AAAI-26
JANUARY 20–27, 2026 | SINGAPORE

TURING

STRIDE-QA: Visual Question Answering Dataset for Spatiotemporal Reasoning in Urban Driving Scenes

AAAI 2026 Oral

Keishi Ishihara, Kento Sasaki, Tsubasa Takahashi, Daiki Shiono, Yu Yamaguchi

Turing Inc.

General VLMs Struggle with Scene Dynamics

TURING

General VLMs trained on static web corpus fail at 4D spatiotemporal reasoning

t = -2.0

t = -1.5

t = -0.5

t = 0

Where will **Region [0]** be relative to the ego vehicle after 1 second, in terms of distance in meters and bearing angle in degrees?

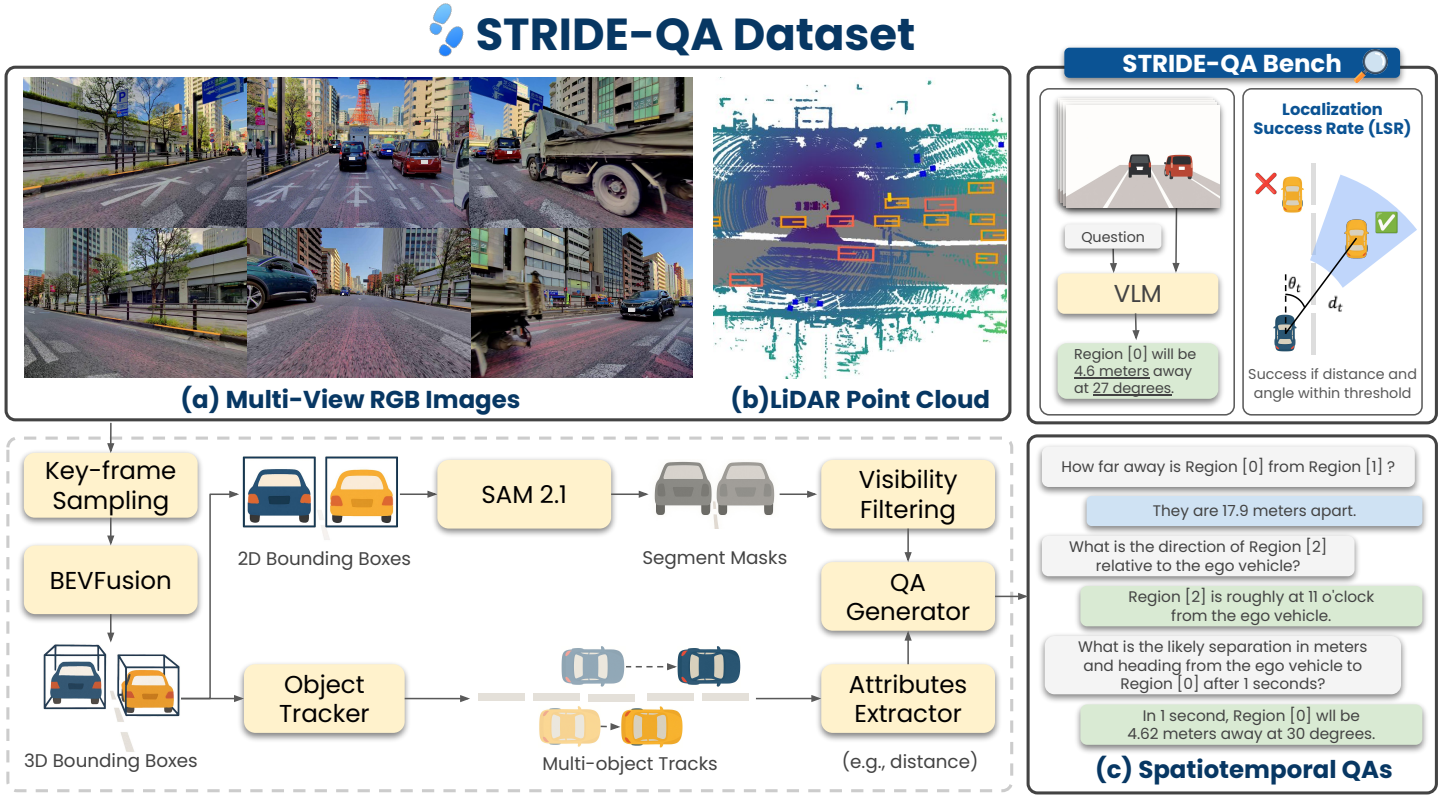
By 1 second, the separation reaches around **3.87 meters** at **-66 degrees** position.

In 1 second, they are **8.5 meters** at **5 degrees**.

Introducing STRIDE-QA Dataset

TURING

Large-scale VQA Dataset for Ego-centric Spatiotemporal Reasoning



Comparison with Existing Datasets

TURING

STRIDE-QA is large-scale (**100 hrs, 16M**), has Spatial and Spatiotemporal QAs

Dataset	# Video	# QA	Viewpoint		QA Type			
			Obj.	Ego	S-Q	S-N	ST-Q	ST-N
Spatial-VQA (Chen et al. 2024a)	—	200 M	✓		✓	✓		
Open Spatial Dataset (Cheng et al. 2024)	—	8.7 M	✓		✓	✓		
Refer-KITTI (Wu et al. 2023a)	6 h	818		✓	✓		✓	
ToD3Cap (Jin et al. 2024)	5.5 h	468 K	✓	✓	✓			
nuScenes-QA (Qian et al. 2024)	5.5 h	460 K	✓	✓	✓		✓	
NuPrompt (Wu et al. 2025)	5.5 h	87.3 K		✓	✓		✓	
nuPlanQA (Park et al. 2025)	119 h	1 M	✓	✓	✓		✓	
TUMTraffic-VideoQA (Zhou et al. 2025)	RGB	≤33.3 h	87.3 K	✓	✓	✓		✓
STRIDE-QA (Ours)	100 h	16M	✓	✓	✓	✓		✓

Ego-centric QA is important for applications like self-driving

Existing Datasets do not cover Numerical Spatiotemporal QAs

QA Type

- S-Q: Spatial and Qualitative
- S-N: Spatial and Numerical
- ST-Q: Spatiotemporal and Qualitative
- ST-N: Spatiotemporal and Numerical

Three VQA Tasks

TURING

Relative to Region [0], where is Region [2] ?

A

Region [2] appears around 2 o'clock from Region [0].

Is the ego vehicle is bigger than Region [1]?

B

Yes, the ego vehicle is bigger in size than Region [1].

What will be the distance and bearing angle of Region [3] from the ego vehicle after 1 second?

C

At 1 second, Region [3] is about 5.99 meters away, at -75 degrees.

A) **Object-centric Spatial QA**
Assessing spatial relations b/w non-ego agents

B) **Ego-centric Spatial QA**
Describing agents' distance, orientation, etc from the ego

C) **Ego-centric Spatiotemporal QA**
Predicting how agent relations evolve over time

Three VQA Tasks

TURING

Relative to Region [0], where is Region [2] ?

A

Region [2] appears around 2 o'clock from Region [0].

Is the ego vehicle is bigger than Region [1]?

B

Yes, the ego vehicle is bigger in size than Region [1].

What will be the distance and bearing angle of Region [3] from the ego vehicle after 1 second?

C

At 1 second, Region [3] is about 5.99 meters away, at -75 degrees.

A) **Object-centric Spatial QA**
Assessing spatial relations b/w non-ego agents

B) **Ego-centric Spatial QA**
Describing agents' distance, orientation, etc from the ego

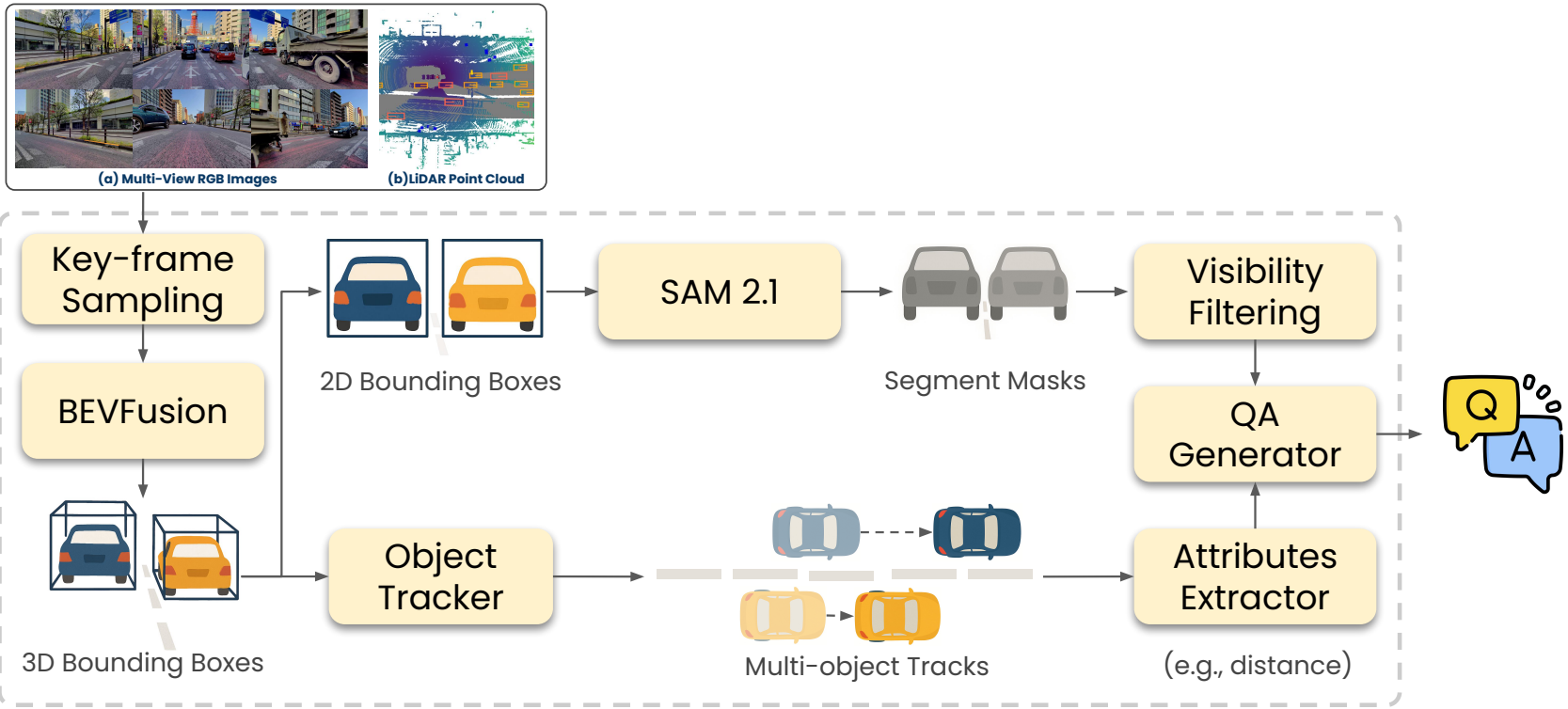
C) **Ego-centric Spatiotemporal QA**
Predicting how agent relations evolve over time

Answering these questions require physical grounding and short-term predictive reasoning

Dataset Construction Pipeline

TURING

Annotation (from images to QA texts) is **fully automated** by our dataset pipeline



Introducing STRIDE-QA Bench

TURING

Benchmark Composition

409 Unique Scene Groups

5,317 QA Pairs

6 Dynamic Interactions

Task Definition

Input:
4 front camera images with marked single object

Output:
distance, heading, and velocity at t = {0, 1, 2, 3} s

GT:

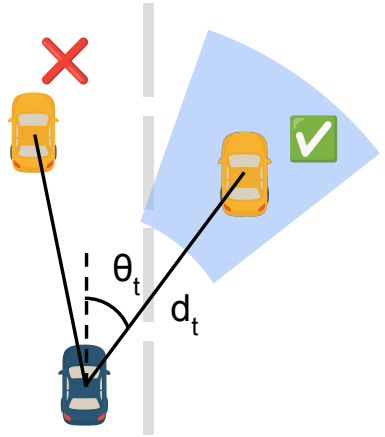
By 1 second, the separation reaches around **3.87 meters** at **-66 degrees** position.

STRIDE-QA Bench Metrics

TURING

STRIDE-QA Bench defines two metrics for evaluation

Localization Success Rate (LSR)



Success if distance and angle within threshold

Localization Success Rate (LSR)

- Evaluates simultaneous accuracy in distance and heading
- Success criteria:
  - Distance error < 25%
  - Heading error < 10°

Temporal Localization Consistency (TLC)

- Measures LSR stability over time t = {0, 1, 2, 3} s
- Requires successful localization across **all 4 timesteps** in a sequence

STRIDE-QA Dramatically Improves Reasoning

TURING

- General-purpose VLMs perform poorly at predicting future interactions
- STRIDE-QA fine-tuned models:
  - **Nearly perfect** at localizing surrounding agents
  - **Dramatically improves** short-term spatiotemporal prediction

General-purpose VLMs

Fine-tuned on STRIDE-QA

Model	LSR ↑				MLSR ↑	TLC ↑
	0s	1s	2s	3s		
GPT-4o	18.1	6.6	6.1	7.6	9.6	0.7
GPT-4o mini	4.6	2.0	0.7	0.7	2.0	0.0
InternVL2.5-8B	2.4	1.0	1.7	0.7	1.5	0.0
Qwen2.5-VL-7B-Instruct	1.0	3.4	4.4	1.0	2.4	0.0
SpatialRGPT-VILA-1.5-8B	0.5	0.2	0.2	0.0	0.2	0.0
Senna-VLM	1.0	0.0	0.2	0.0	0.3	0.0
Cosmos-Reason1-7B	1.5	3.2	2.0	1.5	2.0	0.0
STRIDE-Qwen2.5-VL-7B	96.3	**46.2**	**38.4**	**38.9**	**55.0**	**28.4**
STRIDE-Cosmos-Reason1-7B	**96.8**	43.5	37.4	36.2	53.5	25.4

Analysis of Prediction Patterns

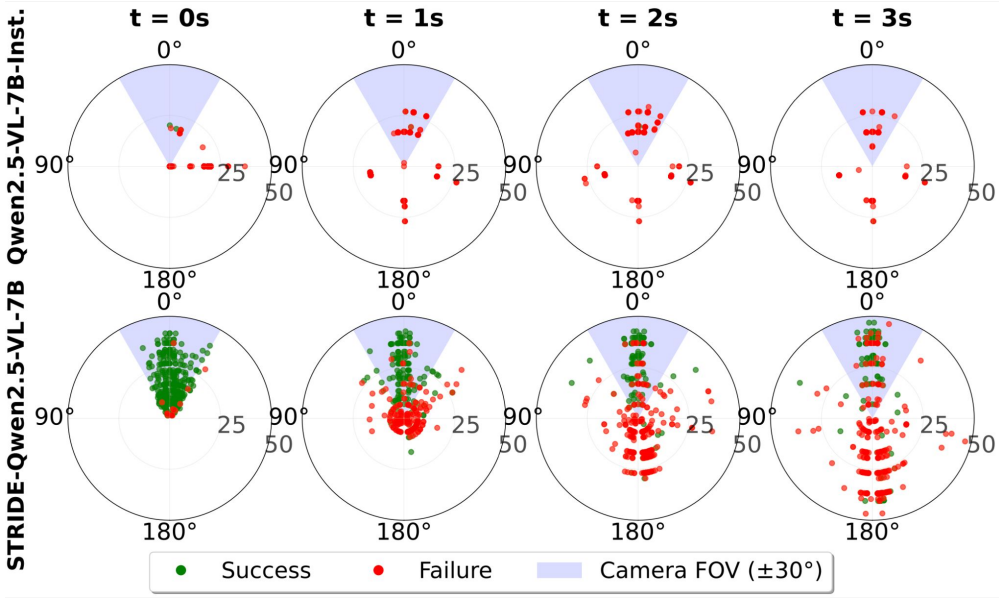
TURING

Baseline (Top Row)

- **Sparse & Biased**
  Predictions cluster in fixed areas
- **Failure**
  Lacks temporal consistency

STRIDE-QA fine-tuned (Bottom Row)

- **High Precision (t=0)**
  Near-perfect localization at t=0
- **Grounded Reasoning**
  Shows clear intent to track motion

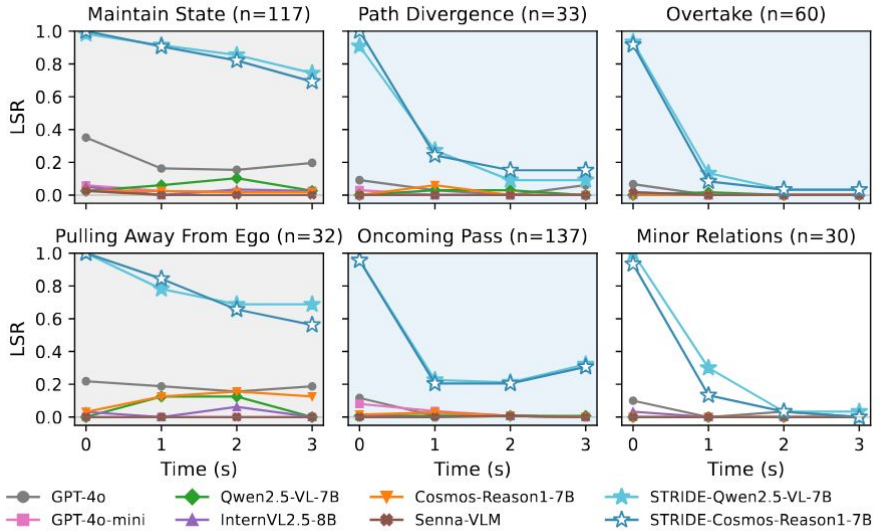


Analysis of Dynamic Interactions

TURING

- **In-View Scenarios**: Performance (LSR) declines gracefully, maintaining reasonable tracking capabilities
- **Out-of-View (OOV) Scenarios**: Performance degrades sharply, indicating a struggle to reason about invisible trajectories

Scenario	OOV Rate
Maintain State	0.05
Pulling Away From Ego	0.19
Oncoming Pass	1.00
Overtake	0.98
Path Divergence	0.94
Minor Relations	0.90
Total	0.65



Summary

TURING

Summary

- **Bridge to Physical AI with STRIDE-QA**
  Introduced a large-scale dataset (**16M pairs**) to enable physically grounded reasoning
- **Proven Effectiveness on New Benchmark**
  Established a rigorous spatiotemporal benchmark, where fine-tuned models achieved **dramatic gains** (near-zero to 28% TLC), proving the dataset value

Future Work

- **Multi-Camera Configuration**
  Tracking **Out-of-View** targets remains challenging with single-camera inputs, limiting long-term consistency
- **Extension to VLA**
  Future work involves extending to **Vision-Language-Action (VLA)** models to verify effectiveness in real-world driving

Thank you!

https://turingmotors.github.io/stride-qa/

Data | Benchmark | Weight

AAAI-26
JANUARY 20–27, 2026 | SINGAPORE

TURING

STRIDE-QA: Visual Question Answering Dataset for Spatiotemporal Reasoning in Urban Driving Scenes

AAAI 2026 Oral

Keishi Ishihara, Kento Sasaki, Tsubasa Takahashi, Daiki Shiono, Yu Yamaguchi

Turing Inc.