A Additional problem motivation

In safety-critical scenarios, and for tasks requiring reasoning about uncertainty, user-set confidence levels $(1-\alpha) \in (0,1)$ are often used to construct *prediction regions*[7] containing $(1-\alpha)\%$ of the predictive probability mass. For example, one might plan so that the 90% prediction regions do not intersect any known obstacles, or plan so that the region's size is reduced, to incentivize safety and to minimize uncertainty, respectively. However, prediction regions generated from $\tilde{f}$ are generally *uncalibrated*; e.g. a region containing 90% of the predictive probability mass may not contain 90% of the true future states, which evolve according to f, not $\tilde{f}$. This lack of calibration can produce catastrophic consequences. Overconfident models may lead to execution-time collisions despite apparently safe plans, while underconfident models may be overly conservative and fail to make task progress. Both cases are undesirable. Approximate models alone are hence insufficient for safety-critical tasks.

B Dynamical system equations

The standard double integrator-dynamics with Heun discretization [9,19] are

$$\begin{bmatrix} p_{t+1} \\ v_{t+1} \end{bmatrix} = \underbrace{\begin{bmatrix} I_{2\times 2} & \Delta t\, I_{2\times 2} \\ 0_{2\times 2} & I_{2\times 2} \end{bmatrix}}_{A} \begin{bmatrix} p_t \\ v_t \end{bmatrix} + \underbrace{\begin{bmatrix} \frac{\Delta t^2}{2} I_{2\times 2} \\ \Delta t\, I_{2\times 2} \end{bmatrix}}_{B} a_t + w_t, \quad w_t \sim \mathcal{N}(0,\, Q)\,, \tag{B.1}$$

where w_t is independently sampled at each time-step, introducing aleatoric perturbations. To represent slipper/lower-friction surfaces, the true system evolves according to the following dynamics in the white regions of the planar and Isaac environments:

$$\begin{bmatrix} p_{t+1} \\ v_{t+1} \end{bmatrix} = \underbrace{\begin{bmatrix} I_{2\times 2} & \mathbf{1.3}\,\Delta t\, I_{2\times 2} \\ 0_{2\times 2} & I_{2\times 2} \end{bmatrix}}_{A_{\text{shift}}} \begin{bmatrix} p_t \\ v_t \end{bmatrix} + \underbrace{\begin{bmatrix} \mathbf{1.3}\,\frac{\Delta t^2}{2} I_{2\times 2} \\ \mathbf{1.3}\,\Delta t\, I_{2\times 2} \end{bmatrix}}_{B_{\text{shift}}} a_t + w_t. \tag{B.2}$$

C Planning Implementation Details

We used Model Predictive Path Integral Control (MPPI) as our trajectory optimizer for all experiments and methods. The objective function used for probabilistically safe plan generation was

$$J := c^{term}_{dist} d_E(\tilde{\mu}_{t+H}) + \sum_{\tau=t+1}^{t+H-1} c^{run}_{dist} d_E(\tilde{\mu}_\tau) + c^{run}_{trace} \mathrm{tr}(\tilde{\mathcal{N}}_{\tau,cal}) + c^{run}_{col} \mathbb{1}_{\{\hat{\mathcal{C}}(X_\tau) \cap \texttt{Unsafe}_t \neq \emptyset\}}, \tag{C.1}$$

where d_E is the Euclidean distance between the predictive Gaussian's mean and the `Goal` region, $\mathrm{tr}(\cdot)$ the trace of the calibrated Gaussian, and the indicator function determines when the $(1-\alpha)$ prediction region of the calibrated Gaussian intersects with the obstacles observed at time t. The trace term is intended to

[7] For Gaussian $\tilde{f}$, prediction regions take the shape of hyperellipsoids.

steer plans towards regions of input space $\mathcal{X}_k$ with lower dynamics uncertainty. For the planar system experiments of Sec. 6.1, we used $c^{run}_{dist} = 0.8$, $c^{term}_{dist} = 1.0$, $c^{run}_{trace} = 1.5$, $c^{run}_{col} = 1,000$, with an MPPI temperature $\lambda = 0.25$, 2048 sampled action-trajectories per step, control noise $I_{2\times 2}$, and horizon of $H = 9$. Horizons that are too short can lead to myopic behavior and long horizons may make all trajectories infeasible (due to the successive scaling performed in our heuristic propagation). For the Isaac Sim trajectory optimization experiments, we used a similar cost function with $c^{run}_{dist} = 1.2$, $c^{term}_{dist} = 1.2$, $c^{run}_{trace} = 2.5$, $c^{run}_{col} = 1,000$, with $\lambda = 0.3$, 4096 samples per step, $I_{2\times 2}$ MPPI control noise, and horizon of $H = 8$. Since a binary collision term can degrade plan quality when many samples collide, we additionally included a finer-grained collision cost term penalizing known-obstacle penetration into the calibrated Gaussian confidence ellipse. This enables differentiating plans that are barely unsafe from those whose uncertainty ellipses deeply overlap with obstacles. This improved performance across methods.

D Additional results

D.1 Planar robot

The test-cases used for numerical coverage validation were collected by spanning the Cartesian grid given by $(p_x, p_y) \in \text{LIN}(map_x^{\min}, map_x^{\max}, 8) \times \text{LIN}(map_y^{\min}, map_y^{\max}, 8)$, $(v_x, v_y) \in \text{LIN}(-1.8, 1.8, 4)^2$, and $(a_{x,t}, a_{y,t}) \in \text{LIN}(-0.9, 0.9, 4)^2$. Again, only non-colliding transitions were considered for both calibration dataset collection and test-case evaluation. Table 4 expands Table 1 to include both ablations to **OCULAR**. The ablation with access to test-environment data has a more relevant dataset, which is nevertheless smaller than our general D_{cal}. We observe similar performance to **OCULAR**, which suggests that our perception information encoding enables an efficient calibration, even when using data from environments different from the evaluation environment. The ablation without perception information, Ablation w/o $\text{ENCODE}(o'_t)$, can perform velocity-action-dependent calibration, but cannot distinguish how uncertainty might vary between ID and OOD regions. Consequently, it is significantly over-conservative while ID, which is undesirable. This ablation's OOD coverage does not always meet the user-set requirement.

In Fig. 5 below, we compare the trajectories generated by each algorithm across the four tested environments (we ran 30 evaluations per environment-method combo). The trajectories of *LUCCa* and **OCULAR** are visually similar, even though our method *does not require any data from the tested environment*. This is numerically supported by the metrics shown in Table 5. The ablation with access to test environment data produces results comparable to **OCULAR**. It appears that using data from other environments (i.e., **OCULAR**) might not necessarily lead to a noticeable performance drop, compared to using environment-specific data. This supports the *utility of our approach for cross-environment local conformal calibration*. The ablation without perception information is visibly over-conservative in the ID regions, leading to longer average time-to-goal. No method got stuck due to high uncertainty.

Table 4: Test-cases results across four planar environment maps.

Metric	Method	Tested map **not** in D_{cal}?	Map S ID	Map S OOD	Map L ID	Map L OOD	Map H ID	Map H OOD	Map U ID	Map U OOD
Marginal Coverage (%)	NoCP	N/A	89.9	6.4	90.0	6.4	90.0	6.4	89.9	6.4
	SplitCP	✗	100.0	69.7	100.0	69.1	100.0	62.3	100.0	71.6
	LUCCa [19]	✗	91.4	93.7	91.1	93.8	90.9	93.1	90.9	93.9
	Ablation w/ test map data	✗	91.5	90.8	90.4	93.3	90.5	93.5	92.6	91.3
	Ablation w/o ENCODE(o'_t)	✓	100.0	89.3	100.0	90.0	100.0	90.7	100.0	88.9
	OCULAR (ours)	✓	90.6	93.4	90.8	93.7	91.2	91.0	90.5	93.8
Median $\hat{\mathcal{C}}$ volume (wrt oracle) ↓	NoCP	N/A	0.99	0.02	1.00	0.02	1.00	0.02	0.99	0.02
	SplitCP	✗	50.26	0.84	49.66	0.83	41.02	0.68	53.32	0.89
	LUCCa [19]	✗	1.11	1.15	1.06	1.14	1.09	1.10	1.07	1.16
	Ablation w/ test map data	✗	1.07	1.12	1.00	1.10	1.05	1.11	1.07	1.11
	Ablation w/o ENCODE(o'_t)	✓	57.86	0.97	59.37	0.99	60.56	1.01	58.38	0.98
	OCULAR (ours)	✓	1.00	1.11	1.05	1.15	1.00	1.13	1.05	1.15

red : coverage below $(1-\alpha) = 0.9$. $\hat{\mathcal{C}}$ volume reported as ratio relative to an oracle using the minimum scaling ξ required to achieve 90% coverage per transition. Test transitions #: S 4,096; L 2,816; H 4,096; U 6,656.

Table 5: Planning results across four planar environment maps (30 runs each).

Method	Tested map **not** in D_{cal}?	Success (%) ↑ S	L	H	U	Steps to completion (mean±std) ↓ S	L	H	U
NoCP	N/A	33.3	0	6.7	0	214.8±39.3	–	155.5±0.7	–
SplitCP	✗	0	0	0	0	–	–	–	–
LUCCa [19]	✗	100	100	100	100	171.1±12.1	211.2±6.5	203.2±6.6	283.3±8.1
Ablation w/ test map data	✗	100	100	100	100	170.9±5.3	206.7±6.4	206.0±8.5	280.3±6.8
Ablation w/o ENCODE(o'_t)	✓	100	100	100	100	181.7±4.7	229.1±10.0	205.5±7.1	293.6±8.4
OCULAR (ours)	✓	100	100	100	100	177.6±7.0	213.6±7.5	199.7±7.7	278.1±7.6

Success = reaching all subgoals without collisions.

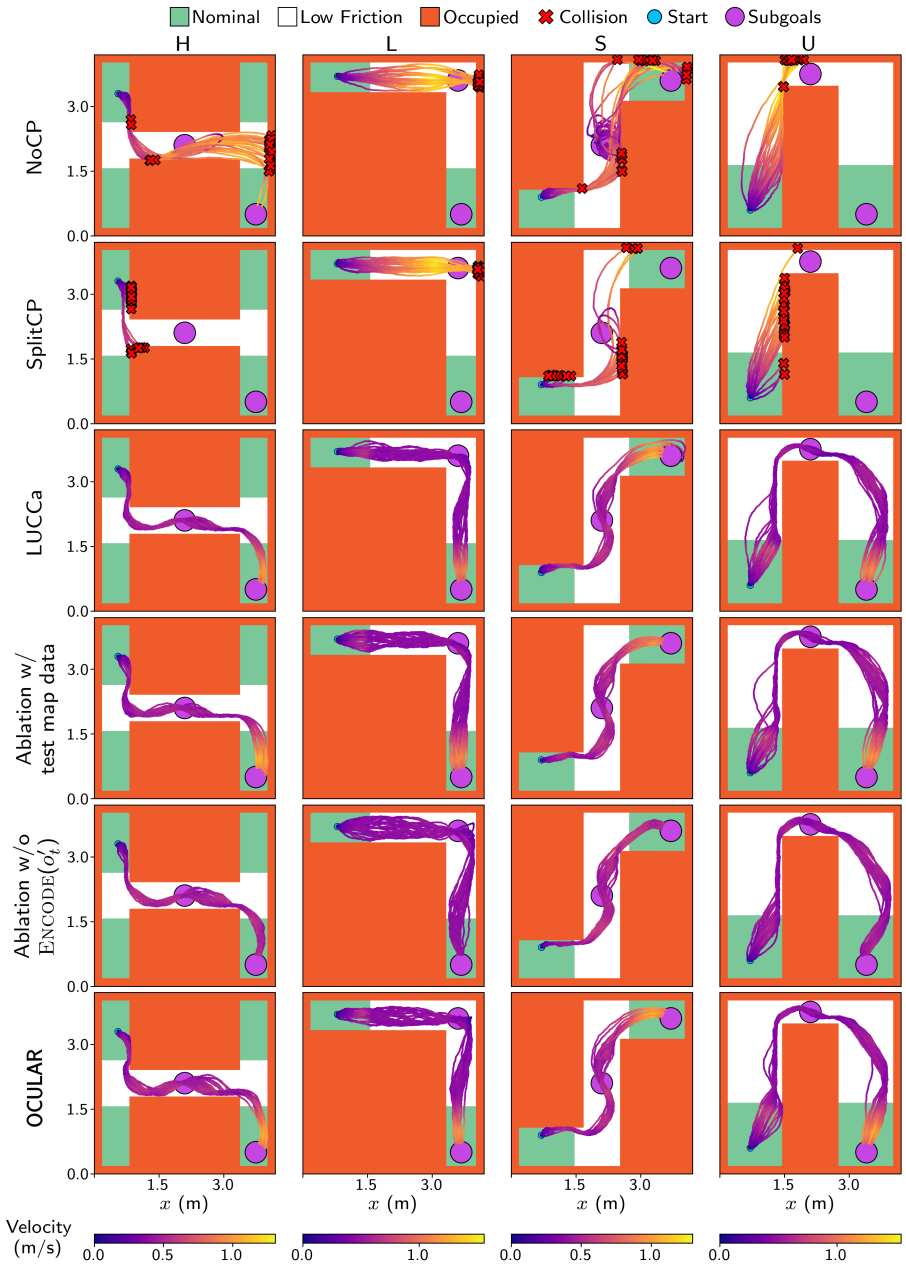

Fig. 5: Trajectories of **OCULAR** and baselines in the planar environment, across 30 runs for each map-method combo. The uncalibrated baseline gains too much momentum in the low-friction region, leading to collisions. *SplitCP* is too conservative, with all its sampled plans being in collision. This results in goal-chasing behavior and an inability to avoid obstacles. *LUCCa* and **OCULAR** have comparable performance, slowing down in high-uncertainty regions, and reaching both subgoals without colliding. However, *LUCCa* uses data specific to each tested map, while *our method produces safe plans without any data from the executed map* (e.g., for map S: *LUCCa* uses data collected in S; our method uses data from H, U, L). The ablation with access to test-time data performs comparatively to **OCULAR**, and the ablation without perception information is slower when ID, moving slower in all settings.

D.2 Isaac Sim

As in the planar experiment, calibration data was collected by spanning the Cartesian grid given by $(p_x, p_y) \in \text{LIN}(map_x^{\min}, env_x^{\max}, 16) \times \text{LIN}(map_y^{\min}, env_y^{\max}, 16)$, $(v_x, v_y) \in \text{LIN}(-1.8, 1.8, 10)^2$, and $(a_{x,t}, a_{y,t}) \in \text{LIN}(-0.9, 0.9, 4)^2$. Additionally, we excluded from the calibration and test data points that were close to the map edges and oriented outwards. This was done because our implementation does not maintain ground truth labels outside the map bounds. See Fig. 6 for the three environments used.

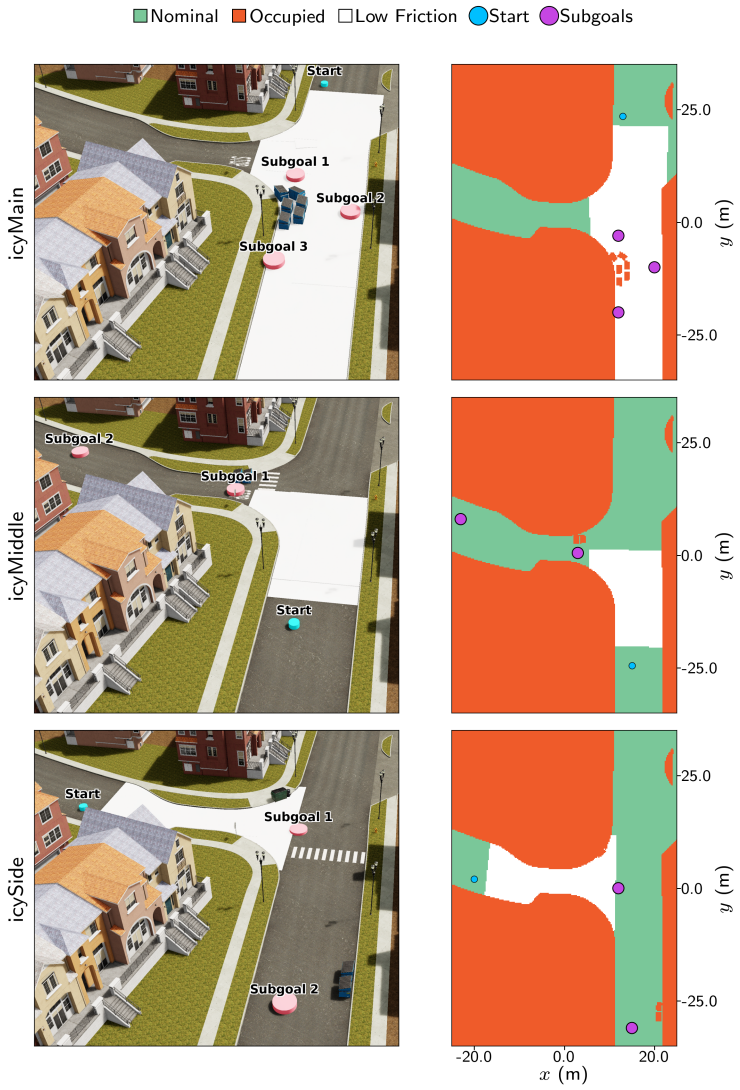

Fig. 6: The three tested Isaac Sim environments (based off Rivermark).

Fig. 7 demonstrates an example observation and the perceived map after a single step (lighter color shades). This map is updated as new observations are received and used for trajectory optimization by all methods.

Table 6 expands on Table 3 by including the ablations. As in the planar setting, the ablation with access to test-time data has comparable performance to **OCULAR**, being sometimes slightly more or less efficient. The ablation without perception information, Ablation w/o ENCODE(o'_t), can considerably undercover when OOD and overcover when ID. This is unsurprising since, without access to observations, it does not have enough information to distinguish both scenarios.

Table 7 reports the results from the safe motion planning experiments. The executed trajectories can be seen in Fig. 8. *NoCP* gains significant momentum leading to collisions. *SplitCP*, depending on the distribution between OOD and ID data in D_{cal}, can become over-conservative and stuck in a narrow passage,

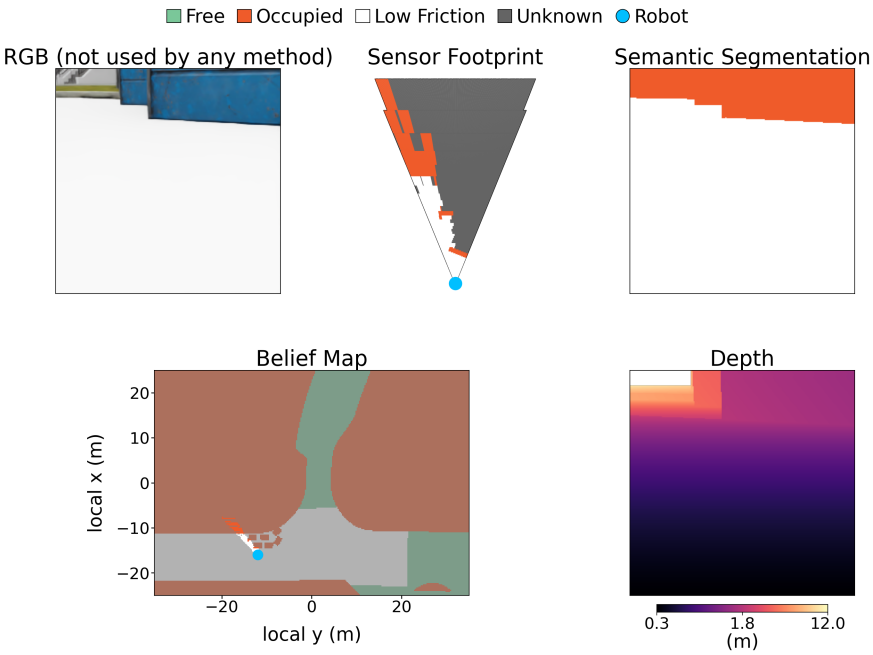

Fig. 7: Example observation and observed map in the Isaac Sim environment. RGB is shown for clarity and is not used by any method. The depth and semantic segmentation image are converted into a planar sensor footprint, as detailed in Sec. 5.2. At the start of each episode, the planner starts from a fully unknown map. The darker shade cells represent the true underlying map semantics, which are unknown to the optimizer. Then, after capturing each observation, the planar semantic map is updated using o'_t as detailed in Sec. 5.4. This map, represented by the lighter shade colors in the bottom left image, is then used for MPC planning at each step. Refer to Sec. 5.4 and the planning videos shown on the project webpage for more details.

or become over-confident and gain unsafe amounts of speed. *SplitCP* was the only method to time-out, i.e., take longer than 700 steps to reach all subgoals. This occurred on all its icySide runs and 20% of its icyMiddle runs, with the remaining runs ending in collision. Both *LUCCa* and **OCULAR** appear to safely navigate all environments, slowing down as necessary to maintain the next-state uncertainty manageable. The ablation with only access to test-time data performs similarly to our approach, being slightly faster or slower in some maps. This indicates that using cross-environment data might not lead to noticeable performance degradation comparatively to using environment-specific data. The ablation without perception information is overly conservative when OOD, as it cannot distinguish between nominal and low-friction regions. In icyMain, this ablation gains too much momentum leading to collisions. This indicates that access to perception information is important, as it could help determine how state-action can vary across different environmental regions.

The test-cases and planning results appear to indicate that our method is capable of producing *volume-efficient* and *adaptive* dynamics uncertainty calibrations, and consequently *safe* motion plans when using more realistic perception information, even *without access to any data from the execution environment.*

Table 6: Test-cases results across three Isaac Sim roads, including ablations.

Metric	Method	Tested map **not** in D_{cal}?	icySide ID	icySide OOD	icyMain ID	icyMain OOD	icyMiddle ID	icyMiddle OOD
Marginal Coverage (%)	NoCP	N/A	90.0	56.7	90.0	56.7	90.0	56.7
	SplitCP	✗	99.5	89.6	99.8	93.0	99.1	85.9
	LUCCa [19]	✗	91.1	91.5	90.1	91.4	90.1	90.9
	Ablation w/ test map data	✗	91.3	91.1	92.2	91.3	90.7	90.2
	Ablation w/o ENCODE(o'_t)	✓	97.2	80.0	96.0	75.6	97.4	82.1
	OCULAR (ours)	✓	91.5	90.1	90.4	91.0	91.1	90.6
Median $\hat{\mathcal{C}}$ volume (wrt oracle) ↓	NoCP	N/A	1.00	0.28	1.00	0.28	1.00	0.28
	SplitCP	✗	3.73	1.03	4.66	1.29	3.07	0.85
	LUCCa [19]	✗	1.08	1.13	1.02	1.10	1.02	1.13
	Ablation w/ test map data	✗	1.08	1.11	1.15	1.12	1.06	1.04
	Ablation w/o ENCODE(o'_t)	✓	2.28	0.65	1.88	0.54	2.39	0.71
	OCULAR (ours)	✓	1.03	1.02	1.02	1.15	1.06	1.06

red : coverage < 0.9. Volume reported as ratio relative to an oracle using the minimum ξ to achieve 90% coverage. Test transition #: icySide 4,464; icyMain 4,464; icyMiddle 4,464.

Table 7: Planning results across three Isaac Sim roads (30 runs each), including ablations.

Method	Tested map **not** in D_{cal}?	Success (%) ↑ icySide	icyMain	icyMiddle	Steps to completion (mean±std) ↓ icySide	icyMain	icyMiddle
NoCP	N/A	0	0	0	–	–	–
SplitCP	✗	0	0	0	–	–	–
LUCCa [19]	✗	100	100	100	339.1±8.7	332.1±13.1	288.4±7.8
Ablation w/ test map data	✗	100	100	100	211.1±5.9	280.0±6.1	305.9±28.5
Ablation w/o ENCODE(o'_t)	✓	100	0	100	282.3±4.9	–	317.1±7.4
OCULAR (ours)	✓	100	100	100	208.1±3.1	311.3±4.7	278.6±6.4

Success = reaching all subgoals without collisions.

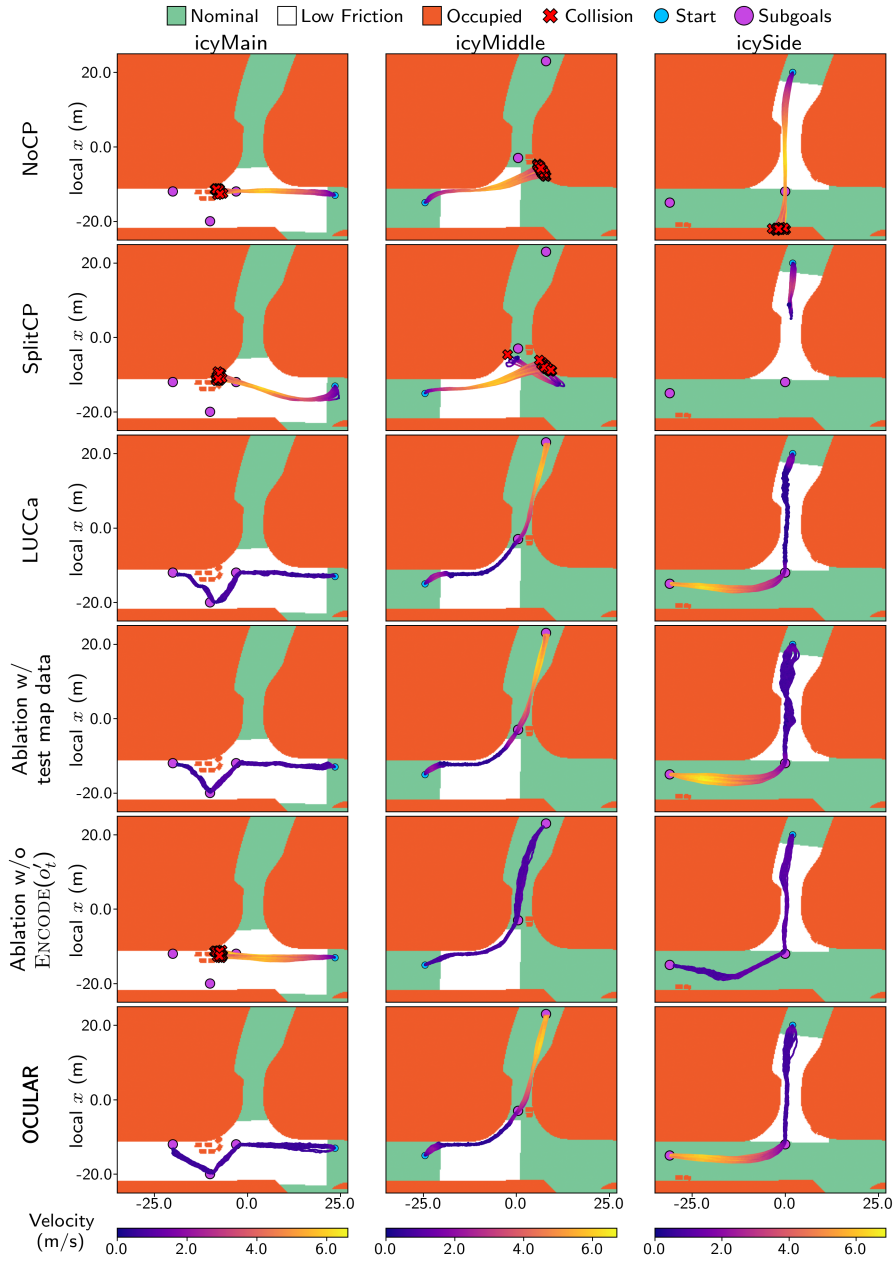

Fig. 8: Comparison of all methods in the Isaac environment. *NoCP* gains significant momentum when OOD, leading to collisions. *SplitCP* can be over-conservative leading to time-outs in tight regions (e.g., icySide) or overconfident over ice (e.g., icyMain), since it cannot distinguish between observations-velocity-actions leading to lower or higher next-state uncertainty. *LUCCa* and **OCULAR** have comparable performance, slowing down in high-uncertainty regions, and reaching subgoals safely. Yet, *LUCCa* uses data specific to each tested environment, while our method produces safe and efficient plans *without any data from the executed environment* (e.g., for map icyMain: *LUCCa* uses data collected in icyMain; our method uses data collected from icyMiddle and icySide). The ablation with access to test-environment data performs comparatively with our method. The ablation without perception information can become overly conservative over nominal terrain (icyMiddle, icySide) or overly optimistic when OOD (icyMain), as it cannot adapt to how the same actions can result in different amounts of uncertainty across regions.