THALES
Building a future we can all trust

NEURAL INFORMATION PROCESSING SYSTEMS

PASSAGE

A Real-Terrain Benchmark for Trustworthy Constrained Path Planning

Geoffrey Delhomme
François Trap
Fateh Kaakai

www.thalesgroup.com



Learned Planners Need Trust Evidence

NEURAL INFORMATION PROCESSING SYSTEMS

MOTIVATION Embedding a planner in a safety-critical system, such as a helicopter or an aircraft, needs more than speed

THE PASSAGE CHALLENGE

As good as A*, and faster? Prove it.

From rescue helicopters to rovers, drones and robots, any machine that plans a route over real terrain needs evidence it can be trusted.



> What trust evidence is needed

Data
complete, correct and representative, within a declared scope

Route quality
against an exact reference, under the same objective

Timing
of the complete request, on target hardware

SAFETY-CRITICAL CONTEXT | Emerging ML assurance guidance points the same way: draft SAE ARP6983 / EUROCAE ED-324, whose recognition EASA proposes in NPA 2025-07, and the FAA Roadmap for AI Safety Assurance (2024).

Motivation: rescue-helicopter terrain routing. The benchmark itself is static 2D grid routing. **Research benchmark, not a certification claim.**

> PASSAGE provides this evidence for learned planners, as a research benchmark.

This document may not be reproduced, modified, adapted, published, translated, in any way, in whole or in part or disclosed to a third party without the prior written consent of Thales © 2026 THALES. All rights reserved.

From Global Terrain to Six Reference Routes

NEURAL INFORMATION PROCESSING SYSTEMS

1 · DATASET Trust need: data complete, correct and representative, within a declared scope

A REAL-TERRAIN PROVENANCE



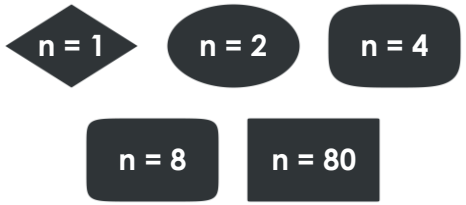
AW3D30 tile coverage

23,993 1° × 1° tiles

~30 m sampling

Original data supplied by JAXA: ALOS World 3D–30m (AW3D30).

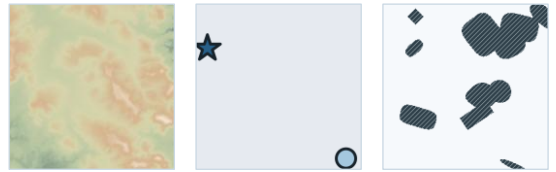
B ONE GEOMETRY FAMILY



Lamé super-ellipses

C THREE-CHANNEL INPUT

Elevation | Markers | Obstacles



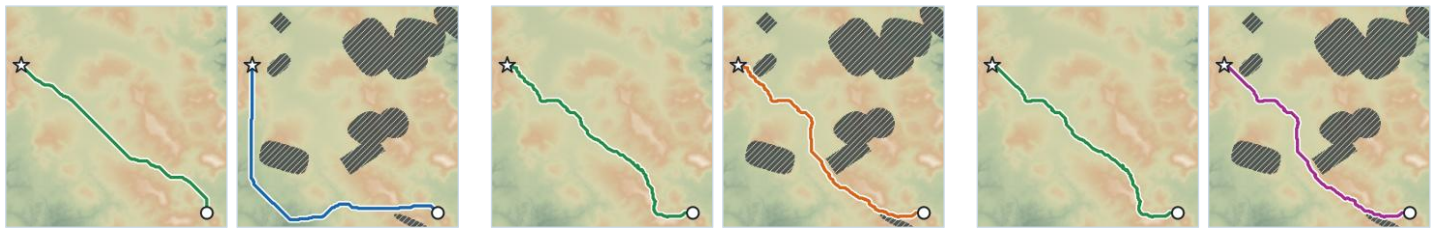
$X : [N, N, 3]$

D SIX STORED A* LABELS

COST MODEL

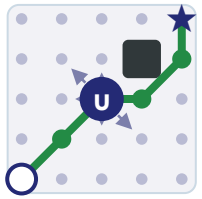
ELEVATION | UPHILL SLOPE | ENERGY PROXY

FREE | OBSTACLES | FREE | OBSTACLES | FREE | OBSTACLES



Each scene stores six exact A* routes: three cost models, with and without obstacles.

E OUR GRID-NATIVE A* REFERENCE



Off-the-shelf graph-based A* was not efficient enough for exact 4096² labels, so we built a grid-native, Numba-compiled A* on an implicit 8-neighbour grid.

40 % less time

than NetworkX A*, the fastest off-the-shelf solver tested

(paper Table 4: one 256² scene, NetworkX graph prebuilt; 27.3 vs 45.6 ms)

> **One real scene, synthetic obstacles, six exact A* labels: 4.4 M scenes and 26.4 M reference routes.**

This document may not be reproduced, modified, adapted, published, translated, in any way, in whole or in part or disclosed to a third party without the prior written consent of Thales © 2026 THALES. All rights reserved.

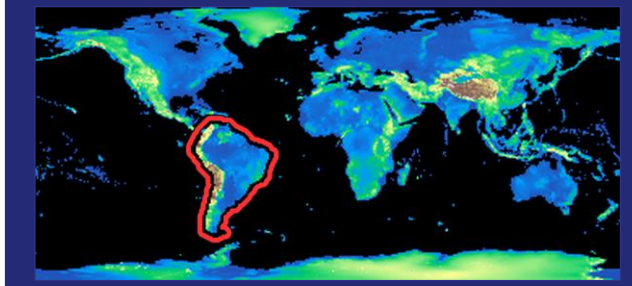
Release, Splits and Trustworthiness by Design

NEURAL INFORMATION PROCESSING SYSTEMS

1 · DATASET Completeness, correctness, representativeness, turned into release checks

F RELEASE IN NUMBERS

- **4.4 M** real-terrain scenes
- **26.4 M** exact A* routes
- **7** grid sizes, $64^2 \rightarrow 4096^2$
- **≈ 1.27 TB** compressed Parquet



DATASET SPLITS

TRAIN	VALID	CALIB	TEST
90 %	4 %	2 %	4 %
fit weights	choose model and loss	fix decoder and fallback settings	score every request; South-American hold-out

G TRUSTWORTHINESS BY DESIGN

12

Twelve data-quality attributes, each with a release check

Attributes inspired by the draft SAE ARP6983 / EUROCAE ED-324

PROVENANCE	COVERAGE	CORRECTNESS	INDEPENDENCE AND FAIRNESS
Source suitability	Completeness	Accuracy	TEST independence
Schema	Representativeness	Integrity	Fairness
Traceability	Sufficiency		Bias
Temporality			

> **4.4 M scenes with a geographic hold-out, and twelve data-quality attributes, each with a release check.**

This document may not be reproduced, modified, adapted, published, translated, in any way, in whole or in part or disclosed to a third party without the prior written consent of Thales © 2026 THALES. All rights reserved.

One Contract: Score the Route, Time the Request

NEURAL INFORMATION PROCESSING SYSTEMS

2 · EVALUATION CONTRACT Trust need: exact-reference route quality; full-request timing on target hardware

A EXECUTION HARNESS · ONE CONTRACT FOR EVERY PLANNER

artifact · edge target · host

```
plan(request: PlanRequest) → PlanResult
```



References never enter the request: vision, graph-neural, RL or classical planners all fit.

B METRICS: SCORE THE EMITTED ROUTE

computed on the final s → g route

COST REGRET ↓ median, p90

$$\text{regret}_i = \frac{c_i^{\text{pred}} - c_i^*}{c_i^*}$$

a failure counts as $K_i = 10^6\, c_i^*$

ASSD ↓ normalised by $|P^*|$

$$\text{path_assd}_i = \frac{\text{ASSD}(\hat{P}_i, P_i^*)}{|P_i^*|}$$

mean nearest-path distance, both ways

H95 ↓ pixels

$$\text{H95}_i = \max\left(q_{95}(\hat{P}_i \to P_i^*), q_{95}(P_i^* \to \hat{P}_i)\right)$$

larger directional 95th-percentile distance

SUCCESS ↑ B_i: declared budget

$$\text{succ}_i = \mathbb{1}\left[\hat{P}_i \text{ legal}, c_i^{\text{pred}} < \infty, T_i \le B_i\right]$$

a late legal path is a timeout failure

OBSTACLE VIOLATION ↓

$$\text{obstv}_i = \mathbb{1}\{\exists (r,c) \in \hat{P}_i : O_i(r,c) = 1\}$$

rates: means over all requests

LATENCY T_{e2e} p50, p95, p99

raw in-memory request arrays → emitted path, beside the same-target CPU A*

ASSD: average symmetric surface distance · H95: 95th-percentile symmetric Hausdorff distance

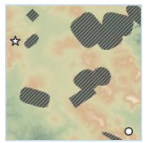
> **Every metric is computed on the final route, and the complete request is timed.**

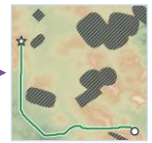
This document may not be reproduced, modified, adapted, published, translated, in any way, in whole or in part or disclosed to a third party without the prior written consent of Thales © 2026 THALES. All rights reserved.

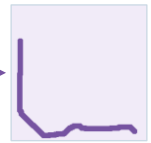
Our Candidates: SEG and REG

NEURAL INFORMATION PROCESSING SYSTEMS

3 · CANDIDATES Trust need: explicit supervision, frozen decoders, declared settings


Input X


A* route P*


SEG mask Y*$_s$

SEG LOSS

$$\mathcal{L}_s = \mathcal{L}_{\mathrm{BCE}}(z, Y_s^*) + 0.006\,\mathcal{L}_{\mathrm{bottleneck}} + 0.01\,\mathcal{L}_{\mathrm{structured}}$$

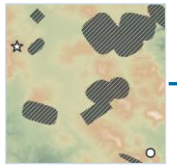
BCE on the route mask + bottleneck and route-ranking terms

B SEG TRAINING

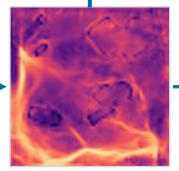
A INFERENCE: TWO FROZEN PIPELINES

path logit z

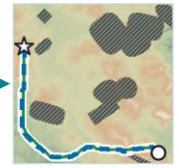
Both decoders receive only the field, O, s and g.


Input X

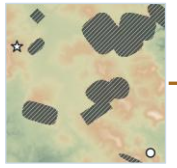
ConvNeXt V2 Atto + UPerNet
frozen · GPU



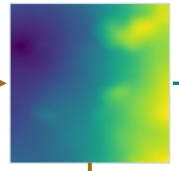
Weighted A* · CPU
on a cost built from the logit
(w, η) = (0.25, 1.5), selected on CALIB



SEG route: 356 cells
+2.5 %
elevation cost vs A*, this scene


Input X

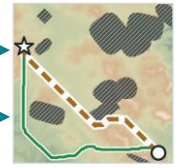
RepViT M1.0 + UPerNet
frozen · GPU
start ○ · goal ★


cost-to-go $\hat{V}$

Bellman descent · CPU
downhill steps on the predicted cost-to-go
κ = 8, selected on CALIB

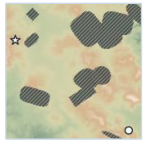
goal reached (this scene)

if descent stalls:
goal-rooted Dijkstra reconnection
not needed on this scene

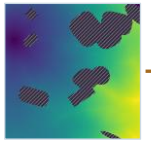


REG route: 229 cells
+16.2 %
elevation cost vs A*, this scene

A* reference (green): 352 cells


Input X

same graph + goal **reverse Dijkstra**


REG cost-to-go Y*$_r$

REG LOSS

$$\mathcal{L}_r = \mathcal{L}_{\mathrm{rec}}(\hat{V}, Y_r^*) + 0.5\,\mathcal{L}_B + 0.1\,\mathcal{L}_{\mathrm{policy}} + \lambda_{\mathrm{opt}}\,\mathcal{L}_{\mathrm{opt}}, \lambda_{\mathrm{opt}} \simeq 0.00747$$

reachable-cell MSE + Bellman consistency, policy and optimal-set terms

B' REG TRAINING

> **Weights learned on TRAIN, decoder settings selected on CALIB then frozen for TEST.**

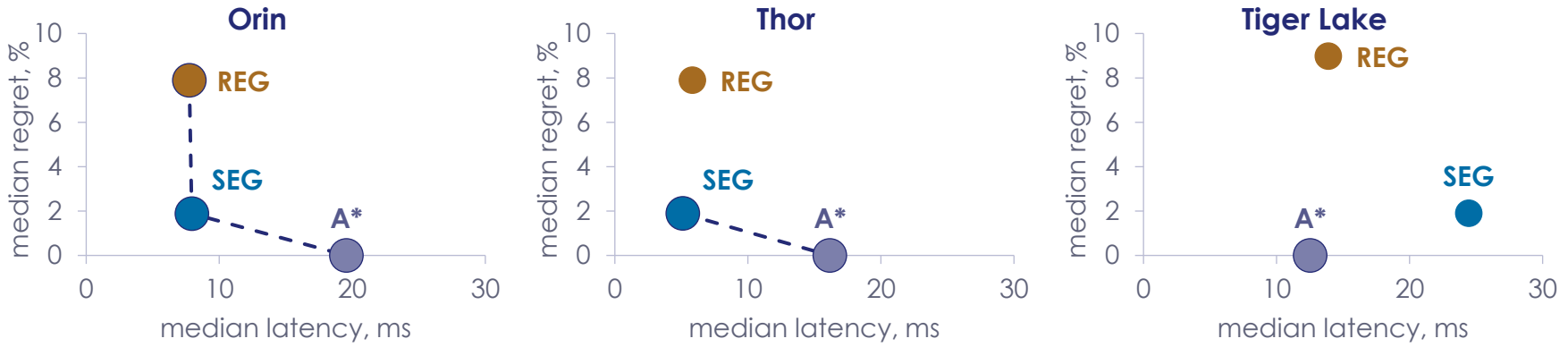
This document may not be reproduced, modified, adapted, published, translated, in any way, in whole or in part or disclosed to a third party without the prior written consent of Thales © 2026 THALES. All rights reserved.

256^2 Results: All Metrics and the Pareto Front

NEURAL INFORMATION PROCESSING SYSTEMS

A **4 · RESULTS** Paper Table 3 · TEST 256^2, elevation objective (w = 62.4), 40,000 scenes per session

Target	Planner	Regret med. % ↓	Regret p90 % ↓	ASSD ↓	H95 px ↓	Succ. ↑	Obst. viol. ↓	p50 ms ↓	p95 ms ↓	p99 ms ↓
Orin	CPU A*	—	—	—	—	1.000	0.000	19.553	42.473	52.027
	SEG	**1.888**	**7.123**	**0.03265**	**12.561**	1.000	0.000	7.940	**13.191**	**17.491**
	REG	7.896	23.422	0.06597	20.810	1.000	0.000	**7.747**	16.228	23.824
Thor	CPU A*	—	—	—	—	1.000	0.000	16.153	35.448	44.084
	SEG	**1.889**	**7.127**	**0.03266**	**12.562**	1.000	0.000	**5.102**	**8.921**	**12.865**
	REG	7.897	23.432	0.06599	20.822	1.000	0.000	5.834	14.374	20.721
Tiger Lake	CPU A*	—	—	—	—	1.000	0.000	**12.503**	27.061	30.843
	SEG	**1.895**	**7.140**	**0.03274**	**12.583**	1.000	0.000	24.442	26.211	27.843
	REG	8.978	25.101	0.06716	21.478	1.000	0.000	13.894	**19.044**	**21.095**



PARETO FRONT

Axes: route quality (median regret) vs speed (median latency), the trade-off PASSAGE measures; lower is better on both.

Orin: REG, SEG, CPU A*

Thor: SEG, CPU A*

Tiger Lake: CPU A*

● **CPU A*** ● **SEG** ● **REG**

Dashed: computed from Table 3

> **Every request succeeds; SEG pays 1.9 % median extra cost; on Orin and Thor, SEG and REG cut median latency by 59–68 % vs CPU A* (Tiger Lake: CPU A* fastest).**

This document may not be reproduced, modified, adapted, published, translated, in any way, in whole or in part or disclosed to a third party without the prior written consent of Thales © 2026 THALES. All rights reserved.

Scaling to 4096 × 4096: Timing the Complete Request

NEURAL INFORMATION PROCESSING SYSTEMS

B **4 · RESULTS** Evaluation on edge GPU + CPU for the candidates and edge CPU for the A* reference

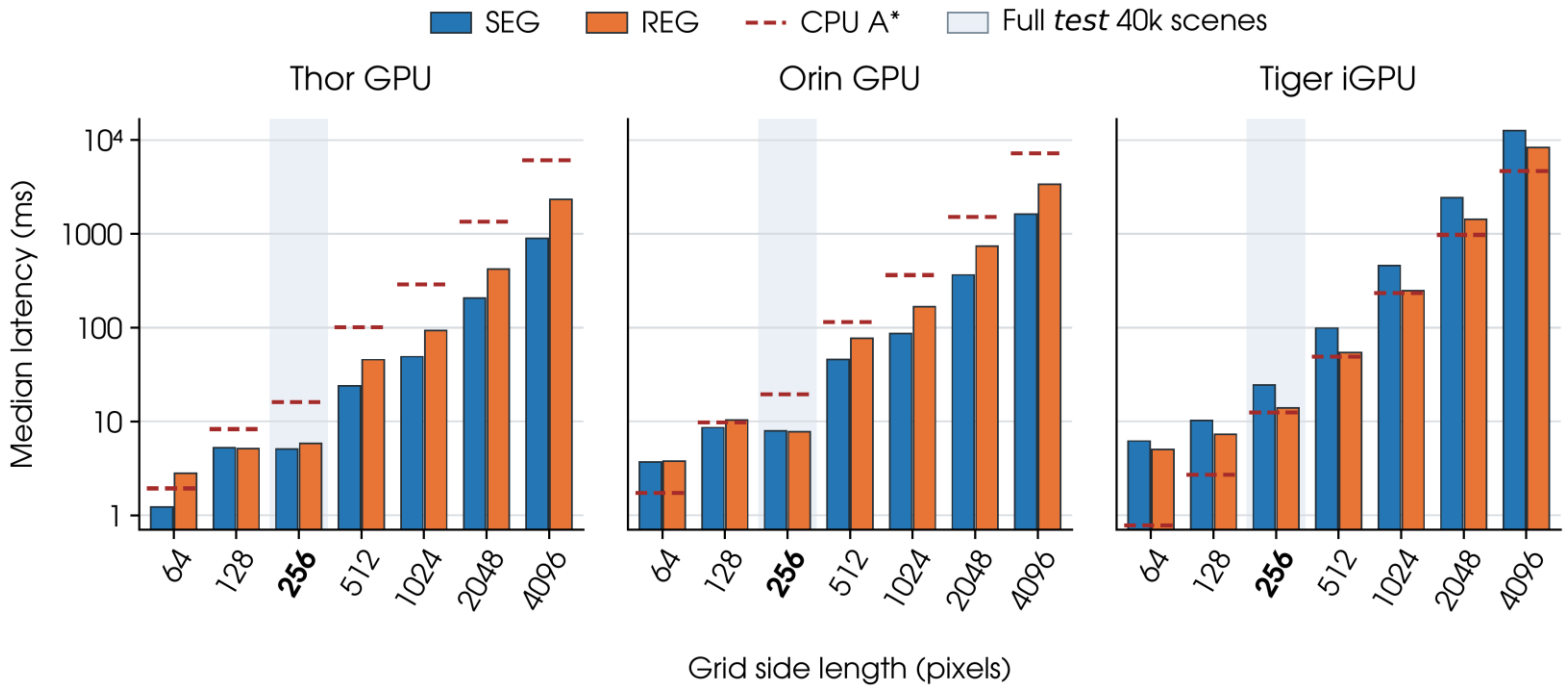
COMPUTE ROUTE

```
plan(request: PlanRequest)
→ PlanResult
```

PATH REQUEST (IN MEMORY) → **INSIDE THE TIMER · incl. transfers and sync**: ENCODE / PREPROCESS → NN / SEARCH → DECODE / POSTPROCESS → PATH RETURN (IN MEMORY)

OUTSIDE: runtime initialisation · input delivery · logging · host validation and scoring

Persistent runtime · 16 warm-up calls · batch size 1



Orin and Thor: both neural medians below CPU A* from 256² upwards.

Tiger Lake: CPU A* has the lowest median at every measured grid size.

HARDWARE (PAPER TABLE 13)

Orin Jetson AGX Orin: Ampere GPU, Arm CPU

Thor Jetson AGX Thor: Blackwell GPU, Arm CPU

Tiger Lake Core i7-1185GRE + Iris Xe iGPU

> **The best trade-off depends on the hardware: time the complete request on the target device.**

This document may not be reproduced, modified, adapted, published, translated, in any way, in whole or in part or disclosed to a third party without the prior written consent of Thales © 2026 THALES. All rights reserved.

Conclusion, Future Work and the Challenge

NEURAL INFORMATION PROCESSING SYSTEMS

C 4 · RESULTS — TAKE-AWAYS Controlled data, exact references, one contract

As good as A*, and faster? Prove it.



> Take-aways

- **4.4 M real-terrain scenes**, exact A* labels, 7 grid sizes
- **Twelve data-quality attributes**, each with a release check
- **One contract**: final route **scored**, full request **timed**
- **SEG**: lower regret than REG; both **below** CPU A* median latency on **Orin and Thor**
- **Best trade-off** depends on the **hardware**

> Future work

- **Graph-neural, RL and hybrid planners**, same contract
- A separate **hidden-test ranking**
- **Irregular obstacles**, hazard layers, other objectives and weights
- **Kinodynamics**, moving obstacles, multi-agent tasks
- Adapt to the **final SAE ARP6983 / EUROCAE ED-324**
- Extend to **robotics and autonomous navigation**

TAKE THE CHALLENGE

huggingface.co/datasets/thalesgroup/passage

DATASET + LEADERBOARD

github.com/ThalesGroup/passage

CODE

Research benchmark, not a certification claim. Terrain: JAXA/EORC ALOS World 3D 30 m (AW3D30), via PASSAGE.

> **> Learned planners can be fast; PASSAGE measures what that speed costs.**

This document may not be reproduced, modified, adapted, published, translated, in any way, in whole or in part or disclosed to a third party without the prior written consent of Thales © 2026 THALES. All rights reserved.