[FILE SUBMISSION: RSS WORKSHOP IN ROBOT WORLD MODELS]

FF-JEPA: LONG-HORIZON PLANNING IN WORLD MODELS WITH LATENT PLANNERS

Sergi Masip, Jonathan Swinnen, Yutong Hu,
Renaud Detry, Tinne Tuytelaars

KU LEUVEN

SCAN FOR PROJECT SITE

// 01 · MOTIVATION

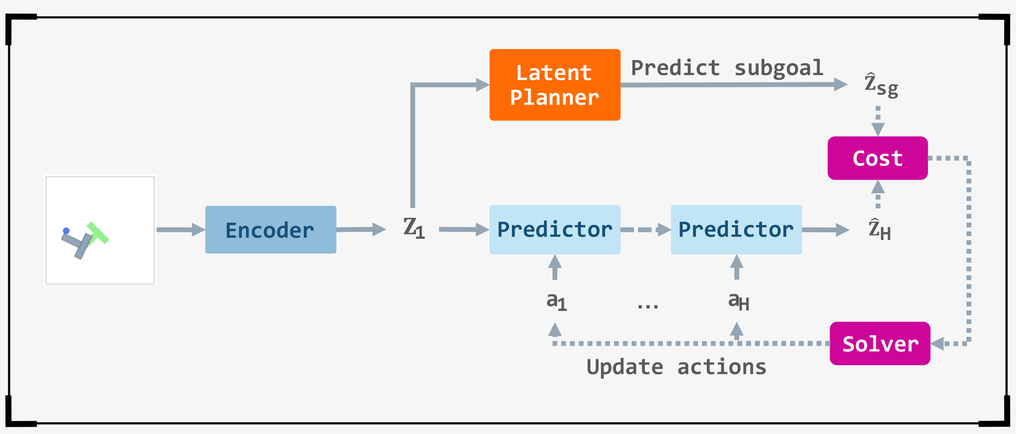
Flat world models can plan with CEM, **but with limitations**.

- Long-horizon planning is expensive
- Long-horizon planning is prone to error compounding
- Most methods need an explicit goal image upfront

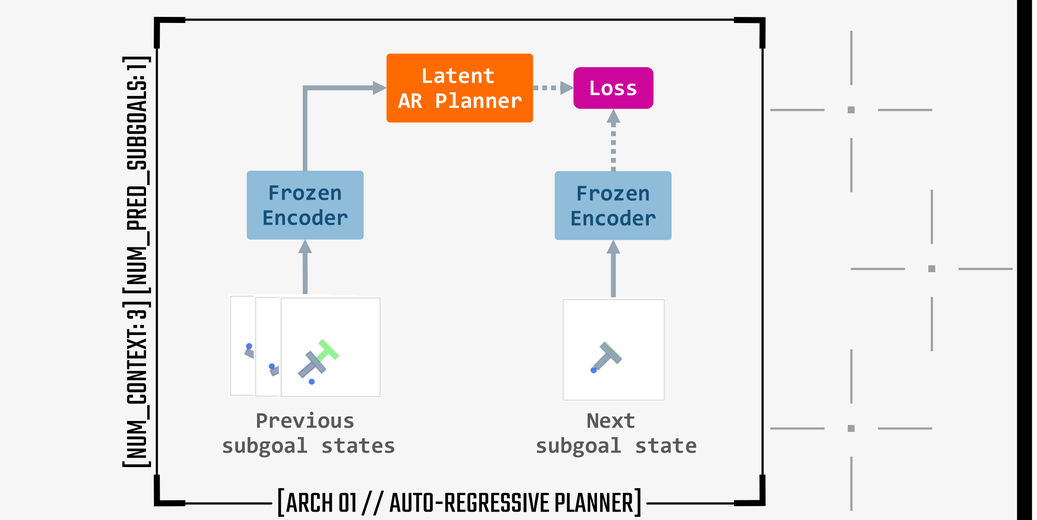
// 02 · OUR APPROACH

We tackle these limitations by proposing Forward-Forward JEPA (FF-JEPA), which introduces an action-free forward dynamics model to JEPA world models that predicts next subgoals.

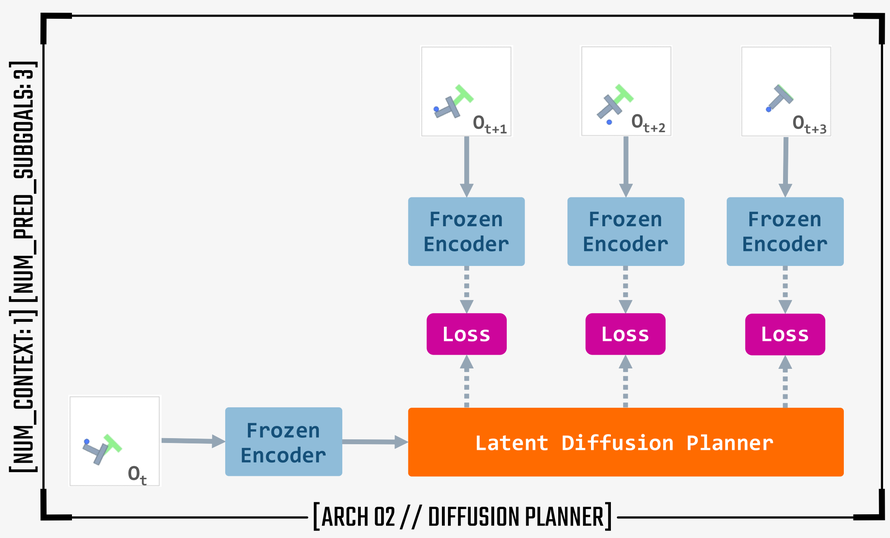
FF-JEPA enables long-horizon and goal-image-free planning.



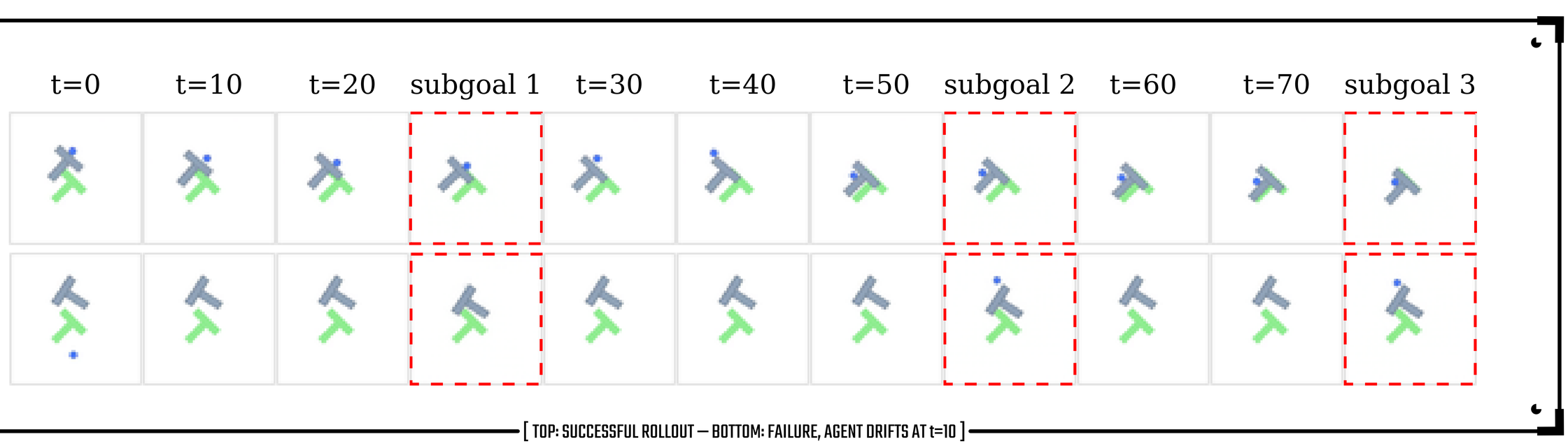
// 03 · LATENT PLANNER ARCHITECTURES



[ARCH 01 // AUTO-REGRESSIVE PLANNER]



[ARCH 02 // DIFFUSION PLANNER]



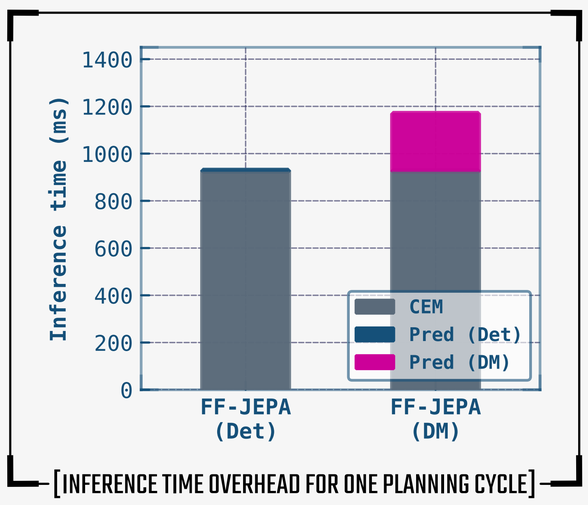
[TOP: SUCCESSFUL ROLLOUT — BOTTOM: FAILURE, AGENT DRIFTS AT t=10]

// 04 · RESULTS

EXPERIMENT VER. 512790...

91.8% vs 3.5% — FF-JEPA vs flat world model at long horizon (t=75) in Push-T.

Model	Short ($t = 25$)	Long ($t = 75$)	Random Init.
Baselines			
DINO*	84.0%	17.0%	—
DINO (Hierarchy)*	89.0%	61.0%	—
LeWM	94.53%	3.52%	0.00%
Ours			
FF-JEPA (Det)	76.95%	88.67%	81.25%
FF-JEPA (DM)	**96.09%**	**91.80%**	**82.42%**



[INFERENCE TIME OVERHEAD FOR ONE PLANNING CYCLE]

- Flat world models collapse under long horizons and random starts (3.5% → 0.0% success)
- FF-JEPA holds up across settings
- Unlike every baseline, FF-JEPA needs no goal image at inference time
- The deterministic planner makes for a good lightweight choice

* not directly comparable, different eval protocol

// 05 · CONCLUSIONS

FF-JEPA fixes the long-horizon and cost problems of flat CEM without the need for a goal image.

- 91.8% success at long horizon vs. 3.5% for flat world models
- Deterministic planner is a lightweight option when speed matters

TAKEAWAY action-conditioned world models + latent planners trained on unlabeled demos are a promising path to scalable, goal-image-free robot learning.

// ACKNOWLEDGEMENTS

This paper has received funding from the Flemish Government under the Methusalem Funding Scheme (grant agreement n° METH/24/009).