

Jornadas de Automática

Predicción de trayectorias en cola larga con visión-lenguaje-acción y segmentación

Javier Borau Bernad^{a,*}, Diego Caballero García-Alcaide^a, Raúl Fernández Matellán^a, Martín Santiago Soto^a, José María Armingol Moreno^a, Araceli Sanchis de Miguel^b

^aLaboratorio de Sistemas Inteligentes, Universidad Carlos III de Madrid, Avda de la Universidad, 30, 28911, Leganés, Madrid, España.

^bLaboratorio de Control, Aprendizaje y Optimización de Sistemas, Universidad Carlos III de Madrid, Avda de la Universidad, 30, 28911, Leganés, Madrid, España.

To cite this article: Borau-Bernad, J., Caballero García-Alcaide, D., Fernández Matellán, R., Santiago Soto, M., Armingol Moreno, J.M., Sanchis de Miguel, A. 2026. Long-Tail Trajectory Prediction with Vision-Language-Action Models and Semantic Segmentation. Jornadas de Automática, 47. <https://doi.org/10.17979/ja-cea.2026.47.13800>

Resumen

La conducción autónoma sigue fallando en situaciones raras y críticas, la llamada *cola larga*, donde los modelos genéricos tienen poca cobertura de datos. Este trabajo aborda la predicción de trayectorias condicionada por instrucciones mediante dos familias de métodos, evaluadas con la puntuación multi-maniobra (*Multi-Maneuver Score*, MMS), que premia el seguimiento de la instrucción a través de varios futuros válidos, no la proximidad a una referencia. La primera induce a un modelo de lenguaje visual (VLM) general a emitir waypoints con un *prompt* de razonamiento estructurado, restringiéndolo a la calzada mediante la segmentación semántica (SAM3) y una reparación determinista. La segunda adopta un modelo visión-lenguaje-acción (VLA) específico de conducción, con una selección multi-hipótesis guiada por un sustituto de la métrica y un desempate conservador por viabilidad de calzada. El enfoque VLA eleva la MMS de 4,34 a 5,15 y, con la selección, hasta 5,52, la más alta de la clasificación pública a fecha de redacción.

Palabras clave: Vehículos autónomos, predicción de trayectorias y planificación de rutas, integración de sensores y percepción, aprendizaje y adaptación en vehículos autónomos, aprendizaje automático, sistemas inteligentes de transporte.

Long-Tail Trajectory Prediction with Vision-Language-Action Models and Semantic Segmentation

Abstract

Autonomous driving still fails in rare, safety-critical situations, the so-called *long tail*, where generic models have little data coverage. This work addresses instruction-conditioned trajectory prediction through two families of methods, evaluated with the *Multi-Maneuver Score* (MMS), which rewards following the instruction across several valid futures rather than proximity to a single reference. The first prompts a general-purpose visual language model (VLM) to emit waypoints via structured reasoning, constraining it to the road with semantic segmentation (SAM3) and deterministic repair. The second adopts a driving-specific vision-language-action (VLA) model with multi-hypothesis selection guided by a surrogate of the official metric and a conservative road-feasibility tie-break. The VLA approach raises the MMS from 4,34 to 5,15, nearly halving the L2 error, and selection increases it to 5,52, the highest score on the challenge's public leaderboard at the time of writing.

Keywords: Autonomous Vehicles, Trajectory and Path Planning, Sensor integration and perception, Learning and adaptation in autonomous vehicles, Machine Learning, Intelligent transportation systems.

*Autor para correspondencia: jborau@pa.uc3m.es

Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)

1. Introducción

La conducción autónoma de extremo a extremo ha progresado notablemente al escalar el tamaño de los modelos y de los conjuntos de datos Caesar et al. (2020); Ettinger et al. (2021); Wilson et al. (2023), pero su comportamiento sigue siendo frágil en escenarios de *cola larga*: cruces poco habituales, adelantamientos, zonas de obras o condiciones meteorológicas adversas (lluvia intensa, nieve, conducción nocturna). Estas situaciones son, a la vez, las más relevantes para la seguridad y las menos representadas durante el entrenamiento, por lo que constituyen un banco de pruebas exigente para cualquier sistema de planificación.

Una formulación reciente de este problema es la *predicción de trayectorias condicionada por instrucciones*, que materializa el reto *KITScenes-LongTail* Wagner et al. (2026): a partir del vídeo de la escena, del movimiento reciente del vehículo ego (21 waypoints a 5 Hz, 4,0 s de pasado) y de una instrucción en lenguaje natural —«girar a la izquierda», «adelantar al camión por la derecha» o «seguir recto»—, el sistema debe predecir los siguientes 5 s de movimiento como 25 waypoints a 5 Hz. El reto aporta solo 3 escenarios etiquetados para entrenamiento y 400 sin etiquetar para evaluación, y los puntúa con la *Multi-Maneuver Score* (MMS): en lugar de medir la distancia a una única referencia, compara la predicción con varios futuros de referencia ocultos y premia que se parezca al experto y ejecute la maniobra indicada mientras se aleja de los incorrectos; los empates se deshacen por la coherencia semántica del razonamiento.

En este trabajo mostramos que estas trayectorias pueden predecirse combinando modelos multimodales con restricciones geométricas, y analizamos dos familias complementarias. La primera induce a un VLM de propósito general a emitir waypoints numéricos mediante un razonamiento estructurado, y garantiza trayectorias viables mediante la segmentación semántica y una reparación determinista. La segunda emplea un VLA específico de conducción y desplaza el esfuerzo hacia la selección de la trayectoria emitida según el objetivo real de la tarea. Como apenas hay futuros etiquetados, el error medio de desplazamiento (ADE) sobre el entrenamiento es solo una señal indicativa, y las decisiones se apoyan en señales que no requieren ground truth.

Las aportaciones de este trabajo son: (i) un análisis del diseño del prompt y de los mecanismos de seguridad basados en SAM3 que permiten inducir trayectorias viables con GEMMA4; (ii) la constatación empírica de que la métrica premia la ejecución de la maniobra indicada, lo que justifica optimizar un objetivo sustituto en lugar del ADE; y (iii) un método basado en Alpamayo 1.5 con una selección multi-hipótesis y un desempate por viabilidad de calzada que alcanza una MMS de 5,52, la puntuación más alta de la clasificación pública del reto en el momento de redactar este trabajo, con una ventaja de 0,31 sobre la mejor entrada del resto de equipos.

2. Estado del arte

2.1. Modelos de lenguaje (visual) aplicados a conducción

Los modelos de lenguaje de gran escala y sus variantes multimodales, habitualmente denominados modelos visión-lenguaje (VLM), han demostrado capacidades emergentes de

razonamiento que se han trasladado progresivamente a tareas de conducción autónoma. El *prompting* con cadena de pensamiento (*chain-of-thought*) Wei et al. (2022) mejora el rendimiento en tareas que requieren razonamiento intermedio; GPT-Driver Mao et al. (2023), DriveGPT4 Xu et al. (2024), DriveVLM Tian et al. (2024) y LMDrive Shao et al. (2024) exploran el uso de modelos de lenguaje y visión-lenguaje para la conducción autónoma, y DriveLM Sima et al. (2024) introduce el razonamiento mediante grafos de preguntas-respuestas para conectar la percepción, la predicción y la planificación.

En la primera familia de métodos de este trabajo empleamos GEMMA4 Google DeepMind (2026) como predictor de trayectorias, inducido a emitir waypoints numéricos a partir de la imagen, el estado reciente del vehículo y la instrucción en lenguaje natural. Esta formulación aprovecha la capacidad del VLM para interpretar la escena y seguir instrucciones, pero presenta una limitación importante: un modelo genérico, aun razonando correctamente sobre la situación de conducción, no ha sido entrenado específicamente para traducir dicho razonamiento a la geometría métrica de las trayectorias.

2.2. Modelos visión-lenguaje-acción (VLA)

Frente a los VLM de propósito general, los modelos visión-lenguaje-acción (VLA) están diseñados para producir acciones o planes, no solo descripciones textuales. Esta diferencia es especialmente relevante en la conducción autónoma, donde la salida final debe ser una trayectoria físicamente plausible y coherente con la instrucción.

La familia Alpamayo de NVIDIA Wang et al. (2025) constituye un ejemplo reciente de VLA aplicado a la conducción. Alpamayo combina el razonamiento causal en lenguaje, denominado *Chain of Causation*, con predicción de trayectorias, integrando información visual, navegación y movimiento futuro. Alpamayo 1.5 se utiliza en este trabajo como generador en la segunda fase de la solución, ya que su diseño está más alineado con la producción de trayectorias que el de un VLM generalista. Frente a arquitecturas clásicas de planificación orientada a la conducción Hu et al. (2023) y a enfoques que conectan el lenguaje con representaciones espaciales como Talk2BEV Choudhary et al. (2024), los VLA incorporan el lenguaje natural y el razonamiento explícito dentro del propio proceso de decisión, lo que los hace especialmente atractivos para escenarios de cola larga condicionados por instrucciones.

2.3. Segmentación semántica como restricción de viabilidad

La familia *Segment Anything* proporciona segmentación de propósito general de alta calidad. El modelo original Kirillov et al. (2023) introdujo la segmentación *promptable*, y SAM3 Carion et al. (2025) extiende esta idea hacia la segmentación de conceptos mediante indicaciones textuales o visuales, permitiendo detectar y segmentar instancias de un concepto abierto, como «asfalto», «acera» o «hierba».

En nuestro caso, SAM3 no predice la trayectoria, sino que aporta una señal de *viabilidad de calzada*. Proyectando los waypoints sobre la imagen frontal se comprueba, punto a punto, si la trayectoria permanece sobre la superficie transitable o invade regiones no permitidas. Esta señal visual se utiliza de dos formas: como mecanismo de realimentación y reparación en la fase basada en GEMMA4, y como criterio conservador de desempate en la fase basada en Alpamayo 1.5.

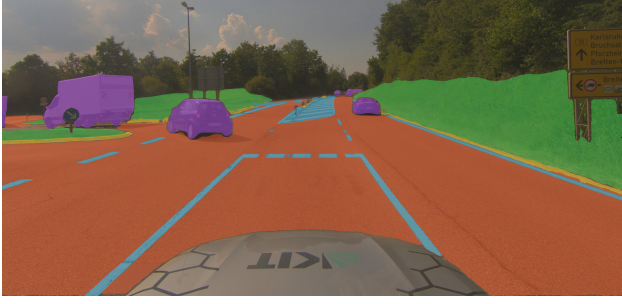


Figura 1: Segmentación de SAM3 sobre un fotograma de KITScenes-LongTail. Las máscaras de superficie (asfalto, acera, hierba, líneas y vehículos) delimitan la región transitable y permiten verificar la trayectoria.

2.4. El conjunto de datos y la métrica de evaluación

El conjunto KITScenes-LongTail Wagner et al. (2026) proporciona vídeo multivista, trayectorias, instrucciones de alto nivel y trazas de razonamiento multilingües anotadas por expertos, centrándose en eventos de cola larga; su partición de entrenamiento solo contiene 3 escenarios etiquetados y la de evaluación, 400. Constituye un banco de pruebas para modelos multimodales, tanto VLM como VLA, que permite evaluar el seguimiento de instrucciones y la coherencia semántica entre las salidas del modelo.

Sus autores proponen la *Multi-Maneuver Score* (MMS), una métrica ligera que puntúa la seguridad, el confort y el seguimiento de instrucciones a través de múltiples futuros válidos, superando las limitaciones de la evaluación por L2 frente a una única trayectoria. La coherencia semántica de las trazas de razonamiento actúa como criterio de desempate. En su estudio base, se observa que la planificación *zero-shot* en escenarios de cola larga es frágil, que el *prompting few-shot* mejora el rendimiento, que el *prompting CoT* en bruto puede producir baja alineación entre el razonamiento y la acción, y que añadir un modelo cinemático de bicicleta sobre las acciones predichas mejora la planificación Wagner et al. (2026). Estos resultados fundamentan varias decisiones de diseño de nuestra solución, como el uso de ejemplos *few-shot*, restricciones cinemáticas, verificación geométrica y trazas de razonamiento.

Como la MMS oficial requiere referencias ocultas, solo puede computarse en el servidor de evaluación. No obstante, el código del reto distribuye un sustituto inspirado en la MMS en el mismo rango $[0, 10]$, calculado como una media ponderada de hasta cinco componentes (Tabla 1). El dato decisivo es que la componente de similitud con la referencia experta, de tipo ADE/FDE, pesa solo 0,10: el resto no necesita referencia anotada y, por tanto, puede calcularse sobre los candidatos de evaluación. Este hecho es la base de la estrategia de selección descrita en la Sección 5.

Tabla 1: Componentes del sustituto inspirado en la MMS. GT: Ground truth

Componente	Peso	¿Necesita GT?
Continuidad	0,25	No
Cinemática	0,25	No
Confort	0,20	No
Instrucción	0,20	No
Similitud experta (ADE/FDE)	0,10	Sí

3. Formulación del problema

A partir de un historial de 21 waypoints a 5 Hz (4,0 s de pasado), una instrucción en lenguaje natural y la información visual de la escena, el sistema debe predecir 25 waypoints a 5 Hz (5,0 s de futuro), donde el primer punto corresponde a $t = 0,2$ s y el último a $t = 5,0$ s. Todas las coordenadas se expresan en el sistema de referencia del vehículo ego, con x hacia adelante, y hacia la izquierda y unidades en metros. Durante la integración se verificó la coherencia de esta convención entre el conjunto de datos, los modelos empleados y las herramientas de evaluación, garantizando que todas las trayectorias se generasen y validasen en el mismo sistema de coordenadas. La predicción puede acompañarse de un objeto de razonamiento con nueve campos de texto que alimenta el desempate en el reto por coherencia semántica.

4. Método I: predicción con modelos de lenguaje visual

La primera familia de métodos trata el problema como una tarea de generación condicionada: GEMMA4 recibe la escena, el estado reciente del vehículo y la instrucción, y emite los waypoints futuros, mientras que un conjunto de mecanismos deterministas verifica que la trayectoria final sea físicamente viable y permanezca sobre la calzada.

4.1. Diseño del prompt de sistema

El prompt de sistema fija el rol del modelo («experto en conducción autónoma»), el formato de salida y, sobre todo, un proceso de razonamiento en cinco pasos que obliga al modelo a anclar su predicción en la geometría visible de la vía antes de emitir cualquier número (Fig. 2). Los cinco pasos son: (1) describir la geometría de la vía (¿cruce, curva, recta?); (2) identificar el límite del asfalto (acera, hierba, bordillo); (3) decidir un perfil de velocidad adecuado; (4) esbozar cualitativamente el movimiento (cuánto avance, cuánto desplazamiento lateral); y (5) solo entonces generar los 25 waypoints numéricos, asegurando que todos permanecen sobre la superficie transitable.

Para este método, hay tres decisiones de diseño clave. Primero, el proceso explícito de cadena de pensamiento fuerza al modelo a razonar sobre evidencia visual antes de numerar. Segundo, una indicación anti-frenado («frena solo si la geometría lo exige») corrige la tendencia, observada en prototipos anteriores, a decelerar en exceso en curvas suaves. Tercero, el uso de etiquetas estructuradas `<reasoning>` y `<trajectory>` hace el *parsing* determinista; cualquier salida sin al menos 25 pares válidos se descarta.

4.2. Mensaje de consulta y ejemplos few-shot

El turno de consulta empaqueta el estado del vehículo, el pasado y la instrucción en un bloque con sintaxis tipo XML. La velocidad se calcula a partir de los últimos waypoints del pasado en lugar de leerse de una etiqueta, lo que hace al modelo robusto frente a discrepancias entre la velocidad anotada y el movimiento real. Antes de la consulta se incluyen tres ejemplos few-shot tomados del conjunto etiquetado, cada uno con su plantilla de estado/pasado/instrucción, un bloque de razonamiento estructurado y la trayectoria experta de referencia. Estos ejemplos anclan al modelo en patrones de conducción humana reales y en el formato de salida esperado, en línea

con el hallazgo de que el prompting few-shot mejora la planificación en la cola larga Wagner et al. (2026).

```
Reason step by step BEFORE generating waypoints:
1. Describe road geometry (intersection? curve?).
2. Identify the asphalt boundary (sidewalk, grass).
3. Decide the speed profile (decelerate? maintain?).
4. Sketch the motion: forward vs. lateral.
5. ONLY THEN emit 25 waypoints, all on asphalt.
Output reasoning in <reasoning>...</reasoning> and
the 25 (x,y) waypoints in
<trajectory>...</trajectory>.
```

Figura 2: Núcleo del proceso de razonamiento en cinco pasos del *prompt* de sistema. El modelo debe anclar su predicción en la geometría visible antes de numerar.

4.3. Restricción de calzada mediante segmentación

La pieza que aporta una señal objetiva es la segmentación con SAM3. Para cada escenario se almacenan en caché máscaras de superficie (asfalto, líneas, hierba, acera, vehículos), de modo que la verificación de calzada no requiere ejecutar segmentación en tiempo de inferencia. Esta señal se emplea de dos formas complementarias:

- Realimentación visual: tras la primera predicción, los 25 waypoints se proyectan sobre la imagen como puntos de color (verde = asfalto, rojo = fuera) y se devuelve al modelo un mensaje que enumera exactamente qué índices son incorrectos, solicitando una trayectoria corregida. El bucle se repite un número limitado de veces y se conserva la trayectoria con menos impactos en zona prohibida.
- Superposición previa: en lugar de esperar a un fallo, la máscara segmentada coloreada se entrega al modelo en la primera llamada, junto con una leyenda del código de colores. Así el modelo dispone de información explícita sobre la superficie transitable desde el inicio, lo que acorta el bucle (media de $\approx 1,3$ llamadas por escenario frente a hasta cinco).

4.4. Reparación determinista y selección de candidatos

Para no depender exclusivamente de que el modelo corrija sus propios errores, introducimos una reparación determinista: los waypoints marcados como fuera de calzada se sustituyen por interpolación lineal entre sus vecinos válidos y el resultado se suaviza con un spline cúbico; la reparación solo se acepta si mejora estrictamente la puntuación de calzada.

Una configuración más elaborada generaliza esta idea a una selección de candidatos: por cada escenario se construyen varias trayectorias (generación primaria, reintento a baja temperatura, líneas base alternativas, reparaciones por interpolación y por ajuste a máscara, variantes con restricción cinemática y un movimiento constante de respaldo) y se conserva la mejor según un orden lexicográfico que prioriza, en este orden, los impactos en zona prohibida y los cruces de línea continua, los puntos fuera de calzada, el déficit de ratio en calzada, el ajuste a la instrucción, la consistencia velocidad-distancia, el confort y la cinemática.

4.5. Trazas de razonamiento

Cada registro incluye un objeto de razonamiento con los nueve campos requeridos, derivados de forma determinista de la trayectoria seleccionada (velocidades por tramo y deriva lateral), de modo que el razonamiento queda alineado con los waypoints emitidos. Esta adición aporta un seguro para el desempate por coherencia semántica sin coste apreciable.

5. Método II: predicción con un modelo visión-lenguaje-acción

El primer método mostró un límite recurrente: un VLM genérico razona bien sobre la escena, pero al convertir ese razonamiento en una trayectoria pierde precisión geométrica. Por eso cambiamos de enfoque y sustituimos el predictor por Alpamayo 1.5, un modelo visión-lenguaje-acción diseñado específicamente para conducción Wang et al. (2025). La diferencia es de fondo: en lugar de pedirle a un VLM que escriba coordenadas, Alpamayo predice trayectorias de forma nativa y acompaña cada una de su razonamiento causal.

5.1. Principio: optimización del objetivo, no de la exactitud

Según se expone en la Sección 2.4, el ADE pesa solo en torno al 10 % de la MMS, mientras que el grueso de la métrica recompensa que la trayectoria ejecute la maniobra indicada. De ahí se sigue una estrategia directa: en vez de emitir una sola predicción, generamos con el modelo un conjunto de hipótesis y nos quedamos con la que maximiza el sustituto inspirado en la MMS. Esto es viable porque la mayoría de las componentes de ese sustituto no necesitan la referencia anotada y pueden calcularse sobre los propios candidatos de evaluación.

La selección termina gobernada, en la práctica, por dos componentes. Las componentes de cinemática y confort están casi saturadas entre las hipótesis del modelo, de modo que apenas sirven para distinguirlas; las que deciden son la continuidad y la instrucción. Conviene además que el término de instrucción se sature ante una maniobra realista: así la selección premia que el coche ejecute la maniobra pedida, pero no recompensa virajes exagerados.

Cualquier intervención que obligue al modelo a seguir la instrucción con más fidelidad mejora la adherencia a la instrucción y eleva el sustituto de la MMS, pero a la vez empeora el ADE respecto a la única referencia anotada. La explicación está en la naturaleza de la cola larga: las maniobras reales suelen ser más sutiles que la instrucción literal. Una orden como «girar a la derecha» puede corresponder, por ejemplo, a una salida suave de una rotonda. En esos casos la predicción libre del modelo queda más cerca de esa referencia concreta, pero sin llegar a ejecutar con claridad la maniobra indicada. Como el conjunto etiquetado es muy reducido, el ADE resulta un selector poco fiable y adoptamos el sustituto como objetivo.

5.2. Desempate conservador por viabilidad de calzada

El sustituto de la MMS valora la continuidad, la cinemática, el confort y la instrucción, pero ignora por completo dónde está la calzada. Esa información es independiente de las demás, así que aporta algo que las otras componentes no ven, y para obtenerla recurrimos a la segmentación. El problema es

que, si pesa demasiado, puede apartarnos de la maniobra correcta, que es justo lo que premia la MMS; por eso la usamos solo para deshacer empates.

El procedimiento es sencillo. Tomamos los candidatos cuyo valor del sustituto está muy cerca del mejor y, entre ellos, elegimos el que mejor se mantiene sobre la superficie transitable: primero el que menos veces pisa acera o hierba y, si hay empate, el que deja menos puntos fuera de la calzada. Como esos candidatos ya eran prácticamente equivalentes según el sustituto, la puntuación de la trayectoria que acabamos emitiendo apenas se mueve, mientras que las invasiones de acera y hierba caen de forma notable. Este paso actúa como un refinamiento conservador: mejora la viabilidad de calzada minimizando el riesgo de degradar la maniobra seleccionada. Dado que solo la cámara frontal está calibrada, la comprobación de calzada se limita a la zona cercana y de frente.

6. Resultados

La Tabla 2 compara las configuraciones de ambos métodos. En la MMS, valores más altos son mejores; en el L2, más bajos son mejores; la coherencia semántica es el desempate.

Dentro del enfoque VLM, la configuración con bucle de realimentación de calzada establece la infraestructura de segmentación con SAM3 y de «permanecer en el asfalto» (MMS 4,20); la variante orientada a eficiencia prioriza el número de llamadas a costa de una ligera pérdida, con su mejor comportamiento en cruces, donde la verificación de calzada resulta más útil; y la selección de candidatos con trazas de razonamiento es la más fuerte del enfoque (MMS 4,34) y la primera con coherencia semántica. El enfoque VLM con GEMMA4 y SAM3 es, por tanto, funcional pero limitado.

El cambio a Alpamayo 1.5 es el punto de inflexión: en muestra única y sin selección, la MMS salta de 4,34 a 5,15 y el L2 medio cae de 4,85 a 3,16. La selección guiada por el sustituto de la MMS la eleva a 5,47 y baja el L2 a 2,50, de modo que la selección mejora a la vez la exactitud y el ajuste de maniobra, no una a costa de la otra. Finalmente, el desempate conservador por viabilidad de calzada con SAM3 lleva la MMS a 5,52 y el L2 a 2,40, con ganancias allí donde la deriva fuera de calzada es más perjudicial: lluvia intensa, adelantamiento e intersección. Esta mejor configuración alcanza la MMS más alta de la clasificación pública del reto¹.

La Fig. 3 ilustra la salida del sistema en dos escenarios de la cola larga: la trayectoria predicha por Alpamayo 1.5 proyectada sobre la cámara frontal, junto con las máscaras de segmentación de SAM3 que verifican la viabilidad de calzada.

La idea de fondo es que importa más optimizar el objetivo de la tarea que la exactitud de la trayectoria. Como el ADE pesa solo $\approx 10\%$ del sustituto y la métrica recompensa la ejecución de la maniobra, lo que más hizo avanzar al enfoque VLA fue la selección hacia ese objetivo, y no tanto el afinado de la predicción en bruto. Esto encaja con un conjunto etiquetado tan pequeño, donde el ADE medido sobre tan pocos ejemplos es un selector poco fiable: por eso apoyamos las decisiones en las componentes del sustituto que no necesitan verdad-terreno y en la realimentación de la métrica oficial.

La elección del modelo apunta en la misma dirección. Para la geometría de trayectorias, uno específico le saca ventaja a uno de propósito general, y de hecho pasar de un VLM inducido a Alpamayo 1.5 ya eleva la MMS en 0,81 y reduce claramente el L2 antes de aplicar cualquier técnica de selección. Por último, las señales de seguridad gratuitas rinden más cuando se usan con cautela: empleada como desempate dentro de un margen, SAM3 limpia la deriva fuera de calzada minimizando el riesgo de degradar el objetivo de maniobra. Y conviene emitir siempre razonamiento, porque protege el desempate de coherencia semántica sin coste apreciable.

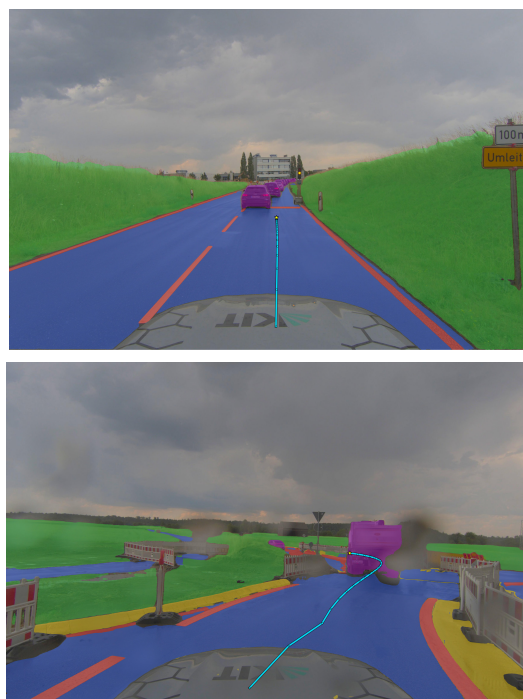


Figura 3: Salida del sistema basado en Alpamayo 1.5, con la trayectoria predicha proyectada sobre la cámara frontal y las máscaras de SAM3. Arriba: tráfico normal con detención ante un semáforo. Abajo: zona de obras. En ambos casos los waypoints permanecen sobre el asfalto, evitando acera y hierba.

7. Conclusión

En este trabajo se ha abordado la predicción de trayectorias condicionada por instrucciones en la cola larga del reto *KITScenes LongTail* a través de dos familias de métodos sucesivas. La primera induce a un VLM de propósito general, GEMMA4, a emitir waypoints mediante razonamiento estructurado, y recurre a la segmentación con SAM3 y a una reparación determinista para asegurar que la trayectoria sea viable. La segunda parte de Alpamayo 1.5, un modelo visión-lenguaje-acción específico de conducción, y traslada el esfuerzo desde generar la trayectoria hacia seleccionarla en función del objetivo real de la tarea.

El resultado más relevante es que en este problema pesa más optimizar el objetivo que la exactitud frente a una sola referencia. Dado que el ADE supone apenas un $\approx 10\%$ del sustituto de la MMS y que la métrica recompensa la ejecución de la maniobra indicada, la mayor mejora no llegó al afi-

¹Clasificación pública del reto KITScenes-LongTail (envío AMv3) consultada el 17/06/2026: MMS =5,52, con una ventaja de 0,31 sobre el siguiente equipo.

Tabla 2: Comparación de las configuraciones de ambos métodos. MMS (↑ mejor), L2 medio (↓ mejor) y coherencia semántica. El cambio al VLA eleva la MMS en 0,81 y reduce claramente el L2; las capas de selección posteriores suben la MMS de forma monótona y terminan reduciendo el L2.

Configuración	Modelo base	Selección / refinamiento	MMS	L2	Coher.
VLM + bucle de calzada (SAM3)	GEMMA4	menos impactos prohibidos	4,20	3,80	0,00
VLM + superposición SAM3 + reparación	GEMMA4	calzada estricta	4,07	4,31	0,00
VLM + selección de candidatos	GEMMA4	multi-candidato lexicográfico	4,34	4,85	0,72
VLA (muestra única)	Alpamayo 1.5	—	5,15	3,16	0,52
VLA + selección por proxy	Alpamayo 1.5	sustituto de la MMS	5,47	2,50	0,54
VLA + selección + desempate SAM3	Alpamayo 1.5	sustituto + viabilidad de calzada	5,52	2,40	0,53

nar la predicción en bruto, sino al seleccionar la hipótesis que mejor cumple ese objetivo. El paso a Alpamayo 1.5 marcó el punto de inflexión, con la MMS subiendo de 4,34 a 5,15 antes de cualquier selección; la selección por sustituto la situó en 5,47 y el desempate conservador por viabilidad de calzada con SAM3 en 5,52, mejorando además el L2 en cada etapa. Esta configuración final ocupa el primer puesto de la clasificación pública del reto a fecha de redacción de este trabajo, superando al anterior líder por 0,31 puntos.

Por último, el recorrido sugiere que, con un conjunto etiquetado tan reducido, las decisiones se sostienen mejor sobre señales que no dependen de la verdad-terreno que sobre el ADE, y que las restricciones geométricas sin coste como SAM3 dan mejor resultado aplicadas con cautela, como refinamiento que limpia la deriva fuera de calzada sin comprometer el objetivo de maniobra. Quedan abiertas varias vías: combinar la selección con un muestreo que refuerce la instrucción ya en la generación, ampliar la comprobación de calzada más allá del cono frontal y promover SAM3 de desempate a término ponderado si la métrica oficial confirma que aporta.

Agradecimientos

Trabajo financiado por la Red Temática IA4MOV (RED2024-153662-T) del Ministerio de Ciencia, Innovación y Universidades (Plan Estatal de I+D+i 2024-2027); por MCI-N/AEI/10.13039/501100011033 (PID2022-140554OB-C32 y PDC2025-165871-C32); por SEGVAUTO 5Gen-CM (TEC-2024/ECO-277) de la Comunidad de Madrid; y por el Programa PIPF-IA de la Universidad Carlos III de Madrid.

Referencias

- Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O., 2020. nusenes: A multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11621–11631. DOI: 10.1109/CVPR42600.2020.01164
- Carion, N., Gustafson, L., Hu, Y.-T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K. V., Khedr, H., Huang, A., et al., 2025. Sam 3: Segment anything with concepts. arXiv preprint arXiv:2511.16719. DOI: 10.48550/arXiv.2511.16719
- Choudhary, T., Dewangan, V., Chandhok, S., Priyadarshan, S., Jain, A., Singh, A. K., Srivastava, S., Jatavallabhula, K. M., Krishna, K. M., 2024. Talk2bev: Language-enhanced bird's-eye view maps for autonomous driving. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). IEEE, pp. 16345–16352. DOI: 10.1109/ICRA57147.2024.10611485
- Ettinger, S., Cheng, S., Caine, B., Liu, C., Zhao, H., Pradhan, S., Chai, Y., Sapp, B., Qi, C. R., Zhou, Y., et al., 2021. Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In:

- Proceedings of the IEEE/CVF international conference on computer vision. pp. 9710–9719. DOI: 10.1109/ICCV48922.2021.00957
- Google DeepMind, 2026. Gemma 4 model card. https://ai.google.dev/gemma/docs/core/model_card_4, accessed: 2026-05-27.
- Hu, Y., Yang, J., Chen, L., Li, K., Sima, C., Zhu, X., Chai, S., Du, S., Lin, T., Wang, W., et al., 2023. Planning-oriented autonomous driving. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 17853–17862. DOI: 10.1109/CVPR52729.2023.01712
- Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al., 2023. Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026. DOI: 10.1109/ICCV51070.2023.00371
- Mao, J., Qian, Y., Ye, J., Zhao, H., Wang, Y., 2023. Gpt-driver: Learning to drive with gpt. arXiv preprint arXiv:2310.01415. DOI: 10.48550/arXiv.2310.01415
- Shao, H., Hu, Y., Wang, L., Song, G., Waslander, S. L., Liu, Y., Li, H., 2024. Lmdrive: Closed-loop end-to-end driving with large language models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15120–15130. DOI: 10.1109/CVPR52733.2024.01432
- Sima, C., Renz, K., Chitta, K., Chen, L., Zhang, H., Xie, C., Beißwenger, J., Luo, P., Geiger, A., Li, H., 2024. Drivelm: Driving with graph visual question answering. In: European conference on computer vision. Springer, pp. 256–274. DOI: 10.1007/978-3-031-72943-0_15
- Tian, X., Gu, J., Li, B., Liu, Y., Wang, Y., Zhao, Z., Zhan, K., Jia, P., Lang, X., Zhao, H., 2024. Drivelm: The convergence of autonomous driving and large vision-language models. arXiv preprint arXiv:2402.12289. DOI: 10.48550/arXiv.2402.12289
- Wagner, R., Tas, O. S., Villa, J., Hauser, F., Shen, Y., Steiner, M., Strutz, D., Fernandez, C., Kinzig, C., Guitierrez-Cabello, G. S., Königshof, H., Immel, F., Schwarzkopf, R., Rack, N. A., Rösch, K., Wang, K., Pauls, J.-H., Lauer, M., Gilitichenski, I., Caesar, H., Stiller, C., 2026. Longtail driving scenarios with reasoning traces: The kitscenelongtail dataset. URL: <https://arxiv.org/abs/2603.23607> DOI: 10.48550/arXiv.2603.23607
- Wang, Y., Luo, W., Bai, J., Cao, Y., Che, T., Chen, K., Chen, Y., Diamond, J., Ding, Y., Ding, W., et al., 2025. Alpamayo-r1: Bridging reasoning and action prediction for generalizable autonomous driving in the long tail. arXiv preprint arXiv:2511.00088. DOI: 10.48550/arXiv.2511.00088
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al., 2022. Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems 35, 24824–24837. DOI: 10.52202/068431-1800
- Wilson, B., Qi, W., Agarwal, T., Lambert, J., Singh, J., Khandelwal, S., Pan, B., Kumar, R., Hartnett, A., Pontes, J. K., et al., 2023. Argoverse 2: Next generation datasets for self-driving perception and forecasting. arXiv preprint arXiv:2301.00493. DOI: 10.48550/arXiv.2301.00493
- Xu, Z., Zhang, Y., Xie, E., Zhao, Z., Guo, Y., Wong, K.-Y. K., Li, Z., Zhao, H., 2024. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. IEEE Robotics and Automation Letters 9 (10), 8186–8193. DOI: 10.1109/LRA.2024.3440097