NEURIPS 2026 · TRACK ON EVALUATIONS AND DATASETS

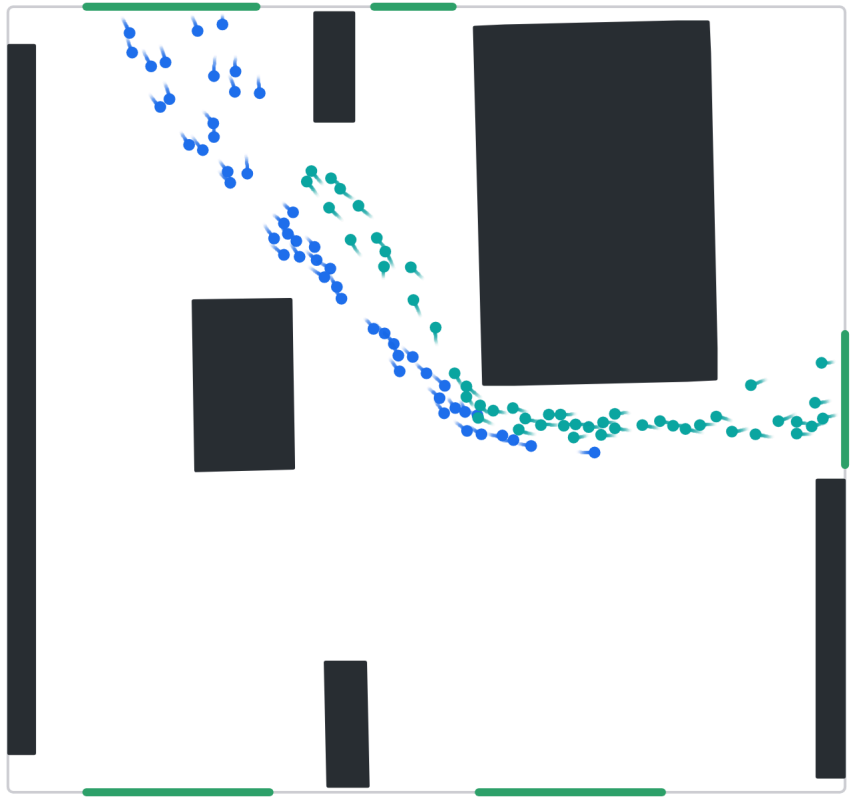
V R D S T

STRIDE

Automated Evaluation of Text-to-Trajectory Alignment across Diverse Contexts

Wanchun Ni[1] · Tao Qi[2] · Leonel Aguilar[1] · Marlene Wagner[1]
Jiugeng Sun[1] · Verena Zimmermann[1] · Mennatallah El-Assady[1]

[1]ETH Zurich [2]Beijing University of Posts and Telecommunications



Illustrative pedestrian simulation

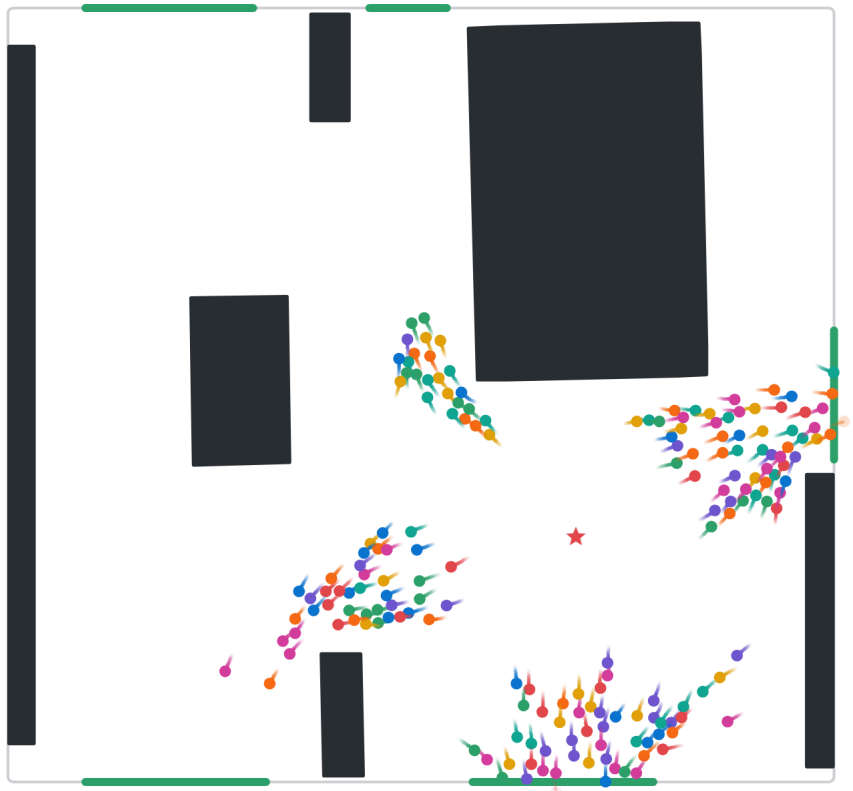
MOTIVATION

Pedestrian trajectories can now be generated from text

SCENARIO DESCRIPTION

A firecracker thrown into a busy plaza causes a brief panic. People flee radially toward the open exits, then slow back down to a calm walking pace.

Text-to-trajectory model



Illustrative pedestrian simulation

e.g. Text-Crowd · CrowdMoGen · LMTraj

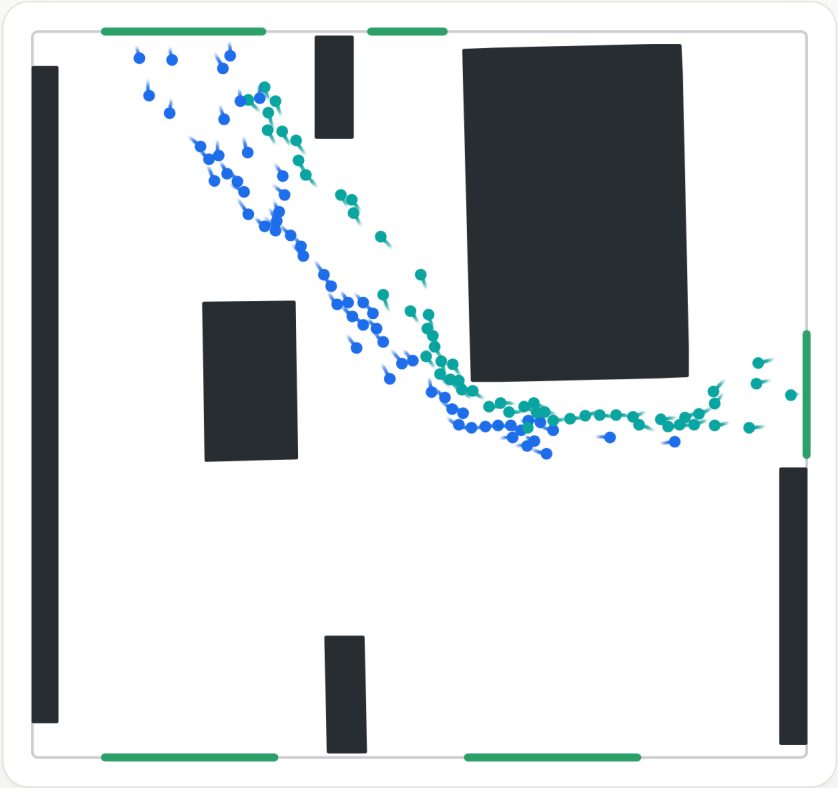
But do the generated trajectories actually follow the text?

MOTIVATION

Same place, different context, different behavior

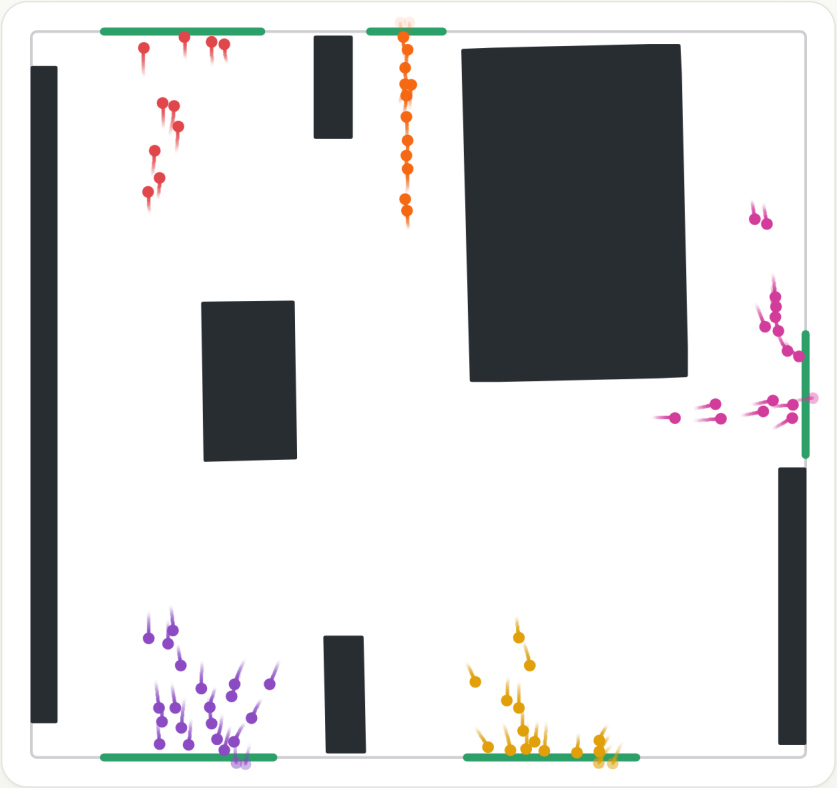
Routine commute

walking speed · bidirectional lanes · avoid obstacles



Emergency evacuation

running · directed outflows to the exits



Generated trajectories must match the scenario — for **individuals**, for the **group**, and for the **environment**.

THE EVALUATION GAP

Evaluating text-to-trajectory alignment is hard

Ground-truth metrics

ADE · FDE · collision rate · EMD

Need real human trajectories for every scenario. Rare events such as evacuations or violence cannot be ethically recorded, and many behaviors are plausible.

Does not scale

LLM-as-judge

A language model rates the trajectories

Opaque and prompt-sensitive. Shown the metrics of an evacuation, GPT-5.2 picked the wrong scenario in 5 out of 5 runs.

Unreliable

Human evaluation

People judge the trajectories

Informative, but slow and expensive across thousands of diverse scenarios.

Costly

Needed: evaluation that is **verifiable**, **automated** and **scalable** — without collecting human trajectories.

OUR APPROACH

STRIDE: break the context down, then verify it with code

1 WHAT TO EVALUATE

VRDST protocol

V R D S T

Five observable axes of pedestrian behavior, grounded in sociology.

2 WHAT TO ASK

Behavioral questions

Are people running in panic?

Do they head for the exits?

An LLM breaks each scenario into checkable questions.

3 HOW TO VERIFY

Deterministic measurements

`mean_speed(traj)` in range? ✓

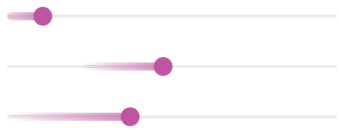
20 functions compute exact statistics from the coordinates.

An **LLM for semantics, code for numbers** → reproducible, interpretable scores.

STEP 1 · EVALUATION SPACE

VRDST: five observable axes of pedestrian behavior

V Velocity



Do pedestrians move at a speed that fits the scenario?

INDIVIDUAL MOTION

R Realism



Is the motion physically feasible — no overlaps, no walking through walls?

PHYSICAL VALIDITY

D Direction



Do pedestrians head where the scenario implies?

INDIVIDUAL MOTION

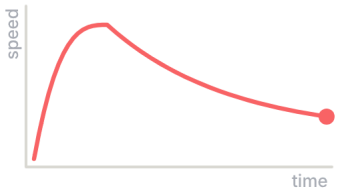
S Spatial



Are pedestrians in the right places — gathered, spread out, near exits?

SCENE ORGANIZATION

T Temporal



Do the trajectories evolve over time as described?

SCENE EVOLUTION

Grounded in McPhail & Wohlstein's dimensions of gathering behavior and in pedestrian dynamics (Helbing; Hall's proxemics).

STEP 2 · BEHAVIORAL DECOMPOSITION

Each scenario becomes a set of checkable questions

SCENARIO Escaping

A firecracker thrown into a busy plaza causes a brief panic. Pedestrians first gather around the disturbance, then disperse radially toward the open exits before calming to a walking pace.

V R D S T **VRDST protocol**
which aspects to probe

TrajFacts
human-trajectory reference values for calibration

LLM

Q1 V **Are people running in panic?**

`mean_speed`

EXPECTED RANGE

Q2 D **Are people running away from the firecracker?**

`directional_entropy` `flow_alignment` `path_linearity`

EXPECTED RANGE

Q3 T **Do people first run fast, then slow down?**

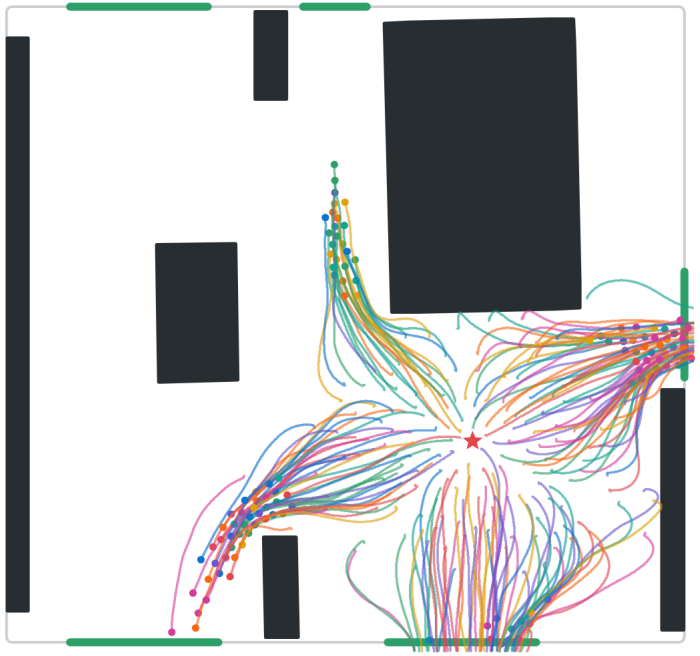
`speed_trend` `density_trend`

EXPECTED RANGE

STEP 3 · DETERMINISTIC VERIFICATION

Answers are verified by code, not by a judge model

GENERATED TRAJECTORIES



V **Q1** Are people running in panic? — Q-SCORE 1.00

mean_speed		✓ 1

D **Q2** Running away from the firecracker? — Q-SCORE 0.67

directional_entropy		✓ 1
flow_alignment		✓ 1
path_linearity		✗ 0

T **Q3** First fast, then slowing down? — Q-SCORE 1.00

speed_trend		✓ 1
density_trend		✓ 1

SCENE SCORE

0.89

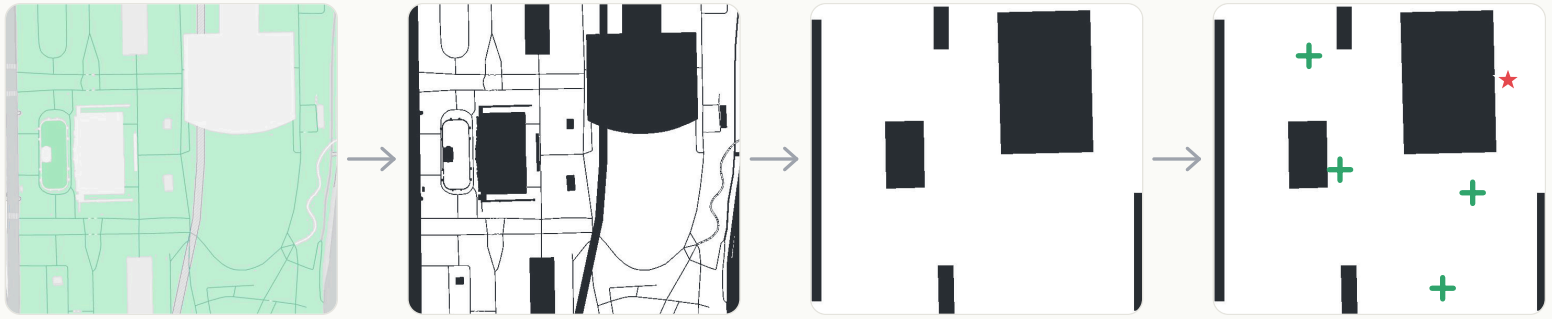
```
= (1.00 + 0.67
  + 1.00) / 3
```

Averaged over measurements, questions, then scenes.

- ✓ Traceable to exact values
- ✓ No LLM at scoring time (benchmark mode)

BENCHMARK

STRIDE-Bench: pedestrian scenarios at scale



Real venue map | Walkable area | Obstacle map | Goals & event center

936 scenarios

6,633 behavioral questions

11,696 measurements

20 measurement tools

11 scenario categories

Berlonghi's typology · safety-critical categories are oversampled

Ambulatory · Disability · Cohesive · Expressive · Participatory · Aggressive · Demonstrator · **Escaping** · **Dense** · Rushing · **Violent**

Benchmark mode

Fixed scenarios, questions and answer ranges. Fully deterministic scoring.

Open mode

An LLM writes the questions for any new scenario description you bring.

VALIDATION

STRIDE agrees with real trajectories — and with people

Real trajectories score high

Real human trajectories, scored with the same pipeline

0.94

Fête des Lumières, Lyon
12 festival recordings

0.91

ETH / UCY
5 everyday scenes

People agree with STRIDE

50 annotators · 15,750 annotations

80%

annotator agreement with STRIDE-Bench

Annotators vs. STRIDE-Bench	κ = 0.73
Annotators vs. each other	α = 0.70

Not tied to one LLM

Agreement with the human consensus label

DeepSeek-V3	0.74
Claude Sonnet 4.6	0.74
Qwen3.6-Plus	0.70
GPT-5.5	0.70
Humans (leave-one-out)	0.69

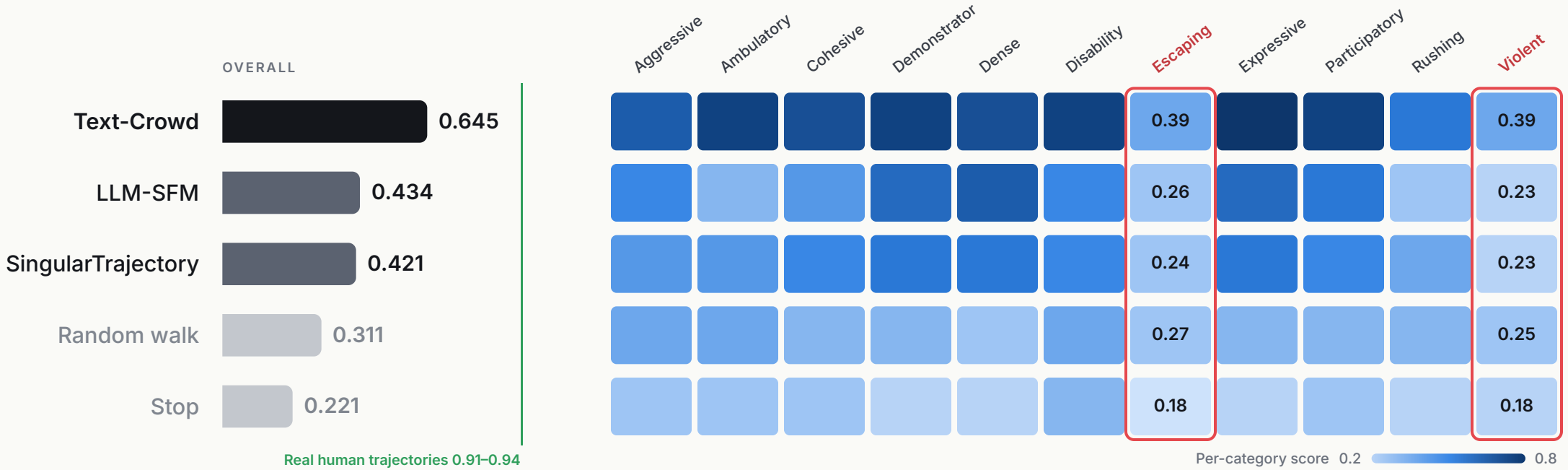
All four fall within the human range.

κ: Cohen's kappa · α: Krippendorff's alpha — both correct for agreement by chance.

EVALUATION

Today's models are far from context-aligned

STRIDE score on STRIDE-Bench (936 scenarios) · higher is better



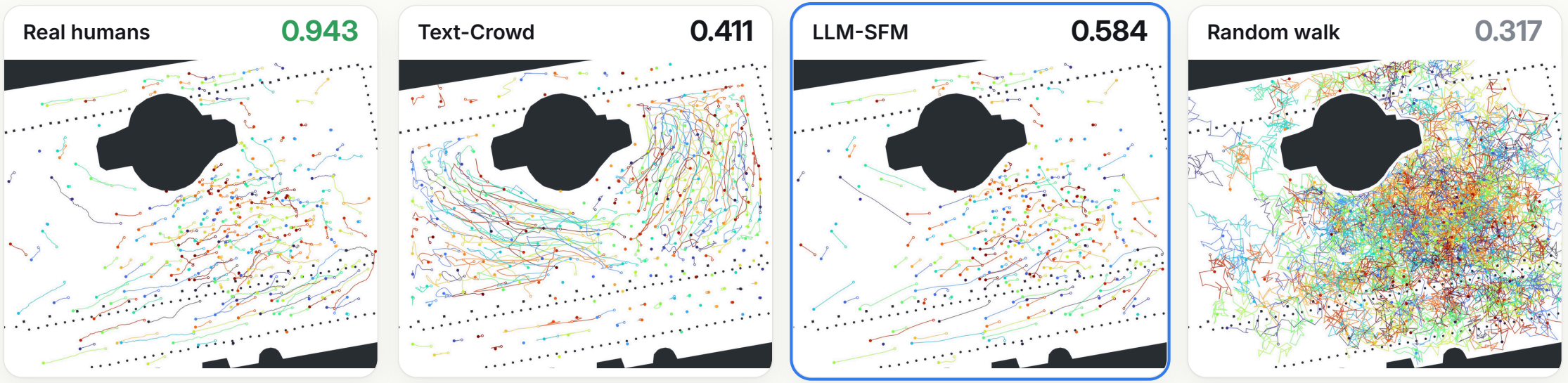
0.645 Best model (Text-Crowd), far below real human trajectories (0.91–0.94).

< 0.40 Escaping and Violent are hardest, even for the best model.

⇄ Rankings shift with rollout length: LLM-SFM leads up to 2 min, Text-Crowd beyond.

EVALUATION · REAL FESTIVAL SCENES

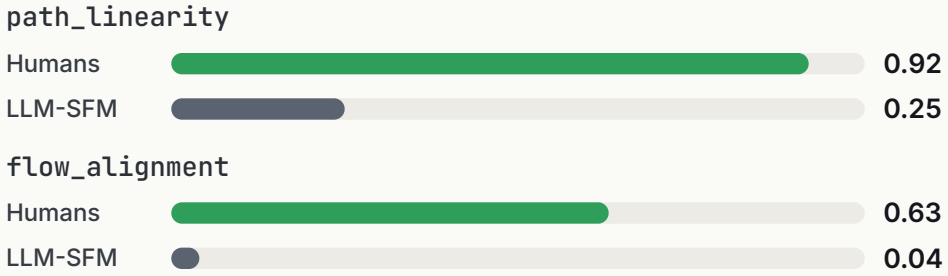
Looking similar is not behaving similarly



Fête des Lumières, Lyon · average STRIDE score over 12 real recordings

STRIDE pinpoints why. LLM-SFM looks human-like at a glance, yet its paths are wiggly and poorly aligned with the main flow.

Measurement values from one Lyon scene.

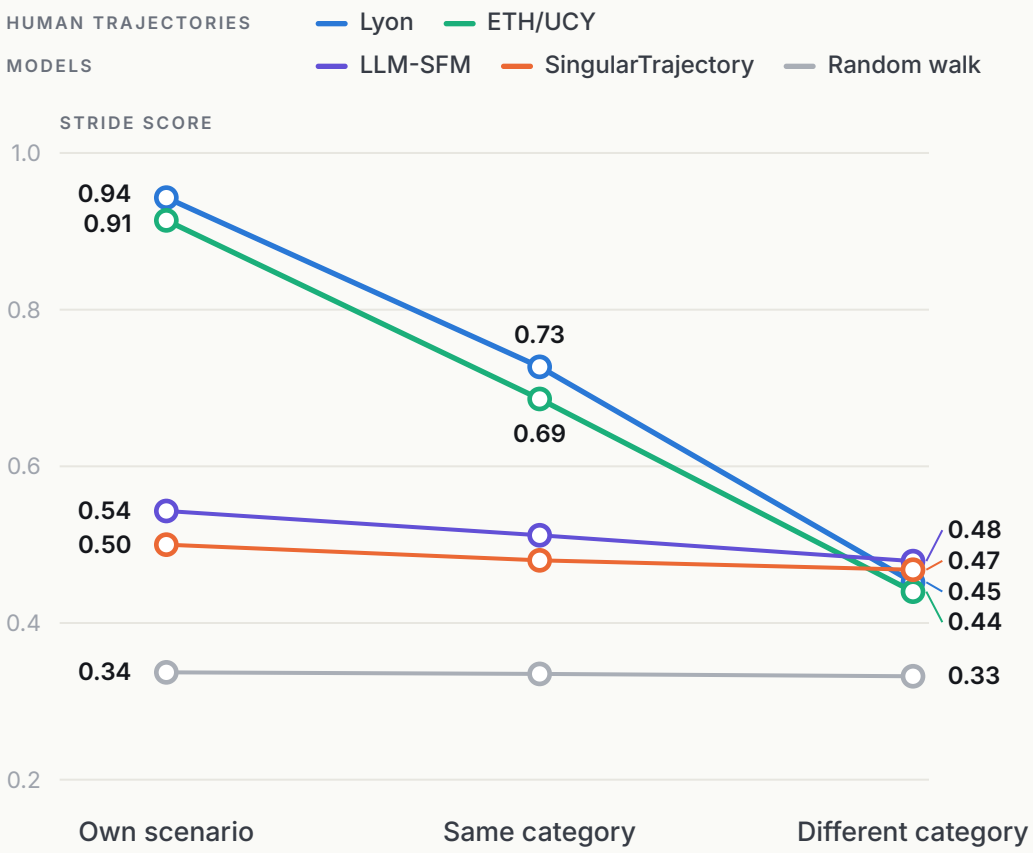


EVALUATION · TEXT-SWAP ABLATION

STRIDE measures context — not just plausible motion

1. Keep the trajectories fixed
2. Swap in another scenario's description
3. Score them again

Real human trajectories lose **about half** of their score under the wrong context. Model scores drop only slightly, and random motion stays flat.



LIMITATIONS

Limitations and future work

Calibration data

Expected ranges are calibrated with real-world event data, which is scarce for rare events such as escape and violence. TrajFacts will grow as more data becomes available.

Measurement coverage

The 20 measurement functions do not yet capture every behavioral factor, such as social relations and group dynamics.

Individual factors

STRIDE-Bench targets macro-level pedestrian scenarios; the effect of individual demographics is not yet covered.

Maps and baselines

Venue maps are simplified to polygons, and baselines run on our maps and text, which may not be their ideal setting.

STRIDE-Bench is built for community-driven extension toward richer scenarios and measurements.

V R D S T

STRIDE

Verifiable, automated and scalable evaluation of text-to-trajectory alignment

Code `github.com/ETH-IVIA-Lab/STRIDE-Bench` **Data** Hugging Face · `eth-ivia-lab/STRIDE-Bench`

Wanchun Ni · Tao Qi · Leonel Aguilar · Marlene Wagner · Jiugeng Sun · Verena Zimmermann · Mennatallah El-Assady

ETH Zurich · Beijing University of Posts and Telecommunications | NeurIPS 2026 · Track on Evaluations and Datasets