Assessing Scene Generation Techniques for Testing COLREGS-Compliance of Autonomous Surface Vehicles

An Experience Report

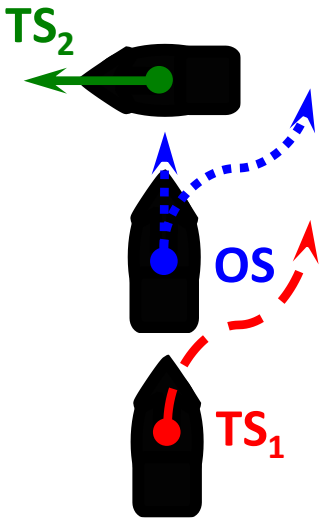
Dominik Frey, Ulf Kargén and Daniel Varró

LINKÖPING UNIVERSITY

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Motivation

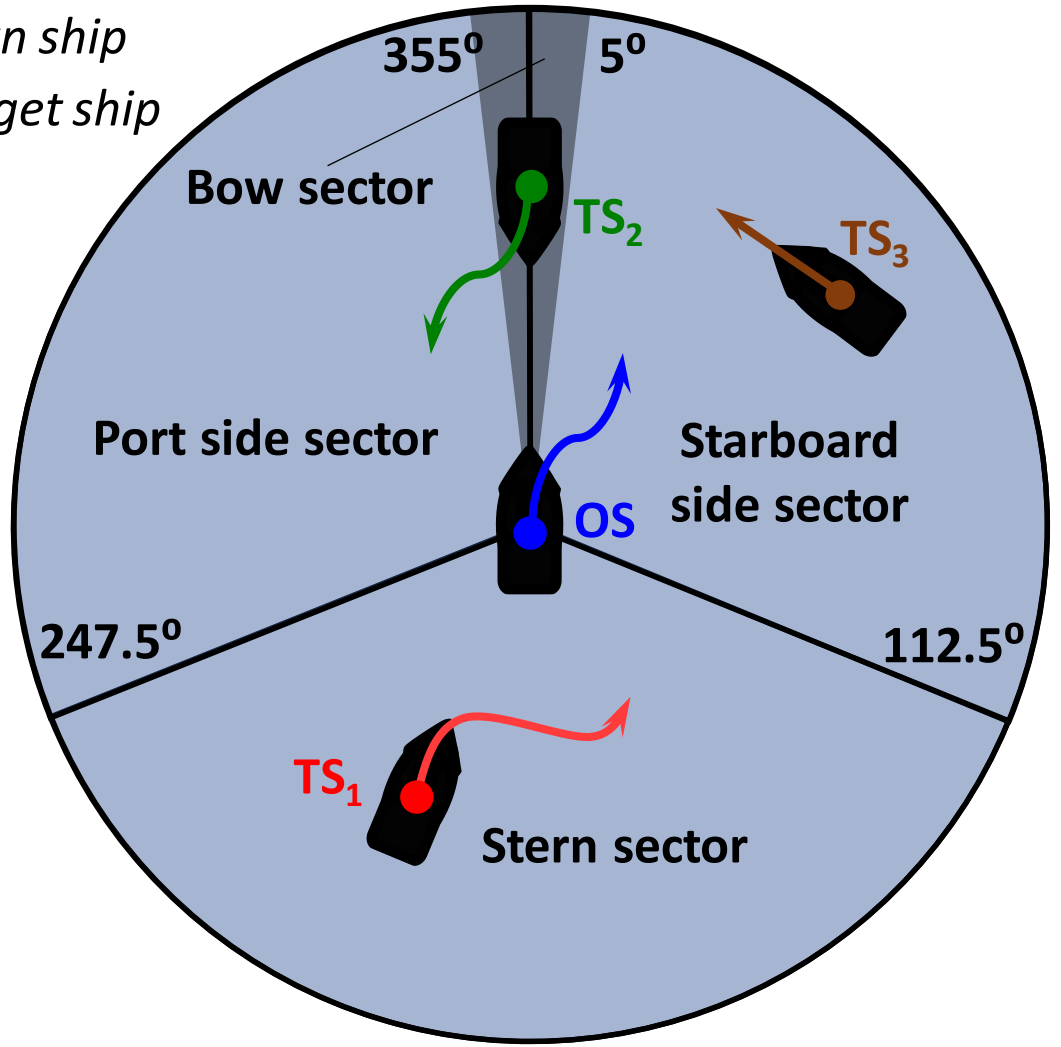
- **Autonomous surface vehicles** (ASVs) increasingly important for both civilian and military applications
  - need to complete their mission autonomously in maritime traffic
- **International Regulations for Preventing Collisions at Sea (COLREGS)** (by the International Maritime Organization)
- COLREGS compliance is critical for the safe operation of ASVs
- But COLREGS are …
  - **underspecified**, formulated with **human operators in mind**



COLREGS situations

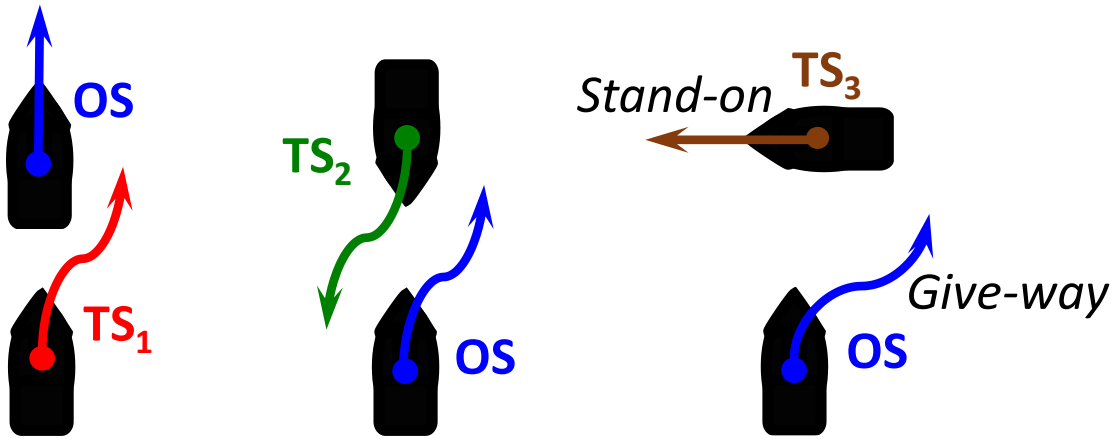
OS : *own ship*

TS : *target ship*



COLREGS apply when

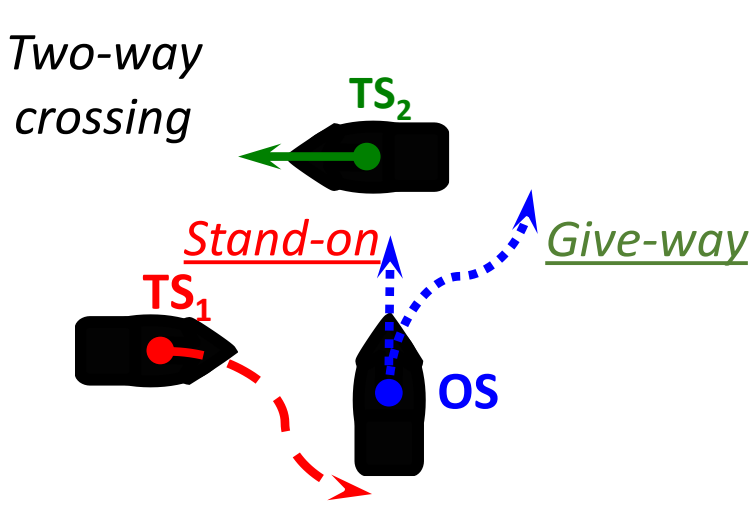
1. Two ships are within visibility distance,
2. on a collision course
3. given one of the following relative bearings:

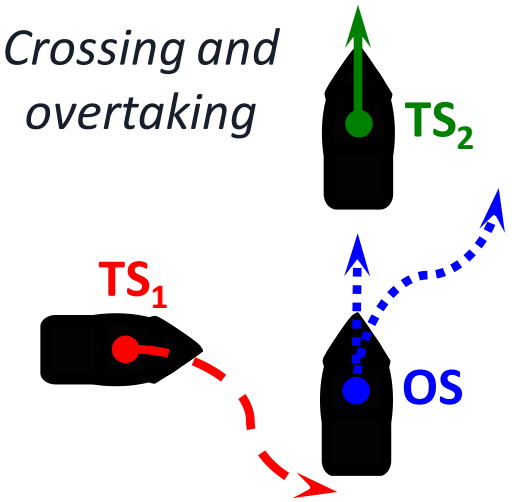


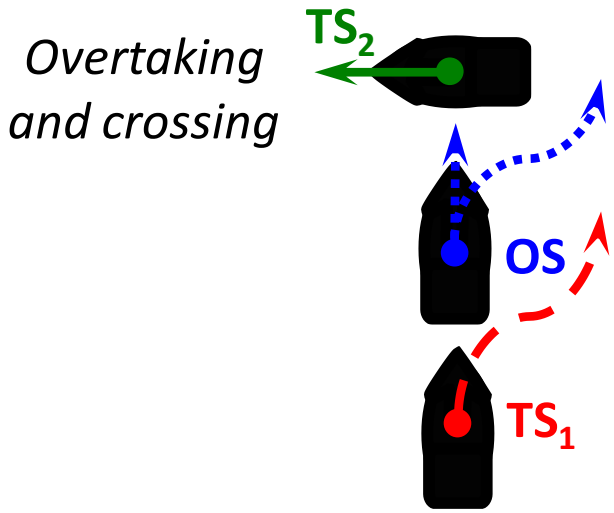
Overtaking **Head-on** **Crossing from port**

Multi-vessel COLREGS scenarios

- **COLREGS scenario:** set of COLREGS situations
- **Ambigious scenario:** give-way + stand-on

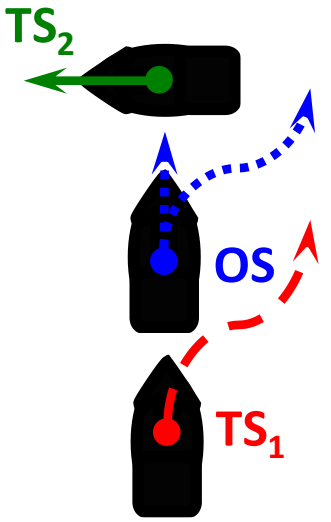






Motivation

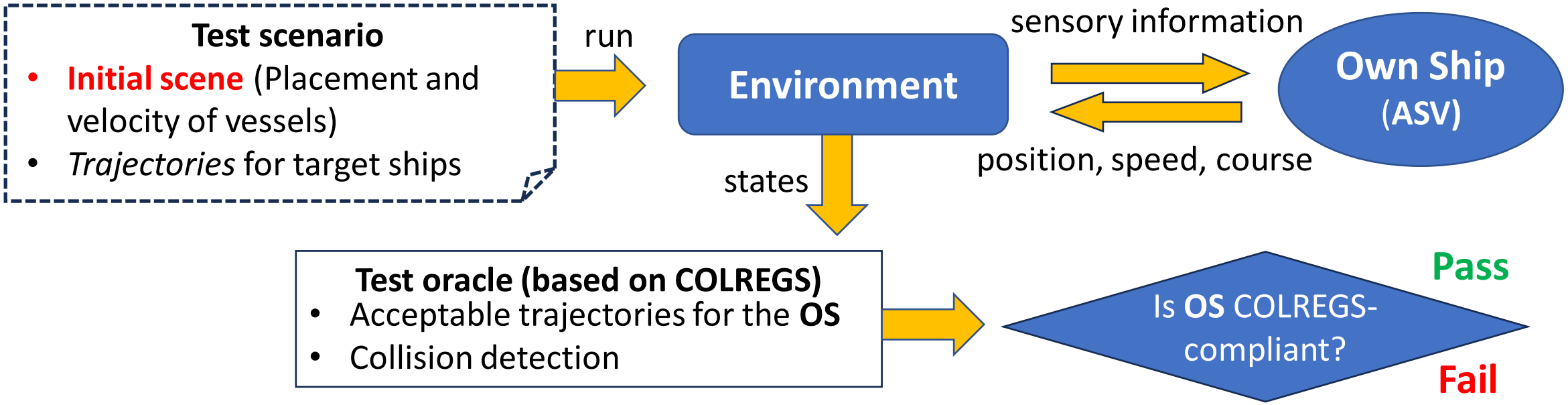
- **Autonomous surface vehicles** (ASVs) increasingly important for both civilian and military applications
  - need to complete their mission autonomously in maritime traffic
- **International Regulations for Preventing Collisions at Sea (COLREGS)** (by the International Maritime Organization)
- COLREGS compliance is critical for the safe operation of ASVs
- But COLREGS are …
  - **underspecified**, formulated with **human operators in mind**
  - **ambiguous** in case of multi-ship encounters



How to ensure safe behavior even in rare *critical edge cases*?

Testing for COLREGS compliance

- Formal methods verify safety at the component level (e.g., model checking)
- System-level safety assurance still relies on **scenario-based testing**



Objectives

Challenge
- Different dynamics
- Open sea geometry
- Diverse actors

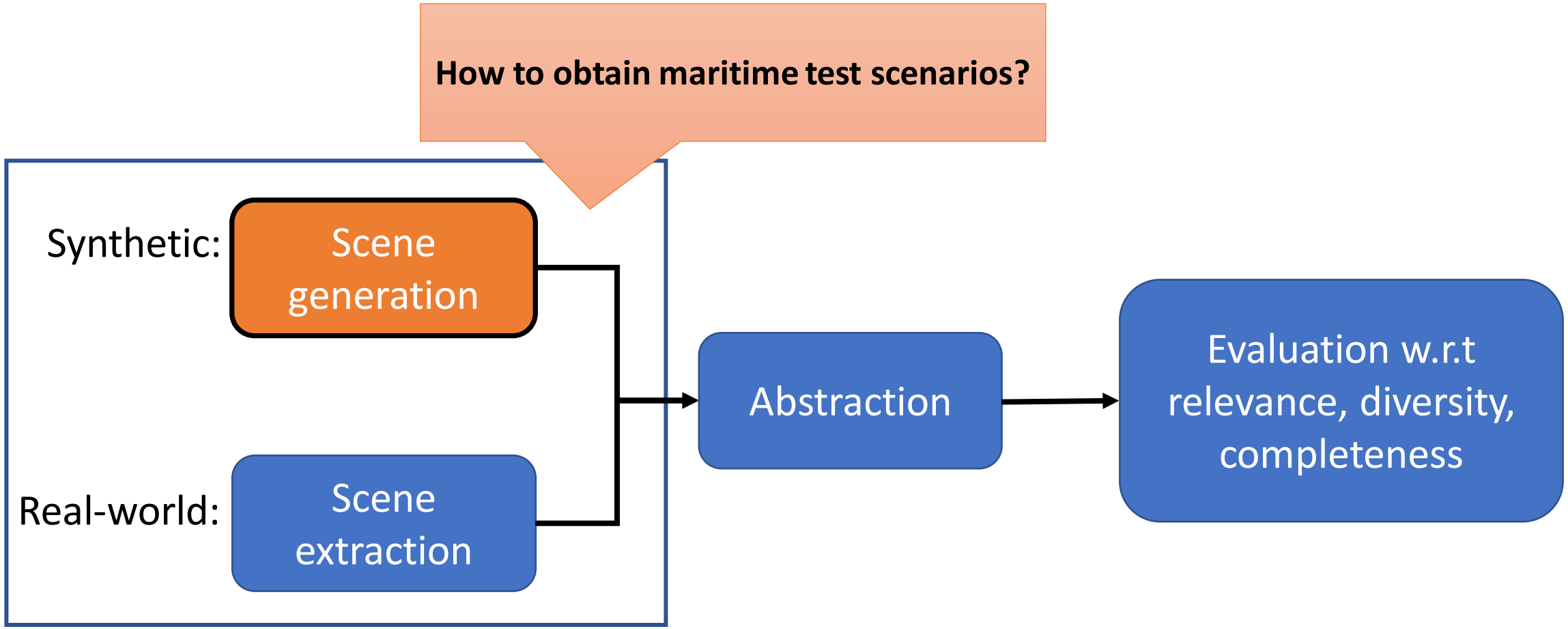
- Adapt two system-level test scenario generation methods used in automotive
- Evaluate their effectiveness in the maritime domain
- Using synthetic and real-world maritime traffic data (AIS)

How to obtain maritime test scenarios?

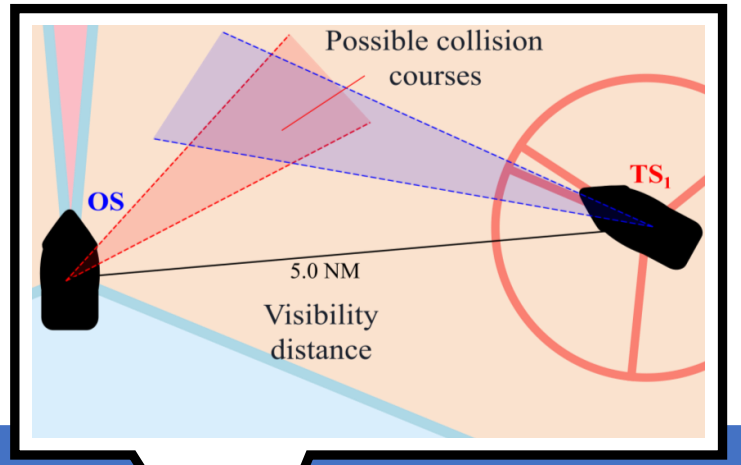
What is an effective test suite?

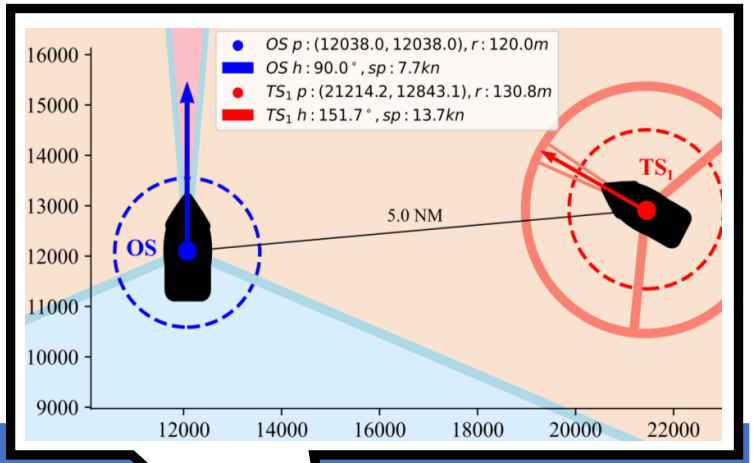
Dominik Frey, Ulf Kargén, Dániel Varró. 2025. Assessing Scene Generation Techniques for Testing COLREGS-Compliance of Autonomous Surface Vehicles. ©

Overview



Scene generation





Geometric constraints (constraint satisfaction problem)

1. ∀ **TS$_i$** is on a collision course with **OS** within visibility rage

2. ∀ (**TS$_i$** and **TS$_j$**) pair within visibility range are not on a collision course.

Constraint satisfaction

Search-based (**SB**)

Rejection sampling-based (**RS**)

Test suite

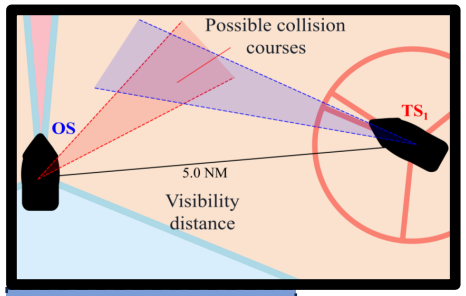
Initial scene$_1$

Initial scene$_2$

⋮

Initial scene$_n$

Search-based constraint satisfaction (**SB**)



CSP model

1 2

Mapping

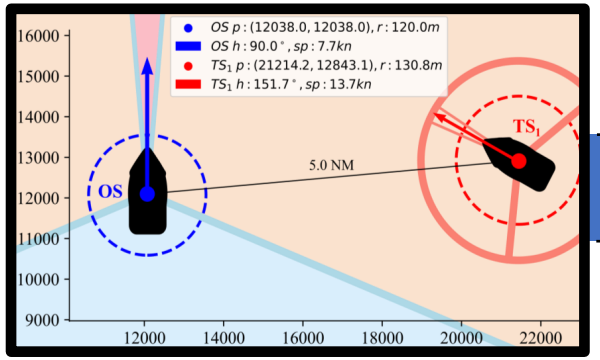
Multi-objective optimization problem

Variables: position, heading, speed, and length

Objective function
geometric distance from satisfying constraints 1 2

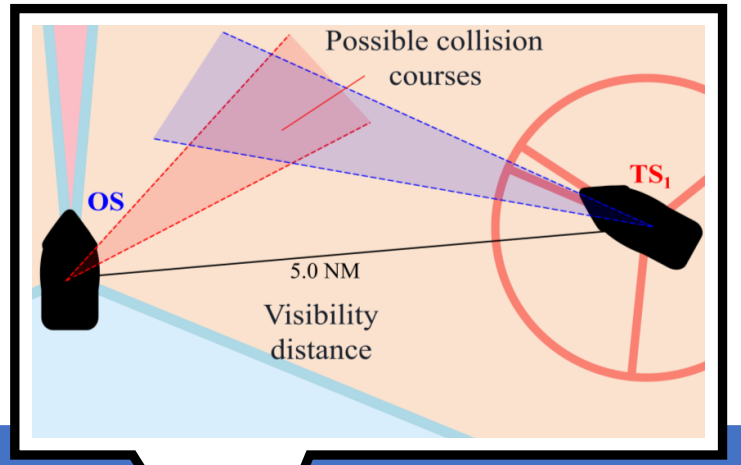
Genetic search

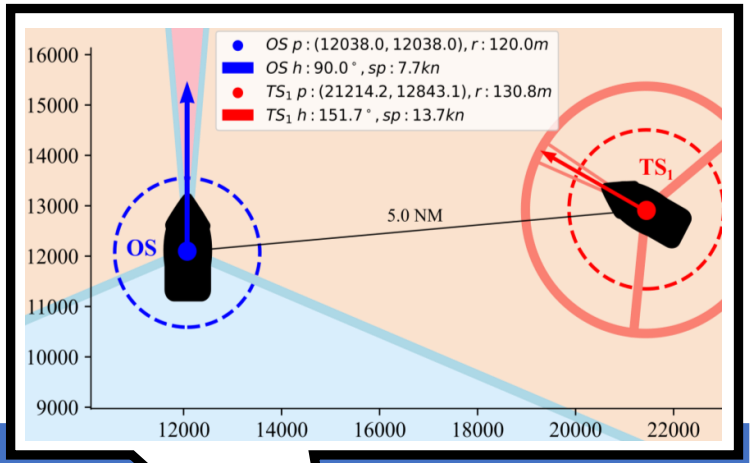
Hyperparameter tuning → **NSGA II**



Initial scene

Scene generation





Geometric constraints (constraint satisfaction problem)

1. ∀ **TS$_i$** is on a collision course with **OS** within visibility rage

2. ∀ (**TS$_i$** and **TS$_j$**) pair within visibility range are not on a collision course.

Constraint satisfaction

Search-based (**SB**)

Rejection sampling-based (**RS**)

Test suite

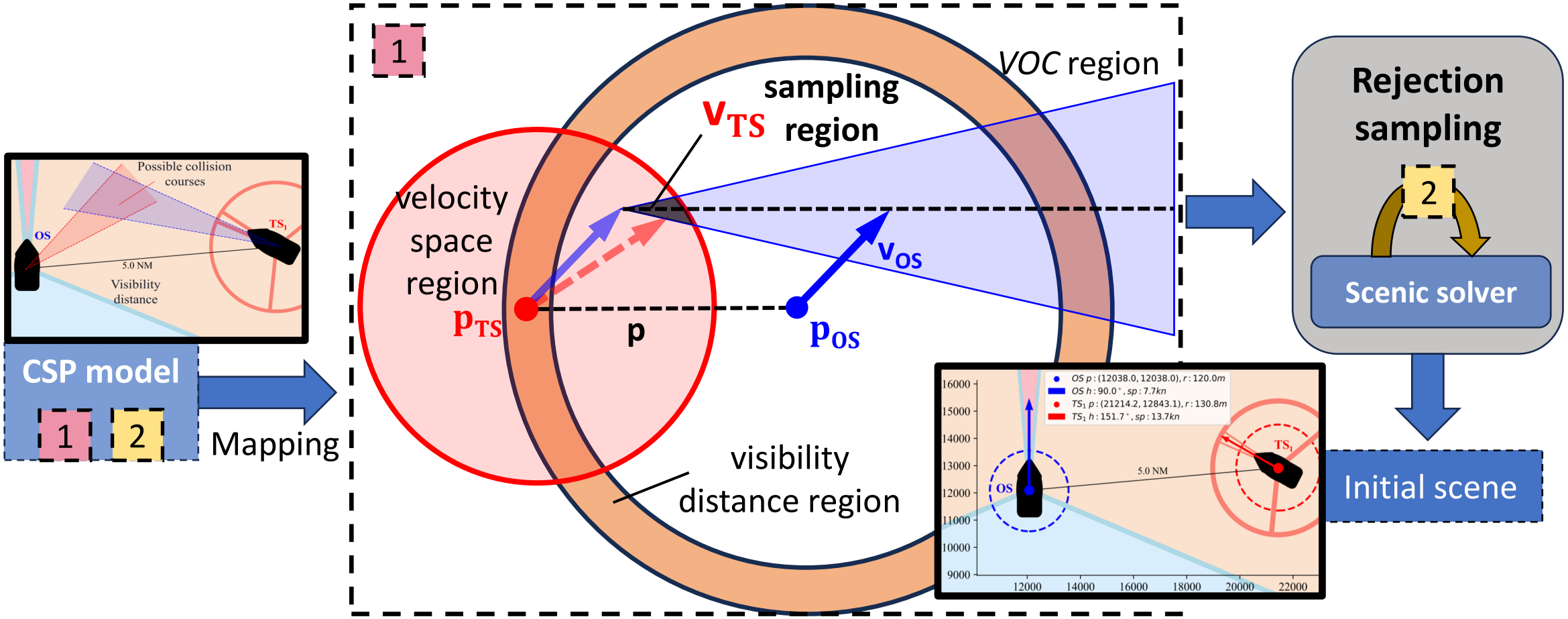
Initial scene$_1$

Initial scene$_2$

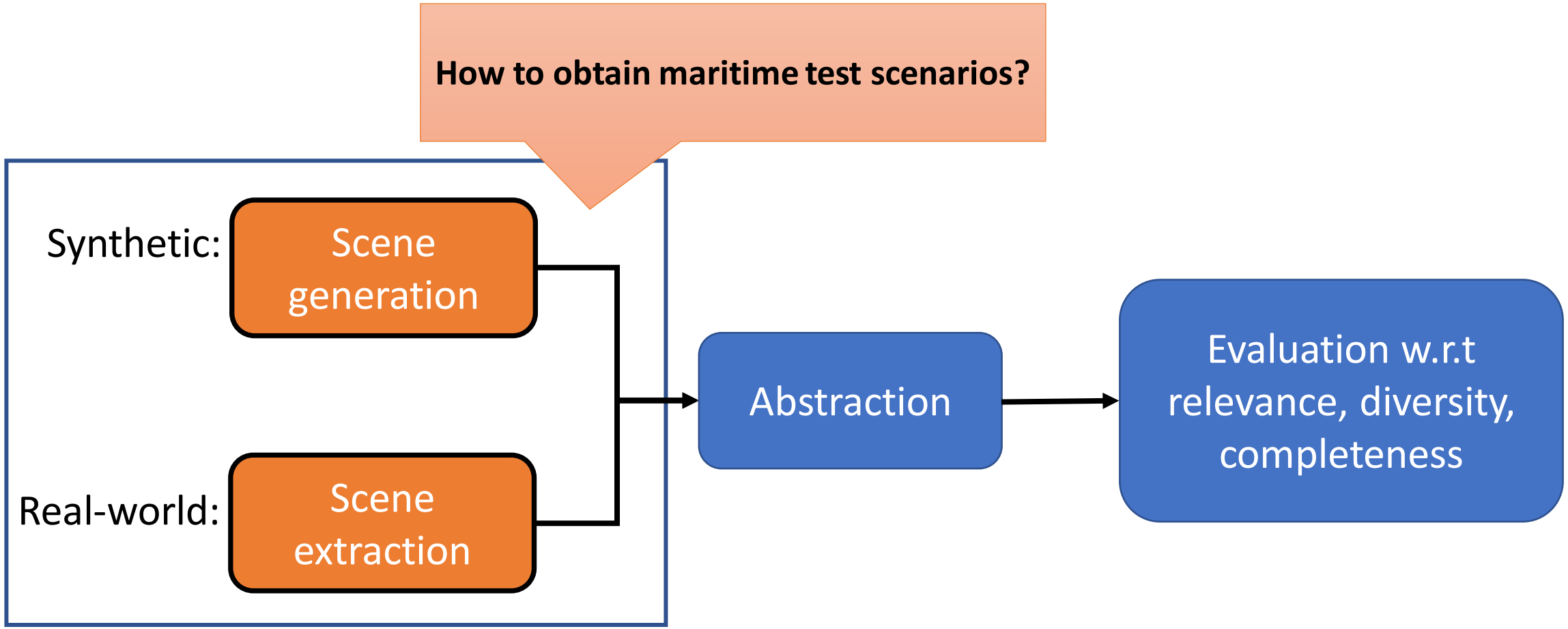
⋮

Initial scene$_n$

Rejection Sampling-based constraint satisfaction (**RS**)



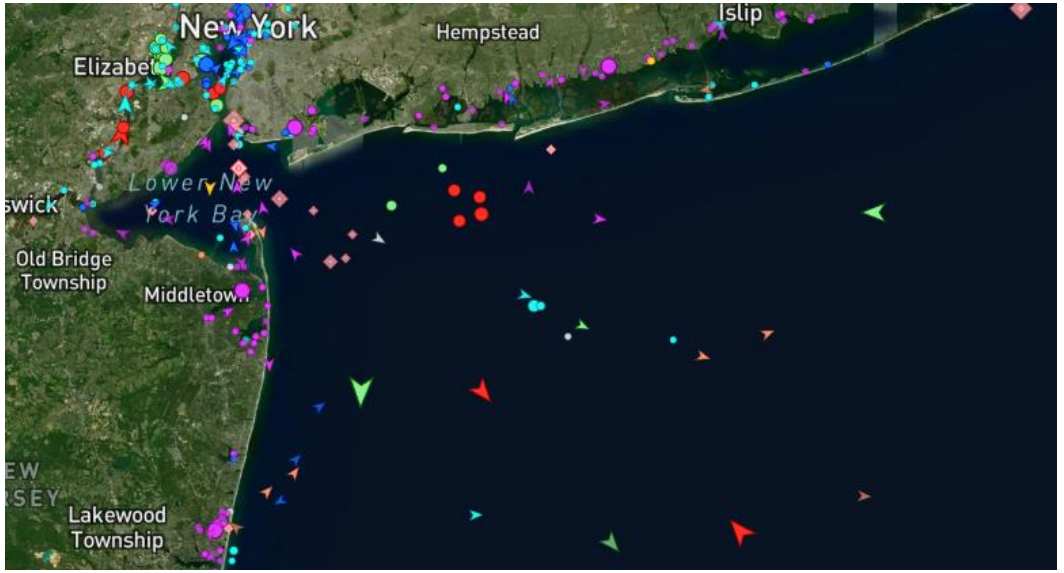
Overview



[2] H. Krasowski & M. Althoff (2022). *CommonOcean: Composable Benchmarks for Ocean Motion Planning.* IEEE ITSC, pp. 1676–1682.

Scene extraction

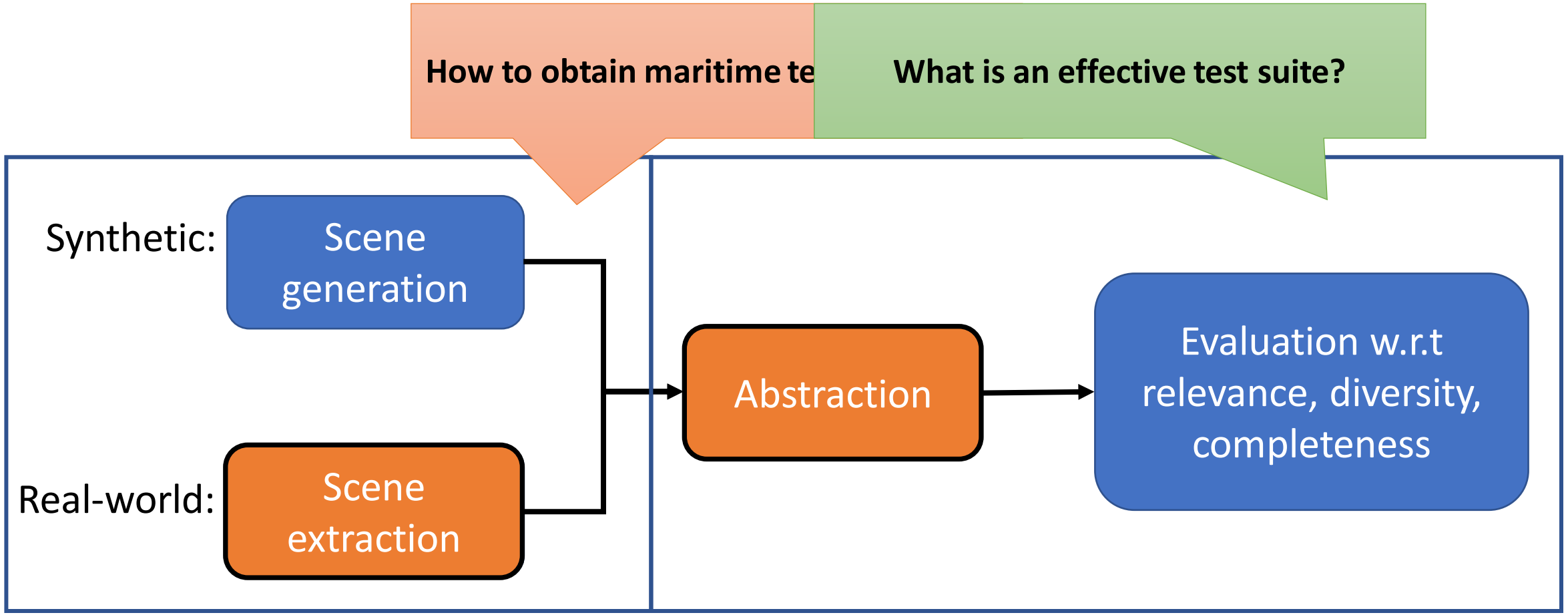
- **Common Ocean benchmark** [2]
- Real-world scenarios extracted from AIS data
- Dominantly 2-vessel encounters (~98%)
- Recorded at real geographical locations
  - Ports of Florida,
  - Middle-east Coast,
  - Upper-west-coast of the US



Source: ww.marinetraffic.com, AIS data

Number of vessels	2	3	4	5	6
Found samples	3222	72	4	0	0

Overview



Characteristics of an effective test suite

1. **Relevance**: represent situations covered by the COLREGS

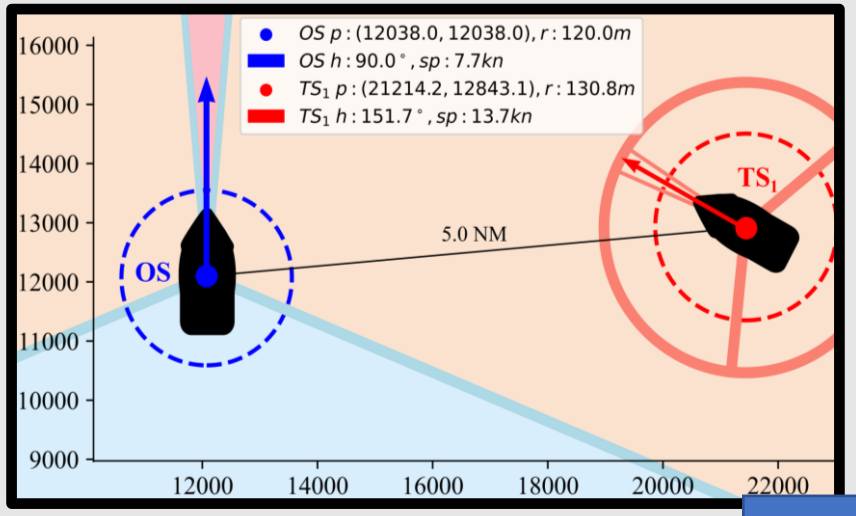
2. **Diversity**: avoid redundant, semantically equivalent scenarios
3. **Completeness**: cover all semantically distinct open sea scenarios, including rare edge cases

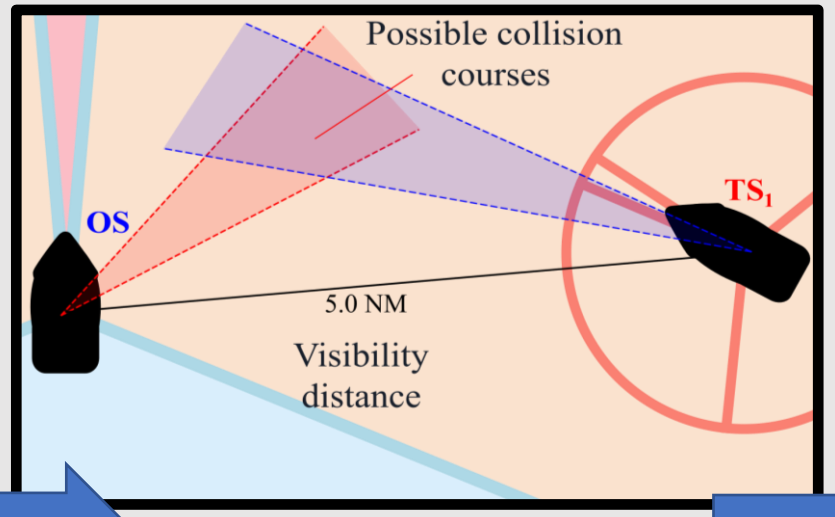
4. **Scalability**: include scenarios with many vessels
5. **Speed**: obtain large numbers of scenarios efficiently

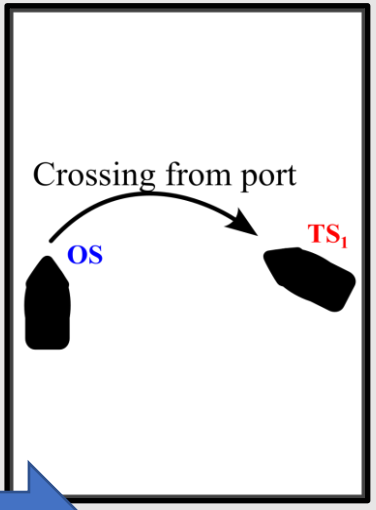
[1] Kargén & Varró (2024), MODELS '24 *Towards Automated Test Scenario Generation for COLREGs Compliance.*

COLREGS scenarios on multiple abstraction levels [1]

Concrete scenarios	Logical scenarios	Equivalence classes
Initial scene (vessel placements)	Geometric constraints	Typed graph



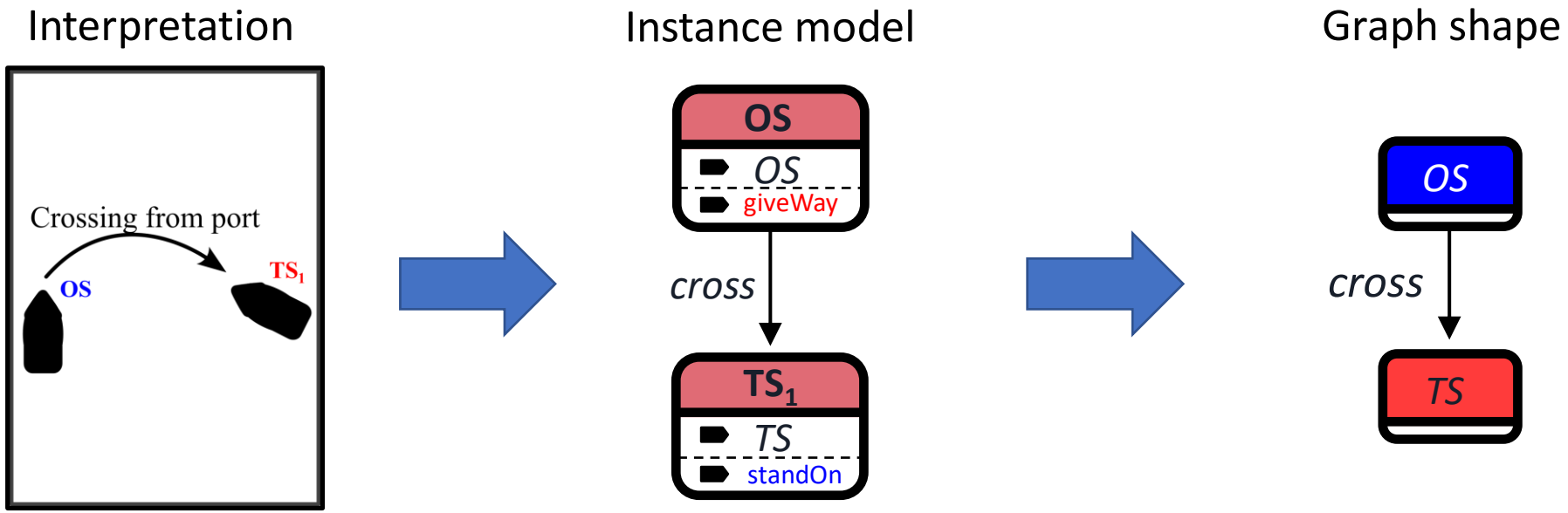




Finding numerical constraints
(collision, distance, relative bearings)

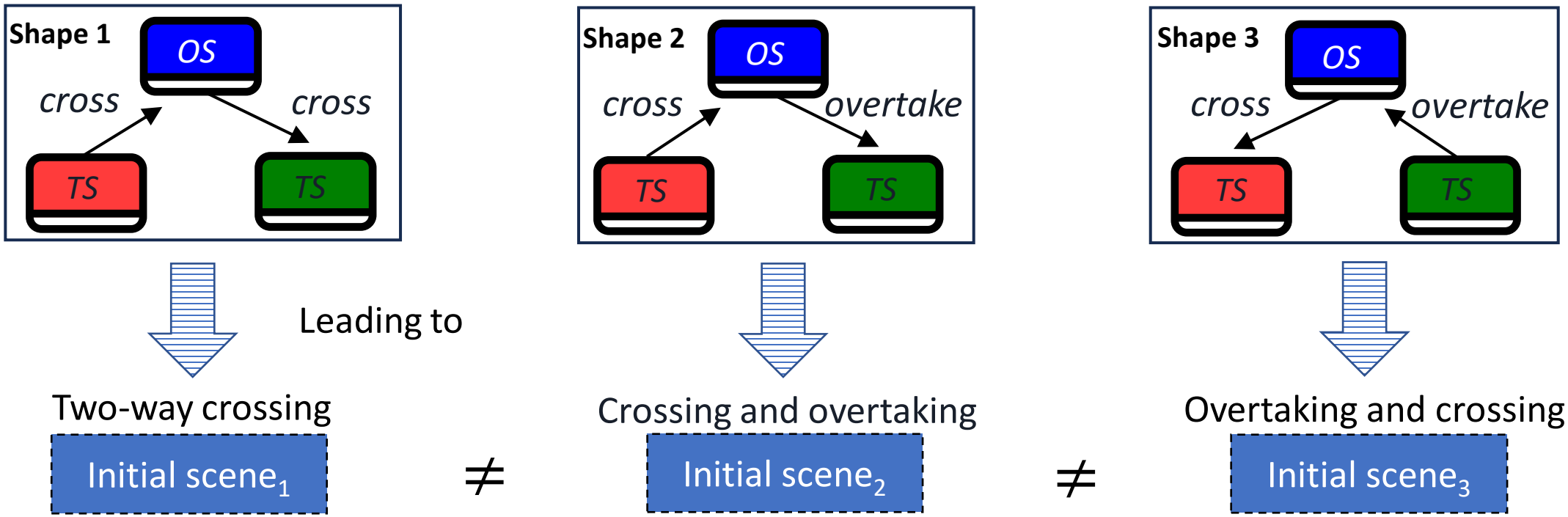
Qualitative abstraction to functional relations

Functional equivalence classes (FECs) as graph shapes



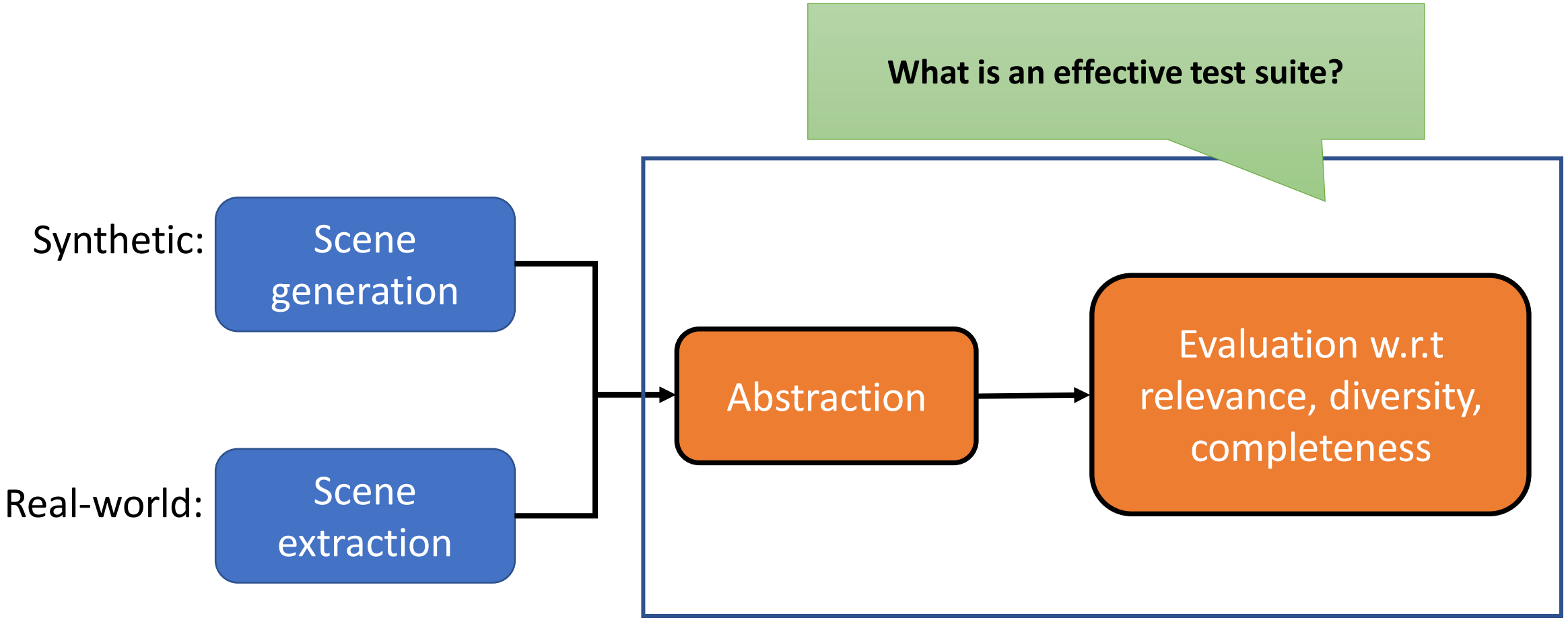
The graph shapes are finitely enumerable abstractions of instance graphs

Functional equivalence classes for test diversity



Covering all equivalence classes gives strong semantic guarantee

Overview



Characteristics of an effective test suite

1. **Relevance**: represent situations covered by the COLREGS

RQ1: How relevant are synthetic and real-world test scenarios for COLREGS compliance?

2. **Diversity**: avoid redundant, semantically equivalent scenarios
3. **Completeness**: cover all semantically distinct open sea scenarios, including rare edge cases

RQ2: How diverse and complete are the test scenes generated by synthetic approaches?

4. **Scalability**: include scenarios with many vessels
5. **Speed**: obtain large numbers of scenarios efficiently

RQ3: How scalable and fast are synthetic scene generation approaches?

RQ3: How scalable and fast are synthetic scene generation approaches?

Search-based Rejection sampling-based

Number of vessels	2		3		4		5		6	
Approach	**SB**	**RS**	**SB**	**RS**	**SB**	**RS**	**SB**	**RS**	**SB**	**RS**
Success rate (%)	100%	100%	100%	100%	100%	9.40%	100%	**0%**	100%	**0%**
Median runtime (s)	0.1 s	0.1 s	0.33 s	11.3 s	0.77 s	**timeout**	2.29 s	**timeout**	7.58 s	**timeout**
Average runtime (s)	0.23 s	0.15 s	1.02 s	16.1 s	2.2 s	572 s	4.9 s	**timeout**	22.6 s	**timeout**
Average time/scene (s)	0.23 s	0.15 s	1.02 s	16.1 s	2.2 s	**5446.8 s**	4.9 s	**timeout**	22.6 s	**timeout**

Answer: - **SB** provides a scene within a scale of seconds, while **RS** does not scale after 3 vessels

Empirical evaluation

RQ3: How scalable and fast are synthetic scene generation approaches?

Answer: - **SB** provides a scene within a scale of seconds, while **RS** does not scale after 3 vessels

RQ1: How relevant are synthetic and real-world test scenarios for COLREGS compliance?

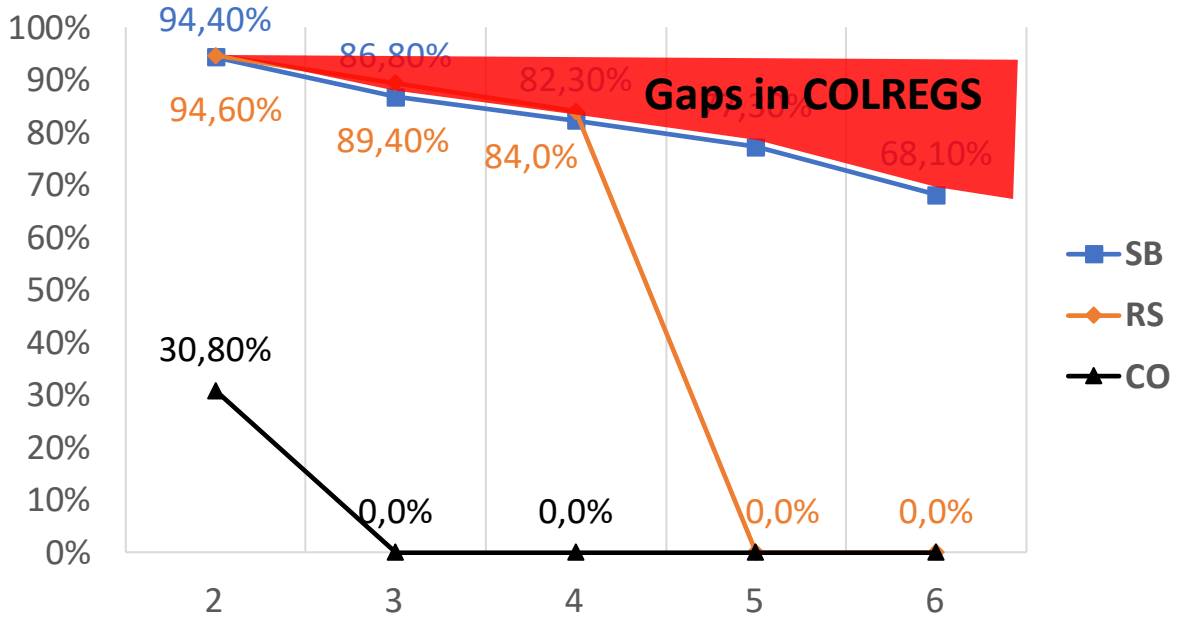
RQ2: How diverse and complete are the test scenes generated by synthetic approaches?

RQ1: How relevant are synthetic and real-world test scenarios for COLREGS compliance?

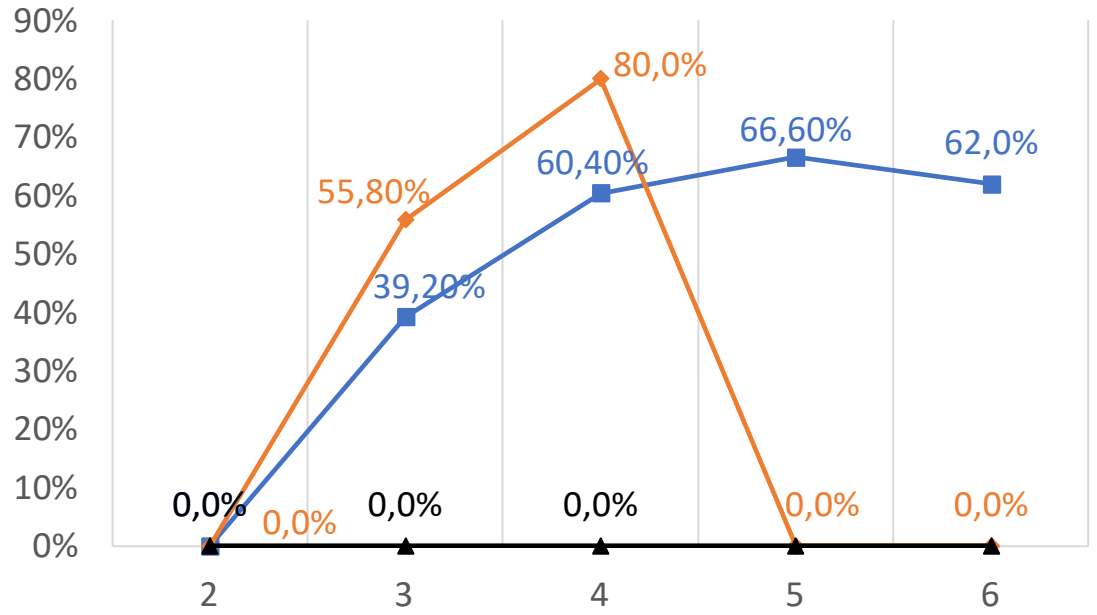
- The ratio of relevant scenarios for the compared approaches:
  - **SB** : Synthetic search-based scene generation
  - **RS** : Synthetic sampling-based scene generation
  - **CO** : Common Ocean benchmark (real-world scenarios)
- 1000 scenes with synthetic approaches for 2—6 vessel scenarios

RQ1: How relevant are synthetic and real-world test scenarios for COLREGS compliance?

The ratio of **relevant** scenes w.r.t the successfully generated scenes



The ratio of **ambiguous** scenes w.r.t the successfully generated scenes



Answer: - **SB** and **RS** generate relevant scenes (with relevance ratios of 68.1% to 94.6%)
- The real-world **CO** dataset has minimal relevance (0% for $K \geq 3$)

Empirical evaluation

RQ3: How scalable and fast are synthetic scene generation approaches?

Answer: - **SB** provides a scene within a scale of seconds, while **RS** does not scale after 3 vessels

RQ1: How relevant are synthetic and real-world test scenarios for COLREGS compliance?

Answer: - **SB** and **RS** generate relevant scenes (with relevance ratios of 68.1% to 94.6%)
- The real-world **CO** dataset has minimal relevance (0% for $K \geq 3$)

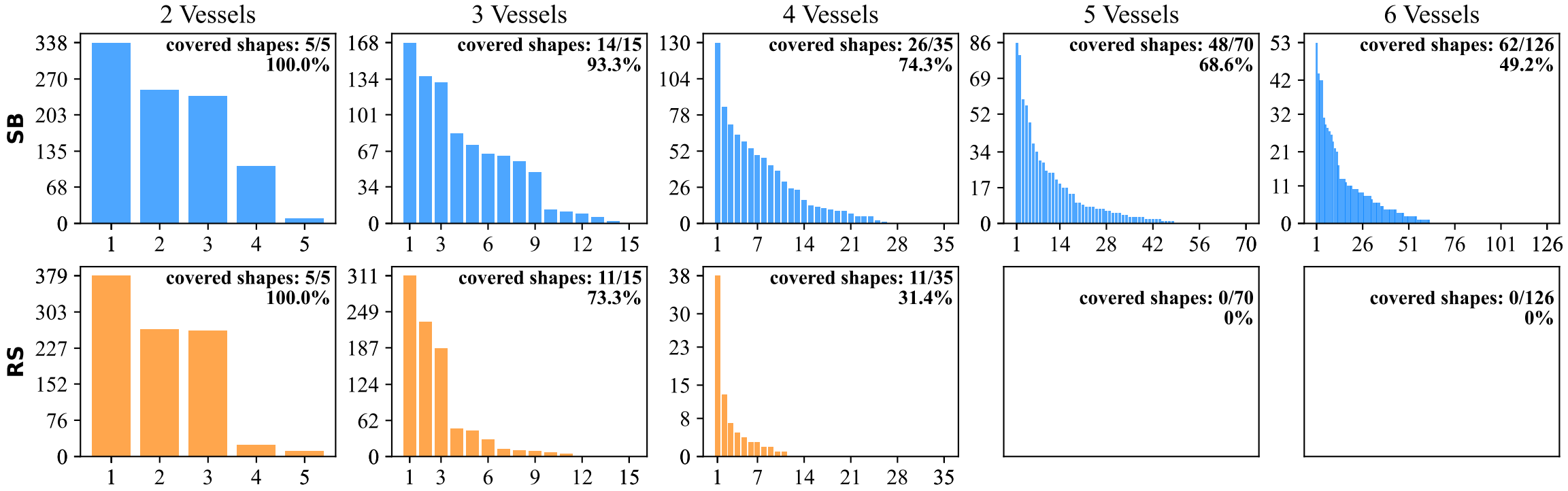
RQ2: How diverse and complete are the test scenes generated by synthetic approaches?

RQ2: How diverse and complete are the test scenes generated by synthetic approaches?

- Deriving all the possible **relevant** FECs using a graph solver.
- Structural coverage metric: $$\frac{|\textit{covered functional equivalence classes}|}{|\textit{total functional equivalence classes}|}$$

RQ2: How diverse and complete are the test scenes generated by synthetic approaches?

The distributions of derived scenes across all relevant equivalence classes for **SB** and **RS**



Answer: - Decreasing coverage for **SB** and **RS** as the number of vessels increase

Empirical evaluation

RQ3: How scalable and fast are synthetic scene generation approaches?

Answer: - **SB** provides a scene within a scale of seconds, while **RS** does not scale after 3 vessels

RQ1: How relevant are synthetic and real-world test scenarios for COLREGS compliance?

Answer: - **SB** and **RS** generate relevant scenes (with relevance ratios of 68.1% to 94.6%)
- The real-world **CO** dataset has minimal relevance (0% for $K \geq 3$)

RQ2: How diverse and complete are the test scenes generated by synthetic approaches?

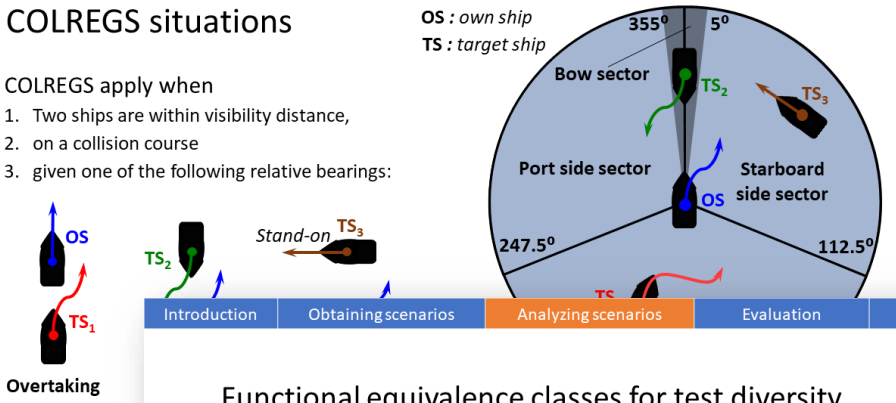
Answer: - Decreasing coverage for **SB** and **RS** as the number of vessels increase

Conclusions

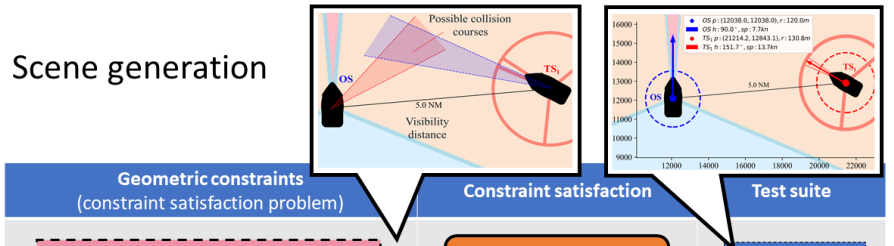
COLREGS situations

COLREGS apply when

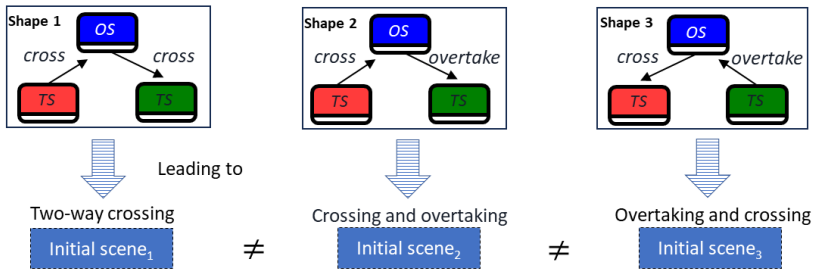
1. Two ships are within visibility distance,
2. on a collision course
3. given one of the following relative bearings:



Scene generation



Functional equivalence classes for test diversity



Covering all equivalence classes gives strong semantic guarantee

Empirical evaluation

RQ3: How scalable and fast are synthetic scene generation approaches?

Answer: - **SB** provides a scene within a scale of seconds, while **RS** does not scale after 3 vessels

RQ1: How relevant are synthetic and real-world test scenarios for COLREGS compliance?

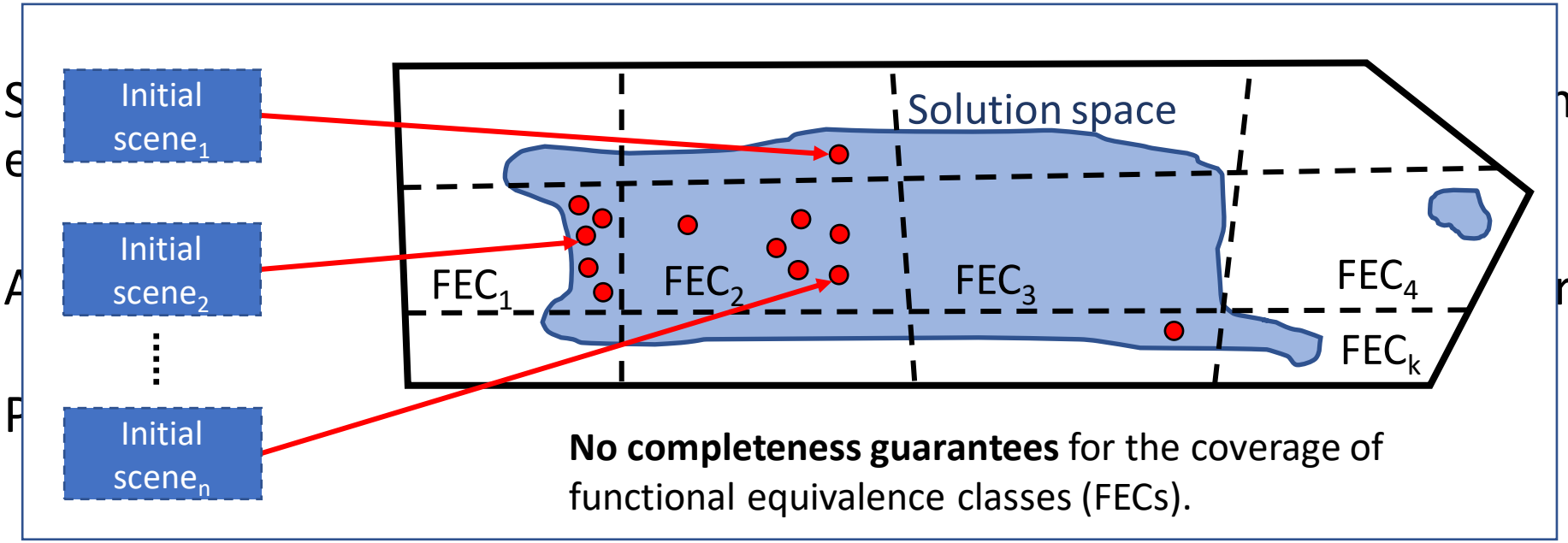
Answer: - **SB** and **RS** generate relevant scenes (with relevance ratios of 68.1% to 94.6%)
- The real-world **CO** dataset has minimal relevance (0% for $K \geq 3$)

RQ2: How diverse and complete are the test scenes generated by synthetic approaches?

Answer: - Decreasing coverage for **SB** and **RS** as the number of vessels increase

Key findings

1. S
   e
2. A
3. F

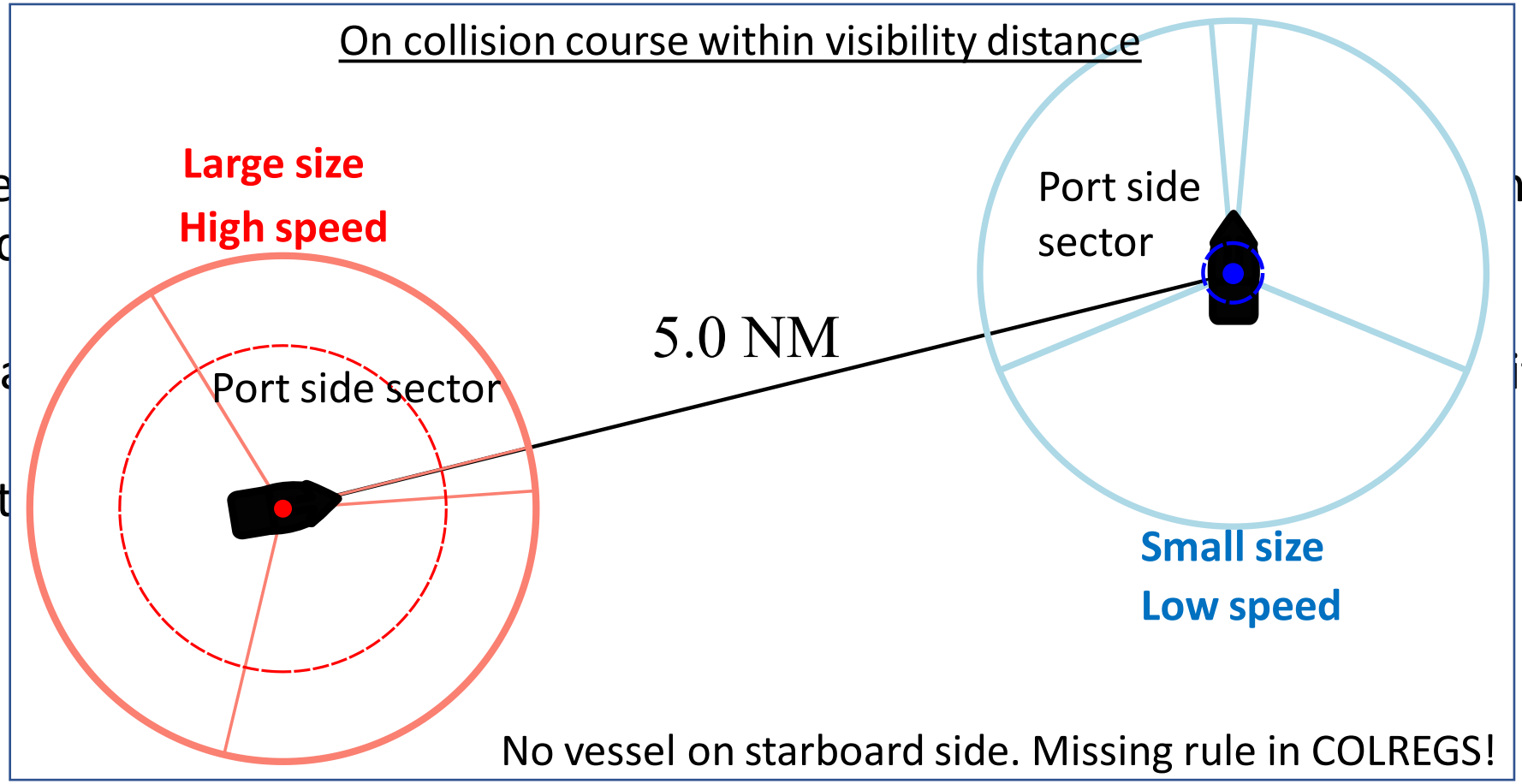


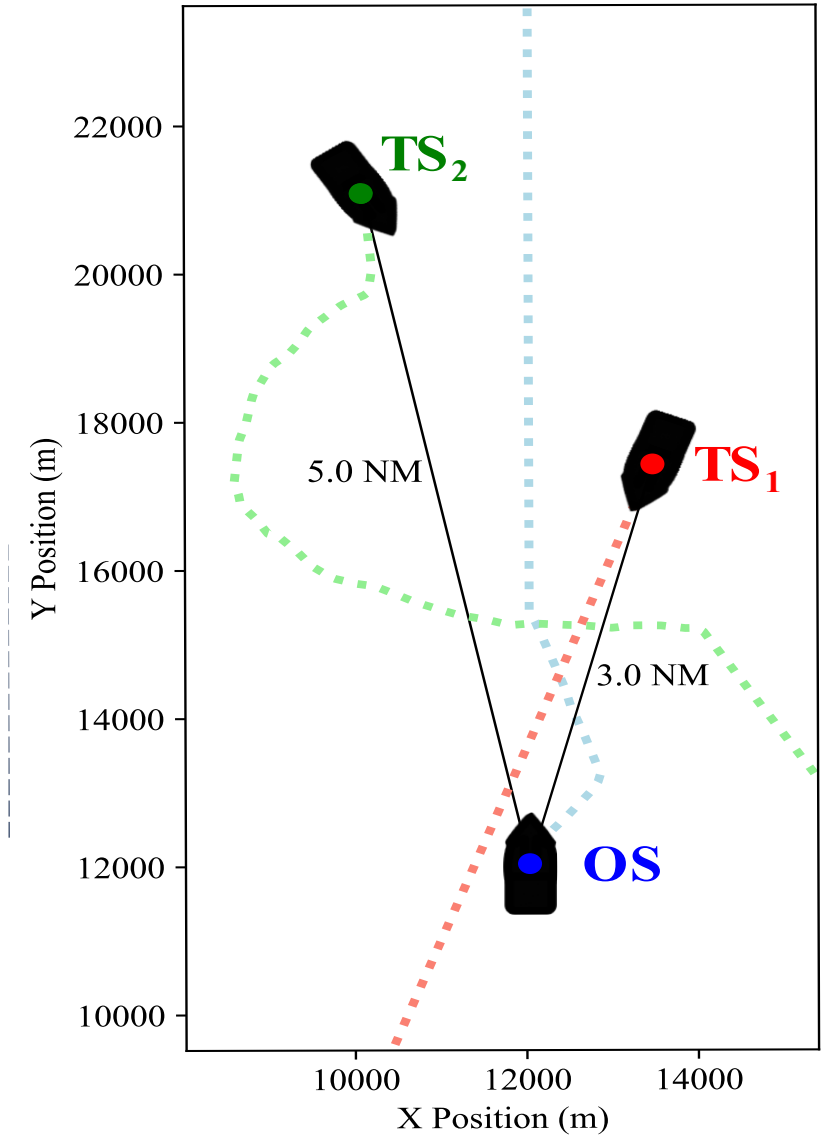
multi-ship

ritime setting.

Key findings

1. Sce … nulti-ship enc…
2. Ada… itime setting.
3. Pot…





compliance

' at the component level (e.g., model checking)

e still relies on **scenario-based testing**

run

Environment

sensory information

position, speed, course

Own Ship (ASV)

states

ased on COLREGS)
jectories for the **OS**
tion

Is **OS** COLREGS-compliant?

Pass

Fail