# الدراسة المرجعية

يعرض هذا الفصل بعض الأبحاث المشابهة لهذا العمل

## مقدمة

إن مجال أنظمة القيادة المعتمدة على النماذج اللغوية الكبيرة مجال جديد، وذلك بسبب التطور الكبير في النماذج اللغوية الكبيرة

التي أظهرت آفاقاً جديدة والتي توفر لنا التفكير المنطقي في المسائل ومعالجة أنواع مختلفة من المعلومات كالصور والنصوص وبالإضافة إلى أهمية التواصل مع المركبات ذاتية القيادة حيث أن التحدث إلى السيارة يساعد في ضمان سلامة الجميع وتزيد من ثقة الناس بها نتيجة إعلام الراكب بالتبرير المنطقي للفعل الذي اتخذته المركبة .

سنستعرض في الفقرات التالية الأفكار التي تم الاعتماد عليها ضمن هذا المجال

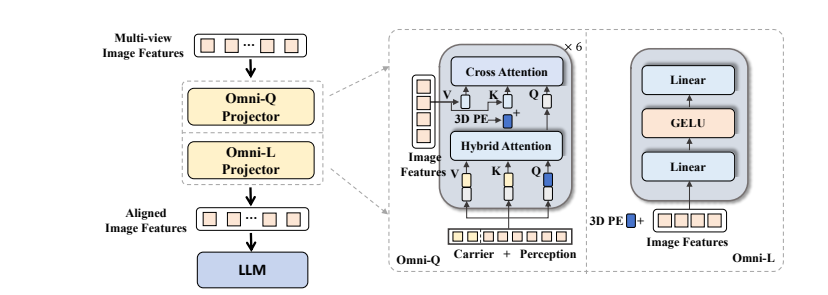
## إطار عمل OmniDrive للقيادة الذاتية

يهدف هذا الإطار إلى تعزيز عملية اتخاذ القرار من خلال تقييم سيناريوهات القيادة المحتملة ونتائجها وذلك من خلال سؤال النموذج اللغوي عن الفعل الواجب اتخاذه ويقوم النموذج بالإجابة والتي هي عبارة عن الفعل الواجب اتخاذه والسبب هذا الفعل.

بدايةً تم فرز البيانات إلى مجموعات بما يضمن اختيار البيانات الأكثر تمثيلاً دلالياً والتي تغطي عناصر حركة المرور الديناميكية والثابتة وكل مجموعة تعكس سلوك قيادة معين مثل الانعطاف يساراً أو التباطؤ....

يتكون هذا الإطار من مكونين أساسيين وهما Omni-Q و Omni-L.

Omni-Q: هو نموذج VLM مع تعديل على Q-Former لتعزيز قدرات تحديد المواقع في نماذج VLM من خلال توظيف أفكار من نماذج الإدراك ثلاثية الأبعاد مثل Stream PETR في بنية Q-Former. مما يحسن الفهم المكاني والربط بشكل أفضل بتوليد اللغة.

Omni-L: هذا المكون مسؤول عن محاذاة اللغة البصرية في النموذج اللغوي حيث يقوم بربط الميزات البصرية متعددة المشاهد بنموذج اللغة الكبير من خلال MLP بسيط لإسقاط الميزات البصرية في مساحة تضمين نموذج اللغة الكبير أي يمكن النموذج من معالجة المعلومات البصرية الواردة من وجهات نظر متعددة في وقت واحد.

تم التدريب على مرحلتين:

* التدريب المسبق ثنائي الأبعاد: يتم تزويد مكونات ربط الصورة بالنموذج اللغوي بفهم أساسي لكيفية ربط المعلومات البصرية بالأوصاف والمفاهيم النصية

البيانات: مجموعات كبيرة من الصور الثنائية الأبعاد المقترنة بنص

الهدف: توليد النصوص استناداً إلى الصور

النتيجة: تهيئة الأوزان للمكونين Omni-L وOmni-Q

* الضبط الدقيق ثلاثي الأبعاد: يتم تكييف النموذج المدرب مسبقاً لفهم تعقيدات البيئات ثلاثية الأبعاد

البيانات: مجموعات بيانات تحتوي على صور متعددة العرض من سيناريوهات القيادة مرفقة بنصوص

الهدف: الإجابة على أسئلة "ماذا لو" أو "ماذا سيحدث" حول مشهد القيادة والتنبؤ بمسار الحركة

النتيجة: يصبح النموذج متخصصاً في فهم المشاهد القيادة ثلاثية الأبعاد والتفكير فيها

## تعزيز وكلاء القيادة الذاتية المعتمدة على النماذج اللغوية الكبيرة من خلال تجارب مجسدة واجتماعية DriVLMe

ظهر هذا البحث لمعالجة مشاكل التعقيد الحقيقي للقيادة في العالم الواقعي الذي لم تعالجه الدراسات السابقة والتي تعالج مواقف بسيطة ومحدودة فقط ولم تعالج مشكلة التعامل مع الرحلات الطويلة أو الانخراط في محادثات طبيعية ومرنة مع الركاب أو التعامل مع الأحداث المفاجئة مثل تغيير الراكب لوجهته في منتصف الرحلة.

طور الباحثون هذا النموذج DriVLMe لفهم كل ما تراه السيارة وما يقوله البشر بحيث يمكن التواصل الفعال بين البشر والمركبة مما يساعد المركبة على إدراك محيطها والتنقل

يتكون وكيل DriVLMe من ثلاثة أجزاء رئيسية تعمل معًا:

مُرمِّز الفيديو: يُشبه "عيني" الوكيل. يأخذ لقطات فيديو خام لما يراه الوكيل ويحولها إلى تمثيل رقمي مُدمج يفهمه "الدماغ".

العمود الفقري لـ LLM (نموذج اللغة الكبير): يُشبه "دماغ" الوكيل يُعالج المعلومات المرئية (من مُرمِّز الفيديو) وأي تعليمات حوارية (مثل "اذهب إلى المكتبة") لاتخاذ القرارات.

وحدة تخطيط المسار: تُشبه "ملّاح GPS" للوكيل. عندما يحتاج الوكيل إلى الوصول إلى معلم مُحدد، تُحدد هذه الوحدة أفضل مسار باستخدام خريطة.

لنتعمق في كل جزء:

### 1. مُرمِّز الفيديو

الغرض: أخذ سجل لما شاهده الوكيل (سلسلة من إطارات الفيديو) وتحويله إلى صيغة يستطيع النموذج اللغوي معالجتها.

خطوات المعالجة:

أخذ العينات: لا يستخدم كل إطار على حدة. بل يأخذ عينات آخر 40 إطارًا من السجل الحديث، هذا يُنشئ مقطع فيديو V سهل الاستخدام، ومن ثم يتم استخدام CLIPViT-L من أجل استخراج خريطة الميزات

يُحوّل هذا المرمز كل إطار إلى شبكة من الأرقام ، حيث تحتوي كل نقطة في الشبكة على D رقم يصف ذلك الجزء من الصورة، ثم يتم إنشاء نوعين من التضمينات:

* تضمين مُركّز مكانيًا Vs: يأخذ خريطة الميزات لكل إطار ويحسب متوسط ​​الميزات عبر الزمن لكل موقع مكاني. هذا يُعطي تمثيلًا يُبرز مكان وجود الأشياء، بمتوسط ​​40 إطارًا.
* ترميز مُركّز زمنيًا Vt: يأخذ خريطة الميزات ويحسب متوسط ​​المعالم عبر الفضاء لكل إطار. هذا يُعطي تمثيلًا يُبرز كيفية تغير الأشياء بمرور الوقت، بمتوسط ​​المواقع المكانية لكل إطار من إطارات.

يتم دمج التمثيلين للحصول على فديو مدمج V=(Vs,Vt) ليصبح قادراً على التقاط الديناميكيات الزمنية والتفاصيل المكانية ويتم استخدام طبقة إسقاط بسيطة لتحويل بعد الفيديو إلى البعد الذي يمكن للنموذج اللغوي استقباله

### 2. العمود الفقري لنظام إدارة التعلم (LLM)

الغرض: هذه هي وحدة صنع القرار المركزية. تستقبل معلومات الفيديو المُعالجة، وتعليمات الحوار (مثل أهداف المستخدم)، وتُقرر الإجراءات البسيطة التي يجب اتخاذها (مثل "الانعطاف يسارًا"، "التسارع").

تم استخدام Vicuna-7B وهو نموذج لغة مفتوح المصدر ومُحدد.

### ٣. وحدة تخطيط المسار (ملاح GPS)

الغرض: تزويد العميل بخطة تفصيلية (مسار) للوصول إلى الوجهة، باستخدام معرفة الخريطة، يُعد هذا أمرًا بالغ الأهمية لتحقيق الأهداف بعيدة المدى.

حيث تأخذ هذه الوحدة الوجهة كدخل والموقع الحالي والخريطة وتقوم بحساب أقصر مسار ثم يُترجم هذا المسار إلى قائمة اتجاهات الانعطاف بلغة طبيعية (مثل: "انعطف يسارًا عند التقاطع التالي"، أو "استمر مستقيمًا لمسافة مبنيين ثم انعطف يمينًا").

### كيفية عمل الوحدات الثلاثة

يراقب مُرمّز الفيديو البيئة ويُحوّل الفيديو إلى خصائص رقمية، والتي بدورها تُنقل إلى النموذج اللغوي ليقوم بمعالجتها بالإضافة إلى التعليمات الحوارية وفي حال كانت الوجهة بعيدة فقيوم النموذج باستخدام GPS واستدعاء وحدة تخطيط المسار والذي يقوم بتحديد أقصر طريق ويعيد قائمة بالاتجاهات المبسطة مثل مستقيم، يسار، مستقيم، يمين... ويقوم النموذج باستخدام هذه التوجيهات بالإضافة إلى المدخلات المرئية اللحظية من مرمز الفيديو لزيادة دقة القيادة مثل أسرع قليلاً أو أدر عجلة القيادة يساراً الآن

يستخدم نظام إدارة المركبات الآن هذه الاتجاهات، بالإضافة إلى المُدخلات المرئية اللحظية من مُرمز الفيديو، لاتخاذ قرارات قيادة دقيقة وبسيطة (مثل "أدر عجلة القيادة يسارًا الآن"، "سرّع قليلاً")، مما يساعد في التنقل الصحيح حتى لو كانت الوجهة بعيدة.

### التدريب

تم التدريب على مرحلتين

المرحلة الأولى: تم إنشاء الأوزان لطبقة الإسقاط بين النموذج اللغوي و مرمز الفيديو وذلك من خلال 50000 فيديو- نص للقيادة مع توصيف الأفعال المتخذة مثل التوقف والانعطاف... .

المرحلة الثانية:

* ضبط التعليمات الاجتماعية: تم ضبط النموذج اللغوي باستخدام LoRA وذلك من خلال 13000 فيديو- حوار والتي تتضمن:
  + فيديو لسيناريو قيادة وحوار مرتبط به.
  + حوارات بين الإنسان والسيارة: حوارات بين الإنسان والسيارة/الوكيل ، على سبيل المثال: "انعطف يسارًا عند التقاطع التالي"، أو "ما هو حد السرعة هنا؟"
  + أهداف طويلة المدى للمخططين: تعليمات تُحدد هدفًا عامًا، مثل "القيادة إلى السوبر ماركت". يحتاج النموذج إلى تعلم كيفية تقسيم هذا الهدف إلى أفعال أصغر.

مما يجعل النموذج اللغوي يفهم ما يريده الإنسان من المحادثة والهدف العام، ثم تحديد الإجراءات التدريجية (الإجراءات الأولية) اللازمة لتحقيقه.

* ضبط التعليمات المجسد :تم إنشاء بيانات تدريب جديدة من خلال محاكي لإدراك ما تراه السيارة عن محيطها المباشر حيث تم تسجيل البيانات مثل الطقس والأجسام القريبة وبعدها والبعد عن نهاية الطريق الحالي وبعد إشارات المرور وحالتها مثل حمراء أو خضراء ووضعوها في قوالب نصية.

أي في المرحلة الأولى تم تعلم كيفية تحويل الفيديو إلى نص وفي الجزء الأول من المرحلة الثانية يتعلم النموذج اللغوي كيفية فهم النوايا البشرية من خلال المحادثة وفي الجزء الثاني يجعل النموذج اللغوي يرسخ فهمه في الواقع الفعلي للقيادة من منظور السيارة

## التحديات والمشاكل

على الرغم من الأداء الجيد لكل من البحثين السابقين لكن لم يكن قادراً على حل المشاكل التالية:

* وقت استدلال غير مقبول: يستغرق الذكاء الاصطناعي وقتًا طويلاً لمعالجة المعلومات واتخاذ القرارات/الاستجابة. وهذا يضر بالقيادة الآنية.
* بيانات تدريب غير متوازنة: قد لا تغطي البيانات المستخدمة لتعليم الذكاء الاصطناعي جميع المواقف بالتساوي، مما يؤدي إلى تحيزات أو ضعف الأداء في السيناريوهات غير الممثلة تمثيلاً كافيًا.
* فهم بصري محدود: لا "يرى" الذكاء الاصطناعي العالم المرئي أو يفسره بشكل مثالي؛ وقد يغفل أو يسيء فهم الإشارات البصرية المهمة.
* تحديات التفاعلات متعددة الأدوار: يواجه صعوبة في المحادثات الطويلة والمتبادلة، مما قد يؤدي إلى فقدان السياق أو الاتساق.