GTA-NarrativeTraj: Language-Aware Trajectory Prediction from GPS and Dialogue in an Open-World Simulator

Anastasiia Sapeha, Eduard Sariiev, Mykyta Sapeha, Ibrahim Kovan, Subashkumar Rajanayagam, Kirill Karpov, Maksim Gering, Dmitry Kachan and Eduard Siemens
Anhalt University of Applied Sciences, Bernburger Str. 55, Koethen, Germany
{*anastasiia.sapeha, eduard.sariiev, ibrahim.kovan, subashkumar.rajanayagam, kirill.karpov, maksim.gering, dmitry.kachan, eduard.siemens*}*@hs-anhalt.de*

Keywords: Narrative Trajectory Prediction, Language-Aware Forecasting, Next Location Prediction, Multimodal, GPS, Audio-to-Text, Speech-to-Text, Subtitles Alignment, Dialogue Grounding, Spatio-Temporal Knowledge Graph (ST-NKG), Map Matching, Road Graph, Synthetic Dataset, Urban Environment Simulation, Intent Extraction, Named-Entity Recognition (NER), Ontology, Grand Theft Auto V (GTA V).

Abstract: GTA–NarrativeTraj is presented as a simulation framework and dataset for Grand Theft Auto V (GTA V) that couples spatiotemporal trajectories with in-game narrative signals (speech audio, subtitles, speaker identity). A ScriptHookVDotNet-based logger records world coordinates and vehicle state at ≥ 1 Hz and captures dialogue events (subtitle text, speaker tags, soundbank IDs) during story-mode play. The released dataset provides tightly time-aligned GPS-like traces and the complete dialogue stream for full playthroughs, yielding a resource in which coordinates, audio, and text jointly form a narrative constraining and explaining agent motion. The task of narrative-grounded mobility prediction is introduced: given recent GPS and ongoing utterances, infer the agent's near-term path and next waypoint while recovering salient context such as interlocutors (who is speaking to whom), scene-level locations, and dialogue-implicated points of interest. The dataset serves as ground truth for these tasks by pairing GPS histories with contemporaneous narrative cues and future motion outcomes - enabling models that reason simultaneously over movement, interlocutors, and places. Reproducibility, offset stability, and licensing are discussed; the release includes code, logs, transcripts, and time-aligned audio features, while excluding raw copyrighted assets.

1 INTRODUCTION

In mobility related applications, predicting the agent's future location is central to planning and risk-management. Beyond geometry and maps, language provides complementary signals about intent (destinations and goals), constraints (traffic, risk), and social context (who speaks to whom, who is being obeyed). Yet city-scale datasets that align movement with natural, in-situ dialogue are scarce. Open-world games provide a viable testbed: photorealistic, controllable environments with scripted missions and ambient speech tied to locations and tasks. GTA V, in particular, combines large urban scale with a consistent narrative structure (missions and diegetic dialogue) that can be explicitly exploited for joint reasoning over movement and language.

GTA–NarrativeTraj is a lightweight mod–and–dataset for Grand Theft Auto V (GTA V) that synchronizes spatiotemporal trajectories with in-game narrative signals-speech audio, subtitles, and speaker labels-during story-mode play. A C# ScriptHookVDotNet mod logs world coordinates and vehicle state at ≥ 1 Hz and captures dialogue events (subtitle text, speaker tags, soundbank IDs). A Python pipeline normalizes speaker names and writes structured logs that are map-matched to a lane-aware road graph extracted from the game's path network. The resulting corpus is tightly time-aligned, with GPS-like coordinates, audio, and text jointly forming a narrative that constrains and explains motion. This supports narrative-grounded mobility prediction: given recent coordinates and ongoing utterances, the agent's near-term path and next waypoint can be inferred, along with salient context such as interlocutors, scene locations, and dialogue-referenced points of interest (POIs).

The proposed framework supports a range of tasks: dialogue-conditioned route choice, next-POI retrieval, speech-to-goal grounding, and construction of narrative knowledge graphs anchored to the road

network. The mod, logging server, and dataset-story playthroughs with synchronized GPS, transcripts, speaker tags, and time-aligned audio features-are released to facilitate reproducible research and to encourage models that jointly reason about movement, interlocutors, and places.[1]

This work introduces a narrative-grounded GTA V framework and dataset that couple GPS, audio, and text at ≥ 1 Hz; defines tasks for narrative-aware trajectory forecasting and context recovery (social circle, locations, POIs); presents baseline families contrasting GPS-only and multimodal models; and discusses reproducibility (offset stability, map matching) and legal/ethical considerations.

This paper is organized as follows: Section 2 reviews related work; Section 3 details the system overview and dataset construction; Section 4 outlines future directions enabled by this resource.

2 RELATED WORK

Modern commercial game engines have been repeatedly exploited to produce large-scale, richly annotated data for perception and autonomy research, with Grand Theft Auto V (GTA V) becoming a particularly influential source. Richter et al. showed that high-fidelity scenes in GTA V can yield pixel-accurate labels at scale ("Playing for Data") and later curated a multi-task benchmark across video, optical flow, instance segmentation, 3D layout, odometry, and tracking ("Playing for Benchmarks"), establishing GTA V as a credible platform for controlled urban experiments without costly manual annotation [1, 2]. Building on this idea, the PreSIL dataset introduced synchronized images, depth, and LiDAR with precise point-wise semantics generated in GTA V [3], while community frameworks such as DeepGTAV turned the game into a configurable environment for data collection [4]. In parallel, purpose-built urban simulators provide complementary affordances: CARLA offers open assets, controllable sensor suites, and standardized driving tasks [5], whereas SYNTHIA delivers large synthetic corpora with pixel-level labels for segmentation [6]. GTA V is viewed as trading some controllability for a uniquely scripted narrative structure (missions and ambient dialogue), which is explicitly exploited in this work. This trade-off provides access to story-driven interactions and in-world conversations at city scale, supplying signals that are difficult to script faithfully in conventional simulation.

Concurrently, language has become an increasingly central supervision signal for navigation and route choice. Indoors, the Room-to-Room (R2R) benchmark frames visually grounded instruction following in real buildings [7]; outdoors, Touchdown studies natural-language navigation and spatial reasoning across Manhattan Street View [8]; StreetNav formalizes following driving-style directions in Street View at city scale [9]; and Talk2Nav introduces long-range urban navigation with dual attention over landmarks and local directions [10]. In a driver-centric setting, Talk2Car collects free-form passenger commands grounded in street scenes [11]. Unlike these datasets, in which an agent executes given instructions, the present setting observes in-game dialogue and asks whether such dialogue helps forecast where characters will go next. This observational regime complements instruction-following by emphasizing latent intent and social context rather than prescriptive step-by-step guidance.

A small but growing body of work conditions trajectory modeling directly on language or leverages language models for motion prediction. Kuo et al. quantify the information gain from linguistic representations for vehicle trajectories [12]; Bae et al. reformulate pedestrian prediction via language-styled prompts (LMTraj) [13]; NeurIPS work on language-driven interactive traffic generation conditions trajectories on natural-language prompts [14]; and recent efforts such as LangTraj use language-conditioned diffusion for traffic scene simulation [15]. Our formulation is closest in spirit to these trends but remains distinct in exploiting diegetic dialogue from a scripted open world, aligned to GPS and speaker identity.

Technically, aligning spoken content and subtitles with time is well studied in multimodal corpora that couple utterance-level text with audio and word boundaries. The LRS family (LRS2/LRS3-TED) provides large-scale audio-visual speech data with subtitle alignment for sentence-level recognition and lip reading [16], and the How2 dataset pairs instructional videos with English subtitles (and translations), enabling multi-task speech–language modeling [17]. To convert utterances into actionable structure, standard semantic formalisms are employed: Abstract Meaning Representation (AMR) encodes predicate–argument graphs for sentences [18]; PropBank/semantic role labeling supplies shallow but broad predicate–argument layers [19]; and TimeML with HeidelTime captures temporal expressions and event relations [20, 21].

[1] Repository: GitHub (GTA-NarrativeTraj). https://github.com/mntw/GTA-NarrativeTraj
We release code and logs/metadata, not raw copyrighted assets.

3 SYSTEM OVERVIEW AND DATASET CONSTRUCTION

3.1 Ethics and Legal Considerations

We restrict to single-player and local data collection. We release only code, logs, and derived metadata/features; no redistribution of copyrighted audio files or raw assets. Users must own a copy of GTA V and extract features locally. We document modding limitations and respect the game's End User License Agreement (EULA).

3.2 Modding Pipeline

The single-player GTA V process is instrumented using ScriptHookV/ScriptHookVDotNet; `GTA_Logger` is loaded on the main game thread. As shown in Figure 1, on each tick (configurable, ≥ 1 Hz) the logger reads

1) in-game and wall-clock timestamps;
2) player/Non-Player-Character (NPC) world coordinates and vehicle state;
3) dialogue metadata (subtitle text, resolved speaker tag, current soundbank ID).

Each sample is serialized and sent over HTTP `POST` to a local Python service, which normalizes speaker IDs and appends a CSV record (Data.csv) following the schema in Table 1. This preserves raw world positions (meters) for downstream map matching and graph projection; memory reads should be guarded by signature scanning rather than hard-coded offsets to remain robust across game updates.

Table 1: Output log schema written by the Python logger (Data.csv).

Field	Description
char	Active character identity (string).
time_ingame	In-game clock at capture time (string, HH:MM:SS).
time_rw	Wall-clock timestamp at capture time (string, HH:MM:SS).
pos	World position in GTA V coordinates (string formatted as "x,y,z", meters).
vehicle	On-foot/vehicle flag and, if in vehicle, class and model display name (string triplet: isInVehicle,Class,Model).
subtitle	Subtitle text captured on the tick (string; empty if none).
speaker	Resolved dialogue speaker tag (string).
soundfile	Soundbank or line identifier co-occurring with the subtitle (string).

3.3 Map Graph

A directed road graph $G = (V, E)$ is constructed from the game's pathing data. In Grand Theft Auto V, road paths are stored as nodes (points) and links (connections between two nodes) in `paths.ipl`. Depending on the installation, an XML variant may also be present at `common.rpf\data\levels\gta5\paths.xml`; in addition, Rockstar ships tiled grid squares of path data under `common.rpf\x64e.rpf\levels\gta5\paths.rpf` (each grid has its own area id). Overall, `paths.ipl` contains 74,530 nodes and 77,934 links. Coordinates are floating-point meters in the game's world frame (with North Yankton included in the south–west corner), and node ids are integers indexing into the node list (Figure 2).

Each node record consists of 22 values: (x, y, z) followed by 19 flags in a fixed order summarized in Table 2.

Table 2: Node flag layout (22 values per node: (x, y, z) followed by flags 0–18 in this order).

Idx	Field	Description
0	is_enabled	0 = enabled, 1 = disabled.
1	is_water	0 = land, 1 = water.
3	speed	{0,1,2,3}: coarse speed hint (unknown/slow/medium/fast).
4	type	0 = normal; 10/18 = pedestrian; 14 = interior; 15/16/17 = stop; 19 = restricted.
5	density	Local traffic/pedestrian density prior (normalized as raw/15).
6	street_name_hash	Hash of street/segment name (32-bit).
7	hw_or_int	1 indicates highway or interior.
8	no_gps	{0,1}: GPS disabled through this node.
9	is_tunnel	{0,1}: tunnel indicator.
11	cannot_go_left	{0,1}: left turn prohibited.
12	left_turn_only	{0,1}: left-only turn.
13	off_road	{0,1}: off-road segment.
14	cannot_go_right	{0,1}: right turn prohibited.
15	no_big_vehicles	{0,1}: heavy/large vehicles prohibited.
16	keep_left	{0,1}: keep-left indicator.
17	keep_right	{0,1}: keep-right indicator.
18	slip_lane	{0,1}: slip lane present.

Disabled nodes are discarded; the remaining nodes form V, with all node attributes preserved for

char	time_ingame	time_rw	pos	vehicle	subtitles_text	speaker	soundfile
Franklin	08:30:20	13:40:19	-1953.1,-420.2,-420.2	True,Sports,RAPIDGT	Eh, remember, we got to be careful with these rides, homie. 'Cause Simeon ain't about to dock my pay again...	FRANKLIN	FBI1_CNAA
Franklin	08:28:48	13:40:16	-2005.6,-388.9,-388.9	True,Sports,RAPIDGT			

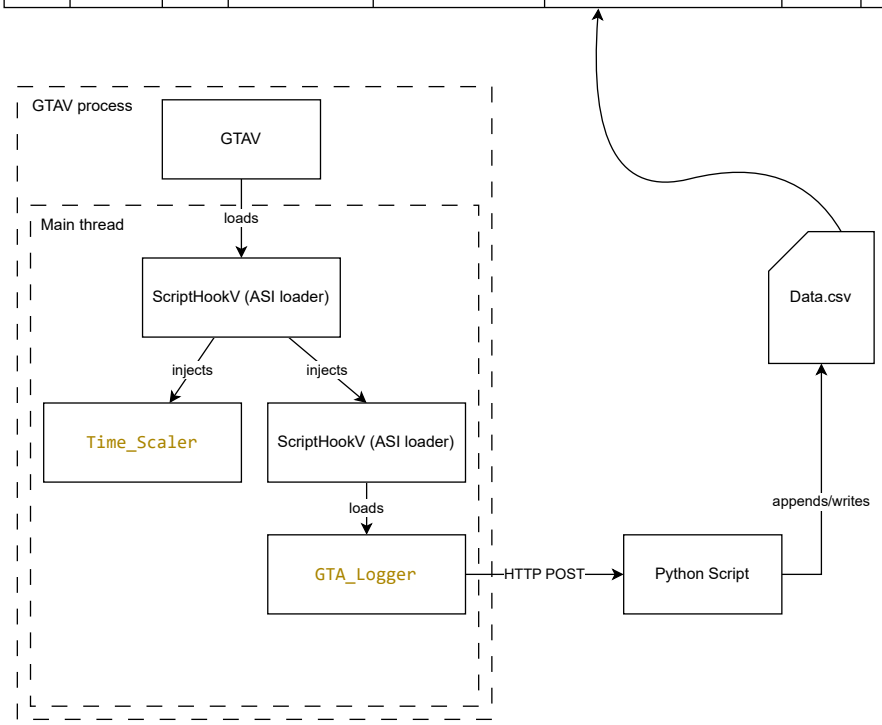

Figure 1: GTA-NarrativeTraj mod architecture.

downstream use (e.g., turn restrictions, tunnel/GPS availability, pedestrian/interior types, density priors). Each link record contains 6 fields — *src*, *dst*, and four flags summarized in Table 3. Directed edges are instantiated from lane counts: if `lanes_in` > 0, the edge $(src \to dst)$ is added with attribute `lanes` $=$ `lanes_in`; if `lanes_out` > 0, the edge $(dst \to src)$ is added with `lanes` $=$ `lanes_out`. Lane values encode one-way/two-way semantics (e.g., `lanes_in` $=$ 2, `lanes_out` $= 0$ denotes a one-way two-lane segment from *src* to *dst*).

Edge lengths are computed as Euclidean distances in (x, y), and the attributes `width`, `lanes`, and `link_code` are retained on edges, yielding a metrized, lane-aware directed graph suitable for routing, map matching, and trajectory forecasting.

3.4 Dataset Description and Statistics

A multimodal dataset was collected from GTA V single-player story-mode gameplay, covering 30 h of real gameplay time dedicated to narrative progression. The corpus aligns three channels: subtitles (text), speech audio (cutscenes and in-world conversations), vehicle telemetry (spatiotemporal trajectories, speeds). The textual channel contains 8,392 non-empty utterances (≈61k tokens); detailed statistics are reported in Table 5.

Vehicle activity is logged with class and model in-

Table 3: Link flag layout (6 values per link: `src`, `dst`, then flags 0–3). Lanes encode one-way/two-way semantics; a common default is `lanes_in`=2, `lanes_out`=0 (one-way, two lanes).

Idx	Field	Description
0	width	{1..10, -1, -10}: nominal segment width/class.
1	lanes_in	{0..6}: lanes from *src* to *dst*.
2	lanes_out	{0..6}: lanes from *dst* to *src*.
3	link_code	(0–5, 8, 9, 10, 17, 18, 19): 8/9 = lane change; 10/17/19 = street change.

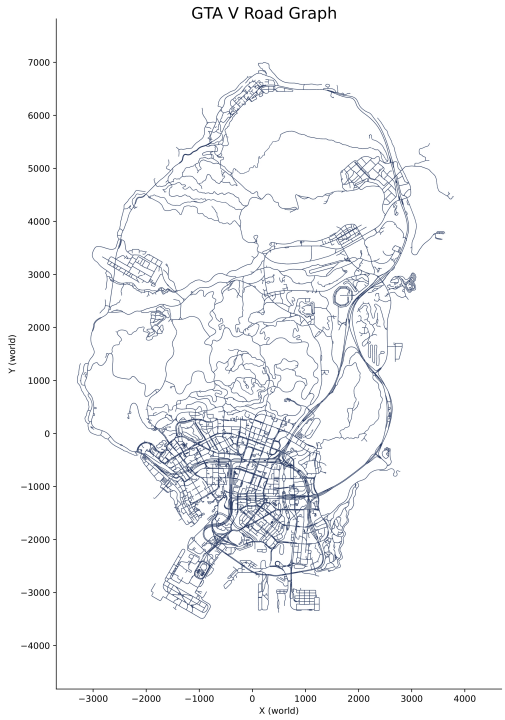

Figure 2: GTA V directed road graph, scale in meters.

formation. The taxonomy includes cars (e.g., super, sports, sedans, muscle, compacts, sports classics, service vehicles), motorcycles, bicycles, air (helicopters, airplanes), water (boats), and trains. Ground-mobility analyses prioritize cars and motorcycles; non-ground classes are considered separately or excluded, depending on task definition.

Trajectories are segmented into rides (trips) associated with the active protagonist under a reproducible policy: onset after a short dwell above a motion threshold in a ground vehicle; termination at sustained near-standstill, player exit, substantial spatial discontinuity, or transition to an excluded class; brief within-ride halts are merged; very short segments are filtered by minimum duration and distance. Aggregates derived from this segmentation yield 132 rides totaling 495.31 km and 38.05 h (in-game) across protagonists (Tables 4 – 5).

Table 4: Trips summary by character (in-game hours).

Character	Distance (km)	Duration (h)	Trips
Michael	205.06	19.34	58
Franklin	147.13	3.19	38
Trevor	143.12	15.52	36
Total	495.31	38.05	132

The spatial footprint of all rides is shown in Figure 3, providing immediate geographic context for subse-quent analyses.

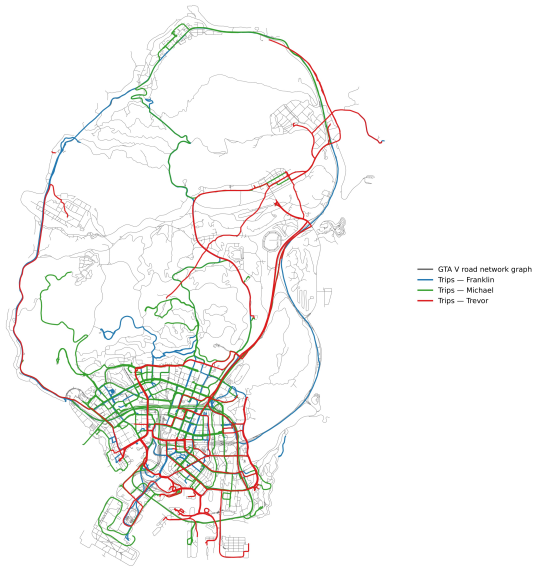

Figure 3: Spatial footprint of all rides by protagonist.

Character coverage reflects the narrative structure: three playable protagonists (Michael, Franklin, Trevor) are tracked alongside 16 primary story characters, with 600+ named entities in the broader universe once minor and episodic NPCs are included. For structural reference, a curated character relationship graph (Figure 4) summarizes inter-character ties with explicit relation labels (e.g., family, friends, colleagues, partners, adversaries, etc.). Beyond listing cast members, the graph encodes direction and type, allowing communities and role clusters to emerge around the protagonists while capturing asymmetric relations. Edge multiplicity and recurrence across missions highlight stable alliances versus transient, mission-specific contacts. This representation supports downstream tasks—entity linking, dialogue attribution, and social role induction—by providing a clean target for comparison against model outputs. In evaluation settings, the labeled ties serve as a validation resource for social-structure or relationship-extraction studies, enabling checks that inferred links are both present and typed correctly in the underlying narrative world.

The audio channel is partitioned into cutscenes and other conversational speech, totaling 2.2 h and 3.2 h, respectively (≈40.9% and 59.1 % of all speech); see Table 5. This split contrasts cleaner scripted material with noisier, overlapping in-world dialogue and enables controlled evaluations in ASR, speaker attribution, and text–audio alignment. Both male and female speakers are represented across multiple age groups.

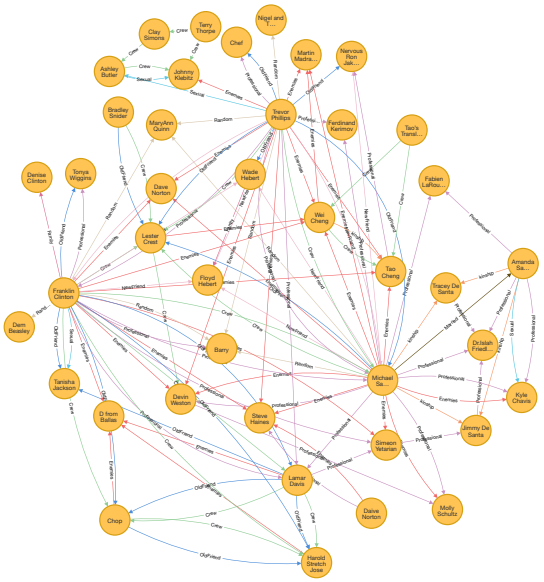

Figure 4: Character social graph. Directed edges encode relation types (e.g., kinship, professional ties, crew membership, enmity) aggregated over story-mode interactions.

Table 5: GTA V dataset - global stats.

Metric	Value
Non-empty utterances	8,392 (≈61,124 tokens)
Sentences	10,642
Playable / main story characters	3 / 16
Total named (game)	600+
Real gameplay time	30 h (real)
Trip duration (sum)	38.05 h (in-game)
Audio: cutscenes	2.2 h
Audio: other speech	3.2 h

4 CONCLUSION AND FUTURE WORK

Speech processing will be extended beyond subtitle capture through integration of robust automatic speech recognition, speaker diarization, and quotation/attribution on in-game audio. Joint parsing and alignment will be developed to reconcile subtitle–audio timing mismatches so that each utterance is reliably attached to the correct speaker and addressee, enabling downstream reasoning about multi-party dialogue and quoted speech.

Geospatial inference will be strengthened with probabilistic and accelerated map-matching, complemented by alternative routing strategies to recover the most plausible paths under noise and occlusion. Destination and route prediction, frequent-pattern mining, next-point-of-interest retrieval, and semantic trajectory representations that couple movement with dialogue-implied intent will be investigated, with sanity checks against public GPS benchmarks to calibrate difficulty and error sources.

A scene-level narrative knowledge graph is planned, linking agents, utterances/addressees, places, missions, and road segments. Graph-based semi-supervised learning and attention message passing will be explored for attribution and for propagating mission goals to candidate waypoints; retrieval-augmented queries (e.g., "Where is the crew likely heading after this line?") will combine trajectory priors with dialogue constraints.

Multimodal pretraining will be examined, including joint audio–text representations with unified masked objectives and geo-acoustic cues (e.g., sirens, surf) for localization. These initializations will support narrative-aware predictors that reason jointly over GPS histories, utterances, and ambient sound with improved data efficiency and robustness to recognition errors.

Dataset scale and use will be broadened through multiple playthroughs with varied driving styles and recording conditions to test cross-run generalization and drift robustness. Mission-level segmentation, POI catalogs, and richer annotations (addressee labels, coreference) will support dialogue-conditioned route choice, next-location retrieval, and checkpoint placement. The pipeline will also be ported to additional open-world environments to assess cross-city transfer while maintaining the same release policy (code, logs, transcripts, and time-aligned features, without raw copyrighted assets).

5 ACKNOWLEDGMENTS

This work was supported by the European Regional Development Fund (ERDF/EFRE) and the State of Saxony-Anhalt within the programme *Sachsen-Anhalt WISSENSCHAFT Forschung und Innovation (EFRE) 2021–2027*, project ReSeDiUm (grant no. ZS/2023/12/182669).

We acknowledge support by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) and the Open Access Publishing Fund of Anhalt University of Applied Sciences.

REFERENCES

[1] S. R. Richter, V. Vineet, S. Roth, and V. Koltun, "Playing for data: Ground truth from computer games," in ECCV, 2016.

[2] S. R. Richter, Z. Hayder, and V. Koltun, "Playing for benchmarks," in ICCV, 2017.

[3] B. Hurl, K. Czarnecki, and S. Waslander, "Precise synthetic image and lidar (presil) dataset for autonomous vehicle perception," in IEEE Intelligent Vehicles Symposium (IV), 2019.

[4] D. Ott and contributors, "Deepgtav: A system to easily extract ground truth from GTAV," https://github.com/David0tt/DeepGTAV, 2018.

[5] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "Carla: An open urban driving simulator," in CoRL (PMLR), 2017.

[6] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, "The SYNTHIA dataset: A large collection of synthetic images for semantic segmentation of urban scenes," in CVPR, 2016.

[7] P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sünderhauf, I. Reid, S. Gould, and A. van den Hengel, "Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments," in CVPR, 2018.

[8] H. Chen, A. Suhr, D. Misra, N. Snavely, and Y. Artzi, "Touchdown: Natural language navigation and spatial reasoning in visual street environments," in CVPR, 2019.

[9] K. M. Hermann, M. Malinowski, P. Mirowski et al., "Learning to follow directions in street view," in AAAI, 2020.

[10] A. B. Vasudevan, D. Dai, and L. V. Gool, "Talk2nav: Long-range vision-and-language navigation with dual attention and spatial memory," International Journal of Computer Vision, 2021.

[11] T. Deruyttere, S. Vandenhende, D. Grujicic, L. V. Gool, and M.-F. Moens, "Talk2car: Taking control of your self-driving car," in EMNLP-IJCNLP, 2019.

[12] Y.-H. L. Kuo and colleagues, "Trajectory prediction with linguistic representations," arXiv:2110.09741, 2022.

[13] I. Bae and coauthors, "Social reasoning-aware trajectory prediction via multimodal language model (lmtraj)," preprint and code, 2024. [Online]. Available: https://github.com/InhwanBae/LMTrajectory

[14] J. Xia and coauthors, "Language-driven interactive traffic trajectory generation," in NeurIPS, 2024.

[15] W. J. Chang and coauthors, "Langtraj: Diffusion model and dataset for language-conditioned trajectory simulation," arXiv:2504.11521, 2025.

[16] T. Afouras, J. S. Chung, and A. Zisserman, "Lrs3-ted: A large-scale dataset for visual speech recognition," 2018.

[17] R. Sanabria, O. Caglayan, S. Palaskar, D. Elliott, L. Barrault, L. Specia, and F. Metze, "How2: A large-scale dataset for multimodal language understanding," arXiv:1811.00347, 2018.

[18] L. Banarescu, C. Bonial, S. Cai, M. Georgescu, K. Griffitt, U. Hermjakob, K. Knight, P. Koehn, M. Palmer, and N. Schneider, "Abstract meaning representation for sembanking," in Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse, 2013.

[19] M. Palmer, D. Gildea, and P. Kingsbury, "The proposition bank: An annotated corpus of semantic roles," Computational Linguistics, vol. 31, no. 1, pp. 71–106, 2005.

[20] J. Pustejovsky, J. M. Castaño, R. Ingria, R. Saurí, R. Gaizauskas, A. Setzer, G. Katz, and I. Mani, "Timeml: Robust specification of event and temporal expressions in text," in AAAI Spring Symposium on New Directions in Question Answering, 2003.

[21] J. Strötgen and M. Gertz, "Heideltime: High quality rule-based extraction and normalization of temporal expressions," in SemEval, 2010, pp. 321–324.