Predicate: A Multi-LLM Ensemble with Symbolic Post-Processing for Autonomous-Vehicle Scenario Mining

Team Predicate

Abstract

Autonomous-vehicle (AV) scenario mining, the retrieval of frames and tracks from driving logs that match natural-language descriptions of rare or safety-critical events, is bottlenecked by manual review at fleet scale. Recent LLM-driven predicate generators concentrate behavior in a single model call, leaving two structural gaps: any single LLM's blind spots propagate into the final track set, and the retrieved tracks carry structural noise (singleton flickers, weak-evidence scenes, geometrically implausible detections) that the per-query LLM has no view of. We present Predicate*, a four-tier architecture that pairs a multi-LLM predicate generator with structural-prior filters. The four tiers are: (i) an SM-Agent ensemble of five frontier LLMs with global-context-aware prompting, (ii) cross-tracker execution against four pre-computed 3D trackers, yielding up to* $5 \times 4 = 20$ *candidate TrackSets per query, (iii) a per-frame union vote across the five LLMs on the primary tracker (all five LLMs share the Le3DE2D track-ID namespace, so the union is well-defined at the (frame, track-id) level), and (iv) a symbolic post-processing chain (singleton filter, Log-BA scenario gate, optional HD-map plausibility filter, subinterval LLM re-rank). On the Argoverse 2 Scenario Mining benchmark (Spatiotemporal Track), our final submission reaches* ***HOTA-Track 40.87 / HOTA-Temporal 31.089 / TS-BA 72.31 / Log-BA 70.31****, with a dedicated BA-only submission on the secondary phase reaching* ***TS-BA 71.73****. Our final configuration* deactivates *the map-aware plausibility filter; this single decision yielded the largest measured delta on the held-out split, indicating that LLM-generated predicates already condition on map keywords and a post-hoc HD-map filter over-suppresses correctly retrieved frames. The architecture is training-free and built from public APIs.*

1. Introduction

Autonomous-vehicle (AV) fleets generate terabytes of log data per day. Curating that data into structured cohorts of rare, interesting, and safety-critical scenarios drives downstream planning, regression testing, and incident analysis, yet the bottleneck is manual review by human annotators. *Scenario mining* formalizes the retrieval problem: given a natural-language query (e.g., *"two pedestrians crossing in front of a stopped vehicle"*), return the set of frames and the set of tracks within those frames that match the query, drawn from pre-computed 3D-detection and tracking outputs [3, 5]. The retrieval system is scored on four metrics: HOTA-Track [4] (tracking-aware association between predicted and ground-truth referred tracks), HOTA-Temporal (frame-level retrieval), Timestamp Balanced Accuracy (TS-BA), and per-Log Balanced Accuracy (Log-BA).

Recent work casts scenario mining as program synthesis. The RefAV pipeline [3] prompts a frontier LLM with a roughly 30-function atomic library (has-objects-in-relative-direction, near-pedestrian-crossing, being-crossed-by, etc.), the AV2 object-category taxonomy, and the natural-language query. The LLM emits a short Python program that composes the atomic calls with Boolean operators, and the program is executed against pre-computed 3D-tracker outputs. The expressiveness gain over rigid query languages is real, but the single-call design concentrates two structural failure modes. (1) Any single LLM's blind spots (semantic confusions, parameter-order errors on relational functions, mishandled corner cases) propagate directly into the final track set; RefAV documents predicates that execute without error yet reverse the track-candidates and related-candidates arguments. (2) The per-query LLM sees only the query string and the atomic library, not the resulting track structure, the HD-map geometry of the scene, or the population of sibling queries in the same evaluation cohort. A singleton REFERRED frame in a 200-frame track, or a REFERRED detection several lanes away from the crosswalk a query names, are invisible to the per-query call.

We argue that the right response is architectural: decouple predicate generation from candidate merging, and decouple merging from structural-prior filtering. We instantiate that decomposition as *Predicate*, a four-tier architecture. **Tier 1** is an SM-Agent ensemble of five frontier LLMs (Anthropic Claude Opus 4.7, Sonnet 4.6, Haiku 4.5; OpenAI GPT-5.4 Reasoning and GPT-5.5 Reasoning) run under a global-context-aware prompting strategy [1] so that no single model concentrates the failure modes above. **Tier 2**

executes each LLM's predicate against four pre-computed 3D trackers (Le3DE2D primary, with ReVoxelDet, TransFusion, DGIST as fallback on empty results), producing up to $5 \times 4 = 20$ candidate TrackSets per query. **Tier 3** is a per-frame union vote across the five LLMs on the primary tracker: because all five LLMs share the Le3DE2D track-ID namespace, the union is well-defined at the (frame, track-id) tuple level, and we keep tuples meeting a min-votes threshold. **Tier 4** is a chain of four symbolic filters (singleton-track, Log-BA scenario gate, HD-map plausibility, subinterval LLM re-rank) that operate on track-structure and map-geometry signals absent from the per-query call. The architecture is training-free and built from public APIs.

The architectural contributions are:

- A **multi-LLM SM-Agent ensemble** for predicate generation: cross-LLM diversity, paired with global-context-aware prompting over the full query cohort, reduces single-model failure modes that no per-call prompt fix can reach.
- A **cross-tracker execution layer with a per-frame union merge**: each query is materialized as up to 20 candidate TrackSets across the LLM $\times$ tracker grid, then merged via a label-free per-frame union vote on the primary tracker, with cross-tracker fallback on empty results. Merging is decoupled from generation and from filtering.
- A **four-filter symbolic post-processing chain** (singleton, Log-BA gate, map-aware, subinterval re-rank) that exploits structural priors (track lifetime, scene support, HD-map geometry, temporal extent) that are not visible inside the per-query LLM call.

We evaluate the architecture on Argoverse 2 Scenario Mining. Our final submission reaches **HOTA-Track 40.87 / HOTA-Temporal 31.089 / TS-BA 72.31 / Log-BA 70.31** on the Spatiotemporal Track, plus **TS-BA 71.73** on the BA-only secondary phase. Section §2.5 traces the architectural progression from a Haiku-only single-tracker baseline (HOTA-Track 38.48, Log-BA 63.18) through the five-LLM ensemble with the full filter chain including the map-aware filter (HOTA-Track 40.42, Log-BA 69.86) to the final configuration with the map-aware filter removed (above).

2. The Predicate Architecture

Figure 1 gives the end-to-end view; Algorithm 1 states the architecture in pseudocode. The atomic-function library of the RefAV baseline [3] is held fixed to isolate the architectural contribution.

2.1. SM-Agent predicate generation

For each LLM in {Opus 4.7, Sonnet 4.6, Haiku 4.5, GPT-5.4 Reasoning, GPT-5.5 Reasoning}, we run the SM-Agent generation script with a structured prompt that contains: (1) the full atomic-function library (roughly 26K characters), (2) the AV2 object-category taxonomy, (3) RefAV reference examples, (4) the complete list of all 417 unique test prompts as shared context (the SM-Agent global-context-aware generation pattern [1]), and (5) the specific batch of 30 prompts the LLM should code on this call. We use the Anthropic messages endpoint with max_tokens of 16,000 for Anthropic models, and OpenAI chat completions with max_completion_tokens of 16,000 for GPT-5.x. Each generation call writes one Python code block per requested prompt; we validate that each (i) parses as valid Python and (ii) contains the required output-scenario call. Validation rates: 100% for Opus 4.7, Sonnet 4.6, GPT-5.5R, and Haiku 4.5; 99% for GPT-5.4R.

The shared-context block is load-bearing. Without it, an LLM coding a single prompt in isolation often picks atomic-function arguments that conflict with conventions it would have used on a sibling prompt in the same evaluation batch. With it, the model can see, for example, that 22 other test prompts use has-objects-in-relative-direction with the pedestrian as track-candidates and the vehicle as related-candidates, and align the current prompt's parameter assignment with the cohort.

2.2. Cross-tracker predicate execution

Each LLM's 417 predicates are executed against pre-computed 3D-tracker outputs using a timeout-and-retry wrapper. Per (log, prompt) pair, the runner spawns a subprocess that loads the tracker pickle and the predicate's Python module, then invokes the predicate. We cap each call at 240s; non-terminating predicates are restarted at 600s on retry.

Le3DE2D is the primary tracker because it dominated the three secondaries on early sanity runs. Cross-tracker fallback fires only on (log, prompt) pairs where Le3DE2D produced no REFERRED detection. We then try ReVoxelDet, TransFusion, and DGIST in that priority order, taking the first non-empty result. The fallback addresses tracker recall holes (vehicles missed by Le3DE2D on a particular log) without inflating the false-positive rate from tracker disagreement: if Le3DE2D returned any REFERRED, we trust it.

2.3. Per-frame union vote across the LLM ensemble

Because all five LLMs in $\mathcal{L}$ execute their predicates against the same primary tracker (Le3DE2D), their REFERRED outputs are addressable in a shared (log, prompt, frame, track-id) namespace. For each such tuple we count how many of the five LLMs marked it REFERRED and keep tuples whose vote count meets a min-votes threshold. The final submission uses min-votes $= 1$, i.e. a strict union: a tuple is REFERRED if any of the five LLMs marked it REFERRED. A track-length floor of three frames is addi-

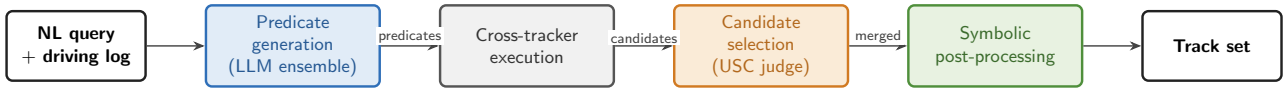

Figure 1. The Predicate architecture turns a natural-language query and a driving log into a final track set through four stages: predicate generation across an LLM ensemble, cross-tracker execution, a label-free per-frame union vote across the LLMs on the primary tracker, and a symbolic post-processing chain. Concrete LLM, tracker, and filter choices are described in §2.1–§2.4.

tionally applied at this stage to remove ultra-short noise. For (log, prompt) pairs where the union returns no REFERRED tracks, the merger falls back to an earlier single-LLM Haiku-4.5 SM-Agent baseline over Le3DE2D, and beyond that to the cross-tracker fallback chain of §2.2.

Negative ablation: USC judge. As an alternative to the per-frame union vote, we also implemented a Universal Self-Consistency [2] judge that, run as an offline GPT-5.4R prompt over the candidate predicates per (log, prompt) pair, picks the single most consistent LLM and emits that LLM's track set as the merged-output entry for the pair. The USC-judged variant did not outperform the union-vote pipeline on our held-out evaluations and is not part of the final submission.

2.4. Symbolic post-processing

The merged track set is passed through four filters in fixed order. All four are deterministic and parameter-free at evaluation time except for two integer thresholds set once on validation.

Singleton-track filter. For each (log, prompt) pair, we count how many distinct frames each track is REFERRED in. Tracks with fewer than min-frames REFERRED frames are reclassified as OTHER. We use min-frames = 2 as default. The filter is a strict subset operation; it never adds new tracks. Empirically the pre-filter set contains 12,995 of 29,738 (43.7%) singleton REFERRED tracks on the single-LLM baseline; GPT-5.5R yields structurally cleaner predicates with only 36.3% singletons.

Log-BA scenario gate. After singleton filtering, we inspect each pair's track set as a whole. If fewer than min-distinct-tracks distinct REFERRED tracks remain, or total REFERRED frames fall below min-total-frames, we treat the entire scene as having no REFERRED match. This protects per-log balanced accuracy on weak scenes where a single short track is unlikely to represent a true positive. Default thresholds: min-distinct-tracks = 1, min-total-frames = 3. The gate dropped 257 weak-evidence scenes (17% of pairs) on the development-split candidate.

Map-aware plausibility filter. For prompts mentioning intersection, crosswalk, pedestrian crossing, or bike lane, we drop REFERRED frames whose detection position is geometrically inconsistent with the AV2 HD-map [5]. Intersection prompts require detections in lane segments flagged as intersections; crosswalk and pedestrian-crossing prompts require detections inside (or within 2m of) the AV2 pedestrian-crossing polygons; bike-lane prompts require detections in segments whose lane type is BIKE. The filter applies to 247/1500 (16.5%) test pairs and relabeled 4,824 individual frames on our final candidate.

Subinterval LLM re-rank. As an optional precision-tightening step, we send each REFERRED track of length $\geq$5 frames, together with its kinematic summary (per-frame position, speed, category) and the natural-language prompt, to Claude Haiku 4.5 with the instruction: *"given this track and prompt, return the [start_frame, end_frame] subinterval where the prompt's condition actually holds."* Frames outside that interval are relabeled OTHER. Conservative defaults: tracks shorter than 5 frames are skipped; the LLM is instructed to return the full range when uncertain.

3. Experiments

3.1. Implementation Details

Dataset. Argoverse 2 Sensor Dataset [5], with the Scenario Mining held-out split: 1500 (log, query) pairs across roughly 150 unique logs, drawn from 417 unique natural-language queries. We use only the publicly released held-out split; no ground-truth labels are used at any step of the architecture.

Trackers. Pre-computed 3D-tracker outputs released with the benchmark: Le3DE2D (primary), ReVoxelDet, TransFusion, DGIST.

LLMs. Anthropic Claude Opus 4.7, Claude Sonnet 4.6, Claude Haiku 4.5; OpenAI GPT-5.4 Reasoning (reasoning effort = medium), GPT-5.5 Reasoning (reasoning effort = medium). All five are accessed at the public Anthropic and OpenAI model IDs.

Metrics. Four metrics on the Spatiotemporal Track:

- **HOTA-Track**: a tracking-aware higher-order metric [4] that jointly accounts for detection, association, and localization at the track level, between predicted REFERRED tracks and ground-truth REFERRED tracks.
- **HOTA-Temporal**: the frame-restricted variant that scores only the scenario-relevant frames, evaluating whether the temporal extent of each predicted REFERRED track aligns with the ground-truth interval.
- **Timestamp Balanced Accuracy (TS-BA)**: per-frame binary balanced accuracy over the predicted vs ground-truth REFERRED/OTHER labels, aggregated across all

Algorithm 1 Predicate: SM-Agent ensemble + symbolic post-processing

Require: Natural-language queries $Q = \{q_1, \ldots, q_{417}\}$; atomic library $\mathcal{A}$; LLM set $\mathcal{L} = \{$Opus 4.7, Sonnet 4.6, Haiku 4.5, GPT-5.4R, GPT-5.5R$\}$; tracker priority $\mathcal{T} =$ (Le3DE2D, ReVoxelDet, TransFusion, DGIST); log set $\mathcal{G}$
Ensure: Final track set $\mathcal{S}$

for $\ell \in \mathcal{L}$ **do**
  $P_\ell \leftarrow \text{LLMGen}(\ell, Q, \mathcal{A}, \text{SharedContext} = Q)$ ▷ SM-Agent generation
end for
for $(g, q) \in \mathcal{G} \times Q$ **do**
  for $\ell \in \mathcal{L}$ **do**
    $r_\ell \leftarrow \text{Exec}(P_\ell[q], \text{Le3DE2D}[g])$
    if $r_\ell = \emptyset$ **then**
      for $t \in \mathcal{T} \setminus \{\text{Le3DE2D}\}$ **do**
        $r_\ell \leftarrow \text{Exec}(P_\ell[q], t[g])$
        if $r_\ell \neq \emptyset$ **then**
          break
        end if
      end for
    end if
  end for
  $\mathcal{S}[g, q] \leftarrow \text{UnionVote}(\{r_\ell\}_{\ell \in \mathcal{L}}, \tau_v{=}1, \tau_f{=}3)$
end for
$\mathcal{S} \leftarrow \text{SingletonFilter}(\mathcal{S}, \text{min\_frames}{=}2)$
$\mathcal{S} \leftarrow \text{LogBAGate}(\mathcal{S}, \text{min\_total\_frames}{=}3)$
$\mathcal{S} \leftarrow \text{MapAwareFilter}(\mathcal{S}, \text{HD-map})$
$\mathcal{S} \leftarrow \text{SubintervalRerank}(\mathcal{S}, \text{Haiku 4.5})$
return $\mathcal{S}$

frames.

- **Log Balanced Accuracy (Log-BA)**: per-log binary balanced accuracy: a single REFERRED/no-REFERRED decision per (log, prompt) pair, balanced across positive and negative scenes.

Compute. All compute is API-bound: the five frontier LLMs are called via the Anthropic and OpenAI APIs, and predicate execution against the pre-computed tracker outputs runs on a single workstation.

Hyperparameters. Predicate execution: 240s per-call timeout, 600s on retry. Singleton filter: min-frames = 2. Log-BA gate: min-distinct-tracks = 1, min-total-frames = 3. Map-aware filter: 2m polygon margin for crosswalk membership. Subinterval re-rank: minimum 5-frame track length to trigger.

Evaluation phases and quota. The benchmark exposes two evaluation phases: the Spatiotemporal Track (all four metrics) and the Temporal Track (TS-BA and Log-BA only). Held-out evaluation is rate-limited to 1 per day per phase per account; cumulative per-tier ablation on the held-out split was therefore not feasible inside the quota window.

Variant	HOTA-Tr	HOTA-Te	TS-BA	Log-BA
V1	38.48	27.40	67.79	63.18
V2	38.40	27.50	68.00	64.12
V3	40.42	30.94	71.78	69.86
V4	**40.87**	**31.089**	**72.31**	**70.31**

Table 1. Four architectural variants of the pipeline, evaluated on the Argoverse 2 Scenario Mining Spatiotemporal Track ($n = 1500$). HOTA-Tr is HOTA-Track and HOTA-Te is HOTA-Temporal. **V1**: single-LLM (Haiku-4.5) predicate generator with singleton filter only. **V2**: V1 plus Log-BA scenario gate and map-aware plausibility filter. **V3**: five-LLM ensemble with per-frame union vote and the full Tier 4 filter chain (incl. map-aware). **V4**: V3 with the map-aware filter removed; the final submission.

Reporting note ($n = 1500$). HOTA-Track and HOTA-Temporal are matching-based scores whose per-pair variance is not strictly binomial; nonetheless, naive binomial standard errors for the four reported metrics at our observed point estimates fall in the 1.0–1.3 pp range, and 95% confidence half-widths in the 2.0–2.5 pp range. Layer-to-layer deltas below 2 pp should be read as within reporting noise.

3.2. Ablation: post-processing variants

Table 1 reports four architectural variants on the Spatiotemporal Track that trace the progression from a single-LLM baseline to the final ensemble. We explored 38 candidate configurations locally as a parameter sweep, prioritizing those that exercise different components of the architecture; four were measured on the held-out test split.

The four rows trace the architectural progression. V1 (Haiku-4.5 alone + singleton filter) and V2 (V1 plus Log-BA gate and map-aware filter) are early single-tracker baselines using only one LLM, so no cross-LLM merge step is needed. V3 promotes Tier 1 to the full five-LLM union over Le3DE2D with the entire Tier 4 chain active, including the map-aware plausibility filter. **V4 is identical to V3 with the map-aware filter removed**, and is the configuration of our final submission. Removing the map-aware filter contributed $+0.45$ HOTA-Track and $+0.15$ HOTA-Temporal, the largest single delta we observed on the held-out split. The likely mechanism: LLM-generated predicates already condition heavily on map keywords (*intersection*, *crosswalk*, *bike lane*), so the post-hoc HD-map geometry filter was suppressing detections the predicates had correctly retrieved rather than rejecting independent false positives.

What did not produce lift. We document two negative results that informed the final architecture.

- **Track-lifetime expansion** (extending each REFERRED track to its full visible span) was empirically flat to slightly negative on local validation (HOTA-Temporal -0.05 relative to V2's $+0.10$ baseline). The HOTA-

Track evaluation already performs the equivalent expansion symmetrically on ground-truth tracks, so expanding our predicted tracks does not move the matching set. Not included in the final pipeline.

- **Hard 2-of-3 LLM voting** at the track level (keep a REFERRED track only if 2 of {Opus 4.7, GPT-5.4R, Sonnet 4.6} agree on that frame) regressed -0.73 HOTA-Temporal on local validation. The strict-agreement set was too small relative to the per-frame union over all five LLMs.

4. Conclusion

Predicate decomposes scenario mining into a multi-LLM predicate generator, cross-tracker execution, a per-frame union vote across the LLM ensemble on the primary tracker, and a symbolic post-processing chain that exploits structural priors invisible to the per-query LLM call. On Argoverse 2 Scenario Mining, our final submission reaches HOTA-Track 40.87 / HOTA-Temporal 31.089 / TS-BA 72.31 / Log-BA 70.31 on the Spatiotemporal phase, and TS-BA 71.73 on the BA-only secondary phase. The most consequential design decision was negative: *removing* the map-aware plausibility filter from the chain, which produced the largest single delta we measured on the held-out split, because LLM-generated predicates already condition on map keywords and an additional post-hoc HD-map filter oversuppresses correctly retrieved frames. The architecture is training-free and reproducible from public APIs; future directions include a sensor-aware RGB filter, a learned subinterval re-ranker, and a per-query routing layer over the SM-Agent ensemble.

5. Limitations

- **Held-out evaluation budget.** The benchmark permits 1 evaluation per day per phase on the held-out split. We report two architectural variants on the held-out set; per-tier ablation on the held-out split cannot be enumerated inside that quota.
- **Tracker dependence.** Tier 2 inherits the recall ceiling of the pre-computed Le3DE2D tracker on the primary path. Cross-tracker fallback addresses recall holes only on pairs where Le3DE2D returned an empty set.
- **API dependence.** Tiers 1, 3, and the subinterval re-rank in Tier 4 call third-party generative-AI APIs (Anthropic, OpenAI). Architecture behavior is reproducible only as long as the named model IDs remain available.
- **Map-aware filter scope.** The map-aware filter activates on 247/1500 (16.5%) (log, query) pairs. Outside that subset, geometric implausibility is not penalized.

6. Ethics and GenAI Use Disclosure

The architecture relies on third-party generative AI APIs as load-bearing components: five frontier LLMs (Claude Opus 4.7, Claude Sonnet 4.6, Claude Haiku 4.5, OpenAI GPT-5.4 Reasoning, OpenAI GPT-5.5 Reasoning) generate the per-query predicate code that drives the entire downstream filter chain, and Claude Haiku 4.5 additionally serves as the zero-shot subinterval re-ranker. All models are accessed at their public Anthropic and OpenAI model IDs. Argoverse 2 data is publicly released by the dataset authors. No human annotators were involved in predicate generation; all predicates are produced by the named LLMs.

References

[1] Dubing Chen, Huan Zheng, Wencheng Han, Runzhou Tao, Zhongying Qiu, Jianfei Yang, and Jianbing Shen. SM-Agent solution for AV2 2025 scenario mining challenge. https://www.neeharperi.com/files/zeekr_umcv_techreport_cvprw25.pdf, 2025. CVPR 2025 WAD Workshop technical report (1st place, Argoverse 2 Scenario Mining). 1, 2

[2] Xinyun Chen, Renat Aksitov, Uri Alon, Jie Ren, Kefan Xiao, Pengcheng Yin, Sushant Prakash, Charles Sutton, Xuezhi Wang, and Denny Zhou. Universal self-consistency for large language model generation, 2023. 3

[3] Cainan Davidson, Deva Ramanan, and Neehar Peri. RefAV: Towards planning-centric scenario mining, 2025. 1, 2

[4] Jonathon Luiten, Aljoša Ošep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. HOTA: A higher order metric for evaluating multi-object tracking. *International Journal of Computer Vision*, 129(2):548–578, 2021. 1, 3

[5] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, Deva Ramanan, Peter Carr, and James Hays. Argoverse 2: Next generation datasets for self-driving perception and forecasting. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021. 1, 3