Vision-Based Safe HRC with Uncertainty Guarantees

Jakob Thumm[1], Marian Frei[2], Tianle Ni[3], Matthias Althoff[4], Marco Pavone[1]
[1]*Stanford University* [2]*RWTH Aachen* [3]*SJTU* [4]*TU Munich*

Motivation

- Robots in human-centered environments require **certifiable safety guarantees**.
- Provably safe HRC relies on **marker-tracked poses** and **conservative motion models**.
- Learning-based pose and motion predictors are less conservative but lack
  - → end-to-end **uncertainty propagation**,
  - → **conformal** prediction guarantees,
  - → handling of **OOD** inputs.
- **Goal**: strong **probabilistic guarantees** for learning-based pose and motion prediction models in **certifiable** safety shields.

Core Concept

In-distribution: trust prediction + conformal set

Out-of-distribution: reuse last valid prediction

Always $\Rightarrow P(\boldsymbol{y} \in \mathcal{S}) \geq 1 - \varepsilon$

- **SLU** score (Sketched-Lanczos) flags OOD inputs at pose and motion level.
- OOD-aware buffer keeps predictions running through **unreliable frames** and resumes after N_{req} valid observations.
- Directly usable in certifiably safe **SARA shield**.

Conformal Prediction Sets

For each joint j and horizon t_k we define the **non-conformity score**

$$A_k^j(\boldsymbol{z}_i) = \frac{\|\boldsymbol{d}_{k,i}^j\|_2}{\sqrt{\lambda_{\max}(\boldsymbol{C}_{k,i}^j)}},$$

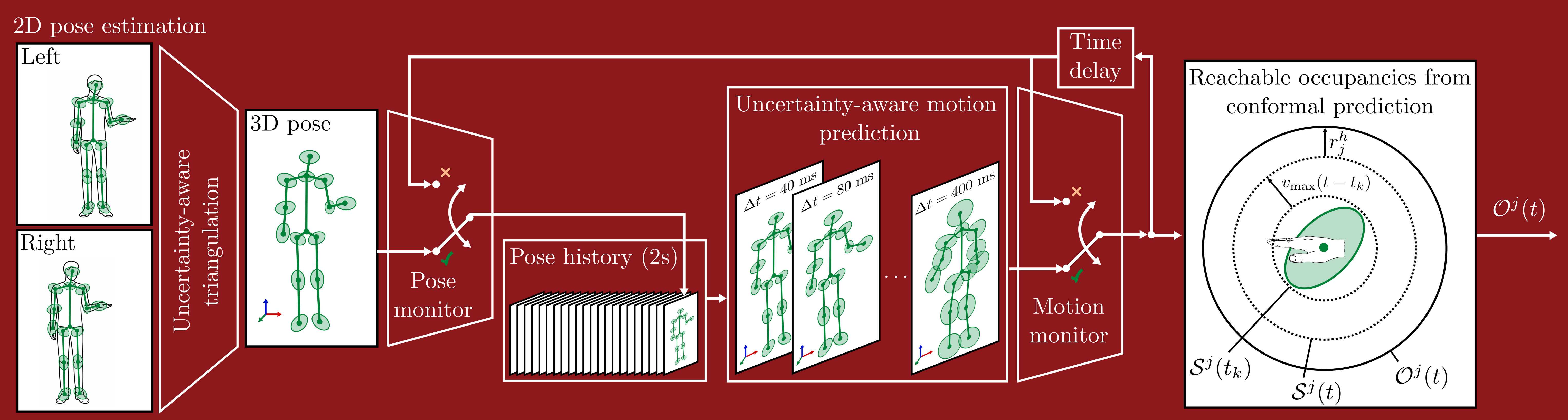
where $\boldsymbol{d}_{k,i}^j$ is the residual and $\lambda_{\max}(\boldsymbol{C}_{k,i}^j)$ is the maximal eigenvalue of the predicted covariance.

Proposition 1 (Conformal set). With α_k^j the $(1-\varepsilon)$-quantile of $\{A_k^j\}$ on calibration data, a ball $\mathcal{S}$ with center $\hat{\boldsymbol{p}}_k^j$ and radius $\alpha_k^j \sqrt{\lambda_{\max}(\boldsymbol{C}_k^j)}$ satisfies $P(\boldsymbol{y} \in \mathcal{S}) \geq 1 - \varepsilon$.

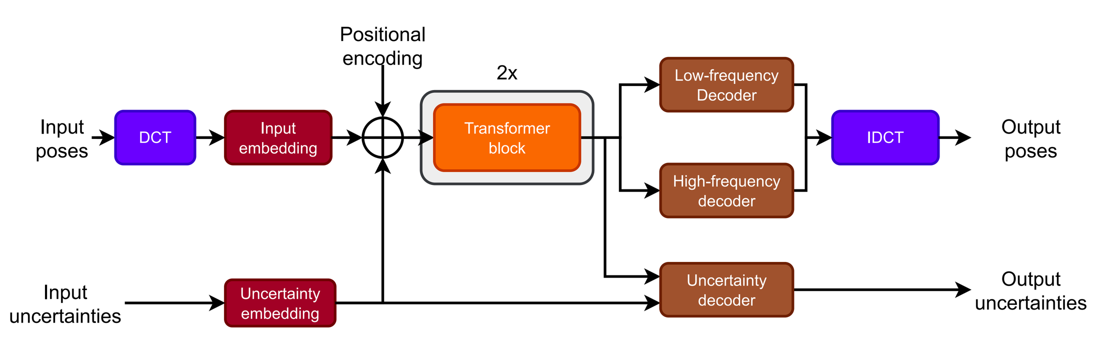
For intermediate times we extend the radius by $\Delta t \, v_{\max}$ with $v_{\max} = 1.6\,\mathrm{m/s}$.

From pixels to certificates:

$11\times$ tighter reachable sets with formal coverage guarantees.



Motion Prediction Architecture



- Input: $K_I = 50$ frames of **poses + covariances**.
- **DCT** representation separates **low- and high-frequency** motion components.
- Dual-path **Transformer** processes each frequency band before re-merging.
- Heteroscedastic covariance head via **Cholesky** $\boldsymbol{C} = \boldsymbol{L}\boldsymbol{L}^\top$ for positive-definiteness.
- Three-phase training:
  1. In: **pose** → out: **pose**
  2. In: **pose** → out: **pose** + **cov**
  3. In: **pose** + **cov** → out: **pose** + **cov**
- Out: $K_P = 10$ future **poses** and **covariances**.

Results

98.25%
conformal coverage
(target $1-\varepsilon = 99\%$)

$11\times$
smaller reachable volume vs. ISO 13855

36%
fewer interruptions via OOD-aware fallback

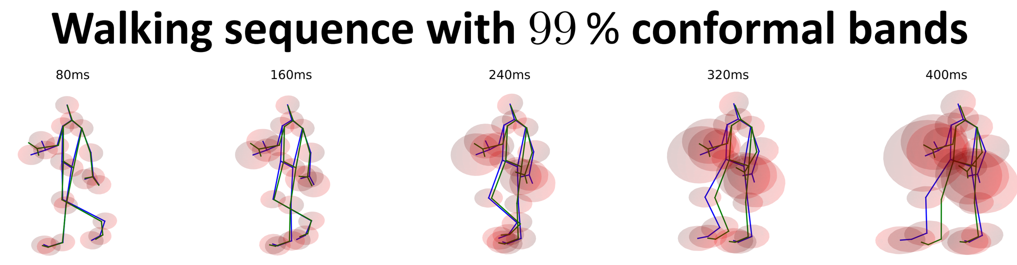
Table I. Motion prediction on H36M, MPJPE (mm) ↓

Method	80 ms	160 ms	320 ms	400 ms
HisRep	10.4	22.6	47.1	58.3
ST-DGCN	10.3	22.7	47.4	58.5
ST-Trans	10.4	23.4	48.4	59.2
SiMLPe	9.6	21.7	46.3	57.3
Ours (stage 1)	**8.7**	**16.7**	**41.1**	**54.5**
Ours (final)	18.4	28.1	53.1	67.2

Table II. Conformal prediction sets on H36M

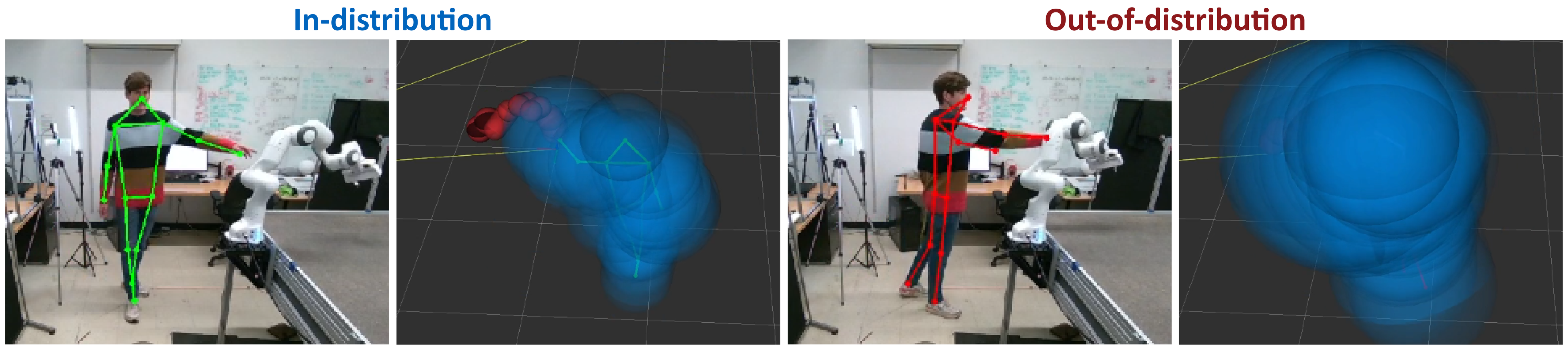
Method	Coverage (%) ↑	Vol. (m^3) ↓
ISO 13855:2010	97.93	0.191
Ours	**98.25**	**0.017**

Walking sequence with 99 % conformal bands



Real-world deployment (Franka + RealSense + SARA shield)

In-distribution **Out-of-distribution**



Certified safe stops with vision-only sensing: **ID**: tight sets; **OOD**: reused predictions → larger sets, operation continues.