segment; the future frame is 3 latent frames ahead.

matches the current frame

matches the future frame

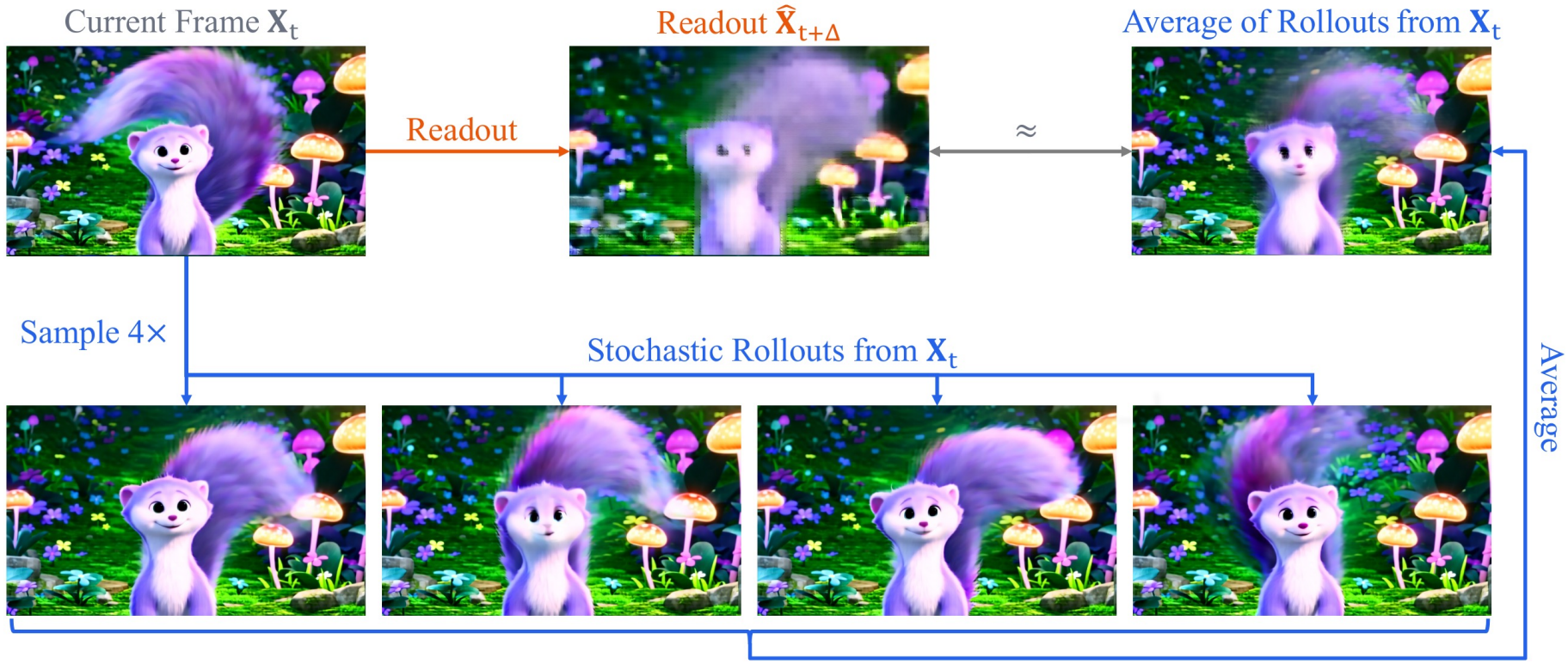
Each frozen generator gets its own MLP readout (same budget) that reconstructs a future frame from the current hidden state alone.

Baseline: closer to the current frame

Video-Mirai: the future position and pose

WHAT THE GENERATOR LEARNS

The readout resembles the average of future rollouts



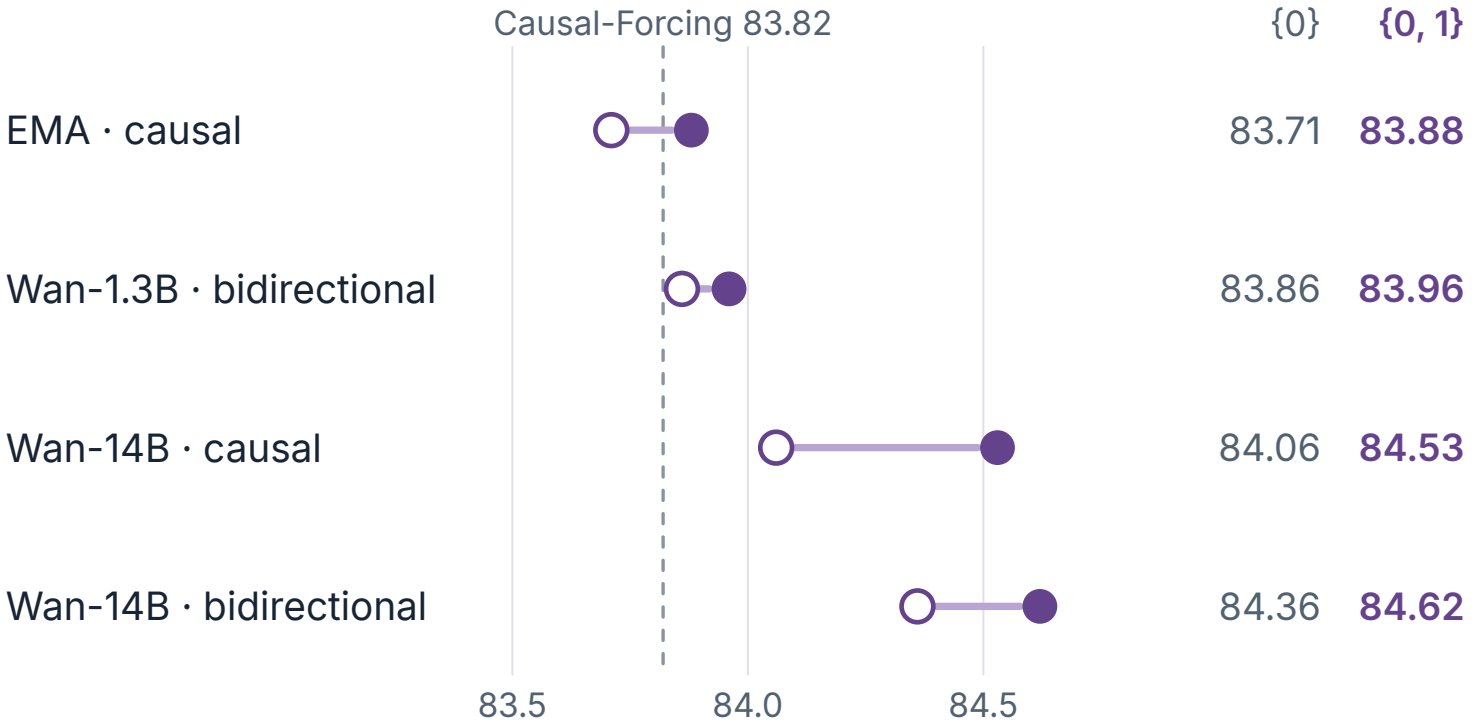
Readout ≈ average of sampled futures. Four rollouts sampled from the same current frame differ, for example in the tail; the one-shot readout from the frozen Video-Mirai state resembles their average, not any single sample.

ABLATIONS

Access to the future improves every encoder

Foresight encoder · VBench Total, 5 s

○ current only, Δ = {0} ● **current + next, Δ = {0, 1}**



Bidirectional encoders see later segments implicitly;
Δ = {0, 1} adds the next segment explicitly.

Best Total Score: Δ = {0, 1}

Δ = {0}	84.36
Δ = {1}	84.15
Δ = {0, 1}	**84.62**
Δ = {0, 1, 2}	84.49

Also improves SkyReels-V2-DF-1.3B

Total	81.82 → **82.89**
Quality	83.09 → 84.26
Semantic	76.72 → 77.42

500 fine-tuning steps, separate protocol

TAKEAWAY

Causality should constrain inference, not representation supervision

1

Training-only

Future frames supervise the hidden states of a causal generator during training.

2

Better rollouts

5 s: Total 83.82 → 84.62
60 s: subject consistency +10.35

3

Zero inference cost

Encoder and predictor are removed; same architecture, compute and KV cache.

Project page **y0uroy.github.io/Video-Mirai**
Code github.com/y0urOy/Video-Mirai
Models huggingface.co/y0urOy/Video-Mirai

Thank you!

UTokyo

NII Inter-University Research Institute Corporation Research Organization of Information and Systems National Institute of Informatics