بسم الله الرحمن الرحيم

دانشگاه شهید چمران اهواز

دانشکده مهندسی

پایان نامه کارشناسی ارشد

گرایش هوش مصنوعی و رباتیک

عنوان:

طراحی و ساخت سیستم هوشمند ارائه توضیحات متنی در خصوص تصمیمات وسایل نقلیه خودران

استاد راهنما:

دکتر محمد جواد رشتی

استاد مشاور:

دکتر سید عنایت اله علوی

نگارنده:

امید مسلمانی ۴۰۱۸۶۶۰۷

پاییز ۱۴۰۴

تقدیم به آنان که مهرشان راهگشا و حضورشان تکیه‌گاه بود؛

پدر و مادر عزیزم،

که با حمایت‌های بی‌دریغ و زحمات خالصانه‌شان، مسیر این راه را برایم هموار ساختند. این دستاورد، ثمره‌ی صبر و فداکاری شماست.

و برادر گرامی‌ام،

که تلاش و پشتکارش در مسیر زندگی، الهام‌بخش راهم، و هم صحبتی‌های صمیمانه‌اش، قوت قلبم بود.

از صمیم قلب، قدردان حضور پرمهرتان هستم.

لازم می‌دانم مراتب سپاس و قدردانی خود را از استاد راهنمای گرانقدر، جناب آقای دکتر محمد جوادرشتی، اعلام نمایم.

راهنمایی‌های دلسوزانه، نکات دقیق و حمایت‌های علمی ایشان در تمامی مراحل انجام این پایان‌نامه، نقشی بی‌بدیل در به ثمر رسیدن آن داشت. دانش، صبر و نگاه هوشمندانه ایشان، همواره الهام‌بخش و راهگشای من در این مسیر پژوهشی بود. برای ایشان از صمیم قلب، سلامتی و توفیق روزافزون آرزومندم.

همچنین بر خود لازم می‌دانم از استاد مشاور گرانقدر، جناب آقای دکتر سید عنایت اله علوی، صمیمانه سپاسگزاری نمایم. راهنمایی‌های ارزشمند، دقت علمی و آموزه‌های سازنده ایشان در طول دوران تحصیل و مراحل انجام این پژوهش، نقش بسزایی در ارتقای دانش و بینش علمی اینجانب داشته است. از درگاه خداوند متعال، سلامتی و کامیابی روزافزون برای ایشان مسئلت دارم.

چکیده

نام خانوادگی: مسلمانی	**نام: امید**	**شماره دانشجویی : ۴۰۱۸۶۶۰۷**
عنوان: طراحی و ساخت سیستم هوشمند ارائه توضیحات متنی در خصوص تصمیمات وسایل نقلیه خودران		
استاد/ اساتیدراهنما: دکتر محمدجواد رشتی		
استاد/ اساتید مشاور: دکتر سید عنایت اله علوی		
درجه تحصیلی: کارشناسی ارشد	**رشته: مهندسی کامپیوتر**	**گرایش: هوش مصنوعی و رباتیکز**
دانشگاه : شهید چمران اهواز	**دانشکده: مهندسی**	**گروه :کامپیوتر**
تاریخ فارغ التحصیلی :		**تعداد صفحه: ۱۷۵**
کلید واژه ها : هوش مصنوعی توضیح‌پذیر، مهندسی دستور مدل‌های زبانی بزرگ، نقشه توجه		

چکیده: با گسترش روزافزون سامانه‌های خودران، شفافیت و توضیح‌پذیری تصمیم‌های این سامانه‌ها به یک ضرورت حیاتی برای جلب اعتماد کاربران، افزایش ایمنی و ارزیابی منصفانه عملکرد تبدیل شده است. این پژوهش یک سامانه ترکیبی پیشنهاد می‌دهد که در آن مدل تصمیم‌گیر ADAPT برای پیش‌بینی کنش رانندگی به‌کار گرفته می‌شود و نقشه‌های توجه آن به‌عنوان پلی میان ورودی تصویری و ماژول توضیح‌گر عمل می‌کنند. در گام بعدی، مدل‌های زبانی بزرگ از جمله خانواده Gemini و GPT ، با دریافت تصمیم نهایی، نقشه توجه و دستور طراحی‌شده، توضیحات متنی کوتاه و قابل‌درک برای انسان تولید می‌کنند. نوآوری اصلی این رویکرد در بهره‌گیری از دانش گسترده LLMهاست که فراتر از داده‌های محدودِ مجموعه داده عمل کرده و توانایی تولید توضیحاتی دقیق‌تر و متنوع‌تر را فراهم می‌سازد .ارزیابی‌ها بر روی مجموعه داده BDD-X و با معیارهای استاندارد حوزه تولید متن نظیر BLEU-٤ و CIDEr-D نشان می‌دهد که رویکرد پیشنهادی مبتنی بر مدل زبانی تنظیمِ‌دقیق‌شده برتری معناداری نسبت به روش‌پایه ADAPT دارد؛ برای مثال، در سناریوی تک مرجع، معیار CIDEr-D از ۱۰۲.۶ به ۱۱۴.۷۳رسید که نشان‌دهنده بهبود ۱۱٪ است. این بهبود در سایر معیارها نیز مشهود بود و ثابت کرد توضیحات تولیدی، هم از نظر زبانی روان‌تر و هم از نظر محتوایی منطبق‌تر با دلایل تصمیم هستند. در مجموع، پژوهش حاضر با ارائه سازوکاری مؤثر برای پیوند توجه بصری با توانایی زبانی LLM ها، گامی در جهت توسعه سامانه‌های خودران توضیح‌پذیرتر و قابل‌اعتمادتر برداشته است.

فهرست مطالب

اهدانامه ..... ب
قدردانی ..... ج
چکیده ..... د
فهرست مطالب ..... ه
فهرست جداول ..... ی
فهرست تصاویر ..... ل

فصل اول

مقدمه ۱

۱–۱– بیان مساله ۱
۱–۲– اهمیت و ضرورت تحقیق ۵
۱–۳– اهداف تحقیق ۸
۱–۳–۱– هدف کلی ۸
۲–۳–۱– اهداف جزئی ۸
۱–۴– نوآوری‌های تحقیق ۹
۱–۵– روش تحقیق ۱۰
۱–۶– نو آوری های تحقیق: ۱۳
۱–۷– ساختار پایان‌نامه ۱۴

فصل دوم

پیش زمینه تحقیق ۱۶

۲–۱– کلیات فصل ۱۶
۲–۲– مفاهیم کلیدی و اصطلاحات ۱۷
۲–۳– مبانی نظری تحقیق ۱۹
۲–۳–۱– هوش مصنوعی توضیح‌پذیر ۱۹

۲–۳–۲– مدل‌های زبانی بزرگ و تفسیر تصمیمات ۲۲
۲–۳–۳– مهندسی دستور ۲۷
۲–۳–۴– مدل‌های تصمیم‌گیر در سیستم‌های خودران ۳۵
۲–۴– پیشینه تحقیق..... ۴۳
۲–۴–۱– پژوهش های داخلی مرتبط..... ۴۳
۲–۴–۲– پژوهش‌های خارجی مرتبط..... ۴۸
۲–۵– جمع‌بندی فصل دوم ۵۸

فصل سوم

روش پیشنهادی ۶۱

۳–۱– کلیات فصل..... ۶۱
۳–۲– معرفی داده‌ها و ابزارها..... ۶۲
۳–۲–۱– معرفی مجموعه داده BDD-X..... ۶۲
۳–۲–۲– ابزارها و کتابخانه‌های مورد استفاده..... ۶۳
۳–۲–۳– مدل‌های زبانی مورد استفاده..... ۶۴
۳–۳– پیش پردازش داده ها..... ۶۵
۳–۳–۱– مراحل پیش‌پردازش ویدیو بر اساس مقاله ADAPT..... ۶۵
۳–۳–۲– پردازش زیرنویس‌ها و تولید زیرنویس‌های افزوده با کمک LLM..... ۶۹
۳–۴– طراحی مدل تصمیم‌گیر..... ۷۱
۳–۴–۱– معماری ADAPT و نحوه استفاده آن در تحقیق حاضر..... ۷۲
۳–۵– طراحی ماژول توضیح‌گر..... ۷۴
۳–۵–۱– نحوه استفاده از attention خروجی ADAPT..... ۷۶
۳–۵–۲– طراحی دستور برای تولید توضیح تصمیم..... ۷۸
۳–۵–۳– معماری کلی سیستم ترکیبی تصمیم‌گیر-توضیح‌گر ۸۰
۳–۵–۴– جمع‌بندی فصل سوم..... ۸۲

فصل چهارم

پیاده سازی و تحلیل نتایج ۸۴

۴–۱– کلیات فصل..... ۸۴
۴–۲– بازتولیدپذیری و محیط اجرا..... ۸۶

۴–۲–۱– پیکربندی سخت‌افزاری/نرم‌افزاری ۸۶
۴–۳– جزئیات دقیق داده و پیش‌پردازش ۸۸
۴–۳–۱– آمار مجموعه داده BDD-X و تقسیم‌بندی دقیق (train / val / test) ۸۸
۴–۳–۲– نمونه‌برداری زمانی و تبدیل‌های Resize/Crop/Normalize ۹۰
۴–۳–۳– پارامترهای تبدیل attention به کانتور قرمز (τ، σ، K، ضخامت) ۹۱
۴–۴– تنظیمات آموزش/استنتاج ۹۳
۴–۴–۱– آموزش ADAPT (روش پایه) ۹۳
۴–۴–۲– استنتاج مدل‌های زبانی ۹۴
۴–۴–۳– بسته‌بندی ورودی‌ها برای LLM ۹۶
۴–۴–۴– باز آموزی ADAPT با زیرنویس های افزوده و تحلیل همگرایی ۹۶
۴–۵– دستور و واریانت‌ها ۹۸
۴–۵–۱– متن کامل دستور ثابت و منطق طراحی آن ۹۸
۴–۵–۲– تنظیمات تولید متن ۱۰۱
۴–۵–۳– آزمایش‌های Fine-Tuning: پروتکل و پیاده‌سازی ۱۰۲
۴–۶– پروتکل ارزیابی ۱۰۴
۴–۶–۱– معیار های ارزیابی ۱۰۴
۴–۶–۲– پیاده‌سازی معیارها و نسخه ابزارها ۱۰۶
۴–۶–۳– دو سناریوی مرجع: تک‌رفرنس و چندرفرنس ۱۰۶
۴–۶–۴– رویه اجرا، تکرار آزمایش‌ها و نکات آماری ۱۰۸
۴–۷– نتایج روش پایه (ADAPT) ۱۰۹
۴–۷–۱– نمرات در سناریوهای تک مرجع و چند مرجع ۱۰۹
۴–۷–۲– بحث کوتاه درباره روش پایه و همخوانی با مرجع ۱۱۰
۴–۸– نتایج مدل‌های زبانی و مقایسه ۱۱۱
۴–۸–۱– مدل Gemini ۲٫۵ Pro ۱۱۲
۴–۸–۲– مدل Gemini ۲٫۵ Flash ۱۱۳
۴–۸–۳– مدل Gemini ۲٫۵ Flash-Lite ۱۱۳
۴–۸–۴– مدل GPT-۵ Mini ۱۱۴
۴–۸–۵– مقایسه تجمیعی با ADAPT در هر دو سناریو ۱۱۴
۴–۹– مطالعه حذف مؤلفه‌ها و تحلیل حساسیت ۱۱۵
۴–۹–۱– اثر تعداد فریم‌های ورودی (n-fractions) بر معیارها ۱۱۵
۴–۹–۲– اثر افزوده‌شدن مراجع LLM (چندمرجع در برابر تک‌مرجع) ۱۱۶
۴–۹–۳– اثر Fine-Tuning ۱۱۷
۴–۱۰– منابع محاسباتی، زمان اجرا و هزینه ۱۲۴
۴–۱۰–۱– جمع‌بندی کمّی منابع و زمان ۱۲۴

۴–۱۰–۲– محرک‌های هزینه و راهبردهای کنترل هزینه ۱۲۵
۴–۱۰–۳– محدودیت‌های مرتبط با منابع ۱۲۶
۴–۱۰–۴– ملاحظات مقیاس‌پذیری ۱۲۷
۴–۱۱– محدودیت‌ها و تهدیدهای اعتبار نتایج ۱۲۸
۴–۱۱–۱– اعتبار سازه‌ای ۱۲۸
۴–۱۱–۲– اعتبار درونی ۱۲۹
۴–۱۱–۳– اعتبار بیرونی و تعمیم‌پذیری ۱۲۹
۴–۱۱–۴– اعتبار استنباطی ۱۳۰
۴–۱۱–۵– سایر ملاحظات ۱۳۰
۴–۱۲– جمع‌بندی فصل چهارم و گذار به فصل پنجم ۱۳۱

فصل پنجم

نتیجه گیری و کار های آتی ۱۳۳

۵–۱– کلیات فصل ۱۳۳
۵–۲– جمع‌بندی یافته‌های کلیدی پژوهش ۱۳۴
۵–۲–۱– روش پایه ADAPT و رفتار معیارها ۱۳۴
۵–۲–۲– نقش LLM ها در تولید توضیح متنی (پیش و پس از Fine-Tuning) ۱۳۴
۵–۳– پاسخ به اهداف/سؤالات پژوهش و میزان تحقق آن‌ها ۱۳۵
۵–۴– دلالت‌های علمی (نظری/روش‌شناختی) ۱۳۷
۵–۴–۱– پیوند «توجهِ ADAPT + توضیح زبانیِ LLM» به‌مثابه سازوکار هم‌ترازی بصری–متنی ۱۳۷
۵–۴–۲– نقش چندمرجع در پایدارسازی و منصفانه‌سازیِ ارزیابی متنی ۱۳۸
۵–۴–۳– ناهمسان‌رفتاری معیارها و پیامدهای آن برای سنجش توضیح ۱۳۸
۵–۴–۴– مقایسه روش پیشنهادی با رویکردهای مبتنی بر RAG در توضیح رانندگی خودران ۱۳۹
۵–۵– دلالت‌های کاربردی و توصیه‌ها برای استقرار در سامانه‌های خودران ۱۴۰
۵–۶– مرزبندی و حدود اعتبار نتایج ۱۴۳
۵–۷– پیشنهادهای پژوهشی آینده ۱۴۴
۵–۸– جمع‌بندی نهایی ۱۴۶

فهرست منابع ۱۴۸

پیوست ها ۱۵۷

(پیوست الف) ۱۵۷

(پیوست ب) ۱۵۸
(پیوست پ) ۱۶۶

فهرست جداول

جدول ۲–۱– تاثیر استفاده از سیگنال های کنترلی در کیفیت زیرنویس های تولید شده به وسیله مقایسه
۳۹ ..... نتیجه معیار های ارزیابی زیرنویس روی خروجی های دریافتی [۶].
۶۰ ..... جدول ۲–۲– بررسی برخی روش های پیشین
جدول ۳–۱–تاثیر تعداد فریم ویدیو بر کیفیت زیرنویس های تولیدی با استفاده از معیار های ارزیابی
۶۶ ..... .ROUGE-L[٦] و METEOR، CIDEr-D، BLEU-٤زیرنویس مانند
۸۷ ..... جدول ۴–۱– پیکربندی محیط اجرا
۸۹ ..... BDD-Xجدول ۴–۲– آمار تفکیک
۹۳ ..... ADAPT[۶]جدول ۴–۳– پیکربندی آموزشی
با زیرنویس های ADAPT سه دوره برتر آموزش captioningجدول ۴–۴– مقایسه مقادیر معیار های
۹۸ ..... در بخش تصمیم گیریADAPTاضافه با روش پایه
۱۰۹ ..... BDD-X(انسانیِ Justificationجدول ۴–۵– نتایج سناریوی تک مرجع (فقط
۱۰۹ ..... LLM(+ چهار بازنویسی BDD-Xجدول ۴–۶– نتایج سناریوی چند مرجع (
۱۱۲ ..... BDD-X(– سناریوی تک مرجع (n=١٦جدول ۴–۷– نتایج نمایه ثابت در
LLM(+ چهار بازنویسی BDD-X - سناریوی چند مرجع (n=١٦جدول ۴–۸– نتایج نمایه ثابت در
۱۱۲ .....
جدول ۴–۹– بهترین نتایج هر مدل در هر معیار (تک مرجع)، اعداد از ۷-۴، ۸-۴ و ۳-۹-۴ استخراج
۱۲۳ .....شده اند.
جدول ۴–۱۰– بهترین نتایج هر مدل در هر معیار (چند مرجع)، اعداد از ۷-۴، ۸-۴ و ۳-۹-۴ استخراج
۱۲۳ .....شده اند.
۱۲۴ ..... Fine-Tuningجدول ۴–۱۱– منابع و زمان

جدول ۲–۱– تاثیر استفاده از سیگنال های کنترلی در کیفیت زیرنویس های تولید شده به وسیله مقایسه
نتیجه معیار های ارزیابی زیرنویس روی خروجی های دریافتی [۶]. ۳۹
جدول ۲–۲– بررسی برخی روش های پیشین ۶۰
جدول ۳–۱–تاثیر تعداد فریم ویدیو بر کیفیت زیرنویس های تولیدی با استفاده از معیار های ارزیابی
زیرنویس مانند BLEU-٤، CIDEr-D، METEOR و ROUGE-L[٦]. ۶۶
جدول ۴–۱– پیکربندی محیط اجرا ۸۷
جدول ۴–۲– آمار تفکیک BDD-X ۸۹
جدول ۴–۳– پیکربندی آموزشی ADAPT[۶] ۹۳
جدول ۴–۴– مقایسه مقادیر معیار های
با زیرنویس های ADAPT سه دوره برتر آموزش captioning
اضافه با روش پایه ADAPT در بخش تصمیم گیری ۹۸
جدول ۴–۵– نتایج سناریوی تک مرجع (فقط
BDD-X(انسانیِ Justification ۱۰۹
جدول ۴–۶– نتایج سناریوی چند مرجع (
چهار بازنویسی BDD-X + LLM(..... ۱۰۹
جدول ۴–۷– نتایج نمایه ثابت در
سناریوی تک مرجع (n=۱٦ – BDD-X(..... ۱۱۲
جدول ۴–۸– نتایج نمایه ثابت در
سناریوی چند مرجع (n=۱٦ – BDD-X چهار بازنویسی + LLM(..... ۱۱۲
جدول ۴–۹– بهترین نتایج هر مدل در هر معیار (تک مرجع)، اعداد از ۷-۴، ۸-۴ و ۳-۹-۴ استخراج
شده اند. ۱۲۳
جدول ۴–۱۰– بهترین نتایج هر مدل در هر معیار (چند مرجع)، اعداد از ۷-۴، ۸-۴ و ۳-۹-۴ استخراج
شده اند. ۱۲۳
جدول ۴–۱۱– منابع و زمان Fine-Tuning ۱۲۴

فهرست تصاویر

۲ ..... *VQA[۳] در مدل «پرسش و پاسخ دیداری» (Grad-Cam تصویر ۱–۱–نمونه ای از کاربرد*

تصویر ۱–۲–روش‌های مختلف تفسیر در خودروهای خودران شامل نقشه توجه، حجم هزینه و زبان

۷ ..... طبیعی می‌باشند [۶].

به آن لایه ی PilotNet تصویر ۲–۱–نمونه ای از مدل کیم و همکاران (۲۰۱۷) که برای بهبود مدل

۳۶ ..... اضافه نموده است [۵۴].LSTM

۳۸ ..... ADAPT[۶] تصویر ۲–۲–معماری مدل

۶۳ ..... و اطلاعات موجود در آن BDD-X[۵۵] تصویر ۳–۱– بخشی از مجموعه داده

۶۸ .... تصویر ۳–۲– شبه کد مراحل پیش پردازش فریم های ویدیویی برای استفاده در مدل پیشنهادی

۷۲ ..... تصویر ۳–۳– بلوک دیاگرام مدل تصمیم گیر

۷۴ ..... تصویر ۳–۴– بلوک دیاگرام مدل توضیح دهنده

تصویر ۳–۵– بلوک دیاگرام پردازشات انجام شده روی نقشه توجه جهت استفاده در مدل توضیح دهنده

۷۶ .....

۷۷ ..... تصویر ۳–۶– نمونه آستانه گذاری در نقشه توجه

۷۸ ..... تصویر ۳–۷–بلوک دیاگرام کلی مدل با تمرکز بر بخش توضیح دهنده

۸۰ ..... تصویر ۳–۸ بلوک دیاگرام معماری سیستم پیشنهادی (بخش های قرمز رنگ مثال است)

۸۴ ..... تصویر ۴–۱–بلوک دیاگرام فصل چهارم

۹۲ ..... تصویر ۴–۲–شبه کد اجرایی جهت تبدیل نقشه توجه به کانتور های قرمز رنگ

با زیرنویس های اضافی تولید ADAPT تصویر ۴–۳– نمودار همگرایی تابع هزینه حین آموزش مدل

۹۷ ..... شده

۹۹ ..... تصویر ۴–۴–متن کامل دستور استفاده شده جهت تولید توضیح (خروجی مدل توضیح دهنده)

۱۰۰ ..... تصویر ۴–۵– ترجمه فارسی متن دستور استفاده شده جهت تولید توضیح

۱۰۲ ..... FT فلوچارت مراحل انجام تصویر ۴–۶–

۱۱۸ ..... Gemini ۲٫۵ Pro مدل Fine Tuning تصویر ۴–۷– همگرایی تابع زیان در حین
۱۱۹ ..... Gemini ۲٫۵ Pro در مدل Fine Tuning تصویر ۴–۸– نمودار دقت مدل در حین
۱۲۱ ..... GPT-٤o مدل Fine Tuningهمگرایی تابع زیان در حین تصویر ۴–۹–
۱۲۱ ..... GPT-٤o در مدل Fine Tuningنمودار دقت مدل در حین تصویر ۴–۱۰–
۱۵۹ .....چراغ قرمز / توقف (موفق) - A تصویر ۱–۱– مثال نمونه
۱۶۰ ..... - پیروی از خودروی جلوییِ کند (موفق)Bتصویر ۱–۲– مثال نمونه
۱۶۱ ..... - اولویت عبور عابر پیاده (موفق)Cتصویر ۱–۳– مثال نمونه
۱۶۲ ..... - عدم تطابق بصری-متنیDتصویر ۱–۴– مثال ضدنمونه
۱۶۳ ..... - ابهام نوری/شب و کلی‌گویی متنEتصویر ۱–۵– مثال ضدنمونه
۱۷۱ ..... همگرایی تابع زیان در مدل تغییر یافته کیم و همکاران ۲۰۱۷ حین آموزش تصویر ۱–۶–

۲.....) *VQA[۳] در مدل «پرسش و پاسخ دیداری» (Grad-Cam*تصویر ۱–۱–*نمونه ای از کاربرد*
تصویر ۱–۲–روش‌های مختلف تفسیر در خودروهای خودران شامل نقشه توجه، حجم هزینه و زبان
۷.....طبیعی می‌باشند [۶].
به آن لایه ی PilotNetتصویر ۲–۱–نمونه ای از مدل کیم و همکاران (۲۰۱۷) که برای بهبود مدل
۳۶ ..... اضافه نموده است [۵۴].LSTM
۳۸ ..... ADAPT[۶]تصویر ۲–۲–معماری مدل
۶۳ ..... و اطلاعات موجود در آن BDD-X[۵۵]تصویر ۳–۱– بخشی از مجموعه داده
۶۸ ....تصویر ۳–۲– شبه کد مراحل پیش پردازش فریم های ویدیویی برای استفاده در مدل پیشنهادی
۷۲ ..... تصویر ۳–۳– بلوک دیاگرام مدل تصمیم گیر
۷۴ .....تصویر ۳–۴– بلوک دیاگرام مدل توضیح دهنده
تصویر ۳–۵– بلوک دیاگرام پردازشات انجام شده روی نقشه توجه جهت استفاده در مدل توضیح دهنده
۷۶ .....
۷۷ .....تصویر ۳–۶– نمونه آستانه گذاری در نقشه توجه

تصویر ۳–۷–بلوک دیاگرام کلی مدل با تمرکز بر بخش توضیح دهنده ۷۸
تصویر ۳–۸ بلوک دیاگرام معماری سیستم پیشنهادی (بخش های قرمز رنگ مثال است) ۸۰
تصویر ۴–۱–بلوک دیاگرام فصل چهارم ۸۴
تصویر ۴–۲–شبه کد اجرایی جهت تبدیل نقشه توجه به کانتور های قرمز رنگ با زیرنویس های اضافی تولید ADAPT ۹۲
تصویر ۴–۳– نمودار همگرایی تابع هزینه حین آموزش مدل شده ۹۷
تصویر ۴–۴–متن کامل دستور استفاده شده جهت تولید توضیح (خروجی مدل توضیح دهنده) ۹۹
تصویر ۴–۵– ترجمه فارسی متن دستور استفاده شده جهت تولید توضیح ۱۰۰
تصویر ۴–۶– فلوچارت مراحل انجام FT ۱۰۲
تصویر ۴–۷– همگرایی تابع زیان در حین Fine Tuning مدل Gemini ۲٫۵ Pro ۱۱۸
تصویر ۴–۸– نمودار دقت مدل در حین Fine Tuning در مدل Gemini ۲٫۵ Pro ۱۱۹
تصویر ۴–۹– همگرایی تابع زیان در حین Fine Tuning مدل GPT-٤o ۱۲۱
تصویر ۴–۱۰– نمودار دقت مدل در حین Fine Tuning در مدل GPT-٤o ۱۲۱
تصویر ۱–۱– مثال نمونه A - (موفق) توقف / چراغ قرمز ۱۵۹
تصویر ۱–۲– مثال نمونه B(موفق) - پیروی از خودروی جلوییِ کند ۱۶۰
تصویر ۱–۳– مثال نمونه C(موفق) - اولویت عبور عابر پیاده ۱۶۱
تصویر ۱–۴– مثال ضدنمونه D - عدم تطابق بصری-متنی ۱۶۲
تصویر ۱–۵– مثال ضدنمونه E - ابهام نوری/شب و کلی‌گویی متن ۱۶۳
تصویر ۱–۶– همگرایی تابع زیان در مدل تغییر یافته کیم و همکاران ۲۰۱۷ حین آموزش ۱۷۱

فصل اول

مقدمه

۱–۱– بیان مساله

در سال‌های اخیر، پیشرفت‌های چشمگیر در حوزه هوش مصنوعی، به‌ویژه یادگیری عمیق و مدل‌های زبانی بزرگ[۱]، مسیر توسعه سیستم‌های تصمیم‌گیر خودکار را متحول کرده‌اند. از جمله این کاربردهای پیشرو، توسعه وسایل نقلیه خودران است که با بهره‌گیری از مدل‌های یادگیری انتها به انتها، قادرند محیط پیرامون خود را درک کرده و تصمیم‌گیری کنند. اما با پیچیده‌تر شدن ساختار درونی این مدل‌ها، پرسش‌هایی اساسی درباره‌ی دلایل تصمیم‌گیری‌های آن‌ها مطرح شده است.

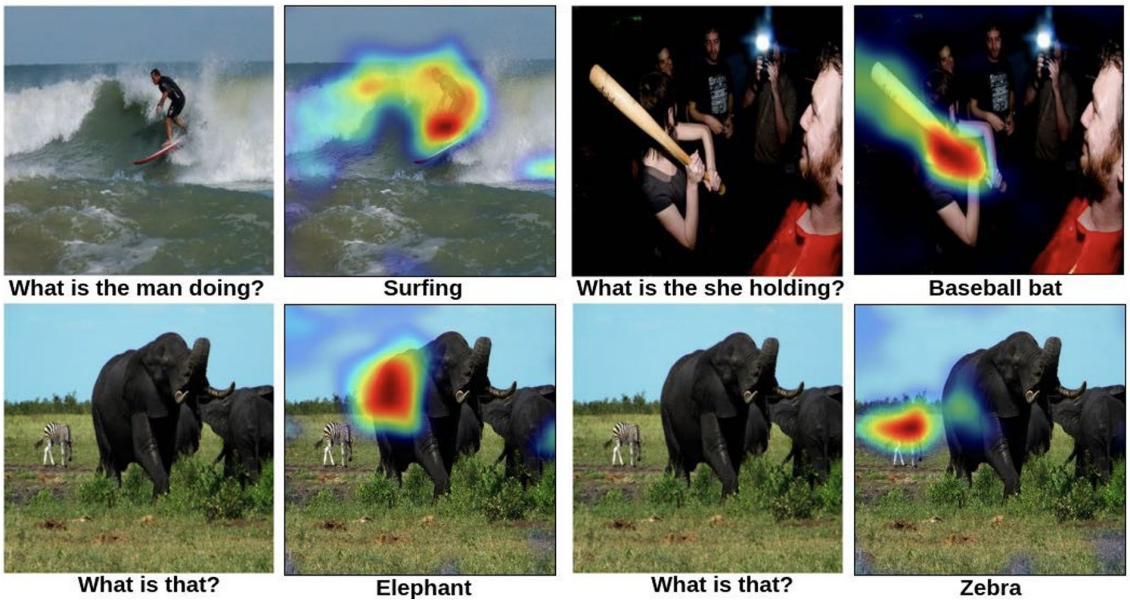
مدل‌های یادگیری عمیق به‌طور معمول مانند «جعبه سیاه» عمل می‌کنند؛ یعنی کاربر نهایی یا حتی توسعه‌دهنده به‌روشنی نمی‌داند چرا یک مدل تصمیم خاصی گرفته است. این موضوع در زمینه‌هایی همچون وسایل نقلیه خودران که تصمیمات آن‌ها مستقیماً با ایمنی کاربران در ارتباط است، چالش‌برانگیزتر می‌شود [۱, ۲]. در پاسخ به این چالش، حوزه‌ای به نام هوش مصنوعی توضیح‌پذیر[۲] شکل گرفته است که هدف آن فراهم کردن راهکارهایی برای توضیح نحوه تصمیم‌گیری مدل‌ها است [۱]. تا کنون روش‌هایی مانند Grad-CAM[۳]، LIME[٤] و SHAP[۲] برای ارائه توضیحات بصری

---

[۱] Large Language Model (LLM)

[۲] Explainable AI (XAI)

یا عددی از مدل‌ها معرفی شده‌اند. این حال، این روش‌ها عمدتاً برای استفاده متخصصان طراحی شده‌اند و برای کاربران عادی، به‌ویژه در محیط‌های عملیاتی مانند جاده‌ها، چندان قابل درک نیستند [۵]. تصویر ۱–۱ نمونه ای از این رویکرد ها را نشان می دهد. این مدل پرسش و پاسخ دیداری[۳] علاوه بر پاسخ که خروجی اصلی مدل است، نقشه توجه را نیز ارائه می دهند که نشان دهنده نقاطی از تصویر است که پاسخ بر اساس آنها تولید شده است. همانطور که گفته شد، اگرچه در مسائل ساده مانند مثال مورد بحث این نقشه های توجه کارآمد هستند ولی توانایی ارائه توضیحات قابل فهم برای تمام مسائل را ندارند [۵].



تصویر ۱–۱–نمونه ای از کاربرد Grad-Cam در مدل «پرسش و پاسخ دیداری» (VQA) [۳]

در سال‌های اخیر چندین تصادف جدی مربوط به حالت خودران[۴] خودروهای تسلا گزارش شده است. برای مثال واشنگتن‌پست گزارش می‌دهد که از سال ۲۰۱۹ تاکنون دست‌کم ۷۳۶ تصادف در ایالات متحده با دخالت سیستم Autopilot تسلا رخ داده که ۱۷ مورد آن منجر به مرگ شده است

[۳] Visual question and answering (VQA)

[۴] Autopilot

[۶]. چنین آمار هشداردهنده‌ای نشان‌دهنده کاستی‌های ایمنی این فناوری است. به‌علاوه، وقوع این حوادث جنجالی باعث کاهش اعتماد عمومی به خودروهای خودران شده است؛ مطالعات نشان می‌دهند بروز تصادفات شاخص می‌تواند به افزایش شک و تردید نسبت به آمادگی این فناوری منجر شود [۷].

یکی از مشکلات مهم در این حوادث، فقدان شفافیت کافی در داده‌ها و عملکرد سیستم است. برای نمونه، در ژانویه ۲۰۱۹ خودروی تسلا مدل ۳ هنگام فعال بودن Autopilot به طور ناگهانی منحرف شد و در یک مسیر فرعی کناری واژگون شد. پس از این حادثه مالک خودرو درخواست دسترسی به داده‌های ثبت‌شده توسط سامانه خودران را مطرح کرد، اما تسلا از ارائه این اطلاعات خودداری کرد. در عمل «جعبه سیاه» خودرو داده‌های مربوط به عملکرد Autopilot را ثبت نمی‌کند و اطلاعات این سیستم تنها روی سرور مرکزی شرکت ذخیره می‌شود[۸]. این عدم شفافیت نه تنها امکان تحلیل مستقل علت حادثه را از بین برده، بلکه نگرانی‌ها و بی‌اعتمادی کاربران را نیز تشدید می‌کند.

در پاسخ به چالش «جعبه سیاه» که در آن منطق درونی مدل‌های یادگیری عمیق پنهان است، تلاش‌های پژوهشی متعددی در حوزه هوش مصنوعی توضیح‌پذیر صورت گرفته است [۱, ۲]. بخشی از این تلاش‌ها در قالب رویکردهای یکپارچه[۵] متمرکز شده‌اند. در این معماری‌ها، که نمونه برجسته آن مدل ADAPT است[۹]، یک مدل واحد و انتها به انتها مسئولیت هر دو وظیفه پیچیده، یعنی تصمیم‌گیری (مانند تعیین زاویه فرمان) و تولید توضیح (مانند ارائه دلیل متنی برای آن تصمیم) را بر عهده دارد. اگرچه این مدل‌ها در نمایش ارتباط مستقیم بین ورودی و خروجی موفق بوده‌اند، اما اغلب از انعطاف‌پذیری کمی برخوردارند. همان‌طور که در بخش نوآوری‌های این پژوهش نیز تأکید شده، در چنین ساختارهایی بهبود یا تغییر در ماژول توضیح‌گر نیازمند بازآموزی کل سیستم است و ارزیابی مستقل کیفیت توضیحات نیز در آن‌ها دشوار می‌باشد.

---

[۵] Integrated Approaches

در کنار تلاش‌های موجود برای ارائه‌ی توضیحات بصری (مانند Grad-CAM[۳] یا LIME[٤])، در سال‌های اخیر تمرکز تازه‌ای بر استفاده از مدل‌های زبانی بزرگ برای تولید توضیحات متنی از تصمیمات مدل‌های پیچیده پدید آمده است [۵, ۱۰]. این مدل‌ها به دلیل قدرت بالای زبانی، قابلیت این را دارند که با دریافت دستور مناسب (Prompt) ، توضیحاتی شبیه به نحوه‌ی بیان انسان‌ها تولید کنند. این رویکرد، که ترکیبی از مهندسی دستور[۶] و مدل‌های LLM است، می‌تواند تفسیر تصمیمات را برای کاربران غیرمتخصص نیز امکان‌پذیر کند.

از سوی دیگر، در حوزه وسایل نقلیه خودران نیز پژوهش‌هایی در حال انجام است که هدف آن‌ها نه‌تنها بهبود دقت تصمیم‌گیری بلکه ارتقاء سطح شفافیت آن‌ها از طریق توضیحات متنی می‌باشد. با ظهور مدل‌های زبانی بزرگ نظیر ChatGPT و Google Gemini، امکان جدیدی برای تولید توضیحات متنی طبیعی درباره رفتار و تصمیمات مدل فراهم شده است [۱۰, ۱۱]. این مدل‌ها توانایی درک و تولید زبان انسانی را دارند و می‌توانند با استفاده از روش‌هایی مانند مهندسی دستور، دلایل تصمیم‌گیری مدل‌های پیچیده را در قالب متن قابل فهم برای انسان ارائه دهند [۱۲, ۱۳]. این امر، به‌ویژه زمانی اهمیت می‌یابد که بخواهیم از سامانه‌های خودران در محیط‌های واقعی و انسانی استفاده کنیم؛ جایی که اعتماد کاربران به تصمیمات سیستم، نقشی کلیدی در پذیرش آن ایفا می‌کند [۱۱ ,۵].

در عین حال، استفاده از LLM به‌تنهایی برای تبیین تصمیمات مدل‌های خودران نیز چالش‌برانگیز است. این مدل‌ها لزوماً به منطق درونی مدل تصمیم‌گیرنده اصلی دسترسی ندارند و ممکن است صرفاً بر اساس شباهت‌های زبانی خروجی‌هایی تولید کنند که الزاماً با فرایند واقعی تصمیم‌گیری هم‌راستا نباشد [۱۴, ۱۵]. بنابراین، **پرسش های اصلی** این پژوهش عبارتند از:

۱. چگونه می‌توان با بهره‌گیری همزمان از یک مدل تصمیم‌گیر تخصصی و یک مدل زبانی بزرگ، توضیحات متنی دقیق‌تر و قابل‌درک‌تری برای تصمیمات سامانه‌های خودران تولید کرد؟

---

[۶] Prompt Engineering

۲. نقشه‌های توجه[۷] چه نقشی در ایجاد پیوند میان ورودی‌های تصویری مدل تصمیم‌گیر و خروجی متنی مدل زبانی ایفا می‌کنند و تا چه حد به همخوانی بصری–متنی کمک می‌نمایند؟
۳. آیا می‌توان از مدل‌های زبانی برای تولید زیرنویس[۸]های اضافی و متنوع، به‌عنوان روشی جهت گسترش داده‌ها[۹]، استفاده کرد و این کار چه اثری بر ارزیابی و کیفیت توضیحات تولیدی دارد؟
٤. فرایند تنظیم دقیق[۱۰] مدل‌های زبانی بر روی داده‌های غنی‌شده چه تأثیری بر بهبود کیفیت و پایداری خروجی توضیحات دارد؟

بدین ترتیب، در این پژوهش تلاش می‌شود تا با ترکیب یک مدل تصمیم‌گیر برای رانندگی و یک مدل زبانی بزرگ برای تولید توضیح، یک سامانه‌ی ترکیبی و قابل تفسیر برای وسایل نقلیه خودران طراحی و پیاده‌سازی شود. این سامانه علاوه بر دقت در تصمیم‌گیری، قادر خواهد بود دلایل آن را به‌صورت متنی و قابل فهم برای انسان بیان کند.

۱–۲– اهمیت و ضرورت تحقیق

افزایش اتکای سیستم‌های هوشمند به مدل‌های یادگیری عمیق، ضرورت شفاف‌سازی عملکرد این مدل‌ها را بیش از پیش آشکار کرده است. این مسئله به‌ویژه در کاربردهایی مانند وسایل نقلیه خودران که تصمیمات مدل می‌تواند جان انسان‌ها را تحت‌تأثیر قرار دهد، اهمیت حیاتی دارد. در چنین شرایطی، کاربران، توسعه‌دهندگان و نهادهای قانون‌گذار، نیازمند درک دلایل تصمیمات سامانه هستند تا بتوانند به عملکرد آن اعتماد کرده و مسئولیت‌پذیری آن را ارزیابی کنند [۱, ۲].

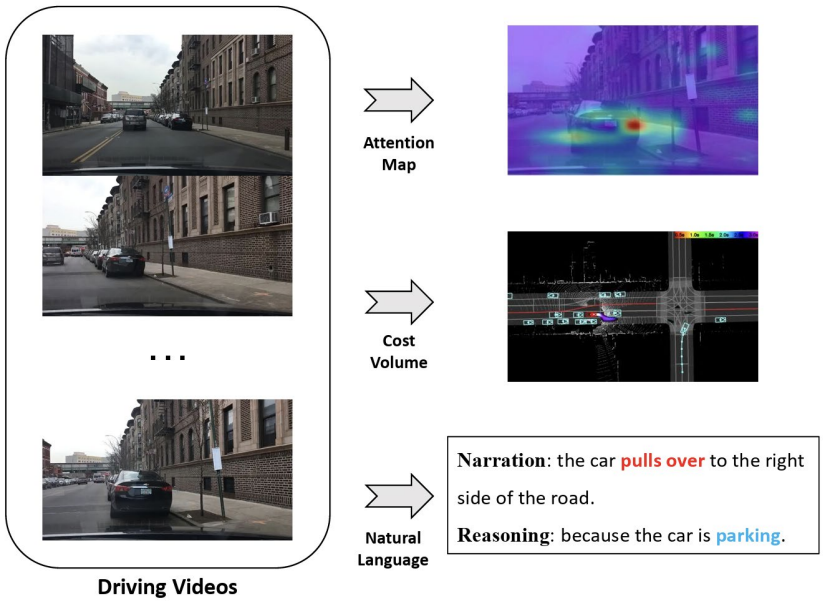
یکی از ابعاد محوریِ لزوم شفافیت در سامانه‌های هوش مصنوعی، الزام قانونی‌ای است که در قلمرو اتحادیه اروپا تعریف شده است. طبق ماده ۱۳ قانون AI اتحادیه اروپا، «سامانه‌های AI با ریسک

---

[۷] Attention Map
[۸] Caption
[۹] Data Augmentation
[۱۰] Fine-Tuning

بالا» باید به گونه‌ای طراحی و توسعه یابند که عملیات آن‌ها "به قدری شفاف باشد که کاربر یا بهره‌بردار بتواند خروجی سیستم را تفسیر کند و آن را به‌درستی به‌کار بگیرد". همچنین ماده ۵۰ این قانون، الزام می‌کند که کاربرانِ سامانه‌های هوش مصنوعی «اطلاع یابند که با یک سامانه AI در تعامل هستند»، و در موارد تولید یا دستکاری محتوای مصنوعی، باید برچسب‌گذاری مناسبی صورت گیرد [۱۶]. این مقررات که از ۲ اوت ۲۰۲۶ به اجرا درمی‌آیند، نشان‌دهنده توجه قانون‌گذار اروپایی به موضوع اعتماد، ایمنی و قابلیت توضیح‌پذیری سامانه‌های هوش مصنوعی هستند. بنابراین در مطالعه حاضر که بر شفافیت تصمیم‌گیری در سامانه‌های خودران تمرکز دارد، لازم است که منظرِ قانونیِ این الزام نیز مورد توجه قرار گیرد؛ نه تنها به دلیل الزامات حقوقی، بلکه به‌خاطر اثر آن بر اعتماد کاربران و پذیرش فناورانه.

روش‌های سنتی توضیح‌پذیری مانند LIME، Grad-CAM و SHAP تا حدی در ارائه تفسیرهایی بصری یا ویژگی‌محور موفق بوده‌اند [۲–۴] . اما اغلب این خروجی‌ها برای افراد غیرمتخصص، از جمله رانندگان، سرنشینان و حتی برخی مهندسان کاربردی، قابل درک یا اعتماد نیستند [۵, ۱۷]. از این رو، نیاز به تولید توضیحات متنی انسانی‌فهم در کنار این روش‌ها، به‌عنوان یک گام مهم برای بهبود ارتباط انسان–ماشین، مطرح شده است. تصویر ۱–۲ مثالی از روش های ارائه توضیحات در وسایل نقلیه خودران را نشان می دهد. با وجود کارایی نقشه توجه یا حجم هزینه، تفسیر مبتنی بر زبان برای مسافران عادی کاربرپسندتر است.



تصویر ۱–۲–روش‌های مختلف تفسیر در خودروهای خودران شامل نقشه توجه، حجم هزینه و زبان طبیعی می‌باشند [۹].

همانطور که اشاره شد ظهور مدل‌های زبانی بزرگ، که توانایی درک و تولید زبان طبیعی را با دقت بالا دارند، یک فرصت منحصر به‌فرد برای تولید توضیحات متنی قابل اعتماد و انعطاف‌پذیر فراهم کرده است [۱۰, ۱۱]. این مدل‌ها می‌توانند با استفاده از مهندسی دستور مناسب، دلایل تصمیم‌گیری سیستم را به زبان طبیعی بیان کرده و از این طریق، شفافیت سیستم‌های خودران را به‌شکل بی‌سابقه‌ای ارتقا دهند [۱۲, ۱۳, ۱۸].

در عین حال، این حوزه هنوز نوظهور است و چالش‌هایی مانند نحوه اتصال دقیق خروجی مدل تصمیم‌گیرنده (مثلاً نقشه توجه یا Frame مهم) به مدل زبانی، یا انتخاب بهترین نوع دستور (سخت، نرم، مبتنی بر مثال یا زنجیره افکار) به قوت خود باقی است [۱۴, ۱۵, ۱۹]. همچنین، روش‌های ارزیابی کیفی و کمی توضیحات متنی در این زمینه به‌طور کامل توسعه نیافته‌اند و نیازمند تحقیق هستند [۱۰, ۲۰].

بنابراین، اهمیت این پژوهش در آن است که با ترکیب قابلیت استنتاجی مدل‌های خودران و قدرت زبانی مدل‌هایLLM ، یک سامانه‌ی قابل توضیح و شفاف طراحی کند که بتواند تصمیمات خود را نه‌تنها اتخاذ، بلکه **توجیه** نیز بنماید. این گام می‌تواند نقطه‌ی عطفی در ارتقای اعتماد عمومی به وسایل نقلیه خودران و تسهیل پذیرش گسترده آن‌ها باشد.

۱–۳– اهداف تحقیق

هدف اصلی این تحقیق طراحی و پیاده‌سازی یک سامانه تصمیم گیری توضیح‌پذیر مبتنی بر مدل‌های زبانی بزرگ برای تبیین تصمیمات مدل‌های هوش مصنوعی در وسایل نقلیه خودران است.

۱–۳–۱– هدف کلی

- طراحی و پیاده‌سازی یک سیستم تولید تصمیم گیر و توضیح دهنده متنی برای تصمیمات مدل‌های خودران با بهره‌گیری از LLM ها و روش‌های XAI .

۱–۳–۲– اهداف جزئی

۱. شناسایی مهم‌ترین فریم‌ها و نواحی توجه مدل تصمیم‌گیر با استفاده از نقشه توجه.

۲. بررسی و پیاده‌سازی انواع روش‌های مهندسی دستور (سخت، نرم، مبتنی بر مثال، زنجیره افکار).

۳. ارزیابی کیفی و کمی خروجی‌های متنی مدل زبانی بر اساس معیارهای معناشناسی، انسجام و انطباق با تصمیم واقعی.

٤. مقایسه عملکرد مدل‌های مختلف (LLM مانند ChatGPT وGemini) در تولید توضیحات متنی از تصمیمات خودران.

۵. تحلیل تأثیر نوع دستور بر کیفیت توضیحات تولیدی.

۱–۴– نوآوری‌های تحقیق

این پژوهش با هدف ارتقای شفافیت و توضیح‌پذیری در سامانه‌های خودران، مجموعه‌ای از نوآوری‌های مفهومی و روش‌شناختی را ارائه می‌دهد که آن را از کارهای پیشین متمایز می‌سازد. این نوآوری‌ها به شرح زیر است:

- **طراحی معماری ترکیبی و تفکیک‌شده**: برخلاف رویکردهای یکپارچه، در این پژوهش یک معماری ترکیبی ارائه شده که در آن مدل تصمیم‌گیر (ADAPT) از ماژول توضیح‌گر (LLM) مجزا است. این تفکیک، انعطاف‌پذیری و قابلیت ارتقاء سامانه را در آینده افزایش می‌دهد.
- **استفاده از نقشه توجه به عنوان پل ارتباطی بصری–معنایی**: نوآوری کلیدی این پژوهش، تبدیل خروجی نقشه توجه مدل ADAPT به نشانه‌های بصری صریح (خطوط پیرامونی قرمز) است. این کار به عنوان یک «قید بصری» عمل کرده و توضیح LLM را مستقیماً به همان نواحی از تصویر که برای مدل تصمیم‌گیر اهمیت داشته، ارجاع می‌دهد.
- **غنی‌سازی داده‌های آموزشی با استفاده از مدل‌های زبانی**: برای افزایش تنوع داده‌های متنی، از یک روش افزایش داده مبتنی بر LLM استفاده شده است. به این صورت که برای هر توضیح انسانی، چندین نسخه بازنویسی‌شده با حفظ معنا تولید و به مجموعه داده آموزشی اضافه گردید تا مدل در برابر تنوع بیانی مقاوم‌تر شود.
- **ارائه پروتکل ارزیابی چندمرجعی**: با توجه به اینکه یک صحنه رانندگی می‌تواند چندین توضیح صحیح داشته باشد، یک پروتکل ارزیابی جدید معرفی شد که در آن علاوه بر مرجع انسانی اصلی، از چندین مرجع متنی تولیدشده توسط LLM نیز برای ارزیابی خروجی مدل استفاده می‌شود. این روش، ارزیابی خودکار را به قضاوت انسانی نزدیک‌تر می‌کند.

۱–۵– روش تحقیق

پژوهش حاضر از نوع کاربردی – تحقیقاتی و به‌لحاظ روش اجرا، توصیفی – تحلیلی با رویکرد طراحی و پیاده‌سازی است. هدف آن طراحی سیستمی ترکیبی برای تبیین تصمیمات مدل‌های خودران از طریق تولید توضیحات متنی قابل فهم برای انسان‌ها با کمک مدل‌های زبانی بزرگ و ساختارهای آگاه از توجه[۱۱].

مراحل تحقیق به‌شرح زیر است:

۱. انتخاب و پیاده‌سازی مدل تصمیم‌گیرنده‌ی خودران با ساختار ADAPT:

- به‌جای استفاده از مدل‌های کلاسیک مانند PilotNet [۲۱]، در این پژوهش از معماری پیشنهادی مقاله‌ی ADAPT [۹] بهره گرفته می‌شود؛ این مدل یک ترانسفورمر آگاه از کنش[۱۲] برای تولید توصیف‌های متنی از صحنه‌های رانندگی است.

۲. استخراج اطلاعات کلیدی برای تغذیه مدل توضیح‌گر:

- خروجی ADAPT شامل تصممیم مدل است. همچنین نقشه توجه برای هر تصمیم استخراج می‌شود که به‌عنوان ورودی های مدل توضیح‌گر استفاده خواهند شد.

۳. طراحی و اعمال روش‌های هدایت مدل زبانی:

بر اساس خروجی مدل ADAPT، از دو رویکرد اصلی برای هدایت مدل‌های زبانی بزرگ جهت تولید توضیحات دقیق استفاده می‌شود:

---

[۱۱] Attention-aware

[۱۲] Action-aware

- مهندسی دستور : طراحی یک دستور ثابت شامل دستورالعمل‌های صریح و چند مثال کلیدی[۱۳] به منظور تعریف دقیق وظیفه، تعیین ساختار خروجی و هدایت اولیه مدل.
- تنظیم دقیق[۱۴]: برای انطباق تخصصی‌تر مدل با داده‌های حوزه رانندگی، مدل‌های پایه با استفاده از سرویس‌های ابری (Google Gemini و OpenAI) بر روی مجموعه داده‌های پروژه به صورت دقیق تنظیم می‌شوند تا توانایی آن‌ها در تولید توضیحات مرتبط و صحیح افزایش یابد.

٤. ارزیابی کیفیت توضیحات متنی:

برای سنجش جامع کیفیت توضیحات متنی تولیدشده، ارزیابی چندوجهی بر اساس سه ویژگی کلیدی انجام می‌شود. هر یک از این ویژگی‌ها با استفاده از مجموعه‌ای از معیارهای کمی و کیفی متناسب سنجیده می‌شوند تا تصویری کامل از عملکرد سامانه به دست آید:

- هم‌راستایی با تصمیم مدل و داده‌های تصویری[۱۵]: این ویژگی بررسی می‌کند که آیا توضیح تولیدشده، به دلایل واقعیِ تصمیم مدل (بر اساس نواحی توجه) و شواهد موجود در تصویر وفادار است یا خیر. برای سنجش این هم‌راستایی از روش‌های زیر استفاده می‌شود:
  - معیارهای معنایی[۱۶]: معیارهایی مانند SPICE[۲۲] و CIDEr-D[۲۳] به کار گرفته می‌شوند.

---

[۱۳] Few-shot Prompting

[۱۴] Fine tuning

[۱۵] Faithfulness and Grounding

[۱۶] Semantic Metrics

SPICE با تحلیل اشیاء، ویژگی‌ها و روابط موجود در متن، آن را با محتوای بصری تصویر مقایسه می‌کند.

CIDEr-D نیز با وزن‌دهی به کلمات کلیدی، میزان اجماع محتوایی[۱۷] توضیح را با توصیف‌های انسانی می‌سنجد.

- ارزیابی کیفی: همان‌طور که در فصل چهارم تشریح شد، تحلیل‌های موردی انجام می‌شود تا مشخص گردد آیا دلایل ذکرشده در متن تولیدی، به اشیاء یا نواحی‌ای اشاره دارند که در نقشه توجه[۱۸] مدل تصمیم‌گیر برجسته شده‌اند یا خیر.

- خوانایی، روانی و انسجام زبانی[۱۹]:

این ویژگی کیفیت ساختاری و زبانی متن تولیدی را ارزیابی می‌کند. یک توضیح خوب باید از نظر گرامری صحیح، روان و قابل درک باشد. این ابعاد با معیارهای استاندارد ارزیابی متن‌محور سنجیده می‌شوند:

- معیارهای مبتنی بر هم‌پوشانی واژگان[۲۰]: از معیارهایی نظیر BLEU۴[۲۴]، ROUGE-L[۲۵] و METEOR[۲۶] استفاده می‌شود.
  ۱. BLEU۴ دقت هم‌پوشانی چهارکلمه‌ای را می‌سنجد،
  ۲. ROUGE-L طولانی‌ترین زیردنباله مشترک را شناسایی می‌کند و

---

[۱۷] Consensus

[۱۸] Attention Map

[۱۹] Fluency and Coherence

[۲۰] N-gram Overlap Metrics

۳. METEOR با در نظر گرفتن مترادف‌ها، ارزیابی انعطاف‌پذیرتری از روانی و تطابق معنایی ارائه می‌دهد

۵. تحلیل و مقایسه:

- تحلیل عملکرد مدل‌های LLM و انواع دستور در شرایط مختلف
- بررسی تطابق توضیحات با رفتار واقعی مدل تصمیم‌گیرنده
- پیشنهاد معماری بهینه برای ترکیب ADAPT با مدل زبانی توضیح‌گر

۱–۶– نو آوری های تحقیق:

در امتداد تحلیل روش تحقیق، نوآوری‌های اصلی این پژوهش که آن را از کارهای پیشین متمایز می‌سازد، به شرح زیر است:

۱. طراحی معماری ترکیبی و تفکیک‌شده[۲۱]: برخلاف رویکردهای یکپارچه که یک مدل واحد مسئولیت تصمیم‌گیری و تولید توضیح را بر عهده دارد ، در این پژوهش یک معماری ترکیبی ارائه شده است. در این ساختار، مدل تصمیم‌گیر تخصصی (ADAPT) وظیفه تحلیل صحنه و اتخاذ تصمیم را انجام می‌دهد و یک مدل زبانی بزرگ به عنوان یک ماژول توضیح‌گر مجزا، وظیفه تبیین آن تصمیم را بر عهده دارد. این تفکیک، انعطاف‌پذیری بیشتری را برای به‌روزرسانی هر بخش فراهم می‌کند.

۲. استفاده از نقشه توجه به عنوان پل ارتباطی بصری: نوآوری کلیدی این پژوهش، استفاده از خروجی نقشه توجه مدل ADAPT به عنوان یک ورودی بصری صریح برای هدایت مدل زبانی است. در حالی که پژوهش‌های قبلی از توجه به صورت یک سازوکار درونی بهره می‌بردند ، در این کار نقشه توجه به نشانه‌های بصری مشخص (خطوط پیرامونی قرمز) تبدیل

---

[۲۱] Decoupled

شده تا به عنوان یک «قید بصری» عمل کرده و اطمینان حاصل شود که توضیح LLM مستقیماً به همان نواحی از تصویر ارجاع دارد که برای مدل تصمیم‌گیر اهمیت داشته است.

۳. بهبود داده‌های آموزشی با استفاده از مدل‌های زبانی: برای غنی‌سازی و افزایش تنوع داده‌های متنی، از یک روش افزایش داده مبتنی بر LLM استفاده شده است . به این صورت که برای هر توضیح انسانی موجود در مجموعه‌داده، چندین نسخه بازنویسی‌شده با حفظ معنا تولید و به مجموعه داده آموزشی اضافه گردید. این رویکرد مدل را در برابر تنوع بیانی مقاوم‌تر کرده و به بهبود عملکرد آن کمک می‌کند.

٤. ارائه پروتکل ارزیابی چندمرجعی: با توجه به اینکه یک صحنه رانندگی می‌تواند چندین توضیح صحیح داشته باشد، یک پروتکل ارزیابی جدید معرفی شد که در آن علاوه بر مرجع انسانی اصلی، از چندین مرجع متنی تولیدشده توسط LLM نیز برای ارزیابی خروجی مدل استفاده می‌شود .این روش، ارزیابی خودکار را به قضاوت انسانی نزدیک‌تر کرده و سنجش منصفانه‌تری از کیفیت توضیحات ارائه می‌دهد.

۱–۷– ساختار پایان‌نامه

این پایان‌نامه در پنج فصل تنظیم شده است که هر فصل به بررسی بخشی از روند پژوهش می‌پردازد. ساختار کلی پایان‌نامه به شرح زیر است:

- **فصل اول: کلیات تحقیق**

در این فصل، موضوع پژوهش معرفی و ضرورت آن تبیین شد. بیان مسئله، اهداف، روش تحقیق، سؤالات پژوهش و تعریف مفاهیم کلیدی نیز ارائه گردید.

- **فصل دوم: مبانی نظری و پیشینه تحقیق**

در این فصل مفاهیم نظری مرتبط با هوش مصنوعی توضیح‌پذیر ، مدل‌های زبانی بزرگ ، روش‌های مهندسی دستور و مدل‌های تصمیم‌گیر خودران به‌تفصیل مورد بررسی قرار می‌گیرند. همچنین پژوهش‌های داخلی و خارجی مرتبط مرور می‌شود.

- **فصل سوم: روش‌شناسی تحقیق**

در این بخش، نحوه پیاده‌سازی مدل تصمیم‌گیر مبتنی بر معماریADAPT ، فرایند طراحی دستورها، انتخاب مدل زبانی، و روش‌های ارزیابی توضیحات تولیدی تشریح می‌گردد.

- **فصل چهارم: پیاده‌سازی و نتایج**

در این فصل، جزئیات پیاده‌سازی سامانه توضیح‌گر، نتایج حاصل از آزمایش‌ها، تحلیل عملکرد مدل در سناریوهای مختلف، و مقایسه بین روش‌های مختلف دستور ارائه خواهد شد.

- **فصل پنجم: نتیجه‌گیری و پیشنهادها**

این فصل به جمع‌بندی نتایج، پاسخ به سؤالات پژوهش، ارزیابی فرضیات، و ارائه پیشنهادهایی برای بهبود یا ادامه‌ی مسیر تحقیق اختصاص دارد.

فصل دوم

پیش زمینه تحقیق

۲–۱– کلیات فصل

در هر سامانه مبتنی بر هوش مصنوعی، به‌ویژه در کاربردهای حساس و پیچیده نظیر وسایل نقلیه خودران، درک و تبیین نحوه تصمیم‌گیری مدل از اهمیت بالایی برخوردار است. در سال‌های اخیر، هم‌زمان با پیشرفت چشمگیر الگوریتم‌های یادگیری عمیق، بحث «قابل توضیح بودن» یا **توضیح‌پذیری**[۲۲] مدل‌ها نیز به یکی از محورهای اصلی در تحقیقات هوش مصنوعی تبدیل شده است. این ضرورت به‌ویژه در سیستم‌هایی که به‌طور خودکار و بدون دخالت انسان تصمیم می‌گیرند، ملموس‌تر می‌شود.

یکی از رویکردهای نوین برای افزایش شفافیت مدل‌های هوشمند، استفاده از **مدل‌های زبانی بزرگ** همچون GPT و Gemini برای تولید توضیحات متنی درباره‌ی رفتار مدل‌های تصمیم‌گیر است. این مدل‌ها، در کنار توانایی زبانی بالا، امکان تعامل انسان‌محور با سامانه‌های هوشمند را فراهم می‌سازند.

---

[۲۲] Explainability

از سوی دیگر، طراحی ورودی مناسب برای هدایت این مدل‌ها که تحت عنوان مهندسی دستور شناخته می‌شود، نقشی حیاتی در تولید توضیحات دقیق و قابل اعتماد ایفا می‌کند.

در این فصل، ابتدا مفاهیم نظری مرتبط با هوش مصنوعی توضیح‌پذیر ، مدل‌های زبانی بزرگ، و انواع روش‌های مهندسی دستور بررسی خواهد شد. سپس معماری‌های مختلف مورد استفاده در مدل‌های تصمیم‌گیر خودران، با تمرکز بر ساختارADAPT ، معرفی می‌گردد. در بخش دوم فصل نیز، پیشینه داخلی و خارجی پژوهش‌ها در این زمینه مرور شده و در نهایت، خلأهای موجود و جایگاه پژوهش حاضر در میان آن‌ها تبیین می‌شود.

۲–۲– مفاهیم کلیدی و اصطلاحات

در این بخش، به‌منظور درک بهتر مباحث پژوهش، برخی از مفاهیم و اصطلاحات کلیدی که در ادامه پایان‌نامه به‌صورت مکرر مورد استفاده قرار می‌گیرند، به‌صورت مختصر تعریف می‌شوند:

- **هوش مصنوعی توضیح‌پذیر:**

مجموعه‌ای از روش‌ها و تکنیک‌ها در یادگیری ماشین که هدف آن‌ها تبیین دلایل تصمیم‌گیری مدل‌های پیچیده و کاهش خاصیت «جعبه‌سیاه» بودن آن‌ها است.

- **مدل‌های زبانی بزرگ:**

مدل‌هایی مبتنی بر شبکه‌های عصبی عمیق که با آموزش بر حجم وسیعی از متون، قادر به درک، تولید و استدلال در زبان طبیعی هستند (مانند GPT-٤ ، Gemini).

- **دستور :**

ورودی متنی یا ساختاری که برای هدایت رفتار مدل زبانی به آن داده می‌شود. بسته به نوع آن، می‌تواند شامل دستور صریح، مثال، یا زنجیره استدلال باشد.

- **مهندسی دستور:**

هنر و علم طراحی مؤثر دستورها به‌منظور به‌دست آوردن خروجی مطلوب از مدل‌های زبانی.

- **نقشه توجه:**

تصویری که نشان می‌دهد مدل تصمیم‌گیرنده در زمان پردازش داده‌ی ورودی، به کدام نواحی (مثلاً بخش‌هایی از یک تصویر) بیشترین توجه را داشته است.

- **تنظیم دقیق**:

یک روش کلیدی در یادگیری ماشین است که در آن، یک مدل از پیش‌آموزش‌دیده[۲۳] که بر روی حجم عظیمی از داده‌های عمومی آموزش دیده است، برای یک وظیفه خاص و با استفاده از مجموعه داده‌ای کوچکتر و تخصصی‌تر، بازآموزی یا تطبیق داده می‌شود. این رویکرد در تضاد با آموزش مدل از ابتدا[۲۴] قرار دارد و به مدل اجازه می‌دهد تا دانش گسترده‌ای که در مرحله پیش‌آموزش کسب کرده را به وظیفه جدید منتقل کند و با هزینه محاسباتی کمتر به عملکرد بالاتری دست یابد. در این پایان‌نامه، برای حفظ انسجام و شفافیت، از این پس از عبارت انگلیسی Fine-Tuning یا مخفف آن (FT) برای اشاره به این مفهوم استفاده خواهد شد.

- **ADAPT**:

---

[۲۳] Pre-trained Model

[۲۴] Training from scratch

یک معماری مبتنی بر ترانسفورمر برای تحلیل صحنه‌های رانندگی و تولید توصیف متنی از آن‌ها که به‌صورت آگاه از کنش طراحی شده و تمرکز آن بر ارتباط میان رویدادها و تصمیمات رانندگی است.

- **محدودیت‌های گراندینگ[۲۵]**

قوانینی هستند که مدل‌های هوش مصنوعی را ملزم می‌کنند خروجی‌های خود را با داده‌های ورودی اصلی مطابقت دهند. این محدودیت‌ها تضمین می‌کنند که نتایج پردازش مدل (مانند استخراج اطلاعات یا خلاصه‌سازی) مستقیماً بر داده‌های منبع استوار باشند [۲۷] و از تولید خروجی‌های نامربوط یا خیالی جلوگیری می‌کنند.

- **پیش‌شرط گراندینگ[۲۶]**

شرایط رسمی و الزامات لازم برای هر عمل یا ابزار را تعریف می‌کند و در زمان اجرا این پیش‌شرط‌ها را با روش‌های رسمی بررسی می‌نماید تا اطمینان حاصل شود ابزار انتخاب‌شده با وضعیت فعلی محیط قابل اجرا باشد [۲۸]. به عبارت دیگر، این مفهوم تضمین می‌کند که تمام الزامات لازم برای اجرای یک عمل بر اساس درک کنونی سیستم برآورده باشند و از تلاش برای انجام اعمال غیرممکن جلوگیری شود.

۲–۳– مبانی نظری تحقیق

۲–۳–۱– هوش مصنوعی توضیح‌پذیر

هوش مصنوعی توضیح‌پذیر به مجموعه روش‌هایی اطلاق می‌شود که هدف آن‌ها ارائه توضیحات قابل‌فهم برای خروجی‌های سیستم‌های پیچیده یادگیری ماشین است. اهمیت XAI از آنجا ناشی می‌شود

---

[۲۵] Grounding constraints

[۲۶] Grounding precondition

که تصمیمات مدل‌های پیشرفته نظیر شبکه‌های عصبی عمیق[۲۷] اغلب به شکل «جعبه سیاه» در می‌آیند و تبیین فرآیند تصمیم‌گیری آن‌ها برای کاربران دشوار است.

اهداف و ضرورت XAI در یادگیری ماشین

- **افزایش اعتماد و پذیرش**: انسان‌ها کمتر تمایل دارند به سیستم‌های تصمیم‌گیر جعبه‌سیاه اعتماد کنند. فراهم‌سازی توضیحاتی شفاف و قابل‌فهم باعث می‌شود کاربران بهتر عملکرد مدل را درک کنند و به آن اعتماد نمایند [۱, ۲۹].

- **تشخیص خطا و سوگیری**: توضیح‌پذیری به شناسایی نقاط ضعف، جهش‌های ناخواسته و بایاس‌های نهفته در داده‌ها کمک می‌کند. برای مثال، اگر مدل پیچیده باشد ولی بتوان توضیحی در مورد نقش هر ویژگی ارائه کرد، می‌توان تبعیض احتمالی علیه گروه خاصی را کشف کرد [۱, ۳۰].

- **تضمین انطباق قانونی و اخلاقی**: قوانین نظیر GDPR و «حق توضیح» در اتحادیه اروپا شرکت‌ها را ملزم می‌سازد که مدل‌های تصمیم‌ساز خود را شفاف کنند. XAI این امکان را فراهم می‌کند که سیستم‌ها علاوه بر دقت، جنبه‌های اخلاقی و قانونی را نیز رعایت کنند [۲۹, ۳۰]. افزون بر این، امکان ارائه توضیح پسینی از فرآیند تصمیم باعث می‌شود مسئولیت‌پذیری پس از بروز حادثه آسان‌تر شود؛ به‌عنوان مثال، شرکت مرسدس بنز با اعلام پذیرش مسئولیت قانونی برای تصادفات ناشی از سیستم خودران خود، گامی مهم در جهت شفاف‌سازی و پاسخگویی برداشت [۳۱].

- **کارایی عملی در کاربردهای حساس**: در سامانه‌های پشتیبان تصمیم حوزه‌های حساس (مانند سامانه‌های پزشکی یا خودروهای خودران)، توضیح مدل برای متخصصان امکان کنترل و

---

[۲۷] Deep Neural Network (DNN)

تصحیح بهتر نتایج را فراهم می‌آورد [۱, ۲۹]؛ این ضرورت در فناوری خودروهای خودران، که تصمیم‌گیری لحظه‌ای و ایمن در آن حیاتی است، اهمیت دوچندان دارد. گزارش‌ها نشان می‌دهند وقوع تصادفات متعدد مرتبط با «جعبه‌سیاه» بودن سیستم‌های تصمیم‌گیری، نگرانی‌های ایمنی را تشدید کرده است و انتظار می‌رود تصمیمات مبتنی بر یادگیری عمیق توضیح‌پذیر باشند تا اعتماد جامعه و مراجع نظارتی را جلب کنند [۳۱].

تفاوت میان توضیح‌پذیری[۲۸] و تفسیرپذیری[۲۹]

در ادبیات XAI گاه دو واژه «تفسیرپذیری» و «توضیح‌پذیری» با هم اشتباه گرفته می‌شوند، ولی پژوهشگران به تفاوت‌های مهمی میان آن‌ها اشاره کرده‌اند. به طور کلی، تفسیرپذیری صفت «غیرفعال» یک مدل است که نشان می‌دهد ساختار یا خروجی مدل تا چه حد برای انسان قابل فهم است؛ مثلاً یک مدل درخت تصمیم ساده یا مدل خطی شفاف، ذاتاً تفسیرپذیر تلقی می‌شوند [۱, ۳۲]. به عبارت دیگر، تفسیرپذیری به قابلیت مدل در تولید یک توصیف ساده و قابل درک از عملکرد خود اشاره دارد (به عنوان مثال یک فرمول ریاضی یا مجموعه‌ای از قواعد قابل انسان‌خوانی) [۱, ۳۲].

در مقابل، توضیح‌پذیری صفت «فعالی» است که مستلزم ارائه توضیح یا دلیل برای نتایج مدل می‌باشد. یک مدل توضیح‌پذیر اقدامات یا رویه‌هایی را اجرا می‌کند تا به کاربر نشان دهد چگونه به نتیجه رسیده است؛ مثلاً تولید نگاشت برجستگی ویژگی‌ها، گزارش درصد مشارکت هر ورودی در خروجی، یا پاسخ به سؤال «چرا این تصمیم گرفته شد؟» [۱, ۳۰]. به عنوان مثال، فرض کنید مدلی درختی که ساختارش برای انسان قابل فهم است (بنابراین تفسیرپذیر است)، اما این مدل پاسخ خود را تنها به صورت پیش‌بینی عددی ارائه کند؛ در این صورت اگر توضیحی در مورد منطق پیش‌بینی ارائه نشود، ما هنوز توضیح‌پذیری نخواهیم داشت. برعکس، روش‌هایی مانند LIME یا SHAP پس از

---

[۲۸] Explainability

[۲۹] Interpretability

آموزش مدل عمل می‌کنند و با وزن‌دهی به ویژگی‌ها توضیح محلی[۳۰] درباره یک پیش‌بینی ارائه می‌دهند؛ در اینجا مدل زیرین ممکن است کاملاً شفاف نباشد ولی خودِ روش توانسته است توضیح تولید کند (توضیح‌پذیری) [۳۰, ۳۲]. این تمایز یک نکته کلیدی را آشکار می‌سازد. یک مدل ذاتاً تفسیرپذیر (مانند یک درخت تصمیم ساده) به خودی خود توضیح‌پذیر نیز محسوب می‌شود، زیرا می‌توان مسیر تصمیم‌گیری آن را به سادگی دنبال کرد. با این حال، عکس این موضوع لزوماً صادق نیست. یک مدل پیچیده می‌تواند از طریق روش‌های پسینی[۳۱] «توضیح‌پذیر» شود، بدون آنکه ساختار درونی آن «تفسیرپذیر» باشد. در چنین حالتی، این «روشِ توضیح‌دهنده» است که شفافیت را فراهم می‌کند، نه خودِ مدل زیرین [۱, ۳۰].

در پژوهش حاضر، تمرکز بر ترکیب دو دسته از این روش‌هاست: استفاده از مدل پسینی (LLM) برای تولید **توضیح متنی از خروجی مدل پیچیده تصمیم‌گیرنده** (ADAPT). این ترکیب از قدرت تبیین بصری (مثلاً با استفاده از نقشه توجه) و قابلیت تولید متن توسط LLM بهره می‌برد تا برای انسان قابل فهم و اعتماد باشد.

۲–۳–۲– مدل‌های زبانی بزرگ و تفسیر تصمیمات

مدل‌های زبانی بزرگ مانند GPT-۳ و GPT-٤ در سال‌های اخیر توجه زیادی را به خود جلب کرده‌اند. این مدل‌ها با تبدیل خروجی‌های پیچیده به روایت‌هایی قابل فهم برای انسان، به ایجاد ارتباط بین رفتار داخلی مدل‌های هوش مصنوعی و درک انسانی کمک می‌کنند [۳۳]. علاوه بر این، حجم عظیم داده‌های آموزشی LLM توانایی برجسته‌ای برای روشن‌تر کردن ساختار و فرآیندهای تصمیم‌گیری مدل‌های پیچیده فراهم می‌آورد [۳۳].

تعریف، معماری و تاریخچه مدل‌های زبانی بزرگ

---

[۳۰] Local Explanation

[۳۱] Post-hoc

مدل‌های زبانی بزرگ به شبکه‌های عصبی عمیقِ با پارامترهای بسیار زیاد گفته می‌شود که با آموزش خودنظارتی بر روی حجم عظیمی از داده‌های متنی، قادر به درک و تولید زبان طبیعی هستند [۳۴]. پایه معماری اغلب این مدل‌ها بر مبنای ساختار Transformer است که برای اولین بار در سال ۲۰۱۷ معرفی شد. [۳۵] Transformer با مکانیزم Attention خود می‌تواند وابستگی‌های طولانی‌مدت در ورودی‌های متنی را مدل کند و محدودیت‌های روش‌های پیشین مانند شبکه‌های بازگشتی[۳۲] و کانولوشن[۳۳] را پشت سر می‌گذارد [۳۴]. یک سال پس از معرفیTransformer ، مدل‌های برجسته‌ای مانند BERT و GPT-۱ با مقیاس صدها میلیون پارامتر ارائه شدند که با بهره‌گیری از روند پیش‌تمرین[۳۴] روی مجموعه‌داده‌های گسترده، ویژگی‌های عمیقی از زبان را آموختند [۳۴].

در روش متداول LLM، ابتدا مدل در وظایف عمومی زبانی (مانند پیش‌بینی واژه بعدی) پیش‌تمرین می‌شود و سپس برای وظایف خاص‌تر با داده‌های کم‌تر به صورت Fine-Tuning درنظر گرفته می‌شود [۳۴]. برای نمونه، اولین نسل‌های LLM همچون ELMo ، GPT-۱ و BERT با استفاده از چنین رویکردی توانستند درک و ویژگی‌های پیچیده‌تری از زبان به دست آورند [۳۴]. با این حال، نقطه‌عطف اصلی در پیشرفت این حوزه زمانی رخ داد که مدل‌ها با مقیاس بسیار بزرگ‌تری توسعه یافتند و قابلیت «پرسش و پاسخ مبتنی بر متن[۳۵]» را نشان دادند. به‌عنوان مثال، در فوریه ۲۰۱۹ مدل[۳۶] GPT-۲ با حدود ۱.۵ میلیارد پارامتر معرفی شد که می‌توانست بدون تنظیم اضافه پاسخ‌های معنادار تولید کند [۳۴]. در سال ۲۰۲۰ مدل [۳۷] GPT-۳ به ابعاد قابل توجه ۱۷۵ میلیارد پارامتر رسید که توانست با پدیده‌های جدیدی از جمله قابلیت‌های «ظهور[۳۶]» در پردازش زبان

---

[۳۲] Recurrent Neural Networks (RNN)

[۳۳] Convolutional Neural Network (CNN)

[۳۴] Pre-training

[۳۵] Prompt-based Learning

[۳۶] Emergent Abilities

مواجهه ایجاد کند. [۳۴] پس از آن، مدل‌هایی مانند T۵[۳۸] ، FLAN[۳۹]، ERNIE ۳٫۰[٤۰] و PaLM[٤۱] عرضه شدند که نشان‌دهنده رشد مداوم مقیاس و توانایی LLM ها بودند. [۳۴].

در ادامه روند توسعه، نسل‌های جدیدتر مانند GPT-٤ ارائه شده‌اند. GPT-٤ یک مدل چندرسانه‌ای است که ورودی‌های متنی و تصویری را می‌پذیرد و در آزمون‌های حرفه‌ای مختلف (مانند آزمون حقوق) عملکردی در سطح انسانی نشان داده است [۴۲]. علاوه بر این، گوگل با معرفی خانواده مدل‌های Gemini گامی جلوتر نهاده است. بزرگ‌ترین نسخه‌ی این خانواده (Gemini Ultra) با ارایه توانایی‌های قوی در درک تصویر، ویدئو، صدا و متن، در بیش از ۳۰ بنچمارک مختلف پیشرفت قابل توجهی داشت و برای نخستین بار در آزمون چندموضوعی MMLU عملکردی مشابه متخصصان انسانی کسب کرد [۴۳]. قابلیت‌های جدید Gemini در استدلال میان‌رسانه‌ای و درک زبان پیچیده، زمینه‌ای برای کاربردهای نوینی مانند سیستم‌های چندرسانه‌ای فراهم کرده است [۴۳].

مدل زبانی بزرگ به‌عنوان ماژول تولید توضیح در مدل‌های چندوجهی بینایی-زبان

در مدل‌های ترکیبی دیداری-زبانی[۳۷]، از LLM ها برای تولید توضیحات متنی درباره داده‌های تصویری یا ویدیویی استفاده می‌شود. این مدل‌ها به‌طور کلی ورودی‌های چندرسانه‌ای (مانند تصویر یا ویدئو) را به توکن‌های متنی تبدیل کرده و سپس از LLM برای استنتاج و ارائه پاسخ‌های زبان‌طبیعی بهره می‌برند. به عنوان مثال:

- DriveGPT٤: یک سیستم خودران توضیح‌پذیر است که به وسیله‌ی LLM ساخته شده. در DriveGPT٤ ویدئوی ثبت‌شده از صحنه رانندگی به عنوان ورودی به مدل داده می‌شود و مدل ضمن پیش‌بینی سیگنال‌های کنترلی مانند سرعت و زاویه فرمان، به پرسش‌های کاربران پاسخ می‌دهد و رفتار خودرو را توضیح می‌دهد. برای نمونه، این مدل قادر است «خودرو در این لحظه حرکت می‌کند» را توصیف کرده و دلیل آن را مانند «مسیر مقابل باز است و می‌توان

[۳۷] Vision-language

به سمت راست پیچید» در قالب توضیح ارائه دهد [۴۴]. این قابلیت‌ها با ایجاد یک مجموعه داده دستورالعملی ویژه رانندگی (BDD-X) و Fine-Tuning LLM با استفاده از چت‌بات‌هایی نظیر ChatGPT محقق شده است [۴۴].

- LVX: رویکردی سیستماتیک برای تبیین درونیات مدل‌های بینایی با کمک LLM است. در این روش، از LLM (مانند GPT-٤) به عنوان یک منبع دانش استفاده می‌شود تا سلسله‌مراتبی از ویژگی‌های بصری هر کلاس را تشریح کند و سپس این توضیحات به همراه تصویرسازی متنی با استفاده ازAPI های متن به تصویر تبدیل می‌شوند. نتایج حاصله برای ساختن یک «درخت توضیح» به کار می‌رود که مسیر طی‌شده در آن نشان‌دهنده مفاهیمی است که مدل بینایی به‌خوبی شناسایی می‌کند. خروجی نهایی LVX یک توضیح انسانی‌فهم در قالب یک درخت شامل ویژگی‌های بصری مرتبط با تصمیم مدل است [۴۵]. به این ترتیب، LLM با فراهم کردن دانش متنی به مدل بینایی کمک می‌کند تا تفسیری کاملاً متنی و ساختاریافته از رفتار خود ارائه دهد.

به طور کلی، بسیاری از مدل‌های چندوجهی جدید (مانند InstructBLIP، LLaVA و MiniGPT-٤) نیز از این ایده‌ی ترکیب LLM با encoder های بینایی پیروی می‌کنند و LLM را به عنوان بخش تولید پاسخ‌های متنی توضیحی درباره تصویر یا ویدئو به کار می‌گیرند.

مزایا و چالش‌های استفاده از LLM برای تفسیر مدل‌ها

- مزایا:
  - **انعطاف‌پذیری زبانی**: LLM ها می‌توانند توضیحاتی به زبان طبیعی و گویا تولید کنند که برای کاربران انسان‌محور و غیرمتخصص قابل فهم باشد. این ویژگی، ارتباط بین خروجی پیچیده مدل‌ها و فهم انسان را تسهیل می‌کند [۳۳].

- **توان تعمیم بالا**: با آموزش روی مجموعه‌داده‌های عظیم و متنوع، LLM ها دانش کلی گسترده‌ای کسب می‌کنند و به سرعت می‌توانند به مسائل یا حوزه‌های جدید تعمیم یابند. به‌عبارت دیگر، آن‌ها اطلاعات و قواعد زبان را از مقادیر زیادی متن فرا می‌گیرند و در کاربردهای مختلف از ترجمه تا تلخیص و پاسخ‌گویی عملکرد برجسته‌ای نشان می‌دهند [۳۴].
- **سازگاری با داده‌های متنوع**: بسیاری از LLM های جدید چندرسانه‌ای هستند و می‌توانند ورودی‌های متنی، تصویری، صوتی و حتی داده‌های طولانی‌مدت را یک‌جا پردازش کنند. به عنوان مثال، خانواده مدل‌های Gemini نشان داده‌اند که توانایی‌های چشمگیری در درک تصویر، ویدئو و متن دارند [۴۳]. این چندوجهی‌بودن، LLM ها را برای کاربردهایی مثل سیستم‌های پزشکی یا سامانه‌های ترکیبی دیداری-زبانی بسیار کارآمد می‌کند [۴۳].

- چالش‌ها:

- **توهم**[۳۸]: LLM ها گاهی اوقات پاسخ‌هایی تولید می‌کنند که با واقعیت ورودی منطبق نیست و اطلاعات ساختگی در توضیحات وارد می‌کنند. این خطاها می‌تواند دقت تفسیرها را کاهش داده و به اشتباه کاربران بینجامد [۴۶].
- **محدودیت حافظه‌ی زمینه**: LLM ها تنها می‌توانند تعداد محدودی از توکن‌ها را هم‌زمان پردازش کنند (مثلاً مدل پایه GPT-٤ حدود ۸۱۹۲ توکن را می‌تواند ورودی بگیرد [۴۷]). این محدودیت باعث می‌شود اگر داده‌های توضیحی بسیار طولانی باشند یا به دنبال اطلاعات پیشین باشیم، نتوان به صورت کامل همه آن‌ها را در نظر گرفت.
- **هزینه‌ی محاسباتی بالا**: آموزش و اجرایLLM ها نیازمند منابع پردازشی بسیار قدرتمند است. هزینه‌های پیش‌آموزش و Fine-Tuning این مدل‌ها بسیار زیاد است

---

[۳۸] Hallucination

[۳۴]؛ برای مثال فقط آموزش نسخه‌های بزرگ مدل می‌تواند به مصرف چندصد پتابایت–ساعت انرژی بیانجامد.

- **کنترل‌پذیری کم**: خروجیLLM ها گاهی از چارچوب خواسته‌شده فاصله می‌گیرد و کنترل دقیق محتوای تولیدی دشوار است. مطالعات نشان داده‌اند حتی روش‌هایی مانند پرسش‌های زنجیره‌ای گاهی توضیحاتی ارائه می‌کنند که از نظر محتوا دقیق نیست یا تحت تأثیر سوگیری قرار دارد [۴۶]. این عدم قطعیت در تولید توضیح، نیازمند نظارت و پالایش اضافی است.

۲–۳–۳– مهندسی دستور

مهندسی دستور به طراحی و بهینه‌سازی سیستماتیک ورودی‌ها (دستورات متنی) برای هدایت پاسخ‌های مدل‌های زبان بزرگ گفته می‌شود. این فرایند با هدف افزایش دقت، ارتباط و هماهنگی خروجی مدل انجام شده و برای بهره‌برداری حداکثری از قابلیت‌های این مدل‌ها ضروری است [۴۸, ۴۹]. به عبارت دیگر، با Fine-Tuning قالب و محتوای دستور، می‌توان رفتار مدل را به سمت تولید پاسخ‌های مدنظر هدایت کرد و از خطاهایی مانند تولید اطلاعات غیرمرتبط یا نامعتبر جلوگیری نمود [۴۸, ۴۹].

اجزای کلیدی یک دستور مؤثر

یک دستور مهندسی‌شده و مؤثر، معمولاً از چندین جزء ساختاری تشکیل شده است که هر یک نقش مشخصی در هدایت مدل ایفا می‌کنند. اگرچه همه دستورها لزوماً تمام این اجزا را ندارند، اما شناخت آن‌ها به طراحی ورودی‌های دقیق‌تر کمک می‌کند[۵۰]:

- **نقش**[۳۹]: پرامپت می‌تواند با تعیین نقش یا هویت مشخصی برای مدل زبانی آغاز شود. در این حالت مدل موظف است پاسخ‌ها را از منظر یک شخصیت یا تخصص معین ارائه دهد. برای نمونه، در بسیاری از سیستم‌های گفتاری، یک پیام سیستمی در ابتدای گفتگو قرار می‌گیرد و نقش مدل را مشخص می‌کند (مثلاً «شما یک دستیار مفید هستید») تا لحن و محتوای پاسخ‌ها به آن شکل بگیرد [۵۱]. بر اساس تحلیل‌های قالب پرامپت، تعیین «پروفایل» یا نقش معادل این است که مشخص شود «چه کسی» یا «چه شخصیتی» پاسخ می‌دهد [۵۰] این روش باعث می‌شود مدل پاسخ‌های خود را مطابق با دانش ضمنی و سبک مورد انتظار آن نقش تنظیم کند.
- **زمینه اطلاعاتی**[۴۰]: پس از تعیین نقش، معمولاً اطلاعات زمینه‌ای لازم است تا مدل درکی بهتر از وظیفه داشته باشد. زمینه شامل هر گونه دانش یا داده قبلی است که به مدل داده می‌شود؛ از جمله توضیحات پس‌زمینه، متن مرتبط یا داده‌های پیشین که می‌تواند پاسخ را هدایت کند. با افزودن این اطلاعات، مدل می‌تواند پاسخ‌هایی دقیق‌تر و مرتبط‌تر ارائه دهد. به عبارت دیگر، زمینه همان اطلاعات کمکی است که مدل برای «فهمیدن» بهتر مسئله به آن نیاز دارد [۴۹, ۵۰]. برای مثال، ارائه جزئیات یک مسئله یا توضیح مختصری درباره موضوع، به مدل کمک می‌کند تا پاسخ‌های خود را متناسب با آنچه در بافت مسئله رخ می‌دهد تنظیم کند.
- **وظیفه و دستور صریح**[۴۱]: هسته اصلی هر پرامپت، وظیفه‌ای است که از مدل انتظار می‌رود انجام دهد. این وظیفه معمولاً در قالب یک سؤال یا یک دستور العمل واضح بیان می‌شود. به طور رسمی این بخش «دستور» یا «Directive» نامیده می‌شود و بیانگر نیّت اصلی پرسش است. برای مثال، عباراتی مانند «متن زیر را خلاصه کن» یا «مقاله ذیل را به انگلیسی

---

[۳۹] Persona

[۴۰] Context

[۴۱] Task/Instruction

ترجمه کن» در حکم دستور صریح هستند. تحلیل‌های مربوط به قالب‌های پرامپت نشان می‌دهد که بیش از ٪۹۰ از دستورات به صورت بیان شفاف (مثلاً «خلاصه کن» به جای سؤال) نوشته شده‌اند تا مدل به طور مستقیم متوجه کارکرد مورد نظر شود [۵۰]. وجود یک دستور روشن و دقیق باعث می‌شود مدل هنگام تولید پاسخ، از فراهم کردن اطلاعات نامربوط اجتناب کرده و تمرکز خود را بر انجام همان وظیفه مشخص‌شده حفظ کند.

- **داده ورودی**[۴۲]: علاوه بر دستور، پرامپت ممکن است شامل داده‌های خاص ورودی باشد که مدل باید بر پایه آن‌ها عمل کند. این داده‌ها می‌توانند متن، فهرست سوالات، مقادیر عددی یا هر نوع اطلاعات دیگری باشند که قرار است مدل روی آن کار کند. به عبارت دیگر، ورودی اطلاعاتی است که کاربر یا سیستم در قالب پرامپت ارائه می‌دهد و خروجی مدل باید به آن مرتبط باشد. بررسی‌های انجام‌شده نشان می‌دهد که هر پرامپت معمولاً شامل دستورالعمل و داده ورودی یا مثال‌هایی است که به مدل نشان می‌دهد چه اطلاعاتی باید پردازش شود. [۵۲]به عنوان مثال، اگر پرامپت حاوی یک مقاله کوتاه باشد و دستور «این مقاله را خلاصه کن»، داده ورودی همان مقاله است که باید مدل آن را «ورودی» خود بداند و بر اساس آن عمل کند.
- **قالب و سبک خروجی**[۴۳]: در اغلب پرامپت‌ها فرمت یا ساختار مورد انتظار پاسخ به‌صراحت تعیین می‌شود. این مشخصه «قالب خروجی» نامیده می‌شود و می‌تواند شامل نحوه سازماندهی پاسخ (لیستی، پاراگرافی، جدول و غیره)، زبان یا سبک نگارش (رسمی، غیررسمی و غیره) و هر ویژگی ساختاری دیگری باشد. تعیین قالب خروجی باعث می‌شود پاسخ تولیدی یکنواخت و سازگار باشد؛ به‌طوری که پس‌پردازش یا درک آن برای کاربران یا سیستم‌های بعدی آسان باشد. مطالعات نشان می‌دهد که توسعه‌دهندگان پرامپت معمولاً

---

[۴۲] Input Data

[۴۳] Output Format/Style

قالب دقیق خروجی (مانند «فهرست‌وار بنویس») و محدودیت‌های موردنظر را مشخص می‌کنند تا پاسخ تولیدشده قابل‌پیش‌بینی و پردازش باشد [۵۰]. با تعیین «قالب و سبک خروجی»، می‌توان حتی لحن نگارش (مثلاً رسمی یا محاوره‌ای بودن نوشته) را نیز از پیش تعیین کرد [۴۹].

- **مثال‌های راهنما**[۴۴]: در روش‌های چندشاتی[۴۵]، به پرامپت نمونه‌های ورودی–خروجی از پیش مشخص ارائه می‌شود که به مدل نشان می‌دهند خروجی مطلوب چگونه باید باشد. این نمونه‌ها که «exemplars» نامیده می‌شوند، به مدل کمک می‌کنند تا ساختار و محتوای دلخواه پاسخ را بهتر یاد بگیرد. برای مثال، اگر بخواهیم مدل فرمت خاصی را رعایت کند یا سبک پاسخ‌های خاصی را تقلید کند، چند مثال متناظر با آن سبک را در پرامپت می‌آوریم. استفاده از مثال‌های راهنما در پرامپت این مزیت را دارد که مدل می‌تواند از شیوه نمونه‌ها تبعیت کند و پاسخ‌های خود را در همان چارچوب تولید نماید [۴۹, ۵۰]. به طور خلاصه، مثال های راهنما پیش‌نمایش‌هایی هستند که نمونه «پاسخ درست» را به مدل نشان می‌دهند تا در تولید پاسخ اصلی راهنمایش کنند.
- **محدودیت‌ها**[۴۶]: محدودیت‌ها قیدها و شرایط خاصی هستند که روی خروجی مدل اعمال می‌شوند. این قیدها می‌توانند شامل طول پاسخ (برای مثال تعداد کلمات یا جملات)، دامنه موضوعی (مثلاً اجتناب از موضوعات خاص)، یا سایر شرایط دقیق باشند. درج محدودیت در پرامپت باعث می‌شود پاسخ نهایی با نیازمندی‌های دقیق مسئله همخوان باشد. برای نمونه، ممکن است در پرامپت تأکید شود که «حداکثر ۵۰ کلمه بنویس» یا «از اصطلاحات تخصصی استفاده نکن». تعیین محدودیت‌ها از طریق پرامپت، خروجی تولیدشده را قابل‌پیش‌بینی‌تر کرده و از ایجاد خطاهای نامطلوب جلوگیری می‌کند [۴۹, ۵۰]. به طور

---

[۴۴] Exemplars

[۴۵] few-shot

[۴۶] Constraints

کلی، محدودیت ها تضمین می‌کنند که مدل هنگام تولید پاسخ، قوانین و خواسته‌های خاص مورد نظر طراح پرامپت را رعایت کند.

- **سبک نگارش**[۴۷]: سبک یا لحن نوشتاری پاسخ از عناصر تکمیلی یک پرامپت خوب است. با مشخص کردن سبک موردنظر، می‌توان رفتار زبانی مدل را جهت داد. برای مثال در پرامپتی ممکن است قید شود که جواب باید «رسمی نوشته شود» یا «به زبان ساده و دوستانه مطرح شود». مطالعات و منابع راهنمای مهندسی پرامپت توصیه می‌کنند با تعیین صریح سبک (Genre, Tone, Style) در متن پرامپت می‌توان به شکل موثری خروجی دلخواه را قالب‌بندی کرد [۴۹]. به این ترتیب مدل می‌داند که باید چه سبکی را تقلید کند؛ برای مثال با مثال‌ها یا دستورالعمل‌هایی که نشان دهند پاسخ باید عاطفی، خلاقانه یا فنی باشد. تعیین سبک نگارش مناسب در پرامپت، یکی از روش‌های مؤثر برای کنترل کیفیت و نوع پاسخ تولیدی محسوب می‌شود.

تعریف و انواع دستورها

روش‌های مختلف مهندسی دستور بر اساس قالب ورودی و نحوه تعامل با مدل تفاوت دارند. مهم‌ترین انواع دستور عبارتند از:

- **دستور سخت**[۴۸]: دستورهایی هستند که به صورت متن طبیعی و قابل فهم برای انسان نوشته می‌شوند، شامل دستورالعمل‌ها یا سؤالات صریح می‌شوند[۵۳].
- **دستور نرم**[۴۹]: به جای استفاده از متن واضح، در این روش از بردارهای پیوسته قابل آموزش (آماری) به عنوان دستور استفاده می‌شود. در تمرین دستور نرم، وزن‌های مدل ثابت نگه داشته شده و صرفاً بردارهای ویژه‌ای (معمولاً پیشوند یا پسوند) آموزش داده می‌شوند [۵۳].

---

[۴۷] Style

[۴۸] Hard Prompt

[۴۹] Soft Prompt

- **دستور مبتنی بر مثال**[۵۰]: در این شیوه چند نمونه ورودی–خروجی مرتبط با کار هدف، درون دستور گنجانده می‌شود. ارائه نمونه‌ها باعث می‌شود مدل بهتر با قالب یا سبک مورد انتظار آشنا شود و پاسخ‌های دقیق‌تری تولید کند.
- **دستور زنجیره افکار**[۵۱]: این تکنیک مدل را تشویق می‌کند تا مراحل میانی استدلال خود را تشریح کند. به عبارت دیگر، با افزودن درخواست «توضیح گام به گام» به دستور، مدل مجبور می‌شود پاسخ را در قالب یک زنجیره منطقی ارائه دهد [۵۳].

روش‌های پیشرفته مهندسی دستور

در کنار روش‌های پایه، تکنیک‌های پیچیده‌تری نیز در مهندسی دستور پدید آمده‌اند که دامنه کاربرد را گسترش می‌دهند:

- **Prophet (۲۰۲۳)**: چارچوبی دو مرحله‌ای برای مهندسی دستور در مسئله پاسخ‌دهی به سوالات تصویری وابسته به دانش[۵۲] است [۵۴]. در این روش ابتدا یک مدل VQA عادی روی مجموعه داده مورد نظر آموزش داده می‌شود تا دانش مقدماتی استخراج شود. سپس دو نوع حدس[۵۳] پاسخ شامل «نامزدهای پاسخ» و «نمونه‌های وابسته به پاسخ» از این مدل گرفته شده و در قالب یک دستور به LLM افزوده می‌شود [۵۴]. بدین ترتیب LLM با بهره‌گیری از سرنخ‌های استخراج‌شده، پاسخ دقیق‌تر و مرتبط‌تری می‌سازد. تجربه‌ها نشان داده است که Prophetبا ترکیب GPT-۳ توانسته در چندین مجموعه داده چالش‌برانگیز VQA عملکرد بهتری نسبت به روش‌های پیشین ارائه دهد [۵۴].

---

[۵۰] Example-based
[۵۱] Chain-of-Thought
[۵۲] knowledge-based VQA
[۵۳] Heuristic

- Fine-Tuning **مدل**: اگرچه FT دقیقاً در حوزه «مهندسی دستور» نیست، اما به عنوان روشی جایگزین برای سازگار کردن مدل با کارهای خاص مطرح است. در FT، مدل زبان از قبل آموزش‌دیده با مجموعه داده ویژه‌ای بازآموزی می‌شود تا وزن‌های آن مطابق با دامنه یا سبک خاصی تنظیم گردد [۵۵]. این کار معمولاً دقت و اختصاصی بودن خروجی‌ها را افزایش می‌دهد، ولی مستلزم قدرت محاسباتی بالا و داده‌های آموزشی قابل توجه است [۵۵]. برای کاهش هزینه محاسباتی، روش‌هایی مانند LoRA (افزودن لایه‌های کم‌بعد آموزش‌پذیر) نیز استفاده می‌شوند [۵۵]. در مجموع، فاین‌تیونینگ در مسائل اختصاصی نتایج بهتری می‌دهد، اما مهندسی دستور به عنوان راه‌حلی سریع‌تر و کم‌هزینه‌تر امکان تطبیق مدل بدون تغییر ساختار آن را فراهم می‌کند [۵۵].

کاربرد دستور در کنترل خروجی LLM

از طریق مهندسی دستور می‌توان رفتار مدل را در هنگام تولید پاسخ کنترل کرد. به ویژه در مسئله تولید توضیحات و استدلال (مثلاً توضیح یک پاسخ یا تشریح روش حل) دستور نقش حیاتی دارد. به عنوان مثال، با اضافه کردن عبارت‌هایی مانند «لطفاً استدلال خود را قدم‌به‌قدم توضیح دهید» به دستور، مدل مجبور می‌شود پاسخ نهایی را همراه با مراحل میانی منطقی ارائه کند [۵۶]. در چنین حالتی مدل به جای یک پاسخ خام، یک توضیح جامع همراه با استدلال می‌دهد. افزون بر این، انتخاب قالب و لحن دستور می‌تواند سبک خروجی را تنظیم کند؛ مثلاً می‌توان با تعیین نقش مدل (مانند «شما به عنوان یک متخصص توضیح دهید») یا تنظیم درجه «دمای نمونه‌برداری[۵۴]» میزان خلاقیت و طول پاسخ را تنظیم کرد [۵۵]. به طور کلی، طراحی دقیق دستور به ما امکان می‌دهد خروجی LLM را در فرمت و ساختار دلخواه شکل دهیم و اطمینان حاصل کنیم که جزئیات و شیوه مورد نیاز (نظیر توضیح دقیق) در پاسخ لحاظ شود [۵۶].

---

[۵۴] temperature

مقایسه روش‌های مختلف مهندسی دستور

روش‌های متفاوت مهندسی دستور مزایا و معایب خاص خود را دارند:

- **دستور سخت در مقابل نرم:** دستور سخت به راحتی قابل نوشتن و فهم انسانی است و نیاز به بازآموزی مدل ندارد، اما حساسیت بالایی به نگارش دارد و کیفیت خروجی مستقیماً به کیفیت دستورات متنی وابسته است [۵۳]. در مقابل، دستور نرم با آموزش یک بردار درونی، قابلیت تطبیق بهتری را فراهم می‌کند و از قدرت کامل مدل بهره می‌برد، اما نیاز به داده آموزش و ترفندهای تنظیم دارد و قابل مشاهده یا ویرایش توسط انسان نیست [۵۳].
- **صفر-شات در مقابل چند-شات**: دستورهای صفر-شات (بدون مثال) سریع و ساده هستند ولی ممکن است مدل را در درک دقیق وظیفه محدود کنند. استفاده از نمونه‌های آموزشی (چند-شات) می‌تواند مدل را با قالب موردنظر آشنا کند و عملکرد آن را بهبود بخشد [۴۹] ، اما به دستور طولانی‌تر نیاز دارد و امکان دارد با افزودن نمونه‌های نامناسب، سردرگمی ایجاد کند.
- **زنجیره افکار**: این روش کارایی استدلال مدل را بالا می‌برد، به ویژه در مسائل پیچیده، اما موجب افزایش طول پاسخ و هزینه محاسباتی می‌شود [۵۳]. همچنین طراحی زنجیره افکار مناسب نیازمند دقت و گاهی آزمون‌وخطا است.
- **مهندسی دستور در مقابل فاین‌تیونینگ**: مهندسی دستور به عنوان راهکاری سریع‌تر و کم‌هزینه‌تر قادر است در حوزه‌های متنوع اعمال شود، اما محدودیتش در وابستگی به فرمت دقیق دستور و کیفیت دستور است [۵۳, ۵۵]. از طرف دیگر، فاین‌تیونینگ به تغییر وزن‌های مدل می‌پردازد و معمولاً دقت بالاتری در وظایف خاص ارائه می‌دهد، اما به سرمایه‌گذاری زمانی و محاسباتی بیشتر نیاز دارد [۵۵].

به طور کلی، انتخاب روش مناسب مهندسی دستور بسته به مسئله، منابع در دسترس و نیازمندی‌های نهایی انجام می‌شود. در این بخش روش‌ها به طور کلی معرفی شدند و مقایسه گردیدند؛ نتایج تجربی مربوط به کاربرد این روش‌ها در فصول بعدی آورده خواهد شد.

۲–۳–۴– مدل‌های تصمیم‌گیر در سیستم‌های خودران

در این بخش به بررسی مدل‌های مهم تصمیم‌گیری در سیستم‌های خودران می‌پردازیم. این مدل‌ها شامل معماری‌های کلاسیک یادگیری سرتاسری (از ورودی تصویر تا خروجی فرمان) و همچنین مدل‌های نوین‌تر با توانایی تبیین‌پذیری (ارائه‌ی توضیحات متنی برای تصمیمات) هستند. با پیشرفت این حوزه، روش‌هایی مانند مکانیزم‌های حافظه‌دار (برای درک توالی زمانی) و مکانیزم‌های توجه بصری به‌کار گرفته شده‌اند تا مدل‌ها بتوانند در شرایط پیچیده رانندگی عملکرد بهتری داشته باشند و قابل اعتمادتر باشند. در ادامه به هریک از این رویکردها و نمونه مدل‌های شاخص آن می‌پردازیم.

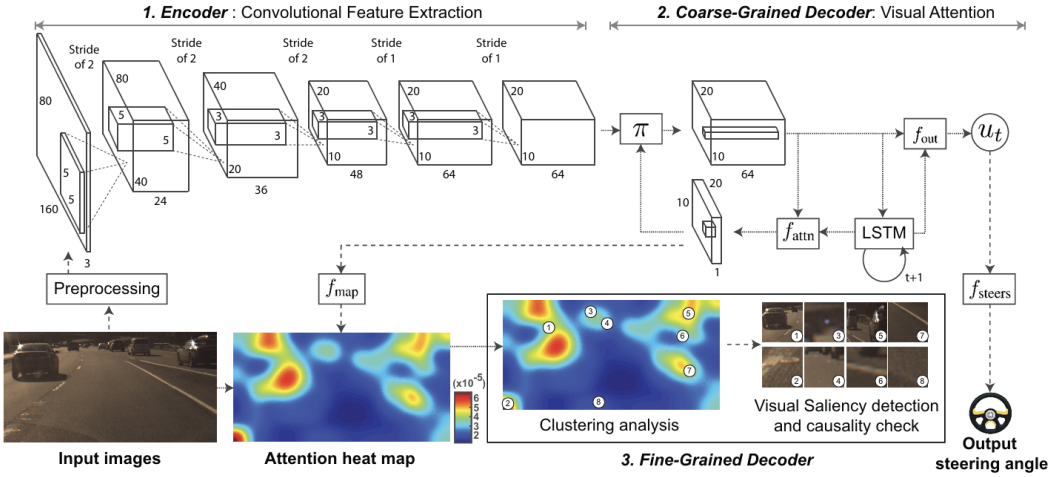
معماری‌های کلاسیک تصمیم‌گیری خودران (از PilotNet تا مدل‌های حافظه‌دار)

یکی از نخستین نمونه‌های موفق یادگیری انتها–به–انتها در رانندگی خودران، شبکه‌ی PilotNet شرکت انویدیا بود [۵۹]. این شبکه یک شبکه‌ی عصبی پیچشی عمیق است که ورودی آن تنها تصویر دوربین جلو (نمای جاده) و خروجی آن فرمان‌های کنترلی رانندگی (مانند زاویه فرمان) می‌باشد[۵۹] . PilotNet با یادگیری از روی داده‌های رانندگی انسان (تصاویر و فرمان متناظر) توانست رانندگی در بین خطوط را در شرایط متنوع (با یا بدون خط‌کشی جاده) به‌خوبی انجام دهد [۵۹]. این موفقیت نشان داد که شبکه‌های عصبی می‌توانند بدون قواعد صریح برنامه‌نویسی‌شده، الگوهای لازم برای کنترل خودرو را از داده بیاموزند [۵۹].

با این حال، یک ضعف مهم PilotNet و مدل‌های تک‌فریمی مشابه، نبود حافظه‌ی زمانی است. این مدل‌ها تنها براساس تصویر لحظه‌ی فعلی تصمیم می‌گیرند و اطلاعات گذشته‌ی نزدیک را در نظر نمی‌گیرند. همین امر در سناریوهای پیچیده (مثلاً هنگام مواجهه با علائم یا حوادثی که چند لحظه قبل

دیده شده‌اند) می‌تواند کارایی مدل را محدود کند [۶۰]. برای رفع این مشکل، پژوهشگران به سراغ مدل‌های حافظه‌دار رفتند. دو رویکرد رایج عبارت بودند از:

- ترکیب CNN با شبکه‌های بازگشتی LSTM برای نگهداری حالت گذشته در مدل.
- استفاده از شبکه‌های سه‌بعدی CNN۳D که توالی فریم‌های ویدئویی را به صورت یک بلوک حجمی پردازش می‌کنند.



تصویر ۲–۱–نمونه ای از مدل کیم و همکاران (۲۰۱۷) که برای بهبود مدل PilotNet به آن لایه ی LSTM اضافه نموده است [۵۶].

این ایده‌ها به مدل امکان می‌دهند وابستگی‌های زمانی کوتاه‌مدت را درک کرده و در مواجهه با رویدادهای پیچیده عملکرد بهتری داشته باشد [۶۰]. به عنوان مثال، محققان نشان داده‌اند افزودن یک لایه‌ی LSTM به انتهای شبکه‌ی بینایی می‌تواند بهبود معناداری در کنترل خودرو در سناریوهای دشوار ایجاد کند، زیرا مدل را نسبت به وقایع چند لحظه‌ی قبل هوشیار نگاه می‌دارد [۶۰]. به طور کلی، معماری‌های کلاسیک مبتنی بر بینایی (مانند PilotNet) با چنین ارتقاءهایی توانستند گام‌های اولیه را در تصمیم‌گیری خودکار خودرو بردارند. هرچند این مدل‌ها در زمان خود عملکرد قابل قبولی داشتند،

اما چالش‌هایی نظیر شفاف نبودن تصمیم‌ها و دشواری تطبیق با موقعیت‌های خارج از توزیع آموزش، انگیزه‌بخش تحقیقات بعدی شد.

معماری ADAPT و تحول در تفسیر تصمیم‌ها

با پیشرفت شبکه‌های عصبی، مدل‌های جدیدتری معرفی شدند که دقت بالاتر و ساختار پیشرفته‌تری برای تبیین تصمیمات داشتند. یکی از برجسته‌ترین آن‌ها معماری ADAPT است که در سال ۲۰۲۳ (توسط جین و همکاران) معرفی شد [۹]. نام ADAPT مخفف Action-aware Driving Caption Transformer بوده و بیانگر ماهیت آن است: یک مدل ترنسفورمر که شرح اقدام رانندگی را به زبان طبیعی تولید می‌کند [۹].

ADAPT یک معماری انتها-به-انتها مبتنی بر ترنسفورمر است که به طور همزمان دو کار را انجام می‌دهد: پیش‌بینی فرمان کنترلی خودرو (مانند مدل‌های پیشین) و تولید توضیح متنی تصمیم همان لحظه [۹]. در این مدل، ویدئوی دریافتی ابتدا به یک نمایش فشرده‌ی ویژگی‌ها تبدیل می‌شود و سپس یک دیکدر ترنسفورمر با استفاده از این نمایش، جمله‌ای تولید می‌کند که هم وضعیت جاری رانندگی را روایت می‌کند و هم دلیل اقدام را بیان می‌نماید [۹]. برای مثال، ADAPT قادر است در لحظه بگوید «خودرو به‌سمت راست تغییر مسیر می‌دهد تا از خودروی پارک‌شده جلوتر عبور کند»، که شامل هم شرح عمل (تغییر مسیر به راست) و هم دلیل (وجود خودروی پارک‌شده در جلو) است.

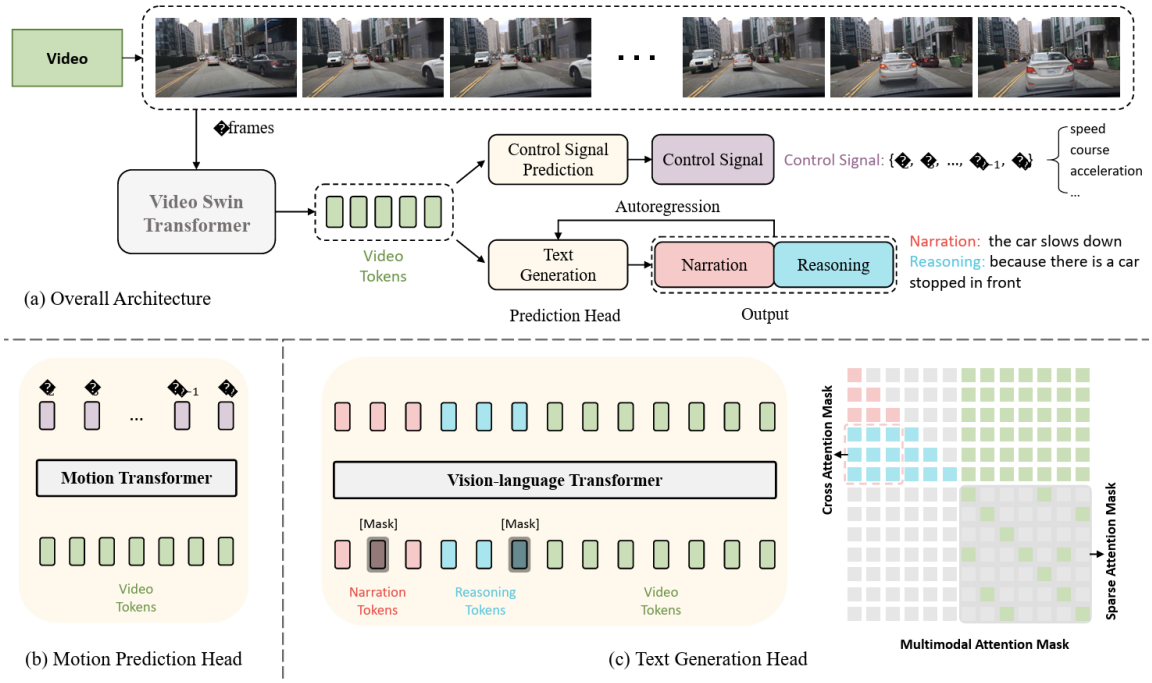
هدف اصلی معماری ADAPT ارتقای شفافیت و قابلیت تبیین در رانندگی خودکار است. این مدل برای هر مرحله از تصمیم‌گیری خودران، یک شرح حرکت[۵۵] به زبان طبیعی به‌همراه دلیل[۵۶] آن ارائه می‌کند [۹]. به عنوان نمونه، ممکن است مدل توضیح دهد: «خودرو به سمت راست جاده می‌کشد (شرح حرکت) چون قصد پارک کردن دارد (دلیل)» [۹]. چنین خروجی زبانیِ دوبخشی کمک می‌کند

---

[۵۵] Narration

[۵۶] Reasoning

تا سرنشینان و ناظران بفهمند چه عملی توسط خودرو انجام می‌شود و چرا آن عمل انجام شده است. معماری ADAPT با تولید این روایت‌ها و استدلال‌ها در کنار سیگنال‌های کنترلی، گامی به سوی تفسیرپذیری بهتر سیستم‌های رانندگی خودکار برداشته است [۹].



تصویر ۲–۲–معماری مدل ADAPT [۹]

ویژگی ممتاز ADAPT ، هم‌آموزی چندوظیفه‌ای[۵۷] آن است. سازوکار آموزش مدل به‌گونه‌ای است که وظیفه‌ی تولید متن توضیحی و وظیفه‌ی پیش‌بینی فرمان کنترل خودرو به‌صورت همزمان و با اشتراک ویژگی‌های ویدیویی انجام می‌گیرند [۹]. این رویکرد چندوظیفه‌ای باعث می‌شود مدل بازنمایی‌های ویدیویی بهتری بیاموزد که هم برای رانندگی و هم برای توصیف زبانی مفید باشند. در مقاله‌ی اصلی ADAPT نشان داده شده است که آموزش مشترک این دو وظیفه کیفیت توضیحات متنی

[۵۷] Multi-task Learning

را بهبود می‌بخشد؛ به طور مشخص، آموزش چندوظیفه‌ای عملکرد تولید شرح و استدلال را به میزان قابل توجهی نسبت به حالتی که مدل تنها متن را یاد بگیرد ارتقاء می‌دهد [۹]. همچنین نتایج آنها حاکی از آن است که داشتن اطلاعات کنترلی در حین آموزش، حتی موثرتر از دادن مستقیم آن اطلاعات به ورودی مدل برای تولید متن است [۹]. به بیان دیگر، همانطور که در جدول ۲–۱ نشان داده شده است، مدل با آگاهی از سیگنال‌های کنترلی (نظیر سرعت و جهت حرکت) و یادگیری همزمان آنها، توضیحات دقیق‌تر و مرتبط‌تری تولید می‌کند؛ حذف هر کدام از این سیگنال‌های کنترلی منجر به افت قابل ملاحظه‌ای در کیفیت شرح و استدلال خروجی می‌شود [۹]. بر همین اساس، در پژوهش حاضر ما نیز بر حفظ این مؤلفه‌های اطلاعاتی تأکید کرده‌ایم تا از مزایای آن در بهبود کیفیت خروجی بهره‌مند شویم.

جدول ۲–۱– تاثیر استفاده از سیگنال های کنترلی در کیفیت زیرنویس های تولید شده به وسیله مقایسه نتیجه معیار های ارزیابی زیرنویس روی خروجی های دریافتی [۹].

استفاده از سیگنال های کنترلی	تصمیم گیری				دلیل آوری			
	BLEU-4	CIDEr-D	METEOR	ROUGE-L	BLEU-4	CIDEr-D	METEOR	ROUGE-L
خیر	۳۳.۲	۲۳۸.۹	۲۹.۷	۶۲.۰	۸.۶	۸۹.۷	۱۴.۱	۳۱.۴
بله	۳۴.۶	۲۴۷.۵	۳۰.۶	۶۲.۸	۱۱.۰۴	۱۰۲.۶	۱۵.۲	۳۲.۰

عملکرد ADAPT بر روی مجموعه‌داده‌ی BDD-X بسیار چشمگیر گزارش شده است. این مدل توانست بهترین نتایج را در هر دو معیارهای خودکار ارزیابی متن (مانند BLEU-٤/CIDEr-D برای کیفیت شرح‌های تولیدی) و ارزیابی انسانی کسب کند [۹]. به بیان دیگر، هم از نظر آماری توضیحات دقیق‌تری نسبت به مدل‌های پیشین تولید کرد و هم برای انسان‌ها قابل فهم‌تر و قابل قبول‌تر بود [۹]. افزون بر این، پژوهشگران یک سامانه‌ی زمان–واقعی مبتنی بر ADAPT ساخته‌اند که ویدئوی زنده‌ی دوربین خودرو را گرفته و به‌صورت بلادرنگ توضیحات و اقدامات را اعلام می‌کند [۹]. وجود چنین

سیستم عملیاتی نشان می‌دهد که معماری ADAPT از لحاظ پیاده‌سازی نیز کارا بوده و می‌تواند در خودروهای واقعی به کار گرفته شود [۹].

در مجموع، ADAPT نمایانگر نسل جدیدی از مدل‌های تصمیم‌گیر خودران است که شفافیت در تصمیم‌گیری را به‌طور جدی مدنظر قرار داده‌اند. این مدل نشان داد که با بهره‌گیری از ترنسفورمرها و یادگیری همزمان کنترل و زبان، می‌توان به سامانه‌ای رسید که نه تنها تصمیمات درستی در رانندگی می‌گیرد، بلکه قادر است هر تصمیم را به شکلی قابل فهم برای انسان توضیح دهد [۹]. چنین قابلیت‌هایی گام مهمی در جهت افزایش اعتماد کاربران به خودروهای خودران و تعامل بهتر انسان با سیستم‌های هوشمند به شمار می‌رود.

نقش توجه بصری در تفسیر رفتار مدل‌های خودران

مکانیزم توجه به یکی از اجزای کلیدی در مدل‌های تصمیم‌گیر خودران تبدیل شده که نقش مهمی در تفسیرپذیری رفتار مدل ایفا می‌کند. ایده‌ی اصلی توجه بصری آن است که مدل به جای پردازش یکنواخت کل تصویر، بخش‌های مهم‌تر صحنه را برجسته‌تر مد نظر قرار دهد؛ درست شبیه انسانی که هنگام رانندگی چشمان خود را بر عوامل مهم (مانند خودروهای جلویی، علائم راهنما یا عابران) متمرکز می‌کند. خروجی مکانیزم توجه را می‌توان به صورت یک نقشه‌ی حرارتی روی تصویر نشان داد که مشخص می‌کند مدل به کدام نواحی تصویر بیشتر توجه کرده است.

استفاده از نقشه‌های توجه، از دو جنبه حائز اهمیت است: نخست آن‌که می‌تواند دقت تصمیم‌گیری مدل را بهبود بخشد، زیرا تمرکز آن را روی اطلاعات مرتبط نگه می‌دارد؛ و دوم این‌که به عنوان یک ابزار شفافیت عمل می‌کند که به توسعه‌دهندگان و کاربران نشان می‌دهد مدل چگونه به وضعیت ورودی می‌نگرد [۵۷]. به بیان دیگر، attention یک پل ارتباطی بین دنیای پیچیده‌ی ویژگی‌های درونی شبکه و درک انسان فراهم می‌کند. برای نمونه، وقتی یک مدل توجه خود را بر ناحیه‌ی چراغ راهنمایی قرمز متمرکز کرده و سپس فرمان توقف می‌دهد، مشاهده‌ی این نقشه‌ی توجه

برای ما توضیح می‌دهد که چراغ قرمز علت توقف بوده است [۵۷]. برعکس، اگر مدل به ناحیه‌ی نامربوطی توجه کرده باشد (مثلاً به تابلوهای تبلیغاتی کنار جاده) ولی ترمز کند، این عدم تطابق می‌تواند نشانه‌ی عملکرد نادرست یا نیاز به اصلاح باشد.

مدل‌های ترنسفورمر برای درک روابط میان عناصر ورودی از سازوکار «توجه» استفاده می‌کنند. در این مکانیزم، هر بردار ورودی ابتدا به سه فضای برداری نگاشت می‌شود: پرسش[۵۸]، کلید[۵۹] و مقدار[۶۰].

توجه به‌صورت تابعی از شباهت میان پرسش و کلیدها تعریف می‌شود و معمولاً با فرمول زیر محاسبه می‌گردد:

$$Softmax\left(\frac{T_{QK}}{\sqrt{k^d}}\right)V = Attention(Q,K,V) \qquad (۲-۱)$$

در این معادله، $Q \in \mathbb{R}^{k^{n\times d}}$ و $V \in \mathbb{R}^{k^{n\times d}}$ به‌ترتیب ماتریس‌های پرسش، کلید و مقدار هستند و k^d بُعد بردارهای کلید است. ضریب نرمال‌سازی $\frac{۱}{\sqrt{k^d}}$ از اشباع تابع softmax جلوگیری می‌کند.

خروجی نهایی حاصل ترکیب خطی مقادیر V با وزن‌هایی است که از شباهت نرمال‌شده بین Q و K حاصل شده‌اند. این وزن‌ها را معمولاً به‌صورت ماتریس احتمال $Softmax\left(\frac{T_{QK}}{\sqrt{k^d}}\right) = A$ نمایش می‌دهند که هر درایه A_{ij} نشان می‌دهد عنصر i از ورودی تا چه حد به عنصر j توجه دارد.

در مدل ADAPT، این سازوکار در دو سطح استفاده می‌شود:

- در سطح Self-Attention برای مدل‌سازی وابستگی‌های درونی دنباله زبانی یا ویژگی‌های تصویری؛

---

[۵۸] Query

[۵۹] Key

[۶۰] Value

- و در سطح Cross-Attention برای هم‌ترازسازی بین ویژگی‌های بصری استخراج‌شده از تصویر و توصیف زبانی تولیدی.

این وزن‌های توجه در عمل می‌توانند به‌عنوان شاخصی از شفافیت مدل به‌کار روند، زیرا با تصویرسازی ماتریس A، می‌توان مشخص کرد کدام بخش از ورودی (مثلاً ناحیه‌ای از تصویر یا کلمه‌ای خاص) بیشترین تأثیر را بر تولید خروجی دارد. از این رو، تحلیل ماتریس توجه یکی از روش‌های رایج برای تبیین تصمیمات مدل‌های ترنسفورمر در حوزه‌های زبانی و چندرسانه‌ای به‌شمار می‌رود.

جمع‌بندی و چالش‌های پیش‌رو

مدل‌های تصمیم‌گیر خودران مسیری طولانی را از شبکه‌های ساده‌ی مبتنی بر تصویر تا سامانه‌های پیچیده‌ی چندوظیفه‌ای طی کرده‌اند. امروزه ما مدل‌هایی در اختیار داریم که نه تنها می‌توانند در شرایط عادی رانندگی کنند، بلکه قادرند دلایل اقدامات خود را نیز توضیح دهند. این پیشرفت‌ها مرهون به‌کارگیری خلاقانه‌ی معماری‌های نوین (ترنسفورمرها)، مکانیزم توجه و داده‌های برچسب‌دار غنی (مانند BDD-X) بوده است.

با این حال، تصمیم‌گیری در محیط‌های پیچیده‌ی رانندگی همچنان یک چالش بزرگ محسوب می‌شود. تنوع بی‌پایان سناریوهای جاده‌ای از شرایط جوی نامساعد گرفته تا رفتارهای غیرمنتظره‌ی رانندگان دیگر ایجاب می‌کند مدل‌های ما بتوانند عمومیت‌پذیری[۶۱] بالایی داشته باشند و از پس موقعیت‌های جدید برآیند. این امر ممکن است نیازمند ترکیب یادگیری داده‌محور با منطق‌های ایمنی یا اطلاعات پیشینی باشد. افزون بر آن، تبیین‌پذیری نیز باید همگام با پیچیدگی سیستم‌ها تقویت شود؛ توضیحات مدل باید واقعاً بازتاب‌دهنده‌ی علل تصمیم‌ها باشند و از ارائه‌ی توجیه‌های صرفاً ظاهری پرهیز شود. [۵۷] پژوهش‌های بسیار جدید نشان می‌دهند که همراستا کردن بیشتر خروجی زبانی با

---

[۶۱] generalization

حالت‌های درونی سیستم (حتی در سطوح ماژول‌های میانی مانند نقشه‌های ادراک و برنامه‌ریزی) می‌تواند گام بعدی در افزایش دقت و صحت تبیین‌ها باشد [۵۸]. به عنوان مثال، مدل مفهومی Hint-AD (۲۰۲٤) تلاش دارد توضیحات زبانی را مستقیماً به خروجی‌های ماژول‌های درونی (شناسایی اشیاء، پیش‌بینی حرکت، برنامه‌ریزی مسیر و غیره) گره بزند تا هر جمله‌ی تولیدی کاملاً منطبق بر روند تصمیم‌گیری ماشین باشد. [۵۸] نتایج اولیه‌ی این رویکرد نشان داده که می‌توان به توضیحات منسجم‌تر و دقیق‌تری نسبت به روش‌های صرفاً دیداری–زبانی دست یافت [۵۸].

در مجموع، مدل‌های تصمیم‌گیر خودران به سوی هوشمندتر و شفاف‌تر شدن پیش می‌روند. هر نسل جدید از این مدل‌ها، علاوه بر بهبود در عملکرد رانندگی، تلاش می‌کند قابلیت اعتماد و تعامل‌پذیری بیشتری را برای انسان‌ها فراهم کند. با ادامه‌ی این روند، انتظار می‌رود خودروهای خودران آینده نه تنها ایمن و کارآمد باشند، بلکه همچون یک راننده‌ی ماهر انسان، قادر به توضیح منطقی اعمال خود بوده و اعتماد سرنشینان را جلب نمایند.

۲–۴– پیشینه تحقیق

۲–۴–۱– پژوهش های داخلی مرتبط

پژوهش‌های داخلی در حوزه سیستم‌های خودران و بینایی ماشین در رانندگی

پژوهش‌های متعددی در ایران به بررسی جنبه‌های گوناگون سامانه‌های خودران و بینایی ماشین در رانندگی پرداخته‌اند. در سال‌های پایانی دهه‌ی ۱۳۸۰ و اوایل ۱۳۹۰، تمرکز پژوهش‌ها عمدتاً بر روی سامانه‌های کمک‌راننده و تشخیص الگوهای بصری در جاده بود. برای نمونه، در سال ۱۳۸۸ یک مقاله‌ی کنفرانسی به روش‌های کلاسیک شناسایی تابلوهای راهنمایی با فیلتر گابور و شبکه عصبی احتمالاتی پرداخت [۵۹]. همچنین لهونی‌فرد و همکاران (۱۳۹۰) مهم‌ترین روش‌های تشخیص علائم راهنمایی و

رانندگی را مرور و مقایسه کردند [۶۰]. که نشان‌دهنده‌ی توجه زودهنگام محققان ایرانی به سامانه‌های بینایی ماشین خودرو بود.

با پیشرفت فناوری، پژوهش‌های داخلی به سمت ابداع روش‌های هوشمندتر برای بینایی ماشین خودروهای خودران حرکت کردند. به عنوان نمونه، فنی و خدایاری (۱۳۹۷) یک الگوریتم هوشمند مبتنی بر منطق فازی برای تشخیص علائم رانندگی ارائه دادند که شامل مراحل پیش‌پردازش تصویر، تشخیص ناحیه‌ی علامت و شناسایی نوع علامت بود [۶۱]. این روش توانست در شرایط واقعی دقتی در حدود ٪۹۲.۶۸ در شناسایی علائم به دست آورد که کارایی بالای آن را نسبت به روش‌های پیشین نشان می‌داد [۶۱]. پژوهشگران تأکید کرده‌اند که این الگوریتم فازی می‌تواند در طراحی سیستم‌های کمکی راننده و خودروی هوشمند به کار رود [۶۱].

در سال‌های اخیر، با نفوذ یادگیری عمیق در بینایی ماشین، تحقیقات پیشرفته‌تری در ایران انجام شده است. برای مثال، گودرزی و همکاران (۱۴۰۲) یک شبکه‌ی عصبی سبک و سریعی مبتنی بر معماری MobileNet برای تشخیص خطوط جاده ارائه کرده‌اند [۶۲]. روش آن‌ها به‌جای تفکیک پیکسل به پیکسل (که محاسبات سنگینی دارد)، از رویکرد ردیف‌محور برای شناسایی خطوط در هر ردیف تصویر بهره می‌گیرد و با استفاده از سه تابع هزینه‌ی متفاوت، هم جزئیات محلی خطوط و هم اطلاعات کلی صحنه را در نظر می‌گیرد [۶۲]. آزمایش این روش روی مجموعه‌داده‌ی استاندارد TuSimple عملکرد مطلوب آن را از نظر دقت و به‌ویژه سرعت نشان داده است [۶۲]. چنین پژوهش‌هایی بیانگر تلاش محققان داخلی برای بهبود ادراک بصری خودروهای خودران با روش‌های به‌روز و بهینه است. از سوی دیگر، هویدافرد و همکاران (۱۴۰۲) به سراغ مدل‌های تصمیم‌گیری مبتنی بر ویدیو برای رانندگی خودکار رفته‌اند. آن‌ها با ترکیب شبکه‌های عصبی Dense و Convolutional در یک شبیه‌ساز رانندگی (پلتفرمUdacity) ، نشان دادند که می‌توان فرآیند آموزش و طراحی خودروهای خودران را سرعت بخشید [۶۳]. نتیجه‌ی پیاده‌سازی ایشان تطابق بالایی میان عملکرد

خودروی کنترل‌شده توسط انسان و خودروی خودران آموزش‌دیده نشان داد [۶۳]، که حاکی از اثربخشی یادگیری عمیق در تصمیم‌گیری‌های رانندگی است.

به طور کلی، وضعیت پژوهش‌های داخلی در حوزه‌ی خودران نشان می‌دهد محققان ایرانی اکثر چالش‌های کلیدی سیستم‌های خودران را در مقیاس آزمایشگاهی بررسی کرده‌اند: از ادراک محیط (تشخیص خطوط، علائم و موانع) گرفته تا تصمیم‌گیری خودکار. خلأ موجود بیشتر در یکپارچه‌سازی این دستاوردها و پیاده‌سازی عملی در مقیاس وسیع دیده می‌شود. تاکنون اغلب کارها در حد شبیه‌سازی یا نمونه‌های آزمایشگاهی بوده و گزارش مستندی از یک خودروی خودران کاملاً عملیاتی در جاده‌های کشور در منابع علمی فارسی به چشم نمی‌خورد. این امر نشان‌دهنده‌ی نیاز به پروژه‌های جامع‌تر (ترکیب بینایی ماشین، کنترل و هوش مصنوعی) و نیز داده‌های بومی برای آزمودن الگوریتم‌ها در شرایط واقعی ایران است. همچنین موضوعاتی مانند ایمنی، استانداردها و مسائل حقوقی خودروهای خودران کمتر در ادبیات پژوهشی فنی داخلی مطرح شده‌اند که می‌تواند حوزه‌ی پژوهشی آینده باشد.

پژوهش‌های داخلی در حوزه‌ی هوش مصنوعی توضیح‌پذیر

در سال‌های اخیر، مفهوم هوش مصنوعی توضیح‌پذیر نیز مورد توجه پژوهشگران ایرانی قرار گرفته و آثار علمی اولیه‌ای به زبان فارسی در این زمینه منتشر شده است. این پژوهش‌ها عموماً بر معرفی مفاهیم، روش‌های متداول و کاربردهای XAI متمرکز بوده‌اند. به عنوان نمونه، دارایی (۱۴۰۳) در یک مقاله‌ی مروری، اهمیت XAI را به عنوان پلی میان پیچیدگی مدل‌های یادگیری عمیق و نیاز به شفافیت در تصمیم‌ها تشریح می‌کند [۶۴]. وی اشاره می‌کند که XAI با روش‌هایی نظیر LIME، SHAP و DeepLIFT تلاش دارد عملکرد مدل‌های جعبه‌سیاه را قابل‌فهم کرده و اعتماد کاربران به خروجی سیستم‌های هوشمند را افزایش دهد [۶۴]. این کار علاوه بر مرور روش‌ها، به کاربردهای گسترده‌ی XAIدر صنایع مختلف از جمله پزشکی، امور مالی و حتی خودروهای خودران می‌پردازد [۶۴]. چنین

آثاری نشان می‌دهد که پژوهشگران داخلی به خوبی از ترندهای جهانی XAI آگاه بوده و در جهت ترویج آن کوشیده‌اند.

فراتر از مقالات مروری، چندین پژوهش اصیل نیز به بررسی فنی و پیاده‌سازی روش‌های XAI پرداخته‌اند. برای مثال، فریدون‌فر (۱۴۰۴) یک روش توضیح‌پذیری مستقل از مدل موسوم به CASTLE را برای کاربرد در معماری کلان‌داده گسترش داده است [۶۵]. انگیزه‌ی این کار از آنجا ناشی می‌شود که اکثر تکنیک‌های متداول XAI روی مجموعه‌داده‌های کوچک آزمایش شده‌اند و کارایی آن‌ها در مقیاس کلان‌داده (با حجم و سرعت بالا) محل سؤال است [۶۵]. در این پژوهش، با بهره‌گیری از چارچوب آپاچی اسپارک، روش CASTLE برای توضیح تصمیم‌های مدل‌های یادگیری ماشین روی داده‌های عظیم مقیاس‌پذیر شده است [۶۵]. ارزیابی روی پنج مجموعه‌داده‌ی بزرگ نشان داده که رویکرد پیشنهادی ضمن حفظ کیفیت تبیین مدل (مشابه CASTLE اصلی) ، قادر است حجم بسیار بالای داده را به شکل کارا مدیریت کند [۶۵]. دستاورد مهم این پژوهش آن است که زمان اجرای روش توضیح‌پذیری به صورت زیرخطی نسبت به اندازه‌ی داده‌ها افزایش می‌یابد، که برای کاربردهای عملی روی کلان‌داده‌ها حیاتی است [۶۵]. چنین تحقیقاتی حکایت از تلاش پژوهشگران ایرانی برای گسترش مرزهای دانش XAI و تطبیق آن با نیازهای خاص (مانند پردازش داده‌های حجیم) دارد.

از سوی دیگر، کاربردپذیری XAI در حوزه‌های مختلف نیز مورد توجه پژوهشگران داخلی بوده است. نمونه‌ی بارز، تحقیقی در مدیریت بازرگانی است که به بررسی تأثیر توضیح‌پذیری هوش مصنوعی بر پذیرش فناوری توسط مدیران پرداخته است. حسینی‌طاهری و اخوان‌تفتی (۱۴۰۳) در یک مطالعه تجربی نشان دادند که افزودن قابلیت توضیح‌دهی به یک سیستم هوش مصنوعی مدیریت ارتباط با مشتری[۶۲] می‌تواند اعتماد و تمایل مدیران به اشتراک‌گذاری دانش با سیستم را افزایش دهد [۶۶]. در این آزمایش که با مشارکت ۱۴۱ مدیر بازاریابی انجام شد، گروهی که سیستم مجهز به توضیح‌پذیری

---

[۶۲] CRM

را مشاهده کردند، احساس منفی کمتری نسبت به تصمیم‌های سیستم داشتند و همکاری بهتری نشان دادند [۶۶]. هرچند در شرایطی که سطح تهدید فناوری بالا بود، تاثیر XAI چندان معنادار نبود، این پژوهش به روشنی بر نقش مثبت XAI در بهبود پذیرش فناوری‌های هوشمند تأکید می‌کند [۶۶]. همچنین پایان‌نامه‌ها و گزارش‌های پژوهشی دیگری نیز در دانشگاه‌های کشور انجام شده که کاربرد XAI را در حوزه‌هایی مانند تشخیص پزشکی (توضیح‌پذیری تشخیص بیماری‌ها)، سامانه‌های توصیه‌گر و حتی تحلیل داده‌های علوم انسانی کاوش کرده‌اند. این نشان‌دهنده‌ی طیف گسترده‌ای از علاقه‌مندی به XAI در میان پژوهشگران ایرانی است.

با وجود این تلاش‌ها، بررسی وضعیت کلی تحقیقات داخلی در حوزه XAI بیانگر آن است که این عرصه هنوز در مراحل آغازین خود قرار دارد. بسیاری از کارهای منتشرشده بیشتر جنبه‌ی معرفی و امکان‌سنجی داشته و توسعه‌ی روش‌های ابتکاری بومی یا چارچوب‌های استاندارد ارزیابی کمتر دیده می‌شود. خلأهای مهم در این زمینه عبارت‌اند از: نبود مجموعه‌داده‌ها و مسائل بومی برای سنجش الگوریتم‌های توضیح‌پذیری، کمبود ابزارهای بومی‌سازی‌شده برای پیاده‌سازی ساده‌ی XAI، و نیز نیاز به پژوهش در ابعاد اخلاقی و حقوقی توضیح‌پذیری (مثلاً در سیستم‌های خودران یا تصمیم‌یارهای پزشکی) در چارچوب مقررات و فرهنگ کشور. پرداختن به این خلأها می‌تواند موجب تعمیق تحقیقات داخلی و نزدیک‌تر شدن آن‌ها به نیازهای واقعی صنعت و جامعه گردد.

جمع‌بندی

مروری بر پژوهش‌های فارسی در دو حوزه‌ی خودروهای خودران و هوش مصنوعی توضیح‌پذیر نشان می‌دهد که هر دو عرصه در کشور ما مورد توجه محققان قرار گرفته و دستاوردهای ارزشمندی حاصل شده است. در زمینه‌ی خودران، کارهای پراکنده‌ای بر روی اجزای مختلف سامانه (بینایی ماشین، تصمیم‌گیری خودکار و ...) صورت پذیرفته و پایه‌های فنی لازم شناخته شده است. در حوزه‌ی XAI نیز ابتدا مفاهیم و روش‌ها تبیین شده و سپس کاربردها و بهینه‌سازی‌هایی متناسب با نیازهای بومی آغاز

گردیده است. با این حال، پیوند این دو حوزه هنوز جای کار دارد. برای مثال، پژوهش خاصی که توضیح‌پذیری تصمیم‌های یک خودروی خودران را به زبان فارسی بررسی کند، به ندرت گزارش شده است. این امر می‌تواند فرصت مناسبی برای تحقیقات آینده باشد تا با ترکیب دانش خودران و XAI، سامانه‌هایی هوشمند، ایمن و قابل‌اعتماد طراحی شوند که همگام با دانش روز دنیا و نیازهای کشور پیش بروند.

۲–۴–۲– پژوهش‌های خارجی مرتبط

دستاوردهای بینایی رایانه‌ای در سامانه‌های خودران

حوزه‌ی بینایی رایانه‌ای در خودروهای خودران طی دو دهه‌ی اخیر پیشرفت‌های قابل توجهی داشته است. در سال‌های ابتدایی، الگوریتم‌های کلاسیک پردازش تصویر برای تشخیص خطوط جاده، خودروها و عابرین به کار گرفته می‌شد که اگرچه ایده‌های اولیه‌ای برای سامانه‌های کمک‌راننده فراهم کردند اما دقت و تاب‌آوری محدودی در شرایط پیچیده‌ی محیطی داشتند. با ظهور یادگیری عمیق، مدل‌های مبتنی بر شبکه‌های عصبی پیچشی انقلابی در بینایی خودرو ایجاد کردند. کایماک و اوکار (۲۰۱۹) با به‌کارگیری شبکه‌های تمام‌کانولوشن[۶۳] توانستند مسئله‌ی بخش‌بندی معنایی جاده را در تصاویر خودرویی حل کنند؛ [۶۷] هرچند مدل آن‌ها به‌صورت جعبه‌سیاه عمل می‌کرد و جزئیاتی از چرایی تصمیم‌های خود ارائه نمی‌داد [۶۸].

یکی از نخستین و تأثیرگذارترین تلاش‌ها برای طراحی سیستم رانندگی خودکار مبتنی بر یادگیری عمیق، مدل PilotNet بود که توسط شرکت NVIDIA در سال ۲۰۱۶ معرفی شد [۲۱]. این مدل با استفاده از یک شبکه عصبی کانولوشنی ساده، از تصاویر دوربین جلوی خودرو به‌صورت مستقیم زاویه فرمان را پیش‌بینی می‌کرد. نوآوری کلیدی PilotNet در آن بود که از روش‌های کلاسیک تشخیص

---

[۶۳] FCN

ویژگی صرف‌نظر کرده و رویکرد یادگیری انتها-به-انتها[۶۴] را برای هدایت خودرو به کار گرفت. با وجود ساختار نسبتاً ساده، این مدل در مسیرهای از پیش ضبط‌شده عملکرد قابل قبولی نشان داد و زمینه‌ساز تحقیقات گسترده‌تری در مسیر ادغام بینایی ماشین با تصمیم‌گیری شد [۲۱]. تحلیل‌های فنی این مدل در بخش ۲-۳-۴ بررسی شده است.

در ادامه، کولترا و همکاران (۲۰۲۰) یک عامل خودران مبتنی بر یادگیری تقلیدی طراحی کردند که از مکانیزم توجه بهره می‌گرفت تا بخش‌های مهم تصویر ورودی برای تصمیم‌گیری را برجسته کند. [۶۹] این رویکرد با نشان دادن نواحی با اهمیت در تصویر، گامی به سوی تفسیرپذیری برداشت، اما هنوز چگونگی تصمیم‌سازی درون مدل به‌طور کامل شفاف نبود [۶۸]. پس از آن، بهبودهای قابل ملاحظه‌ای در سیستم‌های بینایی خودرو با استفاده از شبکه‌های عمیق حاصل شد؛ به‌طور مثال معرفی شبکه‌ی YOLO منجر به افزایش سرعت و دقت تشخیص اشیاء نسبت به روش‌های پیشین مانند شبکه‌های پیشنهاد ناحیه R-CNN و SSD شد [۷۰]. با این حال، اتکا به شبکه‌های عمیق پیچیده نگرانی‌هایی درباره‌ی قابل اعتماد بودن این سامانه‌ها ایجاد کرد، چرا که سازوکار تصمیم‌گیری آن‌ها برای توسعه‌دهندگان و کاربران قابل درک نبود. این مسئله در گزارش‌های اواسط دهه ۲۰۱۰ نیز مورد توجه قرار گرفت، به طوری که مجله‌ی Nature در سال ۲۰۱۶ با طرح این سوال که «آیا می‌توان جعبه‌سیاه هوش مصنوعی را گشود؟» بر ضرورت تبیین‌پذیری مدل‌های یادگیری عمیق تأکید کرد [۶۸]. چنین دغدغه‌هایی زمینه‌ساز موج جدیدی از پژوهش‌ها در حوزه‌ی هوش مصنوعی قابل توضیح شد.

در مطالعات اخیر، یوان و همکاران رویکرد RAG-Driver[۷۱] را معرفی کرده‌اند که مبتنی بر یادگیری در بستر تقویت‌شده با بازیابی[۶۵] است. در این روش، مدل چندرسانه‌ای زبان بزرگ با بازیابی نمونه‌های نمایشی تخصصی زمینه‌سازی شده و توضیحات زبانی برای عملیات رانندگی تولید می‌کند؛ به‌طوری که ضمن دستیابی به عملکرد بسیار بالا، توانایی تعمیم صفر-شات چشمگیری را نیز نشان

---

[۶۴] End-to-End Learning

[۶۵] Retrieval-Augmented In-Context Learning

می‌دهد [۷۱]. در مقابل، روش پیشنهادی ما بر آموزش بازتنظیمی مدل زبان بزرگ متمرکز است که با بهره‌گیری از نقشه‌های توجه استخراج‌شده از مدل تصمیم مستقل ADAPT، توضیحات تولیدشده را بر شواهد بصری استوار می‌کند. بنابراین، تفاوت اصلی در این است که RAG-Driver با تکیه بر بازیابی مثال‌های خارجی به تعمیم سریع در محیط‌های جدید می‌پردازد، در حالی که روش ما با جفت‌سازی مستقیم داده‌های دیداری و زبانی بر ارائه توضیحاتی انسان‌فهم و مستند تأکید دارد.

کاربرد XAI در سیستم‌های خودران

- ظهور و توسعه‌ی روش‌های هوش مصنوعی قابل توضیح

با افزایش پیچیدگی مدل‌های یادگیری ماشین و به‌کارگیری آن‌ها در سامانه‌های حساس مانند خودروهای خودران، نیاز به روش‌هایی برای توضیح پذیر کردن تصمیم‌های این مدل‌ها شدت گرفت. هدف XAI این است که الگوریتم‌های هوشمند، خروجی یا تصمیم خود را همراه با دلیلی قابل فهم برای انسان ارائه کنند [۷۰]. از اواسط دهه‌ی ۲۰۱۰ به بعد، پژوهشگران روش‌های متعددی را برای تحقق این هدف ابداع کردند. به عنوان نمونه، روش مدل-آگنوستیک LIME در سال ۲۰۱۶ معرفی شد که با ایجاد مدل‌های محلی خطی پیرامون هر پیش‌بینی، دلیل آن پیش‌بینی را به‌صورت قابل درک بیان می‌کند؛ [۴] همچنین روش SHAP در سال ۲۰۱۷ با بهره‌گیری از نظریه‌ی بازی شاپلی، سهم هر ویژگی را در تصمیم نهایی مدل محاسبه کرده و توضیحی سراسری‌تر ارائه می‌دهد [۲]. این قبیل روش‌ها که مستقل از معماری مدل عمل می‌کنند، برای تبیین خروجی هر نوع مدل یادگیری ماشین به‌کار گرفته شده‌اند. در حوزه‌ی بینایی کامپیوتر نیز تکنیک‌های پس‌پردازشی نظیر نقشه‌برداری گرادیان-کلاس[۶۶] و نقشه‌های سالینسی رواج یافتند که با محاسبه‌ی گرادیان‌های مدل نسبت به پیکسل‌های ورودی، بخش‌های تاثیرگذار تصویر در تصمیم مدل را به‌صورت نواحی هایلایت‌شده نمایش می‌دهند [۷۰]. بدین ترتیب، برای اولین بار امکان مشاهده‌ی اینکه یک شبکه‌ی عصبی در یک تصویر به چه چیزی

---

[۶۶] Grad-CAM

نگاه می‌کند تا تصمیم خود را بگیرد فراهم شد. مجموع این تلاش‌ها پایه‌های لازم برای ورود XAI به حوزه‌ی خودروهای هوشمند را فراهم کرد تا ضمن حفظ دقت مدل‌های یادگیری عمیق، شفافیت و اعتمادپذیری آن‌ها نیز افزایش یابد [۷۰].

- روش‌های رایج XAI در سیستم‌های خودران

در کاربردهای بینایی کامپیوتر برای رانندگی خودکار، روش‌های مبتنی بر تحلیل بصری از متداول‌ترین شیوه‌های XAI هستند. برای مثال، Bojarski و همکاران (۲۰۱۶) با استفاده از روش «VisualBackProp» نشان دادند کدام پیکسل‌های ورودی بیشترین تأثیر را در پیش‌بینی فرمان حرکتی شبکه عصبی دارند [۳۱]. به همین ترتیب، Kim و Canny (۲۰۱۷) با بهره از نقشه‌های گرمایی[۶۷] و مدل‌های توجه علّی، نواحی تصویر را که بر فرمان نهایی خودرو اثر دارند نشان دادند [۵۶]. دیگر تکنیک‌های بصری مانند Grad-CAM یا Backpropagation هدایت‌شده نیز برای برجسته‌سازی مناطق مؤثر در تصویر به کار رفته‌اند. روش‌های مبتنی بر مدل جایگزین مانند LIME و نیز کاربرد دارند؛ مثلاً یک مطالعه اخیر برای ارائه توضیحات در زمان واقعی، ترکیبی از LIME و SHAP را به کار برده و به تعادل بین دقت تفسیر و هزینه محاسباتی توجه کرده است [۷۰]. علاوه بر این، تکنیک‌های متنی و زبان طبیعی نیز در حال توسعه‌اند؛ به عنوان نمونه Kim و همکاران (۲۰۱۸) با استفاده از شبکه‌های بازگشتی مبتنی بر ویدیو، توضیحات متنی مبتنی بر ویدیو تولید کردند و یک مجموعه‌داده توضیحی (BDD-X) برای شرح دلایل اقدامات خودران معرفی نمودند [۵۷]. در یک رویکرد جدید، استفاده از مدل‌های ترکیبی زبان–بینایی (مانند پاسخ به پرسش‌های تصویری) نیز مطرح شده تا توضیحی انسانی‌پسند از چرایی تصمیمات رانندگی ارائه شود [۳۱].

- چالش‌های توضیح تصمیمات خودروهای خودران

---

[۶۷] Heatmap

تشریح تصمیمات سیستم‌های خودران به دلیل پیچیدگی محیط و ماهیت چندوجهی داده‌ها با چالش‌هایی همراه است. ورودی‌های این سامانه‌ها معمولاً شامل تصاویر دوربین، داده‌های رادار/لیدار و GPS است که هر یک در ابعاد بالا و بدون معنای صریح هستند؛ این موضوع درک عامل‌های مؤثر بر تصمیم را دشوار می‌کند [۷۲]. علاوه بر این، نیاز به پاسخدهی بلادرنگ، سختی تازه‌ای به کارگیری XAIمی‌بخشد. هر توضیح ارائه‌شده باید ظرف تأخیر قابل‌قبولی تولید شود تا بتواند در تصمیم‌گیری زمان واقعی خودرو مؤثر باشد. در یک مثال کاربردی، روش ترکیبی LIME–SHAP اشاره‌شده برای ایجاد توضیحات به صورت بلادرنگ طراحی شده است [۷۰]، اما گزارش شده است هزینه محاسباتی SHAP هنوز مانعی جدی برای اجرای آن در سیستم‌های خودران است [۷۰]. به طور کلی، پردازش سریع داده‌ها و همگام‌سازی تفسیرها با فریم‌های چندرسانه‌ای محیط، از چالش‌های مهم XAI در رانندگی خودکار هستند.

- مطالعات کاربردی XAI در رانندگی خودکار

در سال‌های اخیر، پژوهش‌های متعددی به طور مشخص بر تلفیق روش‌های XAI با سامانه‌های بینایی خودرو متمرکز شده‌اند. یکی از چالش‌های اساسی در خودروهای خودران، کسب اعتماد کاربران و سرنشینان به تصمیم‌های هوش مصنوعی خودرو است [۷۲]. در این راستا، وانگ و همکاران (۲۰۲۱) سیستمی پیشنهاد دادند که با مشارکت فعال راننده‌ی انسان از طریق ردیابی نگاه (gaze) وی، پیش‌بینی‌های سیستم خودران را بهبود می‌بخشد [۷۳]. هرچند این روش بیشتر جنبه‌ی تعامل انسان–خودرو داشت و خود توضیح مستقلی از تصمیم‌های مدل ارائه نمی‌داد، نتایج آن نشان داد که عدم توجه راننده می‌تواند عملکرد سامانه را مختل کند و در نتیجه بر اهمیت به‌کارگیری اصول توضیح‌پذیری نیز صحه گذاشت [۷۰]. به بیان دیگر، نیاز به XAI حتی در چنین سیستم‌های تعاملی نیز احساس شد.

از سوی دیگر، کوتهاواده و همکاران (۲۰۲۱) رویکرد متفاوتی را معرفی کردند. آن‌ها سامانه‌ای به نام AUTO-DISCERN توسعه دادند که از منطق دانش‌بنیان و استدلال مبتنی بر دانش عمومی[۶۸] برای توضیح تصمیم‌های خودران بهره می‌گیرد [۷۴]. این سیستم با ترکیب برنامه‌نویسی مبتنی بر پاسخ[۶۹] تلاش کرد تصمیم‌های خودرو را به شکلی قابل درک برای انسان تشریح کند. با وجود نوآوری در ترکیب دانش انسانی، محدودیت رویکرد AUTO-DISCERN آن بود که توضیح بصری از فرآیند تصمیم‌گیری شبکه‌ی عمیق زیربنایی ارائه نمی‌کرد [۷۰]؛ در واقع بخش یادگیری عمیق همچنان جعبه‌سیاه باقی می‌ماند و خروجی منطق نمادین صرفاً تصویری از تصمیم نهایی بود.

در سال ۲۰۲۲، تمرکز پژوهش‌ها بیش از پیش به سمت تلفیق مستقیم XAI با بینایی خودرو معطوف شد. سلات و همکاران (۲۰۲۲) با به‌کارگیری ترکیبی از شبکه‌های عصبی کانولوشنی، خودرمزگذارها و ساختارهای هرمی، یک سیستم بخش‌بندی معنایی پیشرفته برای خودروی خودران ساختند که دقت بالایی در سناریوهای مختلف داشت [۷۵]. با این حال، روش آن‌ها نیز صرفاً بر بهبود دقت مدل متمرکز بود و جزئیاتی درباره‌ی چرایی خروجی مدل ارائه نمی‌کرد [۷۰]. در مقابل، در همان سال یک پژوهش دیگر برای اولین بار ادغام صریح XAI در بخش‌بندی معنایی تصاویر خودرویی را محقق ساخت. در این روش، نقشه‌های حرارتی توضیحی (مانند Grad-CAM و سالینسی) بر روی نقشه‌های بخش‌بندی جاده اعمال شد تا نشان دهد مدل در هر تصویر به کدام نواحی توجه کرده است [۷۰]. نتیجه‌ی این کار، تولید توضیحات بصری همزمان با خروجی بخش‌بندی بود که یک گام مهم در افزایش شفافیت سیستم‌های بینایی خودران به‌شمار می‌رود. به بیان دیگر، مدل نه تنها جاده و اجسام را تشخیص می‌داد، بلکه برای هر صحنه دلیلی بصری از تصمیم خود ارائه می‌نمود. این دستاورد نشان داد که می‌توان بدون افت کارایی محسوس، لایه‌ای از توضیح‌پذیری را به سامانه‌های ادراک خودران افزود و بدین ترتیب هم اعتماد کاربران و هم امکان عیب‌یابی توسعه‌دهندگان را بهبود داد.

---

[۶۸] Common Sense Reasoning

[۶۹] Answer Set Programming

سال ۲۰۲۳ را می‌توان نقطه‌ی عطفی در پژوهش‌های XAI خودران دانست، چرا که نتایج چندین مطالعه‌ی مهم در این سال منتشر شد. دونگ و همکاران (۲۰۲۳) بر جنبه‌ی اعتماد و پذیرش اجتماعی خودروهای خودران متمرکز شدند و نشان دادند فراهم کردن توضیحات قابل فهم برای سرنشینان (چه حین رانندگی و چه پس از آن) تأثیر چشمگیری در افزایش اعتماد و همکاری کاربران با خودروی خودران دارد [۷۲]. آن‌ها یک چارچوب مدل یادگیری عمیق تبیین‌پذیر توسعه دادند که قادر است فرآیند تصمیم‌گیری خودرو را به زبان طبیعی توضیح دهد. ایده‌ی محوری در کار آن‌ها این بود که تولید توضیح را به شکل یک مسئله‌ی توصیف صحنه[۷۰] در نظر بگیرند؛ بدین صورت که مدل خودران در کنار انجام تصمیم‌گیری اصلی، یک توصیف متنی کوتاه از وضعیت رانندگی جاری و علت اقدام خود تولید می‌کند [۷۲]. برای این منظور از یک معماری عمیق چند–وجهی مبتنی بر ترنسفورمر استفاده شد که توانست هم‌بستگی بین اطلاعات دیداری (تصویر دوربین خودرو) و زبانی را برقرار کرده و توضیحاتی معنی‌دار و نزدیک به درک انسانی ایجاد کند [۷۲]. این سیستم به نوعی عملکرد یک راننده‌ی باتجربه را شبیه‌سازی می‌کرد که می‌تواند دلایل اقدامات خود را برای دیگران شرح دهد. نتایج نشان داد مدل پیشنهادی آن‌ها علاوه بر حفظ دقت در رانندگی، در مقایسه با روش‌های پایه توانست توضیحاتی تولید کند که درک سرنشینان از تصمیم‌های خودرو را بهبود می‌بخشد [۷۲]. چنین رویکردی نمایانگر حرکت به سوی سامانه‌های خودران قابل توضیح و قابل اعتماد است که هم نیازهای فنی و هم انتظارات کاربران را مدنظر قرار می‌دهند.

علاوه بر پژوهش‌های موردی، در سال‌های اخیر مطالعات مروری نیز به تبیین وضعیت و مسیر آینده‌ی XAI در خودروهای خودران پرداخته‌اند. آتاکیشی‌یف و همکاران (۲۰۲۳) یک مرور جامع بر روش‌های توضیح‌پذیری در رانندگی خودکار منتشر کردند که ضمن جمع‌بندی دستاوردهای روز، یک چارچوب مفهومی برای سیستم‌های خودران تبیین‌پذیر end-to-end پیشنهاد می‌کند [۲۷]. این مطالعه بر عناصر اساسی مورد نیاز جهت رانندگی خودران قابل توضیح تأکید نموده و مسیرهای پژوهشی آینده

---

[۷۰] Image Captioning

برای ارتقای شفافیت، اعتماد و پذیرش اجتماعی خودروهای بدون راننده را ترسیم کرده است [۲۷]. انتشار چنین مرورهای منظمی نشان می‌دهد که بحث XAI در خودروهای خودران اکنون به یک رشته‌ی پژوهشی بالغ بدل شده و توجه گسترده‌ی محققان را به خود جلب کرده است.

مدل ADAPT[۷۱] یکی از نمونه‌های پیشرفته و معاصر ترکیب بینایی رایانه‌ای، زبان طبیعی، و توضیح‌پذیری در سامانه‌های خودران است [۱۸]. این مدل که مبتنی بر ساختار ترنسفورمر چندوجهی طراحی شده، قادر است با دریافت فریم‌های ویدیویی از نمای جلوی خودرو، هم‌زمان دو خروجی تولید کند: یک تصمیم کنترلی (مانند توقف، گردش، حرکت مستقیم) و یک توصیف متنی از وضعیت و علت آن تصمیم. معماری ADAPT با بهره‌گیری از attention map های چندسطحی، نه‌تنها به تصمیم‌گیری دقیق می‌پردازد، بلکه توضیحی نسبتاً شفاف از روند ادراکی خود نیز ارائه می‌دهد [۱۸]. معرفی معماری و قابلیت‌های این مدل در بخش ۴-۳-۲ ارائه شده و پیاده‌سازی عملی آن در فصل سوم بررسی خواهد شد.

در تازه‌ترین تحولات، تمرکز بر بهینه‌سازی روش‌های XAI برای کاربرد بلادرنگ در خودرو افزایش یافته است. برای نمونه، در سال ۲۰۲۴ پژوهشگرانی یک راهکار ترکیبی جدید ارائه کردند که دو روش مطرح LIME و SHAP را برای تبیین تصمیم‌های مدل به صورت ترکیبی به کار می‌گیرد [۲۹]. آن‌ها با بهره‌گیری از دقت عمومی SHAP در سنجش اهمیت ویژگی‌ها و سرعت محاسباتی LIME در ایجاد توضیحات محلی، یک چارچوب ترکیبی سبک مناسب پیاده‌سازی روی خودرو طراحی نمودند [۲۹]. این مدل ترکیبی روی داده‌های واقعی (داده‌ی KITTI) و سه معماری معروف شبکه‌ی عمیق (ResNet-۱۸, ResNet-۵۰, SegNet) آزمایش شد. نتایج چشمگیری به دست آمد به طوری که نرخ وفاداری[۷۲] توضیحات بیش از ۸۵٪ و شاخص تفسیرپذیری بیش از ۸۰٪ گزارش شد [۲۹]. جالب‌تر آن‌که افزودن لایه‌های توضیح‌دهنده تأخیر چندانی ایجاد نکرد؛ زمان پاسخ‌گویی مدل با ResNet-۱۸

---

[۷۱] Action-aware Driving Caption Transformer

[۷۲] fidelity

تنها حدود ۰.۲۸ ثانیه و برای ResNet-۵۰ حدود ۰.۵۷ ثانیه بود که برای عملکرد بلادرنگ روی خودروی واقعی کاملاً مناسب است [۲۹]. این دستاورد که با حفظ موازنه‌ی دقت، سرعت و شفافیت حاصل شده، نشان می‌دهد به‌کارگیری XAI در سامانه‌های ایمنی–حیاتی مانند خودروهای خودران امکان‌پذیر بوده و حتی در مقیاس عملیاتی نیز قابل اجرا است [۲۹]. به طور کلی، روند پژوهش‌های سال‌های اخیر حاکی از آن است که جامعه‌ی علمی در حال نزدیک‌تر کردن دو مسیر مجزای گذشته یعنی بینایی کامپیوتری دقیق و هوش مصنوعی قابل توضیح به یکدیگر است؛ به گونه‌ای که نسل آینده‌ی خودروهای خودران نه تنها از نظر ادراک محیط و تصمیم‌گیری هوشمند بلکه از لحاظ توضیح‌دادن تصمیم‌های خود نیز توانمند و قابل اعتماد خواهند بود.

مدل‌های توضیح‌پذیر در رانندگی خودران (تولید توضیحات متنی)

یکی از مشکلات اساسی در به‌کارگیری شبکه‌های عصبی در خودروهای خودران، قابل‌توضیح نبودن تصمیم‌ها برای انسان است [۹]. یک شبکه‌ی انتها–به–انتها ممکن است فرمانی (مانند ترمز یا تغییر مسیر) صادر کند بدون این که مشخص باشد چرا چنین تصمیمی گرفته شده است. برای افزایش اعتماد کاربران و درک بهتر رفتار مدل، پژوهشگران به سراغ روش‌هایی رفته‌اند که همراه با تصمیم کنترلی، توضیح متنی قابل فهمی نیز ارائه کند [۹].

نخستین تلاش‌ها در این راستا بر استفاده از توجه بصری به عنوان ابزاری برای تبیین تصمیم متمرکز بود. به عنوان مثال، در سال ۲۰۱۷ محققان انویدیا روشی ابداع کردند که با تحلیل شبکه‌ی PilotNet مشخص می‌کرد کدام نواحی تصویر بیشترین تأثیر را بر تصمیم فرمان دارند [۷۳]. نتایج نشان داد که مدل به‌طور خودکار اشیاء مهم جاده (مانند خطوط، لبه‌ی جاده، خودروهای دیگر) را می‌آموزد و به آن‌ها توجه می‌کند [۷۳]. این نوع نقشه‌های توجه یا برجستگی بصری کمک می‌کرد تا حدی بفهمیم شبکه به چه چیزهایی نگاه می‌کند، اما هنوز یک توضیح زبانی صریح ارائه نمی‌داد. به

بیان دیگر، کاربر نهایی (مثلاً سرنشین خودرو) شاید ترجیح دهد به جای دیدن هایلایت روی تصویر، یک جمله‌ی قابل درک بشنود که دلیل اقدام خودرو را توضیح دهد [۵۷].

گام بعدی، ترکیب مدل‌های بینایی و زبانی برای تولید توضیحات متن‌محور بود. در این زمینه، کار مهمی توسط کیم و همکاران (ECCV ۲۰۱۸) انجام شد که یک چارچوب توضیح‌پذیر عمیق برای رانندگی پیشنهاد دادند [۵۷]. در روش آن‌ها، دو مولفه‌ی اصلی به‌صورت همزمان یاد گرفته می‌شد:

- یک شبکه‌ی کنترل رانندگی با مکانیزم توجه فضایی که ورودی آن تصاویر ویدئویی و خروجی آن فرمان‌های خودرو (شتاب و تغییر جهت) بود [۵۷]. این شبکه با استفاده از نقشه توجه مشخص می‌کرد کدام بخش‌های تصویر در تصمیم کنترلی نقش دارند.
- یک مدل ویدئو–به–متن مبتنی بر توجه که توضیح زبان طبیعی مربوط به همان بازه‌ی زمانی را تولید می‌کرد [۵۷]. خروجی این مدل جملاتی بود که توصیف می‌کرد چه کاری انجام می‌شود و به چه علتی، به شکل یک توضیح انسانی (برای مثال: «خودرو سرعت را کاهش می‌دهد چون به یک تقاطع نزدیک می‌شود و چراغ قرمز است.»).

نکته‌ی کلیدی، همترازسازی نقشه‌های توجه میان این دو بخش بود؛ بدین صورت که ناحیه‌های مهم تصویر که در شبکه‌ی رانندگی برجسته شده بودند مستقیماً در تولید جمله‌ی توضیحی نیز دخیل بودند [۵۷]. این تضمین می‌کرد توضیح تولیدشده واقعاً بر پایه‌ی همان عواملی باشد که بر تصمیم مدل اثر گذاشته‌اند، نه صرفاً یک جمله‌ی از پیش حفظ‌شده. نویسندگان این روش، توضیحات متنیشان را از نوع "توضیح درون‌نگرانه[۷۳]" توصیف می‌کنند، زیرا بر وضعیت درونی خود شبکه و علت واقعی تصمیم تکیه دارد [۵۷]. این در مقابل "توجیه پساکنشی[۷۴]" است که ممکن است بعد از صدور فرمان، صرفاً دلیلی ظاهراً موجه بیاورد بدون آنکه لزوماً بازتاب‌دهنده‌ی فرآیند درونی مدل باشد [۵۷]. نتایج نشان

---

[۷۳] Introspective
[۷۴] Rationalization

داد که روش درون‌نگرانه می‌تواند روابط علّت و معلولی بین ورودی و خروجی شبکه را بهتر نشان دهد و در نتیجه برای فهم رفتار مدل مفیدتر است [۵۷].

این رویکرد بر روی مجموعه‌داده‌ی جدیدی به نام BDD-X[۷۵] ارزیابی شد که شامل هزاران کلیپ رانندگی همراه با توضیحات متنی انسان درباره‌ی رفتار راننده است [۵۷]. مدل کیم و همکاران توانست با استفاده از این داده‌ها جملات نسبتاً نزدیک به توضیحات انسانی تولید کند و هر اقدام رانندگی را با ذکر دلیل توجیه نماید [۵۷]. به عنوان نمونه، برای یک ویدئو مدل توضیح می‌دهد «خودرو به جلو حرکت می‌کند زیرا خودروی دیگری در مسیرش نیست» که از نظر کیفی شبیه توضیح انسانی معادل («خودرو در خیابان به پیش می‌رود چون مسیرش خالی است») می‌باشد [۵۷]. این موفقیت اولیه نشان داد که ترکیب دیداری-زبانی همراه با مکانیزم توجه می‌تواند گامی موثر به سوی خودروهای خودران قابل‌توضیح باشد.

۲-۵- جمع‌بندی فصل دوم

بررسی تحلیلی این فصل در زمینه‌ی تلفیق سامانه‌های خودران با هوش مصنوعی تبیین‌پذیر و مدل‌های زبانی بزرگ، خلأهای پژوهشی قابل‌توجهی را آشکار ساخت. این خلأها نشان می‌دهند که علی‌رغم پیشرفت‌های قابل‌توجه در هر یک از این حوزه‌ها به‌صورت مجزا، همگرایی مؤثر میان آن‌ها هنوز محقق نشده‌است. برای مثال، سامانه‌های خودران مبتنی بر مدل‌های پیچیده‌ی یادگیری عمیق اغلب فاقد شفافیت و تفسیرپذیری کافی‌اند و روش‌های کنونی نیز به‌طور کامل از ظرفیت مدل‌های زبانی بزرگ برای تشریح فرایند تصمیم‌گیری این سامانه‌ها بهره‌برداری نکرده‌اند. پیامد چنین کاستی‌هایی، بروز چالش‌هایی در زمینه‌ی اعتماد کاربران به عملکرد سامانه‌های خودران و نظارت بر آن‌ها است. این امر بر ضرورت ارائه‌ی رویکردهای نوین برای رفع نقایص مذکور تأکید می‌کند.

[۷۵] Berkeley DeepDrive eXplanation

پژوهش حاضر دقیقاً با هدف پاسخ‌دهی به خلأهای پژوهشی یادشده طراحی شده‌است. روش پیشنهادی برای نیل به این هدف، با ادغام قابلیت‌های هوش مصنوعی تبیین‌پذیر و توانمندی‌های مدل‌های زبانی بزرگ در معماری سامانه‌های خودران، می‌کوشد شفافیت و تفسیرپذیری تصمیم‌های این سامانه‌ها را بهبود دهد. چنین رویکردی ضمن هدف‌قراردادن کاستی‌های شناسایی‌شده، زمینه‌ساز ارتقای اعتماد کاربران به عملکرد سامانه‌های خودران و افزایش شفافیت کلی آن‌ها خواهد بود. بدین ترتیب، در فصل سوم، روش‌شناسی و چارچوب پیشنهادی این پژوهش که پاسخی عملی به شکاف‌های مذکور محسوب می‌شوند، به تفصیل تشریح می‌گردند.

جدول ۲-۲- بررسی برخی روش های پیشین

منبع	سال	هدف	روش	مزایا	معایب	معیار ارزیابی	مجموعه داده	نوع داده
[۵۶]	۲۰۱۷	تشخیص زاویه فرمان	CNN, LSTM	دقت بالا		MAE	Comma.ai,Udacity , HCE	ویدیو
[۷۴]	۲۰۱۸	توضیح متنی برای تصمیم مدل	شبکه کانولوشن و شبکه عصبی	عدم نیاز به مدل جداگانه برای استخراج ویژگی	توضیحات محدود و نیاز به داده آموزشی زیاد، روش	نظر سنجی	BDD-X	ویدیو
[۹]	۲۰۲۳	توضیح متنی برای تصمیم مدل	Transformer	دقت بالا در تصمیم گیری	دقت پایین در توضیح دهنده	BLEU-٤, ROUGE-L, METEOR, CIDEr-D	BDD-X	ویدیو
[۱۰]	۲۰۲۳	توضیح متنی برای تصمیم مدل	text-davinci-۰۳ GPT-۳,۵ GPT-٤	توضیحات پویا روش ارزیابی الگوریتمی	نیاز به ورودی متنی	LFS USE	IMDB, AG News, SNLI	متن
[۲۰]	۲۰۲۳	توضیح متنی برای تصمیم مدل	توضیح پذیری انتها به انتها	توضیحات متنی کامل و قابل فهم	دقت پایین تصمیم گیری	نظرسنجی (انسان و chat gpt)	DQA	تصویر
[۵]	۲۰۲۴	تبدیل خروجی XAI به توضیحات متنی	GPT-٤	قابلیت استفاده در مدل های متنوع	نیاز به خروجی XAI, روش ارزیابی نامناسب	نظرسنجی	SHAPley, fake news, SMV	تصویر / متن

فصل سوم

روش پیشنهادی

۳–۱– کلیات فصل

هدف نهایی این پروژه توسعه یک سامانه خودکار برای تولید توضیحات متنی دقیق و گویا از صحنه‌های رانندگی است. به عبارت دیگر، این سامانه باید بتواند با استفاده از مدل‌های یادگیری عمیق و مدل‌های زبانی بزرگ، شرح یا زیرنویس‌هایی غنی برای تصاویر یا فریم‌های ویدیویی رانندگی ایجاد کند. اهمیت این هدف در آن است که با فراهم کردن توضیحات شفاف و قابل فهم، می‌توان فهم بهتری از محیط رانندگی را برای راننده یا سامانه‌های کمکی فراهم کرد. این توضیحات اضافی می‌توانند به افزایش قابلیت توضیح‌پذیری سیستم‌های هوش مصنوعی در حوزه رانندگی خودکار کمک کنند و درک تصمیمات مدل را آسان‌تر سازند.

فصل سوم پایان نامه حاضر به تشریح روش‌ها و مراحل پیاده‌سازی اختصاص دارد که در راستای نیل به هدف یادشده انجام شده‌اند. نقش این فصل در ساختار کلی پروژه بسیار کلیدی است، زیرا جزئیات روش پیشنهادی، ابزارهای مورد استفاده، و تنظیمات آموزشی در آن ارائه می‌شود. در واقع، فصل سوم پل ارتباطی بین مبانی نظری (مباحث فصل‌های پیشین) و نتایج عملی (فصل‌های بعدی) است و توضیح می‌دهد که چگونه ایده‌های مطرح‌شده به مرحله‌ی اجرا درآمده‌اند.

این فصل به چند بخش تقسیم شده است تا هر جنبه از روش و پیاده‌سازی به صورت منسجم توضیح داده شود. در ابتدا، داده‌ها و ابزارهای به کار رفته، شامل معرفی مجموعه‌دادهBDD-X ، کتابخانه‌ها و مدل‌های زبانی مورد استفاده، معرفی می‌شوند. پس از آن، مراحل پیش‌پردازش داده‌های ویدیویی و متنی به تفصیل بیان می‌گردد. در ادامه، طراحی دو ماژول اصلی سامانه تشریح می‌شود: نخست، مدل تصمیم‌گیر مبتنی بر معماری ADAPTو یک رویکرد جایگزین ناموفق که آزموده شد ، و سپس، طراحی ماژول توضیح‌گر که از خروجی توجه و مهندسی پرامپت بهره می‌برد. در بخش بعدی، معماری کلی سیستم ترکیبی ارائه شده و در پایان، معیارها و روش‌های ارزیابی توضیحات تولیدی معرفی می‌گردند.

۳–۲– معرفی داده‌ها و ابزارها

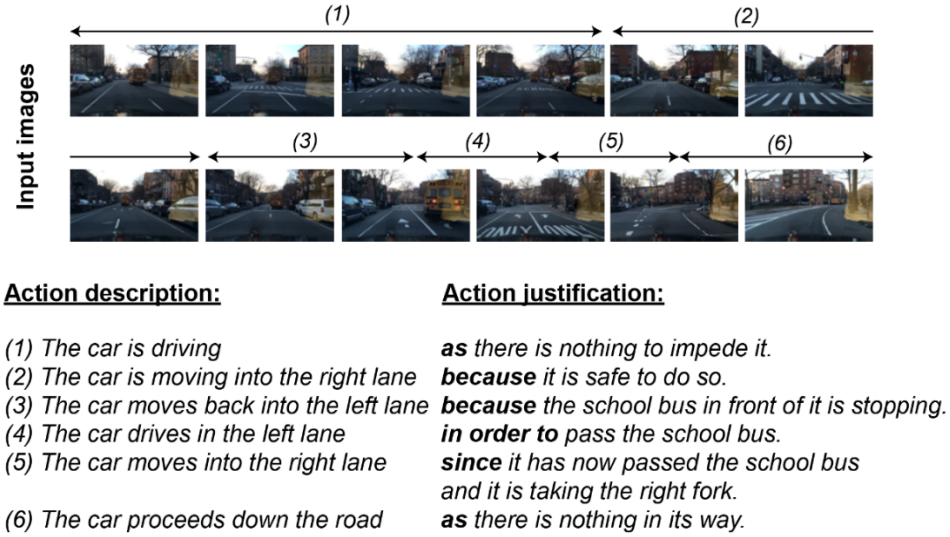
۳–۲–۱– معرفی مجموعه داده BDD-X

تمامی آموزش‌ها و آزمایش‌های مدل بر روی نسخه‌ی رسمی مجموعه‌داده‌ی BDD-X انجام گرفته است [۵۷]. مجموعه‌داده‌ی BDD-X بخشی از پروژه‌ی Berkeley DeepDrive بوده و شامل تصاویر (یا فریم‌های ویدئویی) از صحنه‌های متنوع رانندگی به همراه توضیحات متنی انسان درباره‌ی هر صحنه است. این توضیحات متنی به عنوان زیرنویس‌ها یا شرح‌های مرجع[۷۶] عمل می‌کنند که مدل باید آن‌ها را تا حد امکان بازتولید یا شبیه‌سازی کند.

این مجموعه داده‌ها شامل بیش از ۷۷ ساعت رانندگی در قالب ۶۹۷۰ ویدیو است. این ویدیوها در شرایط رانندگی متنوع، مانند روز/شب، بزرگراه/شهر/حومه شهر، تابستان/زمستان و غیره گرفته شده‌اند. هر ویدیو به طور متوسط ۴۰ ثانیه طول دارد و شامل حدود ۳ تا ۴ عمل است، مانند افزایش

[۷۶] Ground Truth

سرعت، کاهش سرعت، پیچیدن به راست و غیره، که همه آنها با توضیحات و توضیحات حاشیه‌نویسی شده‌اند. این مجموعه داده‌ها شامل بیش از ۲۶ هزار فعالیت در بیش از ۸.۴ میلیون فریم است [۵۷].



تصویر ۳–۱– بخشی از مجموعه داده BDD-X و اطلاعات موجود در آن [۵۷]

۳–۲–۲– ابزارها و کتابخانه‌های مورد استفاده

جهت پیاده‌سازی و اجرای مدل‌های پیشنهادی، از کتابخانه‌ی یادگیری عمیق PyTorch در محیط Google Colab استفاده شده است. دلیل انتخاب این بستر، سهولت دسترسی و امکان استفاده از پردازنده‌های گرافیکی (GPU) قدرتمند بوده است. به طور خاص، از کارت گرافیک قدرتمند NVIDIA A۱۰۰ موجود در Google Colab استفاده گردید که سرعت و کارایی فرآیند آموزش و ارزیابی مدل‌ها را به طور قابل توجهی افزایش داد. کد پایه‌ی استفاده‌شده در این پژوهش، از مخزن رسمی مدل ADAPT اقتباس شده و تغییراتی در آن، به‌ویژه در ورودی‌ها، جهت هماهنگی بیشتر با اهداف پژوهش اعمال شده است.

جهت بهره‌گیری از مدل‌های زبانی بسیار بزرگ مانند Gemini-۲٫۵ ، از رابط‌های برنامه‌نویسی کاربردی[۷۷] رسمی شرکت‌های مربوطه استفاده شده است. به طور مشخص، از API رسمی Google و OpenAI برای دسترسی به مدل‌های زبانی تحت سرویس ابری آن‌ها بهره گرفته‌ایم. این امر اجازه داد تا برخی از مدل‌های بسیار حجیم که روی سخت‌افزار محلی قابل اجرا نبودند، از طریق سرویس‌های ابری مورد استفاده و ارزیابی قرار گیرند.

۳–۲–۳– مدل‌های زبانی مورد استفاده

در این پروژه چندین مدل زبانی بزرگ جهت تولید زیرنویس‌ها و توضیحات مورد آزمایش قرار گرفته اند. مهم‌ترین این مدل‌ها عبارتنداز:

- Gemini-۲٫۵-Pro: یک مدل زبانی بزرگ با اندازه‌ی ۲.۵ میلیارد پارامتر که نسخه‌ی حرفه‌ای Gemini محسوب می‌شود.
- Gemini-۲٫۵-Flash: مدل دیگری از خانواده‌ی Gemini با همان اندازه‌ی پارامتری اما تنظیمات فلش برای سرعت بیشتر.
- Gemini-۲٫۵-Flash-Lite: نسخه‌ی بهینه‌شده و سبک‌تر مدل Gemini-Flash که میزان محدودیت استفاده از آن نسبت به سایر مدل های خانواده Gemini کمتر است.
- GPT-۵-Mini: نسخه‌ای از مدل GPT-۵ متعلق به OpenAI.
- GPT-٤o: نسخه‌ای از مدل GPT-٤ (متعلق به OpenAI) که در این پروژه جهت Fine-Tuning استفاده شده است زیرا تا به امروز این مدل تنها مدل خانواده GPT با قابلیت FT و پشتیبانی از تصاویر در FT است.

مدل Gemini-۲٫۵-Flash-Lite به طور خاص برای تولید توضیحات اضافی (زیرنویس‌های تکمیلی) مورد استفاده قرار گرفت. با این حال، تمامی مدل‌های فوق در طول فرآیند توسعه و آزمایش

---

[۷۷] API

به کار گرفته شدند تا بهترین مدل و تنظیمات برای هدف پروژه انتخاب شود. استفاده از این تنوع مدل‌ها کمک کرد تا مقایسه جامعی میان توانایی مدل‌های مختلف در توصیف صحنه‌های رانندگی صورت گیرد.

۳–۳– پیش پردازش داده ها

در این بخش به روش‌های آماده‌سازی داده‌های خام پیش از ورود به مدل می‌پردازیم. مراحل پیش‌پردازش شامل پردازش ویدیوهای ورودی (استخراج فریم‌ها و کاهش ابعاد تصاویر)، پردازش متن زیرنویس‌های مربوط به ویدیو (شامل تفکیک و توکن‌سازی جملات توضیحی) و نهایتاً سازمان‌دهی داده‌های حاصل برای مرحله‌ی آموزش مدل و ارزیابی عملکرد آن است. در ادامه، هر یک از این مراحل را به تفصیل بررسی می‌کنیم.

۳–۳–۱– مراحل پیش‌پردازش ویدیو بر اساس مقاله ADAPT

۱. نمونه برداری

ویدیوهای خام ثبت‌شده (مثلاً در مجموعه‌داده‌ی BDD-X) دارای طول زمانی و وضوح بالا هستند و مستقیماً قابل استفاده در مدل نیستند [۹]. هر ویدیو به‌طور متوسط ۴۰ ثانیه طول دارد و با نرخ ۳۰ فریم بر ثانیه ضبط شده است (حدود ۱۲۰۰ فریم در هر ویدیو) و وضوح تصویری ۱۲۸۰×۷۲۰ پیکسل دارد [۵۷]. پردازش همه‌ی فریم‌های هر ویدیو هزینه‌بر بوده و حاوی اطلاعات تکراری است؛ لذا نمونه‌برداری یکنواخت از فریم‌های ویدیو به‌عنوان گام نخست پیش‌پردازش انجام می‌شود. در این روش، از هر ویدیو تعداد ثابتی فریم با فواصل زمانی برابر استخراج می‌گردد تا نماینده‌ی کل ویدیو باشند [۹]. به عنوان مثال، اگر هر ویدیو N فریم داشته باشد و بخواهیم T فریم را استخراج کنیم، می‌توان فریم‌ها را با گام تقریباً $\frac{N}{T}$ انتخاب کرد (مثلاً فریم‌های شماره‌ی $\frac{N}{T}, ۲\frac{N}{T}, ۳\frac{N}{T}$ و...)، به‌نحوی که کل طول ویدیو پوشش داده‌شود. این رویکرد نمونه‌برداری یکنواخت تضمین می‌کند که بخش‌های آغازین، میانی

و پایانی هر ویدیو همگی در ورودی مدل نماینده‌ای داشته باشند و مدل تنها بر بخش کوتاهی از ویدیو متمرکز نشود.

تعداد فریم‌های نمونه‌برداری‌شده از هر ویدیو یک ابرپارامتر مهم است که بر دقت مدل و همچنین پیچیدگی محاسباتی تأثیر می‌گذارد. رویکرد ADAPT پیشنهاد می‌کند که هرچه فریم‌های بیشتری به‌صورت متراکم برداشت شود، محتوای بصری کمتری از دست می‌رود و عملکرد مدل در توصیف ویدیو بهبود می‌یابد [۹]. همانطور که درجدول ۳–۱ نشان داده شده است، نتایج آزمایش‌ها در مقاله‌ی ADAPT تأیید می‌کند که افزایش تعداد فریم‌های ورودی از ۲ به ۳۲، کیفیت زیرنویس‌های تولیدی را به طور چشم‌گیری بالا می‌برد.

جدول ۳–۱–تاثیر تعداد فریم ویدیو بر کیفیت زیرنویس های تولیدی با استفاده از معیار های ارزیابی زیرنویس مانند BLEU-٤، CIDEr-D، METEOR و ROUGE-L/*9*.

تعداد فریم	تصمیم گیری				دلیل آوری				هزینه (دقیقه)
	BLEU-4	CIDEr-D	METEOR	ROUGE-L	BLEU-4	CIDEr-D	METEOR	ROUGE-L	
۲	۳۳.۴	۲۲۷.۷	۲۸.۷	۶۱.۰	۸.۷	۶۲.۹	۱۵.۱	۲۹.۸	۲۹۴
۴	۳۲.۹	۲۲۵.۷	۲۹.۰	۶۰.۹	۹.۹	۸۱.۳	۱۴.۹	۳۱.۱	۳۸۲
۸	۳۲.۶	۲۳۶.۱	۲۹.۳	۶۱.۸	۸.۴	۸۳.۷	۱۳.۴	۳۰.۶	۴۴۷
۱۶	۳۲.۵	۲۳۱.۰	۲۹.۵	۶۱.۹	۸.۷	۹۱.۵	۱۳.۸	۳۲.۰	۵۲۸
۳۲	**۳۴.۶**	**۲۴۷.۵**	**۳۰.۶**	**۶۲.۸**	**۱۱.۰۴**	**۱۰۲.۶**	**۱۵.۲**	**۳۲.۰**	۷۹۷

برای نمونه، امتیاز شاخص CIDEr-D مربوط به توضیحات «دلیل» با افزایش فریم‌های نمونه‌برداری از ۲ به ۳۲ فریم، از حدود ۶۳ به ۱۰۳ ارتقاء یافته است که نشان‌دهنده‌ی پوشش بهتر محتوای ویدیو و عدم از دست رفتن جزئیات مهم است. البته افزایش تعداد فریم‌ها هزینه‌ی محاسباتی بیشتری نیز در بر دارد (مثلاً زمان آموزش با ۳۲ فریم در هر ویدیو تقریباً سه برابر حالت ۲ فریم گزارش

شده است) [۹]؛ بنابراین در عمل باید بین دقت مدل و کارایی محاسباتی موازنه برقرار کرد. در پیاده‌سازی ما (بر مبنای نتایج ADAPT)، تعداد معینی فریم (که جزئیات دقیق آن در فصل ۴ گزارش خواهد شد) به‌صورت یکنواخت از هر ویدیو استخراج می‌گردد تا بهترین توازن بین دقت و هزینه حاصل شود.

۲. کاهش بعد[۷۸] و نرمال سازی

پس از نمونه‌برداری فریم‌ها، کاهش ابعاد تصاویر و نرمال‌سازی آن‌ها انجام می‌شود. به منظور تطبیق با ورودی‌های مورد نیاز شبکه‌ی بینایی (ویدیو-ترنسفورمر)، تمامی فریم‌های انتخاب‌شده به اندازه‌ی ثابت ۲۲۴×۲۲۴ پیکسل تغییر اندازه داده شده و در صورت لزوم برش مرکزی[۷۹] می‌خورند [۹]. کاهش رزولوشن تصاویر به ابعاد ۲۲۴ مزایایی دارد: نخست حجم داده‌ی ورودی را به‌شدت کاهش داده و سرعت آموزش را بالا می‌برد، دوم اینکه شبکه‌ی بینایی (نظیر مدل Swin) معمولاً برای این ابعاد از پیش‌آموزش دیده شده و بهترین عملکرد را روی تصاویر با این اندازه ارائه می‌دهد [۹]. پس از تغییر اندازه، مقادیر پیکسل‌های هر فریم نیز نرمال‌سازی می‌شوند (مثلاً تطبیق با توزیع آماری مجموعه داده‌های مرجع مانند ImageNet) تا شبکه‌ی عصبی از نظر مقیاس شدت روشنایی/رنگ دچار چالش نشود. خروجی این مرحله مجموعه‌ای از تصاویر فریم یکنواخت و هم‌اندازه است که آماده‌ی استخراج ویژگی‌های بصری هستند.

۳. استخراج سیگنال های کنترلی

آخرین گام در پیش‌پردازش ویدیو، آماده‌سازی داده‌های جانبی مرتبط با هر فریم است. در مسئله‌ی مورد نظر (توصیف خودران)، علاوه بر ویدیو، سیگنال‌های کنترلی خودرو (نظیر سرعت لحظه‌ای و زاویه‌ی فرمان خودرو) نیز وجود دارند که همزمان با ویدیو ثبت شده‌اند. [۹] برای هر فریم

---

[۷۸] Downsize

[۷۹] Center Cut

انتخاب‌شده، مقادیر سیگنال کنترلی در همان زمان نمونه‌برداری استخراج می‌شود و به‌صورت یک برچسب کمکی ذخیره می‌گردد. به این ترتیب به‌ازای هر ویدیو یک توالی از فریم‌های منتخب و توالی متناظر سیگنال‌های عددی خواهیم داشت. این اطلاعات در مرحله‌ی آموزش برای انجام یادگیری چندکارهٔ مدل (پیش‌بینی همزمان توضیحات متنی و سیگنال‌های کنترلی) مورد استفاده قرار می‌گیرند. پژوهش ADAPT نشان داده است که دخیل کردن سیگنال‌هایی مانند سرعت و زاویه‌ی حرکت[۸۰] در کنار داده‌های ویدیویی، باعث بهبود دقت مدل در تولید توضیحات می‌شود، چرا که این سیگنال‌ها نمایانگر وضعیت دینامیکی خودرو بوده و به تفسیر بهتر صحنه و عمل خودرو کمک می‌کنند [۹].

تصویر ۳–۲ مراحل گفته شده در بالا را به صورت خلاصه و در قالب شبه کد نشان می دهد.

```
Input: I ∈ [۰,۱]^(H×W×۳)   # تصویر RGB نرمال‌شده

     A ∈ [۰,۱]^(h×w)      # نقشه توجه نرمال‌شده

Params: τ = ۰,۷, σ = ۰, thickness = ۳


۱) O ← uint۸(۲۵۵ · I)                # پایه ترسیم

۲) A' ← resize(A, (W, H))            # هم‌ترازسازی با تصویر

۳) T  ← threshold(uint۸(۲۵۵ · A'), ۲۵۵·τ) # ماسک دودویی نواحی مهم

٤) C  ← find_external_contours(T)       # کانتور مؤلفه‌های هم‌بند

٥) draw_contours(O, C, color=red, thickness=۳)  # ترسیم outline

٦) save_as_jpeg(O)
```

تصویر ۳–۲– شبه کد مراحل پیش پردازش فریم های ویدیویی برای استفاده در مدل پیشنهادی

---

[۸۰] Course

۳–۳–۲– پردازش زیرنویس‌ها و تولید زیرنویس‌های افزوده با کمک LLM

هر ویدیو (و هر بخش رفتاری در ویدیو) دارای یک زیرنویس متنی است که عمل رانندگی انجام‌شده توسط خودرو و دلیل آن را توصیف می‌کند [۵۷]. در مجموعه‌داده‌ی BDD-X هر زیرنویس از دو بخش تشکیل شده است: بخش اول شرح اقدام[۸۱] که به زبان ساده می‌گوید خودرو چه کاری انجام می‌دهد (مانند «خودرو متوقف می‌شود») و بخش دوم دلیل آن اقدام (مانند «زیرا چراغ راهنمایی قرمز است»). [۵۷] این دو جمله در کنار هم توضیح کاملی از رفتار خودرو ارائه می‌دهند. پیش از ورود این متون به مدل زبانی، لازم است آن‌ها را پردازش و آماده‌سازی کنیم.

در مرحله‌ی پیش‌پردازش متن، ابتدا تمامی حروف به شکل استاندارد (معمولاً حروف کوچک لاتین برای مجموعه‌دادگان انگلیسی) در می‌آیند و کاراکترهای غیرضروری (علائم نگارشی اضافی، اعداد و ...) در صورت نیاز حذف یا تبدیل می‌شوند. سپس برای نمایش متون به شکل عددی قابل فهم برای مدل، از توکن‌سازی مبتنی بر زیرواژه‌ها استفاده می‌کنیم. ما با الهام از روش به‌کاررفته در مدل‌های زبانی بزرگ مانند BERT، از الگوریتم WordPiece برای شکستن جملات به واحدهای معنایی کوچکتر بهره بردیم. این روش به مدل امکان می‌دهد واژه‌های ناآشنا یا ترکیبی را نیز به‌صورت توکن‌های قابل پردازش درآورد. به عنوان نمونه، کلمه‌ی «stops» در واژه‌نامه‌ی ما به دو توکن «stop» و «##s» تفکیک می‌شود. به این ترتیب ریشه‌ی اصلی کلمه (stop) و پسوند آن s (که نشان‌دهنده‌ی سوم‌شخص مفرد فعل است) جداگانه مد نظر قرار می‌گیرند. استفاده از توکن‌های زیرواژه‌ای باعث می‌شود حجم واژه‌نامه محدود بماند و در عین حال انعطاف مدل در برابر کلمات ناشناخته افزایش یابد. پس از شکستن جمله به توکن‌ها، توکن‌ها به شناسه‌های عددی متناظرشان در واژه‌نامه‌ی مدل نگاشت می‌شوند تا قابل پردازش توسط شبکه‌ی عصبی باشند. همچنین به‌منظور یکسان‌سازی طول ورودی‌های متنی، تمامی جملات تا طول مشخصی لایه گذاری[۸۲] می‌شوند؛ بدین صورت که برای هر جمله حداکثر مثلاً

---

[۸۱] Narration

[۸۲] Pad

۱۵ کلمه/توکن در نظر گرفته می‌شود و جملات کوتاه‌تر با توکن ویژه‌ی خنثی ([PAD]) پر می‌شوند. (در مجموعه‌دادهٔ مورد استفاده، طول اکثر جملات زیرنویس کوتاه بوده و ۱۵ کلمه برای هر بخشِ شرح اقدام یا دلیل کافی است.) علاوه بر این، از نشانه‌های ویژه‌ای همچون [CLS] در ابتدای توالی و [SEP] در انتهای هر جمله استفاده می‌شود تا مدل نواحی مختلف (جمله‌ی اول و جمله‌ی دوم) را تشخیص دهد. در مجموع، با این پردازش‌ها، متن هر نمونه ویدیو به دنباله‌ای از شناسه‌های عددی با طول ثابت تبدیل می‌شود که آماده‌ی ورود به بخش زبانی مدل است [۹].

گام ابتکاری بعدی، افزایش داده‌های متنی با کمک مدل‌های زبانی بزرگ است. با وجود آنکه مجموعه‌داده‌ی اصلی BDD-X شامل نزدیک به ۷۰۰۰ ویدیو و ۲۹۰۰۰ توضیح متنی است [۵۷]، این حجم از داده لزوماً به معنای تنوع زبانی کافی نیست و مدل ممکن است در مواجهه با جملات تازه دچار افت عملکرد شود. این محدودیت از چند عامل کلیدی ناشی می‌شود:

اولاً، ماهیت تکراری سناریوهای رانندگی منجر به توصیفات مشابه و کلیشه‌ای می‌شود. ثانیاً، وجود تنها یک توضیح انسانی برای هر رویداد، تنوع دیدگاهی و بیانی را از بین می‌برد. در نتیجه، این خطر وجود دارد که مدل به جای یادگیری ارتباط مفهومی میان تصویر و زبان، بر روی الگوهای زبانی خاص مجموعه داده دچار بیش‌برازش شود و توانایی تعمیم خود را از دست بدهد.

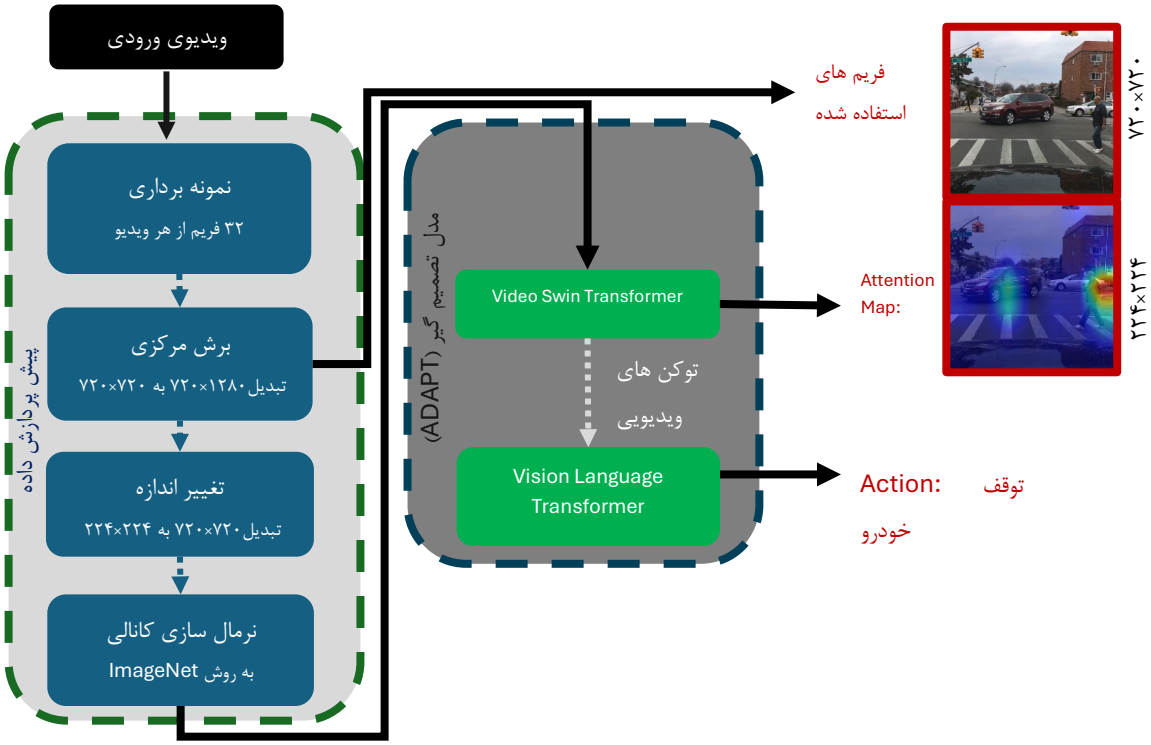
برای رفع این محدودیت، از توان مدل‌های زبان بزرگ نظیر GPT بهره گرفتیم تا زیرنویس‌های افزوده تولید کنیم. ایده اصلی این است که با دادن هر زیرنویس موجود به یک مدل زبانی قدرتمند، از آن بخواهیم همان معنا را با ادبیات یا جزئیات متفاوتی بیان کند. به عنوان نمونه، فرض کنید زیرنویس اصلی برای یک ویدیو این باشد: «خودرو به راست می‌پیچد زیرا عابر پیاده در حال عبور است.» یک LLM می‌تواند چند بازنویسی مختلف از این عبارت تولید کند، مثلاً: «خودرو برای احتیاط به سمت راست منحرف می‌شود چون یک عابر پیاده از خیابان عبور می‌کند.» یا «ماشین به سمت راست متمایل می‌شود، چرا که عابری در حال گذر از عرض خیابان است.». این جملات افزوده از نظر مفهومی معادل

زیرنویس اصلی هستند اما از کلمات و ساختار جمله‌ی متفاوتی استفاده می‌کنند. ما با تنظیم مناسب دستور (دستورالعمل ورودی) به مدل زبانی بزرگ، اطمینان حاصل کردیم که جملات جدید از لحاظ معنا دقیق بوده و اطلاعات اصلی (عمل خودرو و دلیل آن) را حفظ می‌کنند. سپس به‌ازای هر زیرنویس اصلی، یک یا چند زیرنویس بازنویسی‌شده‌ی جدید را به مجموعه‌ی آموزش اضافه نمودیم. بدین ترتیب حجم و تنوع داده‌های آموزشی در بخش زبان به شکل چشمگیری افزایش یافت، بی‌آنکه نیاز به جمع‌آوری دستی توضیحات جدید باشد. پژوهش‌های اخیر تأیید کرده‌اند که بهره‌گیری از داده‌های ساختگی تولیدشده توسط مدل‌های هوش مصنوعی می‌تواند به غنای مجموعه‌داده و بهبود عملکرد مدل بیانجامد [۷۵]. در واقع، مدل زبانی بزرگ با دانش گسترده‌ی زبانی خود جملاتی روان و متنوع تولید می‌کند که مدل ما را برای مواجهه با عبارات مختلف در توصیف یک صحنه‌ی رانندگی آماده‌تر می‌سازد [۷۵]. در فصل ۴، جزئیات بیشتری در مورد نحوه‌ی استفاده از LLM و تعداد زیرنویس‌های افزوده‌شده ارائه خواهیم کرد و تأثیر این روش را بر نتایج مدل مورد ارزیابی قرار می‌دهیم.

۳–۴– طراحی مدل تصمیم‌گیر

تصویر ۳–۳ شمای کلی مدل تصمیم گیر را نشان می دهد. این مدل بر پایه مدل ADAPT است که در بخش ۳–۴–۱ به طور مفصل توضیح داده می‌شود. در بخش پیش پردازش، ۳۲ فریم ویدیو با سایز ۱۲۸۰×۷۲۰ برش مرکزی داده شده و به ۷۲۰×۷۲۰ تبدیل می‌شوند. سپس سایز فریم ها از ۷۲۰×۷۲۰ به ۲۲۴×۲۲۴ کاهش می یابد. این سایز مناسب برای Video Swin Transformer آموزش دیده موجود در مدل ADAPT است. در نهایت نرمال سازی کانالی انجام شده و فریم ها وارد مدل ADAPT می‌شوند. از مدل تصمیم گیر ۳ خروجی مهم استخراج می‌شود:

۱. فریم های خام ۷۲۰×۷۲۰
۲. نقشه توجه ۲۲۴×۲۲۴
۳. تصمیم



تصویر ۳–۳– بلوک دیاگرام مدل تصمیم گیر

۳–۴–۱– معماری ADAPT و نحوه استفاده آن در تحقیق حاضر

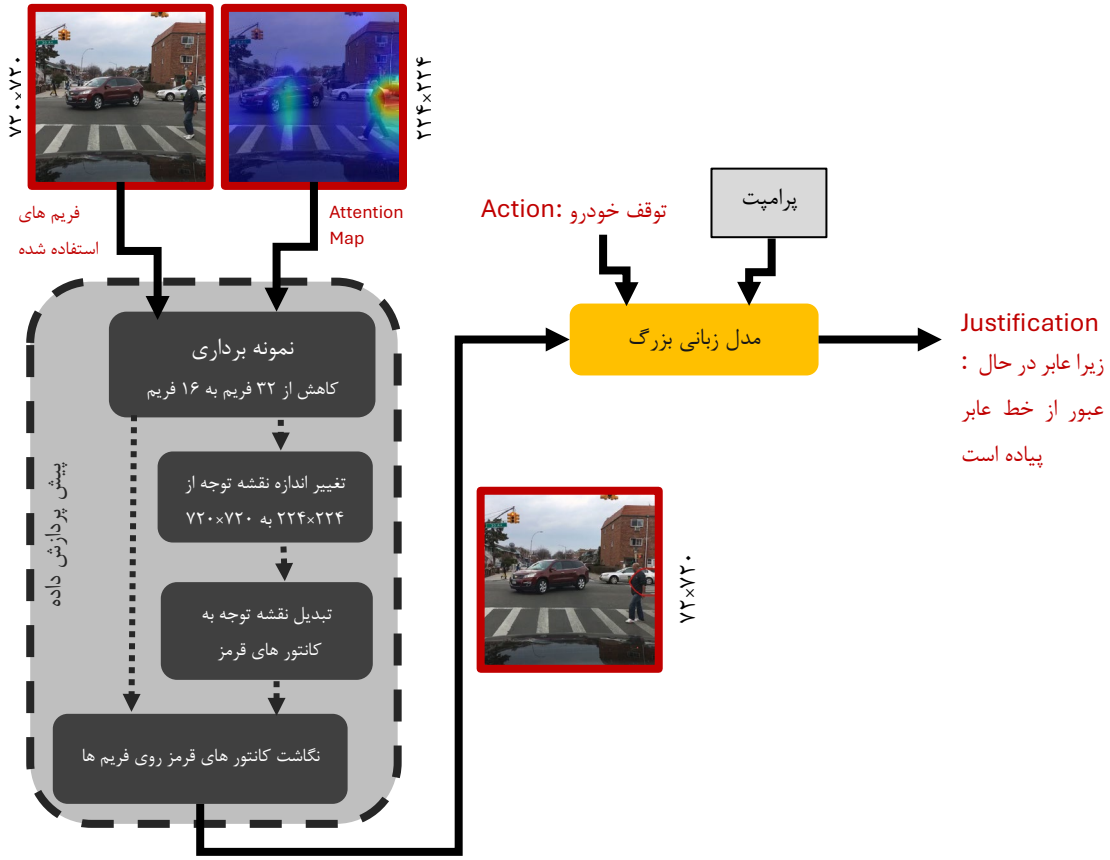
شمای کلی معماری مدل ADAPT در تصویر ۲–۲ نشان داده شده است. در این معماری، ورودی یک ویدیوی نمای جلو از خودرو است و خروجی شامل سیگنال‌های کنترلی وسیله نقلیه (مانند فرمان و سرعت) به‌همراه یک توضیح متنی درباره اقدام فعلی خودرو و دلیل آن می‌باشد. برای استخراج ویژگی‌های ویدیو، از یک ترنسفورمر ویدیویی (مانندVideo Swin Transformer) استفاده می‌شود که فریم‌های نمونه‌برداری‌شده از ویدیو را به توکن‌های ویدیویی تبدیل می‌کند. سپس دو بخش پیش‌بینی موازی بر مبنای این نمایش ویدیویی مشترک عمل می‌کنند: یک سر (head) شبکه وظیفه‌ی پیش‌بینی فرمان کنترلی را بر عهده دارد و سر دیگر متن توضیحی (شامل روایت اقدام و استدلال آن) را تولید می‌کند [۹]. به این ترتیب، مدل ADAPT به صورت همزمان هم رفتار کنترلی خودرو را تصمیم‌گیری می‌کند و هم توضیح قابل فهمی برای آن ارائه می‌دهد.

در پیاده‌سازی این تحقیق، ساختار کلی مدل ADAPT بدون تغییر اساسی از منبع اصلی اقتباس شده است. به جز تطبیق‌های جزیی مربوط به قالب ورودی داده‌ها (برای سازگاری با مجموعه‌داده مورد استفاده)، هیچ تغییری در معماری شبکه انجام ندادیم. به عبارت دیگر، بخش استخراج ویژگی ویدیویی و سرهای دوگانه‌ی تولید فرمان و متن عیناً مطابق معماری اصلی ADAPT حفظ شده‌اند. این مدل تصمیم‌گیر در چارچوب پژوهش حاضر به‌عنوان ماژول تولیدکننده‌ی اقدام و تبیین به‌کار گرفته شده است تا ضمن تصمیم‌گیری‌های کنترلی، توضیحات قابل فهمی نیز برای هر اقدام فراهم شود. این مدل به دلیل معماری نوین و چندوظیفه‌ای آن انتخاب شد که به طور همزمان هم به پیش‌بینی فرمان کنترلی و هم به تولید توضیح متنی برای آن می‌پردازد. علاوه بر این، ADAPTعملکرد برجسته‌ای بر روی مجموعه‌داده BDD-X ، که در این پژوهش نیز مبنای کار است، از خود نشان داده و نتایج آن در ارزیابی‌های خودکار و انسانی برتر بوده است. ساختار مبتنی بر ترنسفورمر و یادگیری مشترک این دو وظیفه، ارتباطی قوی میان ادراک بصری صحنه و دلایل زبانی تصمیم ایجاد می‌کند که مستقیماً با هدف این پژوهش برای دستیابی به تبیین‌پذیری هم‌راستاست.

همانطور که در بخش ۲-۳-۳ بیان شد، برای بهبود بخش تصمیم گیر مدل ADAPT تصمیم گرفتیم از رویکرد گسترش داده‌ها به کمک زیرنویس‌های اضافی بهره ببریم. روش کار بدین صورت بود که برای هر نمونه‌ی ویدیو در مجموعه‌ی آموزش، چند زیرنویس (تصمیم + توضیح) جایگزین تولید کردیم که همان صحنه‌ی رانندگی را با بیان‌های متفاوت توصیف می‌کنند. زیرنویس اولیه‌ی هر ویدیو توسط انسان ارائه شده بود؛ ما به کمک یک مدل زبانی بزرگ (نظیر یک مدل مبتنی بر GPT) چهار زیرنویس دیگر به‌صورت مصنوعی ایجاد کردیم که از نظر معنایی با زیرنویس اصلی هم‌سو هستند اما از لحاظ جمله‌بندی و جزئیات متفاوت‌اند. بدین ترتیب، هر ویدیو حداکثر پنج توصیف متنی مختلف (یک توضیح اصلی به‌همراه چهار توضیح تولیدشده توسط LLM) در مجموعه‌ی داده‌ی آموزش دارد. این زیرنویس‌های افزوده‌شده ممکن است بر جنبه‌های مختلفی از صحنه تاکید کنند یا با واژگان متنوع‌تری همان مفهوم را بیان نمایند. در نتیجه، مدل طی فرایند آموزش مواجهه‌ی گسترده‌تری با

توصیف یک وضعیت رانندگی پیدا می‌کند و کمتر به یک عبارت‌بندی خاص وابسته می‌شود. بهره‌گیری از چنین زیرنویس‌های تولیدشده توسط LLM می‌تواند با افزودن دانش زبانی متنوع به داده‌های آموزش، دقت و قابل‌فهم بودن خروجی‌های مدل را بهبود دهد. به بیان دیگر، مدل با دیدن توصیف‌های گوناگون برای یک موقعیت مشابه، درک عمیق‌تری از ارتباط بین ویدیو و زبان پیدا می‌کند که می‌تواند آن را در مواجهه با موارد جدید یاری رساند.

۳–۵– طراحی ماژول توضیح‌گر



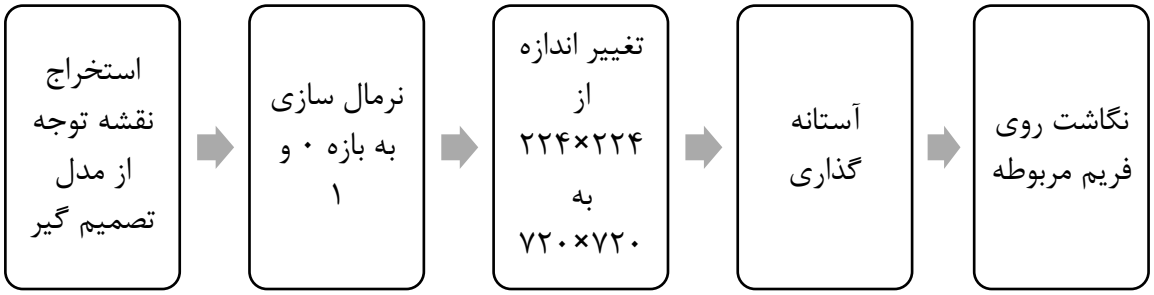
تصویر ۳–۴– بلوک دیاگرام مدل توضیح دهنده

برای طراحی ماژول توضیح‌گر، ابتدا نقشه‌های توجه حاصل از مدل ADAPT از روی دنباله‌ی فریم‌های ورودی استخراج می‌گردد. این نقشه‌ها پس از نرمال‌سازی و آستانه‌گذاری، به‌جای نمایش گرادیانیِ متداول (آبی/زرد/قرمز)، به نشانه‌های بصری صریح تبدیل می‌شوند؛ بدین صورت که پیرامون نواحی با بیشترین میزان توجه، خطوط پیرامونی قرمز ترسیم می‌گردد تا بدون پوشاندن جزئیات تصویر، مکان دقیق نواحی کانونی برای مدل زبانی آشکار شود. متعاقباً، یک فریم شاخص (یا مجموعه محدودی از فریم‌ها) همراه با نشانه‌های توجهِ ترسیم‌شده و نیز «اقدام/تصمیم»[۸۳] تولیدشده توسط ADAPT، در قالب یک بسته ورودی منسجم به مدل زبانی بزرگ[۸۴] ارائه می‌شود. دستور واحد و از پیش‌طراحی‌شده، نقش و قیود پاسخ را به‌روشنی مشخص می‌کند (بیان موجز در ۱ یا ۲ جمله، پرهیز از عبارت‌های زائد و فرضِ صحت تصمیم سامانه)، تا LLM بر مبنای سرنخ‌های فضاییِ توجه و لنگر معناییِ «اقدام»، دلیل تصمیم را به شکلی کوتاه، مستقیم و قابل‌فهم تولید کند. این طراحی از یک‌سو اتصال معناییِ «چه کاری انجام شده است» (اقدام) و «کجا برای مدل مهم بوده است» (توجه) را برای LLM فراهم می‌آورد و از سوی دیگر با ساده‌سازی نمایش توجه به صورت خطوط قرمز، ابهام ادراکی LLM در تفسیر نقشه‌های رنگی پیوسته را کاهش می‌دهد. بدیهی است مقادیر دقیق پارامترهای پردازش (مانند آستانه، تعداد فریم‌های منتخب و ضخامت خطوط) و تحلیل‌های مقایسه‌ای مربوط، در فصل چهارم گزارش خواهد شد.

---

[۸۳] Action
[۸۴] LLM

۳–۵–۱– نحوه استفاده از نقشه توجه خروجی ADAPT



تصویر ۳–۵– بلوک دیاگرام پردازشات انجام شده روی نقشه توجه جهت استفاده در مدل توضیح دهنده

به‌منظور بهره‌گیری مؤثر از اطلاعات توجهِ تصویری[۸۵] تولیدشده در ADAPT و قابل‌فهم‌کردن آن برای ماژول زبانی، جریان زیر اجرا شد:

- استخراج نقشه توجه برای هر فریم

برای هر ویدئو که پس از پیش‌پردازش به توالی T فریم $(I_۱, I_۲, \dots, I_T)$ با اندازه ۲۲۴×۲۲۴ تبدیل شده است، از ADAPT یک نقشه توجه $A_t \in \mathbb{R}^{H\times W}$ برای فریم t به‌دست می‌آید (خروجی لایه‌های توجهِ ماژول بینایی/ترنسفورمر). این نقشه بیانگر توزیعِ اهمیت ناحیه‌ای در تصویر است.

- نُرمال‌سازی و هم‌ترازسازی به ابعاد تصویر

برای هر A_t، ابتدا به بازه [۰,۱] نُرمال‌سازی می‌کنیم:

$$\widehat{\mathrm{A}}_{\mathrm{t}} = \frac{A_t - \min(A_t)}{\max(A_t) - \min(A_t)} \qquad (۳-۱)$$

سپس با درون‌یابی دوسویه $\widehat{\mathrm{A}}_{\mathrm{t}}$ به اندازه تصویر ورودی ۷۲۰×۷۲۰ بزرگ‌نمایی می‌شود تا مختصات نقشه توجه با مختصات پیکسلی فریم هم‌راستا گردد:

$$\tilde{A}_t = resize(\hat{A}_t, ۷۲۰ \times ۷۲۰) \qquad (۳-۲)$$

---

[۸۵] Visual Attention

- آستانه‌گذاری برای جداسازی نواحی شاخص

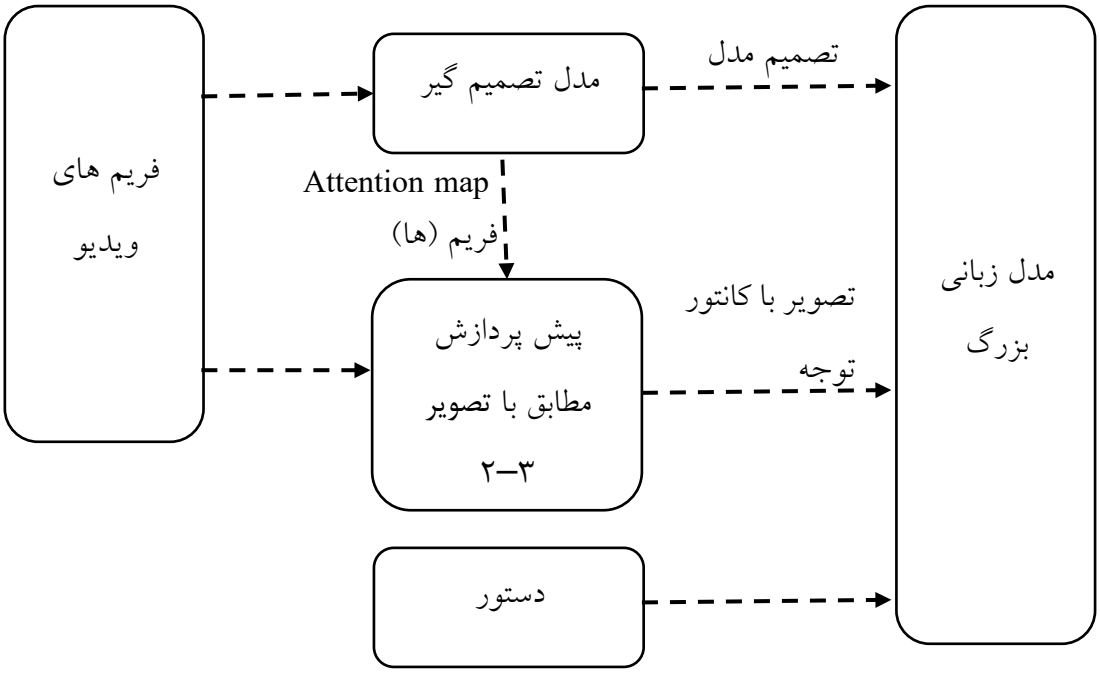
نقشه دودوییِ ناحیه‌های «پر اهمیت» با آستانه τ تولید می‌شود:

$$M_t(i,j) = \begin{cases} ۱, & if\ S_t(i,j) \geq \tau \\ ۰, & otherwise \end{cases} \quad (۳-۳)$$

(پارامترهای τ به‌صورت تجربی انتخاب شده و مقادیر دقیق در فصل ۴ گزارش می‌شود.)

تصویر ۳–۶– نمونه آستانه گذاری در نقشه توجه

در ادامه، به منظور تعیین مرز نواحی پرتوجه، از الگوریتم استخراج کانتور خارجی استفاده گردید تا تنها مرز بیرونی هر مؤلفه هم‌بند در ماسک حفظ شود و حفره‌های داخلی نادیده گرفته شوند. این کانتورها با رنگ قرمز و ضخامت معین بر روی تصویر اصلی ترسیم شدند تا بدون پوشاندن محتوای تصویر، نواحی کلیدی به‌طور واضح و قابل‌درک مشخص گردند. این نمایش بصری جایگزین نقشه‌های رنگی پیوسته‌ی متداول (مانند Grad-CAM) شده و موجب درک ساده‌تر نواحی کانونی توسط مدل زبانی می‌شود.



تصویر ۳–۷–بلوک دیاگرام کلی مدل با تمرکز بر بخش توضیح دهنده

۳–۵–۲– طراحی دستور برای تولید توضیح تصمیم

دستور یکسان برای همه مدل‌ها

به منظور استفاده از مدل‌های زبانی بزرگ در نقش ماژول توضیح‌گر، یک دستور ثابت و از پیش‌طراحی‌شده برای تمام مدل‌ها به کار گرفته شد. این دستور به گونه‌ای طراحی گردید که ضمن تعیین نقش مدل زبانی، چارچوب وظیفه و زمینه اطلاعاتی را مشخص کند، محدودیت‌های محتوایی و ساختاری پاسخ را نیز به وضوح بیان نماید. در تدوین دستور، از روش‌های رایج مهندسی دستور مانند ارائه مثال‌های کوتاه و صریح، تعیین طول مطلوب پاسخ، و اجتناب از عبارات غیرضروری استفاده شد.

هدف از این طراحی، اطمینان از یکنواختی شرایط ورودی برای همه مدل‌های زبانی و کاهش واریانس ناشی از نحوه تفسیر دستورها بود. دستور به‌گونه‌ای ساختاربندی شد که با دریافت فریم‌های

حاوی نواحی پرتوجه (نشانه‌گذاری‌شده با خطوط قرمز) و اقدام تصمیم‌گیر، مدل زبانی بتواند در قالب یک یا دو جمله کوتاه، دلیل تصمیم را به‌صورت مستقیم و بدون حواشی بیان کند. ساختار دقیق و متن کامل دستور در فصل چهارم و هم‌زمان با جزئیات پیاده‌سازی و ارزیابی ارائه خواهد شد.

خلاصه عناصر کلیدی این دستور یکسان:

- دستورِ نقش واحد[۸۶]: همه مدل‌ها به‌عنوان سامانه توصیف سناریوهای رانندگی عمل می‌کنند تا فهم مشترکی از وظیفه داشته باشند.
- قالب ورودی سازگار[۸۷]: دستور، ورودی را یکسان قاب‌بندی می‌کند تا همه مدل‌ها همان زمینه را دریافت کنند.
- قید سخت[۸۸] بر خروجی: خروجی باید حداکثر ۱یا۲ جمله کوتاه باشد؛ بدون زمینه‌پردازی اضافه یا درخواست اطلاعات بیشتر.
- پرهیز از عبارت‌پردازی زائد[۸۹]: آغاز یا پایان‌های کلیشه‌ای صریحاً ممنوع شده تا تمرکز پاسخ روی «دلیل» باقی بماند.
- فرض صحت تصمیم: چون مدل به اطلاعات درونی سامانه دسترسی ندارد، باید تصمیم را صحیح فرض کند و فقط دلیل محتمل را بر مبنای شواهد دیداری بیان کند.
- نمونه‌های راهنما: سه پاسخ نمونه موجز و علّی برای نمایش قالب و لحن مطلوب گنجانده شده‌اند.

با این طراحی و به‌کارگیری یک دستور واحد برای همه مدل‌ها، چه مبتنی بر GPT و چه سایر مدل‌ها (مانند Gemini)، شرایط ارزیابی یکنواخت می‌شود و هر تفاوت کیفیِ مشاهده‌شده به خودِ

---

[۸۶] Single-role prompt

[۸۷] Consistent input format

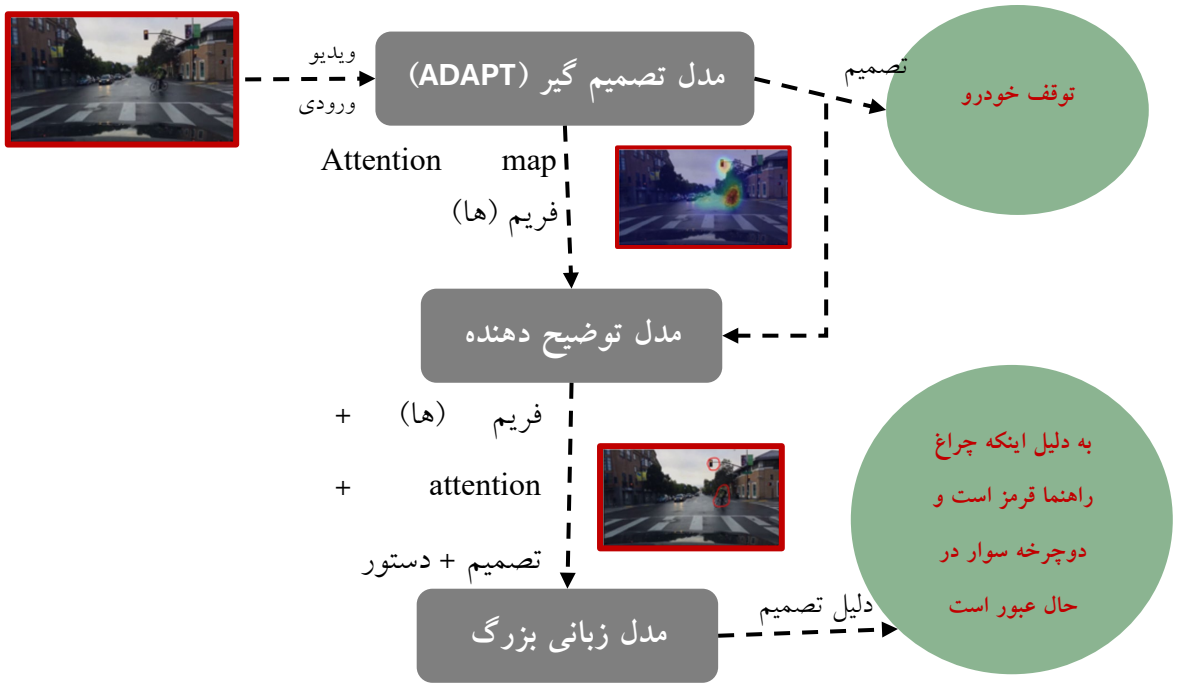
[۸۸] Hard constraint

[۸۹] Redundant phrasing

مدل‌ها نسبت داده خواهد شد، نه به تفاوت در نحوه طرح مسئله. استفاده از یک دستور یکسان، رویه‌ای رایج در مقایسه مدل‌های زبانی است زیرا سوگیری و واریانس ناشی از متن دستور را به حداقل می‌رساند.

۳–۵–۳– معماری کلی سیستم ترکیبی تصمیم‌گیر–توضیح‌گر

همان‌گونه که در بخش‌های پیشین تشریح شد، سیستم پیشنهادی شامل دو جزء اصلی است: مدل تصمیم‌گیر که وظیفه‌ی پردازش ورودی (برای مثال، داده‌ی ویدیویی) و اخذ تصمیم مناسب را بر عهده دارد، و مدل توضیح‌گر که دلیل تصمیم اتخاذشده را به زبان طبیعی بیان می‌کند. در این بخش، نحوه‌ی یکپارچه‌سازی این دو مدل در قالب یک معماری کلی توصیف می‌شود و فرآیند کلی سیستم ترکیبی تصمیم‌گیر–توضیح‌گر به‌صورت مرحله‌ای تشریح می‌گردد. شکل اصلی این بخش نیز نمای کلی معماری پیشنهادی را همراه با یک مثال عملی نشان می‌دهد.



تصویر ۳–۸ بلوک دیاگرام معماری سیستم پیشنهادی (بخش های قرمز رنگ مثال است)

تصویر ۳–۵ نمای کلی معماری سیستم ترکیبی تصمیم‌گیر-توضیح‌گر را به همراه یک سناریوی نمونه نشان می دهد. این معماری شامل مراحل متوالی زیر است:

۱. پردازش ورودی و تصمیم‌گیری: ابتدا داده‌ی خام (مانند تصاویر یک ویدیو) وارد سیستم شده و توسط مدل تصمیم‌گیر پردازش می‌شود. مدل تصمیم‌گیر پس از تحلیل ویژگی‌های مهم ورودی، یک تصمیم نهایی متناسب با وضعیت ارائه می‌دهد (برای مثال، در کاربرد خودران ممکن است تصمیم «توقف» یا «ادامه‌ی حرکت» اتخاذ گردد). هم‌زمان، مدل تصمیم‌گیر اطلاعات تکمیلی مرتبط با مبنای تصمیم خود را استخراج می‌کند؛ از جمله نقشه‌ی توجه مدل نسبت به آن فریم (ها). این اطلاعات پایه‌ی اولیه‌ی توضیح را فراهم می‌کنند.

۲. تولید توضیح توسط مدل توضیح‌گر: در گام بعد، مدل توضیح‌گر با استفاده از داده‌های استخراج‌شده‌ی مرحله‌ی قبل و تصمیم نهایی، به تولید توضیح متنی می‌پردازد. مدل توضیح‌گر که یک مدل زبانی را در بر دارد، اطلاعاتی نظیر فریم کلیدی برجسته‌شده، نتیجه‌ی تصمیم (خروجی مدل تصمیم‌گیر) و سایر ویژگی‌های مهم را به عنوان ورودی دریافت می‌کند. سپس با بهره‌گیری از دانش زبانی خود، توضیحی قابل فهم برای انسان تولید می‌نماید که دلیل تصمیم مدل را تشریح می‌کند. به عنوان نمونه، در سناریوی به تصویر کشیده‌شده، مدل تصمیم‌گیر تشخیص می‌دهد که خودرو باید متوقف شود؛ زیرا چراغ راهنمای ترافیک قرمز است و یک دوچرخه‌سوار در حال عبور از تقاطع می‌باشد. مدل توضیح‌گر بر همین اساس جمله‌ای نظیر «خودرو متوقف شد چون چراغ قرمز بود و دوچرخه‌سواری از مقابل عبور می‌کرد.» تولید می‌کند که به روشنی علت تصمیم را بیان می‌دارد.

۳. خروجی ترکیبی و قابل درک: در نهایت، سیستم ترکیبی خروجی خود را در قالب یک تصمیم همراه با توضیح آن ارائه می‌کند. بدین‌ترتیب هر تصمیم مدل دارای یک توجیه قابل فهم بوده و کاربر نهایی (یا سامانه‌ی نظارتی) می‌تواند ضمن دریافت نتیجه‌ی تصمیم‌گیری، توضیح همراه آن را نیز مشاهده کند. برای مثال در کاربرد خودروی خودران، سیستم اعلام می‌کند «توقف» و

همزمان دلیل آن را توضیح می‌دهد. خروجی ترکیبی تصمیم‌به‌همراه توضیح موجب می‌شود که فرآیند تصمیم‌گیری برای ناظران انسانی شفاف و قابل درک باشد.

این معماری ترکیبی مزایای قابل‌توجهی را به همراه دارد. نخست، نسبت به مدل های سنتی مانند کیم و همکاران[۵۶] تعامل‌پذیری سامانه با کاربر بهبود می‌یابد؛ زیرا کاربر می‌تواند دلیل هر تصمیم را درک کرده و در صورت لزوم بازخورد یا پرسش‌های تکمیلی مطرح کند. دوم، نسبت به سایر روش های توضیح پذیری مانند Grad-Cam قابلیت ارزیابی توضیحات فراهم می‌شود؛ به این معنی که کیفیت و درستی توضیح تولیدشده را می‌توان به‌طور مستقل سنجید و بدین‌ترتیب اطمینان حاصل کرد که مدل نه‌تنها تصمیم درستی می‌گیرد بلکه دلایل موجهی نیز ارائه می‌دهد. سوم، این رویکرد بر خلاف مدل های سنتی باعث تقویت شفافیت مدل تصمیم‌گیر می‌گردد؛ زیرا تصمیم‌های سیستم دیگر به صورت «جعبه‌ی سیاه» نیستند و همراه با استدلال آشکار عرضه می‌شوند که این خود به اعتمادپذیری سامانه می‌افزاید. به طور کلی، ترکیب یک مدل تصمیم‌گیر با یک مدل توضیح‌گر گامی موثر به سوی توسعه‌ی سامانه‌های هوشمند قابل توضیح و قابل اعتمادتر است. در فصل بعدی، به ارزیابی معماری پیشنهادی و تحلیل نتایج به‌دست‌آمده پرداخته خواهد شد و میزان اثربخشی و عملکرد این سیستم ترکیبی در سناریوهای واقعی مورد بررسی قرار خواهد گرفت.

۳–۵–۴– جمع‌بندی فصل سوم

در این فصل، روش‌شناسی پژوهش حاضر با تمرکز بر طراحی و پیاده‌سازی یک سامانه ترکیبی تصمیم‌گیر-توضیح‌گر برای سناریوهای رانندگی شرح داده شد. ابتدا با معرفی کلی داده‌ها و ابزارهای مورد استفاده، مجموعه داده BDD-X به‌عنوان پایگاه اصلی داده تشریح شد و مدل‌ها، کتابخانه‌ها و محیط‌های اجرایی مورد استفاده معرفی گردید. در ادامه، فرایند پیش‌پردازش داده‌ها شامل مراحل پردازش ویدئو بر اساس روشADAPT ، تولید زیرنویس‌های افزوده با کمک مدل زبانی بزرگ و آماده‌سازی داده‌ها برای آموزش و ارزیابی بیان شد.

سپس ساختار و نحوه استفاده از مدل تصمیم‌گیر ADAPT توضیح داده شد و فرآیند بسط داده‌ها با استفاده از زیرنویس‌های افزوده تشریح گردید. تلاش‌های جایگزین، از جمله تبدیل مسئله به طبقه‌بندی سرعت و جهت و پیاده‌سازی مدل‌های Kim et al. (۲۰۱۷) و ADAPT به‌عنوان طبقه‌بند نیز ارائه شد که با وجود آزمون عملی، نتایج مطلوبی به همراه نداشتند.

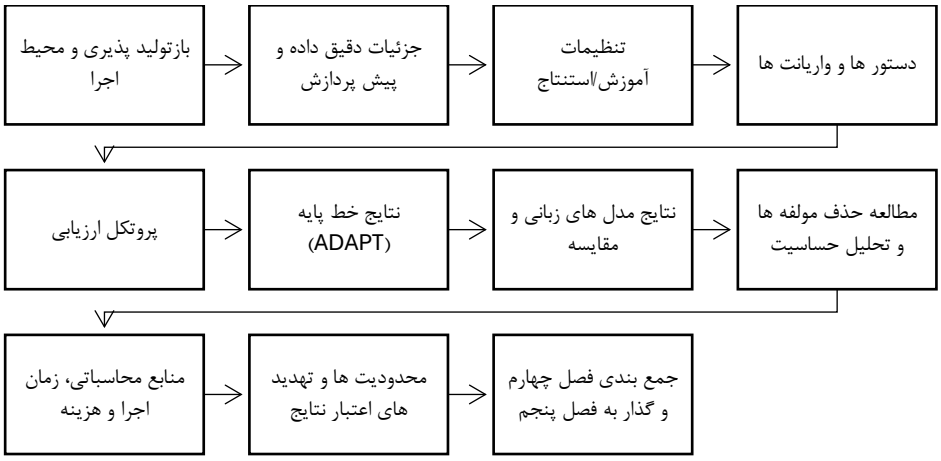
در بخش پایانی، طراحی ماژول توضیح‌گر مبتنی بر LLM و نقش استفاده از خروجی attention مدل ADAPT در تولید توضیحات بررسی گردید. معماری کلی سامانه، به‌عنوان ترکیبی از ماژول تصمیم‌گیر و ماژول توضیح‌گر، در قالب یک نمای کلی تشریح شد و در نهایت معیارها و روش‌های ارزیابی توضیحات مدل معرفی گردید. این فصل با ایجاد تصویری شفاف از اجزای روش پیشنهادی، زمینه را برای ورود به فصل چهارم و ارائه جزئیات پیاده‌سازی و نتایج فراهم می‌کند.

فصل چهارم

پیاده سازی و تحلیل نتایج

۴–۱– کلیات فصل

فصل حاضر به پیاده‌سازی و ارزیابی تجربی رویکرد معرفی‌شده در فصل سوم اختصاص دارد. در فصل قبل، معماری کلی سامانه ترکیبی تصمیم‌گیر- توضیح‌گر، جریان پیش‌پردازش، نحوه بهره‌گیری از توجه مدل ADAPT و طرح مهندسی دستور برای تولید توضیح‌های کوتاه توسط مدل‌های زبانی تبیین شد. در این فصل، همان طرح به‌صورت قابل‌بازتولید پیاده‌سازی می‌شود و نتایج کمی و کیفی حاصل، با پروتکل ارزیابی مشخص و معیارهای استاندارد گزارش می‌گردد.



تصویر ۴–۱–بلوک دیاگرام فصل چهارم

در گام نخست، مشخصات محیط اجرا و بازتولیدپذیری ارائه می‌شود: پیکربندی سخت‌افزاری/نرم‌افزاری (از جمله بستر اجرا، نوع GPU ، نسخه کتابخانه‌ها و وابستگی‌ها)، شیوه مدیریت کد و آزمایش‌ها (کنترل نسخه[۹۰]، بذرهای تصادفی[۹۱] ، ثبت رخدادها و لاگ‌ها) و دسترسی‌های موردنیاز (کلیدهای API و مسیرهای داده). سپس، جزئیات دقیق داده و پیش‌پردازش بیان می‌شود؛ از جمله آمار و تقسیم‌بندی train/val/test برای BDD-X ، نرخ و الگوی نمونه‌برداری زمانی از ویدئوها (سناریوهای n-fraction)، پارامترهای دقیق تغییر اندازه/برش/نرمال‌سازی، و ابرپارامترهای تبدیل نقشه توجه به کانتورهای قرمز (آستانه، هموارسازی، تعداد نواحی منتخب و...) همراه با شبه‌کد اجرایی.

در ادامه، تنظیمات آموزش/استنتاج تشریح می‌گردد: ابرپارامترهای آموزش ADAPT (به‌عنوان روش پایه‌ی تصمیم‌گیر/توضیح‌گر)، و جزئیات استنتاج مدل‌های زبانی مورد ارزیابی (Gemini ۲٫۵ Pro، Gemini ۲٫۵ Flash، Gemini ۲٫۵ Flash-Lite، GPT-٤o، GPT-٥ Mini) شامل نحوه بسته‌بندی ورودی (فریم کلیدی + کانتور توجه + اقدام/تصمیم)، پارامترهای فراخوانی API و تنظیمات تولید متن. سپس متن کامل دستور ثابت و منطق طراحی آن ارائه می‌شود و آزمایش‌های تکمیلی (Prompt Tuning برای GPT و Gemini) از منظر پروتکل اجرا و پیاده‌سازی مستندسازی خواهد شد.

بخش بعدی به پروتکل ارزیابی و پیاده‌سازی معیارهای استاندارد BLEU-٤ ، METEOR، ROUGE-L، CIDEr-D، SPICE اختصاص دارد؛ هر دو سناریو تک‌مرجع (زیرنویس انسانی BDD-X) و چندمرجع (افزودن چهار زیرنویس تولیدشده با LLM به‌عنوان مراجع مکمل) تشریح می‌شوند و رویه اجرای آزمایش‌ها و (در صورت استفاده) آزمون‌های آماری توضیح داده می‌شود. پس از آن، نتایج روش پایه ADAPT و سپس نتایج کمی مدل‌های زبانی یادشده گزارش و مقایسه می‌گردد. به‌علاوه، یک تحلیل ابلیشن و حساسیت ارائه خواهد شد تا اثر عوامل کلیدی (تعداد فریم‌های ورودی، وجود/عدم وجود نشانه توجه در ورودی LLM ، و افزودن مراجع LLM) روشن شود. در ادامه، تحلیل کیفی با

---

[۹۰] Version control

[۹۱] seed

نمونه‌های تصویری موفق و موارد خطا ارائه می‌شود و میزان هم‌خوانی توضیح‌های زبانی با نواحی توجه بررسی خواهد شد. در پایان فصل، نتایج مسیر ناموفقِ طبقه‌بندی سرعت/جهت (بازطراحی Kim ۲۰۱۷و ADAPT به‌عنوان طبقه‌بند) به‌صورت خلاصه و مستند آورده می‌شود و سپس محدودیت‌ها، ملاحظات محاسباتی و تهدیدهای اعتبار نتایج بحث می‌گردد.

تمرکز این فصل بر شماره‌گذاری دقیق تنظیمات، پارامترها و خروجی‌ها است تا هر پژوهشگری بتواند آزمایش‌ها را تکرار کند و به نتایج قابل مقایسه برسد. بدین‌سان، فصل چهارم پیوند میان «طرح روش» و «عملکرد واقعی» را برقرار می‌سازد و بستر لازم را برای جمع‌بندی نهایی و پیشنهادهای آینده فراهم می‌کند.

۴–۲– بازتولیدپذیری و محیط اجرا

این بخش سازوکارهای تضمین بازتولیدپذیری نتایج را تشریح می‌کند: پیکربندی دقیق سخت‌افزاری/نرم‌افزاری، شیوه‌ی مدیریت کد و آزمایش‌ها، و نحوه‌ی دسترسی به وابستگی‌های بیرونی (کلیدهای API، مسیرهای داده و اسکریپت‌ها). هدف آن است که هر پژوهشگری بتواند با تکیه بر این اطلاعات، آزمایش‌ها را با حداقل تفاوت دوباره اجرا کند.

۴–۲–۱– پیکربندی سخت‌افزاری/نرم‌افزاری

- بستر اجرا: تمامی آزمایش‌ها در محیط Google Colab انجام شده است.

پردازنده گرافیکی: در اکثر اجراها از NVIDIA A۱۰۰ (۴۰ GB) استفاده شده و در برخی اجراها NVIDIA L۴ به‌کار رفته است.

- CUDA: نسخه ۱۲.۴
- چارچوب و کتابخانه‌ها: چارچوب اصلی یادگیری عمیق PyTorch ۲٫۶ بوده و از کتابخانه‌های استاندارد پردازش تصویر و آرایه (نظیر OpenCV و NumPy) استفاده شده است.

- ثبت نسخه‌ها: نسخه دقیق اجزا (CUDA/PyTorch و کتابخانه‌های کلیدی) در گزارش محیط هر اجرا ثبت و در جدول ۴–۱ خلاصه شده است. با توجه به ماهیت موقتی Colab، در آغاز هر نشست، همان نسخه‌ها نصب و بررسی می‌شوند تا ناهمگونی محیطی به حداقل برسد.

جدول ۴–۱– پیکربندی محیط اجرا

مولفه	مقدار
بستر اجرا	Google Colab
GPU	NVIDIA A۱۰۰ (۴۰ GB)؛ در برخی اجراها NVIDIA L۴
CUDA	۱۲.۴
PyTorch	۲.۶
کتابخانه‌های کمکی	OpenCV، NumPy و ... (مطابق گزارش محیط هر اجرا)

مدیریت کد و آزمایش‌ها:

- مخزن و شاخه اجرایی: کدها در مخزن GitHub[۹۲] پروژه نگهداری می‌شود (branch: master). برای هر اجرای اصلی، شناسه commit متناظر همراه با خروجی‌ها ثبت شده است تا وضعیت دقیق کد در زمان اجرا قابل ارجاع باشد.پیکربندی آزمایش‌ها: تنظیمات هر آزمایش (ابرپارامترها، مسیر داده‌ها، مدل هدف) در فایل‌های پیکربندی مستقل ذخیره و همراه با خروجی‌ها آرشیو شده‌اند؛ بدین‌ترتیب هر آزمایش با همان تنظیمات قابل تکرار است.
- پیکربندی آزمایش‌ها: تنظیمات هر آزمایش (ابرپارامترها، مسیرهای داده، مدل هدف) در پرونده‌های پیکربندی مستقل ذخیره و به همراه مصنوعات خروجی آرشیو شده‌اند؛ بدین‌ترتیب، هر آزمایش با همان تنظیمات قابل تکرار است.

[۹۲] https://github.com/OmidMsl/explainable_drive

- بذرهای تصادفی[۹۳]: به‌منظور کاهش واریانس، بذرهای تصادفی برای چارچوب‌ها/ماژول‌های مربوط ثابت‌گذاری شده‌اند. با وجود این، به‌طور شفاف اشاره می‌شود که برخی عملیات موازی GPUمی‌توانند اندکی غیرقطعی باقی بمانند؛ لذا اجرای چندباره آزمایش‌های حساس توصیه می‌شود.
- ثبت لاگ و پایش: لاگ‌های آموزشی/استنتاجی به‌صورت فایل ذخیره شده و با TensorBoard پایش می‌شوند. تمامی مصنوعات (مدل‌های ذخیره‌شده، خروجی‌های ارزیابی، جداول و اشکال) به‌صورت ساخت‌یافته بایگانی شده‌اند.

۴–۳– جزئیات دقیق داده و پیش‌پردازش

این بخش، مشخصات کمیِ داده و کلیه مراحل پیش‌پردازش را با ذکر مقادیر نهایی ارائه می‌کند تا اجرای مجدد آزمایش‌ها بدون ابهام امکان‌پذیر باشد. ترتیب ارائه به‌ترتیب شامل آمار و تفکیک مجموعه داده BDD-X، طرح نمونه‌برداری زمانی و تبدیل‌های هندسی/نرمال‌سازی، و در پایان پارامترهای تبدیل attention به نمایش کانتوری (ورودی LLM) است.

۴–۳–۱– آمار مجموعه داده BDD-X و تقسیم‌بندی دقیق (train / val / test)

- نسخه داده: نسخه رسمی مجموعه‌داده BDD-X از مخزن GitHub (شاخه‌ی master، آخرین commit در ۲۰۱۹-۰۱-۱۹).
- واحد نمونه: در تمامی مراحل، واحد پردازش «کلیپ ویدیویی» پس از قطعه‌بندی به ۳۲ فریم یکنواخت (مطابق ۱-۳-۳) است.

---

[۹۳] Seeds

- تفکیک نهایی: داده به سه بخش آموزش[۹۴]، اعتبارسنجی[۹۵] و آزمون[۹۶] تفکیک شده است. آمار دقیق هر بخش (تعداد کلیپ‌ها و جمع فریم‌های مؤثر) در جدول ۴–۳ درج شده است.

جدول ۴–۲– آمار تفکیک BDD-X

بخش	تعداد کلیپ
آموزش	۲۱،۱۴۳
اعتبارسنجی	۲،۵۱۹
آزمون	۲،۵۶۶
جمع	**۲۶،۲۲۸**

- طول زمانی کلیپ‌ها (ثانیه): میانگین ۸.۷ ، میانه ۶، کمینه ۱ و بیشینه ۵۴.
- سیاست تفکیک: بدون هم‌پوشانی صحنه‌ای/زمانی؛ تفکیک بر مبنای «تصمیم‌های متفاوت» هر بخش.
- پاک‌سازی/فیلتر: برای خط لوله اصلی ADAPT و ماژول توضیح‌گر حذفی اعمال نشده است؛ صرفاً در آزمایش‌های fine tuning، کلیپ‌هایی که امکان قطعه‌بندی مناسب نداشتند حذف شده‌اند.
- کنترل نشت اطلاعات: جداسازی دقیق مجموعه‌ها، جلوگیری از هم‌پوشانی صحنه‌ای و نشت زمانی[۹۷]، پیش‌پردازش مستقلِ هر بخش و رعایت پروتکل ارزیابی منصفانه.

---

[۹۴] Train

[۹۵] Validation

[۹۶] Test

[۹۷] Temporal leakage

۴–۳–۲– نمونه‌برداری زمانی و تبدیل‌های Resize/Crop/Normalize

- نمونه‌برداری زمانی
  - برای مدل تصمیم‌گیر (ADAPT): هر کلیپ خام، به ۳۲ فریم با فواصل یکنواخت نگاشت می‌شود.
  - برای ماژول توضیح‌گر : هر کلیپ خام، به ۱۶ فریم با فواصل یکنواخت نگاشت می‌شود.
  - بخش تصمیم گیر با همین ۳۲ فریم محاسبه می شود. سپس ارزیابی با زیرمجموعه‌های ۱، ۴، ۸، ۱۲ و ۱۶ فریمِ نخست هر کلیپ انجام می شود. در تمامی این سناریوها، سایر شرایط ثابت نگه داشته شده‌اند.
- تبدیل‌های هندسی[۹۸]
  - ورودی خام دوربین: ۱۲۸۰×۷۲۰ (نسبت ۱۶:۹).
  - برش مرکزی به مربع: برای جلوگیری از اعوجاج و حفظ میدان دید عمودی، از برشِ افقیِ مرکزی استفاده می‌شود: از تصویر ۱۲۸۰×۷۲۰، ستون‌های کناری جمعاً ۵۶۰ پیکسل حذف شده و ناحیه مرکزیِ ۷۲۰×۷۲۰ نگه داشته می‌شود (هر سمت ≈ ۲۸۰ px).
  - تغییر اندازه: برای مدل تصمیم گیر ناحیه ۷۲۰×۷۲۰ به ۲۲۴×۲۲۴ تغییر اندازه می‌یابد (در عمل با درون‌یابی دوخطی). برای مدل توضیح دهنده تغییر اندازه ای انجام نمی شود.
  - یکنواخت‌سازی شدت: مقادیر شدت پیکسل‌ها به بازه [۰, ۱] نگاشت می‌شود.
- نرمال‌سازی کانالی (به روش ImageNet[76])

---

[۹۸] Geometric Transforms

$$\sigma = [۰.۲۲۵, ۰.۲۲۴, ۰.۲۲۹],\ \mu = [۰.۴۰۶, ۰.۴۵۶, ۰.۴۸۵],\ Normalize(x) = \frac{x-\mu}{\sigma} \quad (۴-۱)$$

این نرمال‌سازی بر همه فریم‌ها (پس از تغییر اندازه) اعمال شده است.

افزایش داده تصویری به‌عمد به‌کار نرفته است؛ تنها افزوده‌سازی در سطح زیرنویس‌های متنیِ مرجع (چهار زیرنویس اضافی به ازای هر نمونه با LLM) انجام شده که شرح آن در فصل ۳ آمده است.

- سازگاری با خط لوله[۹۹] ADAPT
  - ترتیب پردازش فریم‌ها، قالب تنسورها و نرمال‌سازی مطابق نیازهای ورودیِ ماژول ویدیویی ADAPT رعایت شده است؛ به‌این‌منظور، تمامی فریم‌های هر کلیپ پس از تبدیل‌های فوق در قالب بسته‌های ورودی هم‌شکل[۱۰۰] قرار می‌گیرند.

۴–۳–۳– پارامترهای تبدیل attention به کانتور قرمز (τ، σ، K، ضخامت)

برای آن‌که LLM نواحیِ کانونیِ نگاه مدل تصمیم‌گیر (ADAPT) را به‌صورت صریح و غیرمبهم دریافت کند، نقشه attention از قالب پیوسته گرادیانی به نمایش کانتوری قرمز تبدیل می‌شود (نمای بیرونی لکه‌های پُرتوجه). شیوه انجام این کار در ۳–۵–۱ شرح داده شد. در این پژوهش، تبدیل با پارامترهای ثابت زیر انجام شده است:

- آستانه توجه (τ): «آستانه توجه در این پژوهش برابر با ۰.۷ در نظر گرفته شد؛ بدین معنا که هر پیکسل با مقدار توجه بزرگ‌تر یا مساوی این حد، «مهم» تلقی می‌گردد (τ=۰,۷). انتخاب این مقدار بر اساس آزمون چند نمونه‌ی اولیه و بررسی کیفی خروجی‌ها و میزان پوشش بصری حاصل از آن انجام گرفت و مقدار ۰.۷ به‌عنوان سطحی متعادل میان دقت و گستره پوشش انتخاب شد.

---

[۹۹] pipeline

[۱۰۰] Batch-friendly

- تعداد نواحی (K) : همه کانتورهای خارجی به‌جای انتخابTop-K ، تمام مؤلفه‌های عبوری از آستانه ترسیم می‌شوند.
- ضخامت خط کانتور: ۳ پیکسل.
- **حداقل مساحت مؤلفه: اعمال نشده است (همه مؤلفه‌ها رسم می‌شوند).**
- قالب خروجی تصویر: JPEG .

```
# Input: I ∈ [۰،۱]^(H×W×۳) تصویر RGB نرمال‌شده
# A ∈ [۰،۱]^(h×w)  نقشه توجه نرمال‌شده
Params: τ = ۰،۷, σ = ۰, thickness = ۳

# ۱) O ← uint۸(۲۵۵ · I)  پایه ترسیم
# ۲) A' ← resize(A, (W, H))  هم‌ترازسازی با تصویر
# ۳) T ← threshold(uint۸(۲۵۵ · A'), ۲۵۵·τ)  ماسک دودویی نواحی مهم
# ٤) C ← find_external_contours(T)  کانتور مؤلفه‌های هم‌بند
# ٥) draw_contours(O, C, color=red, thickness=۳)  ترسیم outline
```

تصویر ۴–۲–شبه کد اجرایی جهت تبدیل نقشه توجه به کانتور های قرمز رنگ

همانطور که در *تصویر ۴–۲* نشان داده شده است، ترسیم به‌صورت خط پیرامونیِ قرمز بدون پُرشدگی انجام می‌شود تا جزئیات صحنه پوشانده نشود و نشانه مکانی برای LLM واضح باشد. با توجه به σ=۰ ، مرزها بدون فیلتر هموارسازی ترسیم می‌شوند.

۴–۴– تنظیمات آموزش/استنتاج

این بخش، تنظیمات اجرایی لازم برای بازتولید نتایج را به‌صورت منسجم و پایان‌نامه‌ای ارائه می‌کند: نخست پیکربندی آموزش ADAPT به‌عنوان روش پایه‌ی تصمیم‌گیر/توضیح‌گر، سپس تنظیمات استنتاج مدل‌های زبانی و در پایان شیوه‌ی بسته‌بندی ورودی‌ها برای LLM.

۴–۴–۱– آموزش ADAPT (روش پایه)

الگوریتم بهینه‌سازی و دقت محاسباتی. آموزش با بهینه‌ساز AdamW انجام شده است (وزن‌کاهی weight decay = ۰٫۲؛ سایر پارامترها مطابق پیاده‌سازی مرجع ADAPT محاسبات با fp۱۶ و چارچوب DeepSpeed انجام شده و Zero Optimization Stage = ۰ به‌کار رفته است. برشِ گرادیان با نرم $\|g\|_۲ \leq ۱.۰$ فعال بوده و تجمع گرادیان[۱۰۱] برابر با ۱ بوده است.

- ابرپارامترهای پایه آموزش ADAPT با پیکربندی زیر انجام شده است:

جدول ۴–۳– پیکربندی آموزشی ADAPT [۹]

پارامتر	مقدار
اندازه دسته[۱۰۲] فاز آموزش	۴
انداره دسته فاز ارزیابی	۱۲
فریم در هر کلیپ	۳۲
نرخ یادگیری[۱۰۳]	۰.۰۰۰۲
تعداد دوره[۱۰۴]	۴۰
Precision / Engine	fp۱۶ with DeepSpeed (ZeRO stage ۰)

[۱۰۱] Gradient accumulation

[۱۰۲] Batch size

[۱۰۳] Learning rate

[۱۰۴] Epoch

ضریب یادگیریِ بخشBackbone	۰.۰۰۵
احتمال ماسک‌گذاری	۰.۵
Gradient clipping	۱.۰
Gradient accumulation	۱

- زیان[۱۰۵] چندوظیفه‌ای[۱۰۶] و معیار انتخاب مدل: زیان آموزش به‌صورت چندوظیفه‌ای تعریف شده است؛ وزن زیانِ تولید زیرنویس = ۰.۹۵ و وزن زیانِ سیگنال/سنسور = ۰.۰۵است. سیاست توقف زودهنگام به‌کار نرفته است؛ بهترین checkpoint بر مبنای ترکیبی از کمینه‌ی زیان آموزش/اعتبارسنجی و بیشینه‌ی امتیاز معیارهای زیرنویس (در اعتبارسنجی) گزینش شده است. [۹]
- بستر اجرا و هزینه‌ی محاسباتی: آموزش روی Google Colab و با دو پردازنده گرافیکی NVIDIA A۱۰۰ (۴۰GB) و NVIDIA L۴ انجام شده است (بسته به دسترس‌پذیری؛ اکثر اجراها A۱۰۰). هر دوره به‌طور متوسط حدود ۱:۴۵ ساعت به طول انجامیده و زمان کل آموزش در حدود ۶۲۱ ساعت بوده است.

۴-۴-۲- استنتاج مدل‌های زبانی

- مدل‌ها: چهار مدل زبانی ارزیابی شده‌اند: Gemini ۲٫۵ Pro، Gemini ۲٫۵ Flash، Gemini ۲.۵ Flash-Lite و GPT-۵ Mini . عملیات fine tuning با دو مدل برتر هر پلتفرم یعنی Gemini ۲٫۵ Pro و GPT-۴o انجام شده است.

---

[۱۰۵] Loss

[۱۰۶] Multitask

- تنظیمات تولید متن: برای یکسان‌سازی شرایط، temperature = ۰٫۵ تعیین شده است؛ پارامترهای top-k و top-p در مقدارهای پیش‌فرضِ سرویس نگه داشته شده‌اند. طول خروجی به «۱ یا ۲ جمله کوتاه» محدود شده (مطابق طراحی دستور در ۴–۵).
- سناریوی ورودی. استنتاج LLM در پنج سطحِ فریم ورودی انجام شده است: ۱، ۴، ۸، ۱۲ و ۱۶ فریم نخست هر کلیپ (همه با کانتور قرمزِ توجه). بسته به سرویس:
  - برای Gemini: تصاویر به‌صورت مستقیم (تصویر چندوجهی) ارسال شده‌اند؛
  - برای GPT: تصاویر در بدنه‌ی درخواست به‌صورت base٦٤ ضمیمه شده‌اند.

انتخاب سطوح ۱، ۴، ۸، ۱۲ و ۱۶ فریم نخست هر کلیپ با هدف پوشش طیفی از حالات مرزی بین پیش‌بینی آینده[۱۰۷] و تشخیص[۱۰۸] صورت گرفت. بدین معنا که استفاده از تعداد کمتر فریم‌ها (مانند ۱ یا ۴) بیشتر ماهیت پیش‌بینی آینده دارد، در حالی که استفاده از کل ۱۶ فریم، معادل با پوشش کامل ویدیو و در نتیجه نزدیک‌تر به تشخیص است. به این ترتیب، سطوح میانی مانند ۸ و ۱۲ فریم شرایط بینابینی را بازنمایی می‌کنند و امکان بررسی تأثیر میزان اطلاعات ورودی بر کیفیت استنتاج مدل زبانی را فراهم می‌سازند.

- هر درخواست شامل اکشنِ خروجی ADAPT برای همان بازه‌ی زمانی و فریم(های) کانتورگذاری‌شده بوده است؛ سایر فراداده‌ها (از جمله سیگنال کنترلی و توضیحات تولیدی مدل ADAPT) حداقلی نگه داشته شده‌اند تا تمرکز مدل بر نشانه‌های بصری و اکشن باشد.
- داده‌ها برای بهینه‌سازی دستور به این صورت گردآوری شده‌اند: ۲۴۳۱ بخش ویدیو، هرکدام حاوی یک زیرنویس + ۴ فریم نخست آن بخش برای آموزش و ۵۸۵ بخش ویدیو، حاوی یک زیرنویس + ۴ فریم نخست آن بخش برای اعتبارسنجی (فریم‌ها دارای کانتور قرمز هستند). جزئیات روش و نتایج در ۴–۵–۳ ارائه می‌شود.

---

[۱۰۷] prediction

[۱۰۸] detection

۴–۴–۳– بسته‌بندی ورودی‌ها برای LLM

برای تضمین یکنواختی ارزیابی و هم‌خطی اطلاعات، ورودی LLM به‌صورت یک «بسته کمینه» سازمان‌دهی شده است که شامل موارد زیر است:

۱. تصویر/تصاویر کانتورگذاری‌شده: $n \in \{1,4,8,12,16\}$ فریم نخست کلیپ، پس از برش مرکزی و تغییر اندازه به ۲۲۴ × ۲۲۴ و نرمال‌سازی، به‌صورت JPEG با کانتور قرمزِ نواحی پرتوجه (مطابق۳–۳–۴).

۲. اکشن ADAPT: خلاصه تصمیم مدل تصمیم‌گیر برای همان بازه‌ی زمانی (لنگر معنایی توضیح).

۳. فراداده حداقلی: شمار فریم‌های ارسالی و شناسه نمونه، صرفاً برای رهگیری.

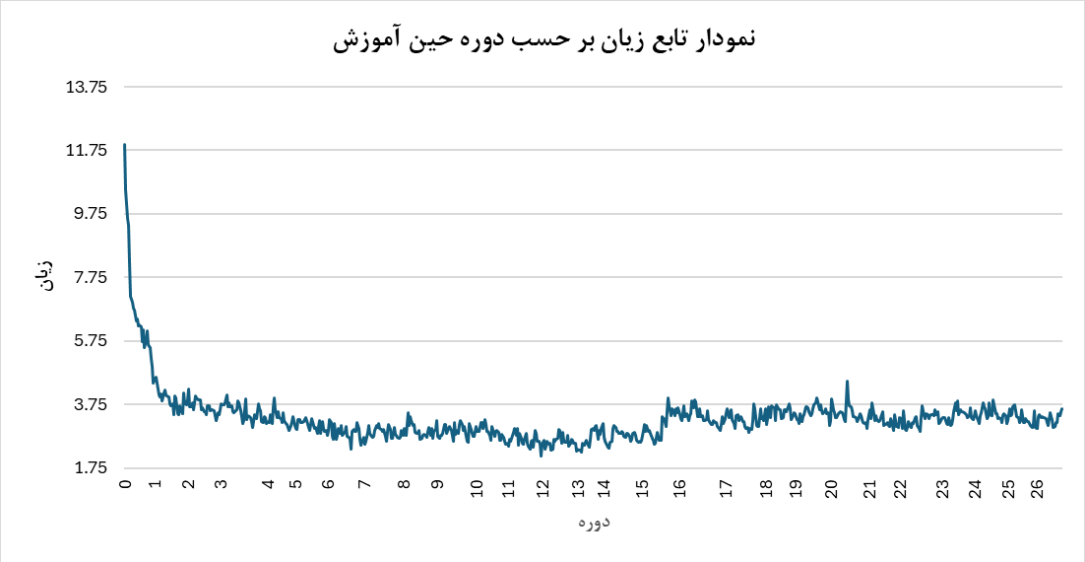
در همه‌ی سناریوها، ترتیب زمانی فریم‌ها حفظ شده و کانتورهای توجه با پنجره‌ی تصمیم‌گیری ADAPT هم‌زمان شده‌اند. این بسته‌بندی، از یک‌سو یکنواختی ورودی را برای همه مدل‌ها تضمین می‌کند و از سوی دیگر امکان مقایسه مستقیم کیفیت توضیحات را در سطوح مختلف اطلاعات بصری (۱ تا ۱۶ فریم) فراهم می‌سازد.

۴–۴–۴– باز آموزی ADAPT با زیرنویس های افزوده و تحلیل همگرایی

به‌منظور بهبود پایداری یادگیری در بخش تصمیم‌گیر و بهره‌برداری از داده‌های زبانیِ غنی‌تر، نسخه‌ای از ADAPT با همان پیکربندی پایه (۱–۴–۴) بازآموزی شد که در آن، علاوه بر زیرنویس‌های اصلی مجموعه BDD-X، برای هر نمونه چهار بازنویسی هم‌معنا (تولیدشده توسط LLM) نیز به‌عنوان مراجع زبانی افزوده به کار رفت. این بازآموزی صرفاً بر سرِ تصمیم‌گیر اعمال شد و سایر مؤلفه‌ها بدون تغییر باقی ماندند. اجرای آموزش بر GPU نوع NVIDIA L٤ انجام شد. از نظر زمان اجرا، این دورِ آموزش در ۵ دوره به انجام رسید و بهترین مدل در همین بازه کوتاه حاصل شد؛ بنابراین، با وجود

آن‌که بهبود معناداری در معیارهای ارزیابی تصمیم‌گیر نسبت به پیکربندی پایه مشاهده نشد، همگرایی آموزشی پایدارتر شده و نوسان زیان در مراحل آغازین کاهش یافت.

برای مستندسازی رفتار یادگیری، منحنی‌های زیان و دقت در آموزش/اعتبارسنجی گزارش می‌شوند (تصویر ۴–۳). همان‌گونه که دیده می‌شود، کاهش زیان آموزشی در ۲–۳ دوره نخست با افت نوسان در منحنی اعتبارسنجی همراه است. کمترین هزینه در دوره ۱۲ (۲.۱۲ ≈) حاصل میشود ولی طبق ارزیابی معیار های captioning در جدول ۴–۵ بهترین نقطه عملکرد زودتر از نسخه پایه و در دوره ۶ به‌دست می‌آید. با این حال، معیارهای اعتبارسنجی نهایی (مطابق ۷-٤) تغییر چشم‌گیری نشان نمی‌دهند و لذا این مسیر به‌عنوان بهینه‌سازیِ پایداری و نه ارتقای دقت نهایی قلمداد می‌شود.



تصویر ۴–۳– نمودار همگرایی تابع هزینه حین آموزش مدل ADAPT با زیرنویس های اضافی تولید شده

جدول ۴–۴– مقایسه مقادیر معیار های captioning سه دوره برتر آموزش ADAPT با زیرنویس های اضافه با روش پایه ADAPT در بخش تصمیم گیری

	BLEU-٤	METEOR	ROUGE-L	CIDEr-D	SPICE	زمان آموزش (دقیقه)
آموزش دوره ۶	**۳۵**	۲۹.۲	۶۱.۶	۲۳۷.۶	**۶۰**	**۶۲۱**
آموزش دوره ۱۶	۳۳.۱	۲۹.۲	۶۱.۶	۲۱۷.۴	۵۷.۶	۱۵۳۶
آموزش دوره ۲۶	۳۳.۸	۲۹.۳	۶۱.۹	۲۳۴.۷	۵۹.۶	۲۸۷۵
روش پایه ADAPT	۳۴.۶	**۳۰.۶**	**۶۲.۷**	**۲۴۶.۷**	۵۹.۷	۷۹۷

یادداشت تجربی (توقف زودهنگامِ مسیر بهینه‌سازی): یک تلاش تکمیلی با تغییر جزئی ابرپارامترها (تنها برای ۱ دوره) نشان داد که هرچند زیان آموزشی کمتر می‌شود، معیارهای اعتبارسنجی کاهش می‌یابند؛ به‌همین دلیل، این مسیر پیگیری نشد تا از بیش‌برازشِ موضعی جلوگیری شود. (گزارش معیارهای نهایی تغییر نکرده و مبنای ارزیابی مطابق نسخه پایه باقی مانده است.)

۴–۵– دستور و واریانت‌ها

۴–۵–۱– متن کامل دستور ثابت و منطق طراحی آن

برای تضمین شرایط ارزیابی یکنواخت، از یک دستور ثابت برای همه مدل‌ها استفاده شد. این دستور نقش مدل، بستر اطلاعاتی، وظیفه و قیود خروجی را به‌صراحت تعیین می‌کند و با چند نمونه کوتاه (few-shot) الگوی پاسخ مطلوب را نشان می‌دهد. محدودیت «۱یا۲ جمله کوتاه»، پرهیز از

عبارات زائد، و فرضِ درستیِ تصمیم سامانه تصمیم‌گیر در متن دستور تصریح شده است. متن نهاییِ دستور که در همه فراخوانی‌ها به‌کار رفته، در تصویر ۴–۴ و ترجمه آن در تصویر ۴–۵ نشان داده شده است:

You are a description system in driving scenarios.

Your task is to capture an image of a driving scenario, the decision, and the area of interest of the decision-making system. Then explain why the decision-making system made that decision.

Your input is {{frames_count}} frames of a camera in front of the vehicle, including the area of interest marked with a red circle.

Explain simply, briefly, and without explanation in only ۱ or ۲ short sentences. Avoid additional sentences, requests for more information, etc., do not explain context or given input information. You do not have access to the full information that the decision-making system does. So you consider that decision correct. Under no circumstances should you write more than ۲ sentences. Avoid writing sentences like 'Based on the provided image ...' or 'In conclusion ...'.

Here are some examples:

* since the truck in front wasn't moving fast.

* because the light is green.

* due to a vehicle in front of it moving slowly as it turns right.

تصویر ۴–۴–متن کامل دستور استفاده شده جهت تولید توضیح (خروجی مدل توضیح دهنده)

شما یک سامانه توصیف‌گر در سناریوهای رانندگی هستید.

وظیفه شما دریافت تصویر یک سناریوی رانندگی، تصمیم اتخاذشده، و ناحیه موردعلاقه سامانه تصمیم‌گیر است؛ سپس توضیح دهید چرا سامانه تصمیم‌گیر آن تصمیم را اتخاذ کرده است.

ورودی شما {{تعداد فریم}} فریم از دوربین جلوی وسیله نقلیه است که در آن ناحیه موردعلاقه با یک دایره قرمز مشخص شده است.

به‌صورت ساده و کوتاه و بدون حاشیه‌پردازی، تنها در ۱ یا ۲ جمله کوتاه توضیح دهید. از افزودن جملات اضافه، درخواست اطلاعات بیشتر و مانند آن خودداری کنید؛ زمینه کلی یا اطلاعات داده‌شده ورودی را شرح ندهید.

شما به تمامی اطلاعاتی که سامانه تصمیم‌گیر در اختیار دارد دسترسی ندارید؛ بنابراین تصمیم اتخاذشده را صحیح فرض کنید.

به‌هیچ‌وجه بیش از ۲ جمله ننویسید.

از نوشتن عبارت‌هایی مانند «بر اساس تصویر ارائه‌شده...» یا «در نتیجه...» خودداری کنید.

نمونه‌هایی از قالب مطلوب:

*چون کامیون جلویی سریع حرکت نمی‌کرد.

*چون چراغ سبز است.

* به‌دلیل خودرویی در جلو که هنگام پیچیدن به راست آهسته حرکت می‌کند.

تصویر ۴–۵– ترجمه فارسی متن دستور استفاده شده جهت تولید توضیح

دستور تصویر ۴–۴ با هدف: (۱) حذف واریانس ناشی از نحوه راهنمایی، (۲) هدایت مدل به پاسخ‌های علّیِ موجز و مستقیماً مرتبط با نشانه‌های بصری (دایره قرمز) و «اکشن» ADAPT، و (۳) تضمین مقایسه‌پذیری میان مدل‌ها تدوین شده است.

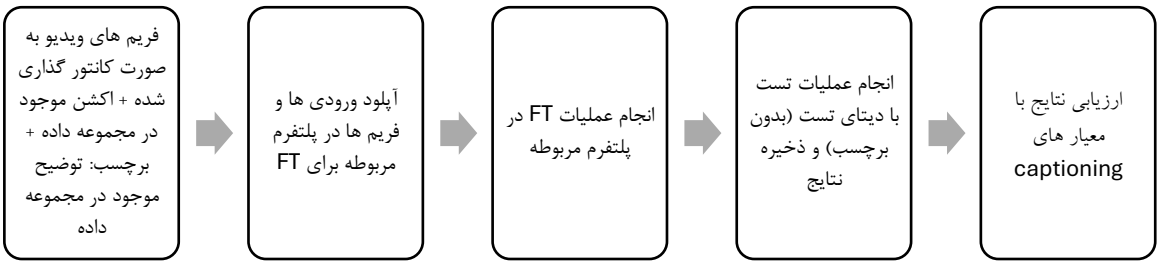
۴–۵–۲– تنظیمات تولید متن

- الگوی نمونه‌گیری: temperature = ۰٫۵؛ پارامترهای top-k و top-p در مقادیر پیش‌فرض سرویس نگه داشته شده‌اند.
- تعداد خروجی در هر درخواست: یک پاسخ.(n=۱)
- کنترل طول/قالب پاسخ: از طریق قید زبانی در دستور (حداکثر ۱یا۲ جمله کوتاه) اعمال شده است؛ از محدودیت‌های عددی حداکثر تعداد توکن برابر با ۱۰۲۴ بوده و stop-sequence اختصاصی استفاده نشده است.
- سیاست پایداری اجرا: در صورت بروز خطا، برای هر نمونه حداکثر ۴ تلاش انجام می‌شود؛ در صورت تداوم خطا، اجرای همان مورد متوقف و ثبت می‌گردد.
- ارسال ورودی چندوجهی:
  - برایGemini ، تصاویر کانتورگذاری‌شده (JPEG) به‌صورت مستقیم با ارجاع ذخیره‌سازی ابری Google Cloud ارسال شده‌اند.
  - برایGPT ، تصاویر به‌صورت Base٦٤ (با لینک میزبانی بیرونی) در بدنه درخواست قرار داده شده‌اند.

یادآوری: بسته ورودی شامل $n \in$ Error! Bookmark not defined.فریم کانتورگذاری‌شده + اکشن ADAPT است (رجوع به ۳–۴–۴).

۴–۵–۳– آزمایش‌های Fine-Tuning: پروتکل و پیاده‌سازی

در راستای بهبود مدل توضیح دهنده، در این پژوهش Fine-Tuning انجام شده است. هدف FT، انطباق بهتر مدل‌های زبانی با «توضیح اقدام» در سناریوهای رانندگیِ BDD-X بوده است.



تصویر ۴–۶– فلوچارت مراحل انجام FT

- مدل‌های پایه FT:
  - Gemini ۲٫۵ Pro: چندوجهی (متن+تصویر)
  - GPT-٤o: چندوجهی (متن+تصویر)
- داده و قالب ورودیِ آموزش FT:
  - هدف/برچسب: جمله Justification (توضیح انسانیِ BDD-X) برای همان بازه.
  - ساخت نمونه آموزشی: برای هر نمونه، ۴ فریم نخست کلیپ (همراه با کانتور قرمزِ توجه) + اکشن ADAPT به‌عنوان ورودی، و Justification انسانی به‌عنوان خروجی هدف استفاده شده است.
  - اندازه مجموعه‌ها: ۲۴۳۱ نمونه برای آموزش و ۵۸۵ نمونه برای اعتبارسنجی.
- ارسال تصاویر در FT:
  - برای Gemini، تصاویر روی Google Cloud میزبانی و در جریان FT ارجاع داده شده‌اند.

- برایGPT-٤o ، تصاویر روی سرویس میزبانی خارجی (مانند Hugging Face) بارگذاری و لینک آن در فرایند FT استفاده شده است.

- ابرپارامترها و معیار انتخاب:
  - ابرپارامترها: روی هر دو پلتفرم در حالت خودکار (پیش‌فرض سرویس) تنظیم شده‌اند.
  - معیار انتخاب بهترین مدل: کمینه زیان روی مجموعه اعتبارسنجی؛ در فرایند FT امکان محاسبه مستقیم معیارهای ارزیابی (BLEU-٤/ROUGE-L/CIDEr-D/SPICE) در همان پلتفرم وجود نداشته است.
  - زیرساخت اجرا: مدیریت منابع/زیرساخت در این پلتفرم‌ها غیرقابل‌تنظیم/نامشهود است.
- تنظیمات استنتاج پس ازFT :
- برای هر دو مدلFT شده، تنظیمات استنتاج مطابق ۲–۵–۴ است: temperature=۰٫۵، top-k/top-pپیش‌فرض سرویس، یک خروجی در هر درخواست، و سیاست ۴ تلاش ← توقف در صورت خطا.
- نکته هزینه/اندازه تصویر (OpenAI/GPT): جهت کاهش هزینه، پلتفرم در زمان اجرا اندازه تصویر را به ۵۱۲×۵۱۲ تقلیل داده است؛ این رفتار در ثبت آزمایش‌ها لحاظ شده است.

جمع‌بندی: این بخش، دستور واحد، سیاست‌های تولید متن، و پروتکل Fine-Tuning چندوجهی برای Gemini ۲٫۵ Pro و GPT-٤o را به‌صورت بازتولیدپذیر مستندسازی کرد. نتایج کمی/کیفی و مقایسه نسخه‌های FT–شده با نسخه‌های پایه و نیز با ADAPT در بخش‌های ۷–۴ تا ۱۰–۴ گزارش خواهد شد.

۴–۶– پروتکل ارزیابی

این بخش نحوه پیاده‌سازی ابزار ارزیابی، سناریوهای مرجع، و رویه اجرای آزمون‌ها را به‌صورت بازتولیدپذیر تشریح می‌کند. تعریف نظری معیارها در فصل سوم آمده است؛ در این‌جا صرفاً جزئیات اجرایی و تنظیمات گزارش می‌شود.

۴–۶–۱– معیار های ارزیابی

برای ارزیابی خودکار کیفیت توضیحات تولیدشده توسط مدل، از معیارهای متداول حوزه‌ی توصیف تصویر شامل BLEU-٤، METEOR، ROUGE-L، CIDEr-D و SPICE استفاده شده است. هر یک از این معیارها به صورت مختصر در ادامه معرفی می‌شوند:

- **BLEU-٤**

این معیار یکی از اولین شاخص‌های خودکار ارزیابی شباهت متنی است[۲۴]. BLEU ابتدا برای سنجش کیفیت ترجمه‌ی ماشینی پیشنهاد شد و برابر با میانگین هندسی دقتn-gram های مشترک بین جمله تولیدشده و جملات مرجع (تا چهارکلمه‌ای) است که یک جریمه‌ی طول جمله نیز برای ترجمه‌های بسیار کوتاه در نظر می‌گیرد [۲۴].

- **ROUGE-L**

معیار ROUGE به‌طور اولیه برای ارزیابی خلاصه‌سازی متن معرفی گردید و مقایسه‌ی n-gramها، دنباله‌های واژگان و جفت‌واژه‌های مشترک بین متن تولیدی و متن مرجع را انجام می‌دهد[۲۵]. گونه‌ی ROUGE-L در واقع طولانی‌ترین زیردنباله‌ی مشترک[۱۰۹] بین جمله تولیدشده و جمله مرجع را اندازه‌گیری می‌کند و به دلیل ماهیت مبتنی برRecall ، به جملات طولانی امتیاز بالاتری می‌دهد [۲۵].

- **METEOR**

[۱۰۹] Longest Common Subsequence

معیار METEOR نیز از حوزه‌ی ترجمه‌ی ماشینی گرفته شده است [۲۶]. این معیار بر اساس میانگین هارمونیک دقت و recall در تطابق‌های تک‌واژه[۱۱۰] بین زیرنویس تولیدی و زیرنویس مرجع تعریف می‌شود و علاوه بر تطابق دقیق واژگان، مترادف‌ها و عبارات بازنویسی‌شده را نیز برای تطابق در نظر می‌گیرد [۷۷]. به بیان دیگر METEOR برخی ضعف‌های BLEU را با در نظر گرفتن ارزیابی مبتنی بر recall و امکان تطابق معنایی کلمات برطرف می‌کند.

- **CIDEr-D**

معیار CIDEr به طور خاص برای ارزیابی توصیف تصویر معرفی شده است [۲۳] و میزان «توافق» یا اشتراک محتوایی بین زیرنویس تولیدشده و زیرنویس‌های مرجع را می‌سنجد [۲۳]. در این روش پس از انجام stemming، هر جمله به مجموعه‌ای از n-gram های ۱ تا ۴ کلمه‌ای تبدیل می‌شود و سپس با محاسبه‌ی هم‌وقوعی این n-gram ها بین جمله مدل و جملات مرجع، یک شباهت مبتنی بر کسینوس محاسبه می‌گردد. برای افزایش دقت، CIDEr-D از وزن‌دهی tf-idf روی n-gram ها بهره می‌گیرد تا به عبارات رایج وزن کمتر و به عبارات خاص‌تر وزن بیشتری اختصاص دهد [۷۷]. به این ترتیب نمره‌ی CIDEr-D منعکس‌کننده‌ی میزان تطابق توصیف مدل با توصیفات مرجع بر پایه‌ی عبارات محتوایی مهم است.

- **SPICE**

معیار SPICE برای ارزیابی شباهت معنایی زیرنویس‌ها مطرح شده است [۲۲]. در این روش هر زیرنویس به یک گراف صحنه (شامل اشیاء، روابط و ویژگی‌ها) تبدیل می‌شود و سپس میزان تطابق عناصر گراف (موجودیت‌ها و روابط) بین زیرنویس تولیدشده و زیرنویس‌های مرجع سنجیده می‌شود. نمره‌ی SPICE برابر با مقدار F_۱ حاصل از مطابقت این روابط و اشیاء بین توضیح مدل

---

[۱۱۰] Unigram

و مراجع انسانی است [۲۲]. همچنین SPICE برای تطابق معنایی دقیق‌تر از مترادف‌یابی با وردنت (مشابه روش METEOR) استفاده می‌کند تا اختلافات جزئی واژگانی را نادیده بگیرد [۷۷].

۴–۶–۲– پیاده‌سازی معیارها و نسخه ابزارها

- ابزار ارزیابی: بسته پایتونی aac_metrics (نسخه ۰.۶.۰) برای محاسبه BLEU-٤، METEOR، ROUGE-L، CIDEr-Dو SPICE به‌کار رفته است.
- سطح محاسبه: همه معیارها در سطح corpus-level گزارش شده‌اند.
- تنظیمات معیارها:
  - BLEU-٤: هموارسازی[۱۱۱] پیش‌فرض بسته (method ۳).
  - ROUGE-L: خروجی F۱ بر مبنای LCS.
  - CIDEr: گونه CIDEr-D (پیش‌فرض ابزار).
  - SPICE: پیکربندی پیش‌فرض.
- پیش‌پردازش متنی و توکن‌سازی: پیش‌پردازش استاندارد aac_metrics شامل کوچک‌سازی حروف[۱۱۲]، حذف علائم نگارشی و نُرمال‌سازی فاصله‌ها اعمال شده است. توکن‌سازی برای اغلب معیارها PTB و برای ROUGE-L مبتنی بر whitespace است؛ SPICE از نشانه ساز[۱۱۳] خاصی استفاده نمی‌کند (مطابق پیش‌فرض‌های ابزار).

۴–۶–۳– دو سناریوی مرجع: تک‌رفرنس و چندرفرنس

ارزیابی مدل‌ها در این پژوهش به دو شیوه‌ی متفاوت انجام شده است:

---

[۱۱۱] smoothing

[۱۱۲] Lowercase

[۱۱۳] Tokenizer

- ارزیابی با یک مرجع: در روش اول، خروجی هر مدل تنها با یک توضیح مرجع انسانی موجود در مجموعه داده BDD-X مقایسه می‌شود. به عبارت دیگر، از آنجا که هر ویدئو در مجموعه‌ی BDD-Xفقط یک توضیح متنی توسط annotator انسانی دارد، معیارهای فوق براساس همان تک‌جمله مرجع محاسبه می‌گردند.
- ارزیابی با مراجع افزوده: در روش دوم، برای ارتقاء کیفیت ارزیابی خودکار، علاوه بر توضیح مرجع اصلیBDD-X ، چهار زیرنویس اضافه که به کمک یک مدل زبان بزرگ برای هر ویدئو تولید شده‌اند نیز به مجموعه‌ی مراجع افزوده می‌شوند. بدین ترتیب هر صحنه مجموعاً پنج توضیح مرجع خواهد داشت که ارزیابی را جامع‌تر می‌کند، زیرا امتیازدهی معیارها بر پایه‌ی تطابق با چندین توصیف ممکن صورت می‌گیرد و در نتیجه با قضاوت انسانی همبستگی بیشتری می‌یابد .طبق مطالعات، بسیاری از معیارهای خودکار در صورت استفاده از مراجع متعدد، همبستگی بالاتری با ارزیابی انسانی نشان می‌دهند [۷۸]؛ بنابراین افزودن زیرنویس‌های متنوع به‌عنوان مرجع می‌تواند سنجش کیفیت توضیحات را منصفانه‌تر و نزدیک‌تر به معیار انسانی کند. در متن دستور تولید رفرنس‌ها، اجتناب از تکرار تصریح شده است (نمونه دستور در پیوست/ضمیمه فصل آورده می‌شود).

دامنه ارزیابی برای مدل‌ها:

- ماژول توضیح‌گر :برای هر مدل، ارزیابی در پنج سطح ورودی $n \in \{1,4,8,12,16\}$ (تعداد فریم‌های نخست با کانتور توجه) انجام می‌شود.
- روش پایه ADAPT: خروجی متنی ADAPT بر مجموعه آزمون واحد و تحت همین پروتکل ارزیابی می‌شود. افزون بر حالت ۳۲ فریم (تنظیم بومی مدل)، در صورت نیاز به تحلیل حساسیت، واریانت‌های کاهش‌یافته نیز گزارش می‌شوند؛ حالت ۱۲ فریم برای ADAPT قابل محاسبه نبوده است.

۴–۶–۴– رویه اجرا، تکرار آزمایش‌ها و نکات آماری

- مجموعه آزمون و هم‌ترازی: همه مدل‌ها (ADAPT و LLM ها) بر همان مجموعه آزمون و نمونه‌های یکسان ارزیابی شده‌اند؛ تنها تفاوت، تعداد فریم‌های ورودی در سناریوهای n-fraction است.
- سیاست اجرا و پایداری: برای هر نمونه، در صورت بروز خطا در فراخوانیLLM ، تا ۴ تلاش انجام می‌شود. مواردی که تا پایان اجرا خروجی معتبر دریافت نکنند، از محاسبه نمره آن مدل حذف و تعداد نمونه‌های مؤثر در گزارش نتایج ذکر می‌شود (به‌منظور شفافیت مخرج محاسبه).
- تکرار آزمایش‌ها: با توجه به ابعاد داده و هزینه ی تکرار آزمایش روی پلتفرم های مدل های زبانی، ارزیابی برای هر پیکربندی به‌صورت single-run انجام شده است (بدون میانگین‌گیری بین اجراها).
- تنظیمات تولید متن: در همه ارزیابی‌ها temperature = ۰٫۵ و سایر پارامترهای نمونه‌گیری (top-k/top-p) در مقادیر پیش‌فرض سرویس قرار داشته‌اند؛ در هر درخواست یک خروجی تولید شده است.(n=۱)
- آزمون‌های آماری: آزمون معناداری آماری (bootstrap / randomization test و ...) به‌کار نرفته است؛ در صورت نیاز، می‌توان آن را به‌عنوان کار تکمیلی پیشنهاد کرد.
- گزارش نتایج: برای هر مدل و هر سطح n، امتیازهای BLEU-٤ ، METEOR، ROUGE-L، CIDEr-D و SPICE در سطح corpus-level گزارش می‌شود. در کنار هر جدول، تعداد نمونه‌های مؤثر (پس از حذف احتمالی نمونه‌های ناموفق) درج می‌گردد تا مقایسه‌ها منصفانه باشد.

جمع‌بندی: بدین ترتیب، ارزیابی روی مجموعه آزمون واحد، با دو سناریوی مرجع (تک‌رفرنس/چندرفرنس) و تنظیمات یکسان نمونه‌گیری زبانی انجام شده و جنبه‌های اجرایی (ابزار، نسخه، پیش‌پردازش، خطای اجرا و شمار نمونه‌های مؤثر) به‌طور صریح مستندسازی شده است.

۴–۷– نتایج روش پایه (ADAPT)

در این بخش، عملکرد ADAPT به‌عنوان روش پایه تولید توضیح بر روی مجموعه آزمون و در دو سناریوی مرجع «تک‌رفرنس» و «چندرفرنس» گزارش می‌شود. تمام امتیازها در سطح corpus-level و با پروتکل ۶-۴ محاسبه شده‌اند؛ هیچ نمونه‌ای حذف نشده و مدل همان پیکربندی آموزشی ۱-۴-۴ را دارد.

۴–۷–۱– نمرات در سناریوهای تک مرجع و چند مرجع

جدول ۴–۵– نتایج سناریوی تک مرجع (فقط Justification انسانیِ BDD-X)

ورودی (n) فریم	BLEU-٤	CIDEr-D	METEOR	ROUGE-L	SPICE
۱ frame	۲.۷۹	۳۶.۲۳	۹.۹۹	۲۳.۴۴	۱۲.۷۸
۴ frame	۷.۷۵	۸۲.۳۴	۱۲.۸۹	۲۹.۷۶	۲۱.۸۶
۸ frame	۸.۱۹	۸۷.۴۹	۱۳.۴۷	۳۱.۳۳	۲۳.۷۰
۱۲ frame	۴.۵۷	۴۶.۹۷	۱۱.۵۵	۲۴.۸۲	۱۴.۶۳
۱۶ frame	۸.۷۰	۹۱.۵۰	۱۳.۸۰	۳۲.۰۰	۲۴.۸۰
۳۲ frame	**۱۱.۴۰**	**۱۰۲.۶۰**	**۱۵.۲۰**	**۳۲.۰۰**	**۲۵.۶۰**

جدول ۴–۶– نتایج سناریوی چند مرجع (BDD-X + چهار بازنویسی LLM)

ورودی (n) فریم	BLEU-٤	CIDEr-D	METEOR	ROUGE-L	SPICE
۱ frame	۱۰.۴۵	۴۶.۰۵	۱۵.۹۵	۳۷.۴۹	۷.۶۷
۴ frame	۱۶.۲۳	۸۱.۹۳	۱۹.۷۶	۴۴.۱۸	۱۶.۰۹
۸ frame	۱۶.۶۰	۸۶.۰۶	۲۰.۵۸	۴۵.۶۶	۱۷.۲۲
۱۲ frame	۱۲.۱۹	۵۳.۴۹	۱۶.۹۳	۳۹.۵۹	۹.۲۹

۱۶ frame	۱۷.۶۷	۹۱.۰۳	۲۱.۰۹	۴۷.۰۸	۱۸.۱۵
۳۲ frame	**۱۸.۴۵**	**۹۵.۱۳**	**۲۱.۲۰**	**۴۷.۰۵**	**۱۸.۲۲**

یادآوری: در ADAPT ورودی بومی ۳۲ فریمی است؛ بااین‌حال به‌منظور تحلیل حساسیت، خروجی‌های ADAPT برای زیرمجموعه‌های زمانی نیز محاسبه شده‌اند.

۴–۷–۲– بحث کوتاه درباره روش پایه و همخوانی با مرجع

۱. روند وابسته به اطلاعات زمانی: افزایش تعداد فریم‌های ورودی به‌طور کلی با بهبود تدریجی امتیازها همراه است (به‌ویژه در BLEU-۴، METEOR و ROUGE-L). این الگو، انتظار روش‌شناختی را تأیید می‌کند: دسترسی بیشتر به متنِ تصویریِ صحنه (اطلاعات زمانی/زمینه‌ای) به توضیح‌های نزدیک‌تر به مرجع انسانی منتهی می‌شود. افتِ موضعی در عملکرد با ۱۲ فریم ورودی، ریشه در معماری مدل پایه (ADAPT) دارد. همانطور که قبلا گفته شد، این مدل از یک ترنسفورمر ویدیویی به عنوان ستون فقرات بهره می‌برد که دارای ساختاری سلسله‌مراتبی و چندمرحله‌ای است. در این نوع معماری‌ها، عملیات نمونه‌برداری و ادغام ویژگی‌ها در بعد زمان، زمانی به بهینه‌ترین شکل خود انجام می‌شود که تعداد فریم‌های ورودی توانی از عدد ۲ باشد (مانند ۸، ۱۶ یا ۳۲ فریم که مدل بر روی آن آموزش دیده است). عدد ۱۲، که توانی از ۲ نیست، این تقسیم‌بندی‌های متقارن و متوالی را در لایه‌های عمیق‌تر مدل مختل می‌کند. این ناهمگونی معماری می‌تواند منجر به استخراج ویژگی‌های کمتر بهینه و نمایش ضعیف‌تری از محتوای زمانی کلیپ شود و در نتیجه، عملکرد مدل را در مقایسه با ورودی‌های سازگارتر مانند ۸ یا ۱۶ فریم، کاهش دهد.

۲. اثر افزودن مراجع (چندرفرنس): در سطح ۳۲ فریم، گذار از تک‌رفرنس به چندرفرنس موجب افزایش محسوس METEOR (۱۵.۲۰→۲۱.۲۰) و ROUGE-L (۳۲.۰۰→۴۷.۰۵) و

همین‌طور BLEU-٤ (۱۸.۴۵←۱۱.۴۰) شده است؛ یعنی با در دست داشتن چند بیان هم‌معنا از یک حقیقت، هم‌پوشانی بین خروجی ADAPT و مراجع تقویت می‌شود. در مقابل، CIDEr-D و SPICE در چندرفرنس نسبت به تک‌رفرنس کمی کمتر گزارش شده‌اند. این رفتار، هرچند خلاف انتظار کلاسیک است، اما می‌تواند ناشی از توزیع واژگانی و ساختاری متفاوت در بازنویسی‌های LLM باشد که وزن‌های tf-idf (در CIDEr-D) و استخراج روابط صحنه‌ای (در SPICE) را به‌شکل نامتوازن تحت تأثیر قرار داده است؛ به بیان دیگر، بازنویسی‌ها اگرچه شباهت سطح‌واژگانی/نحوی را افزایش می‌دهند، اما لزوماً شبکه مفهومیِ اشیاء–روابط را به همان میزان تقویت نمی‌کنند. این مشاهده در فصل تحلیل (۴–۹ و ۴–۱۰) با نمونه‌های کیفی بازبینی می‌شود.

۳. همخوانی با مرجع انسانی: در هر دو سناریو، بهترین وضعیت ADAPT در ورودی ۳۲ فریم حاصل می‌شود؛ به‌ویژه رشد METEOR و ROUGE-L در چندرفرنس، همخوانی مناسبی با انتظار ما از تعبیرپذیری زبانی نشان می‌دهد. بنابراین، این نتایج، تصویری شفاف و معیارمند از روش پایه فراهم می‌کند تا در بخش‌های بعدی، عملکرد مدل‌های زبانی (پایه و FT–شده) به‌صورت منصفانه با آن سنجیده شود.

۴–۸– نتایج مدل‌های زبانی و مقایسه

در این بخش، عملکرد مدل‌های زبانی ارزیابی‌شده بر روی مجموعه آزمون، تحت دو سناریوی تک مرجع و چند مرجع (مطابق ۶-۴) گزارش و تحلیل می‌شود. برای اختصار و مقایسه منصفانه، یک «نمایه ثابت» از نتایج هر مدل در n=۱۶ فریم ارائه می‌گردد و سپس نکات روندیِ هر مدل بیان می‌شود.

جدول ۴–۷– نتایج نمایه ثابت در n=۱۶ – سناریوی تک مرجع (BDD-X)

مدل	BLEU-٤	CIDEr-D	METEOR	ROUGE-L(F۱)	SPICE
Gemini ۲,۵ Pro	۷.۰۱	۵۸.۴۷	۱۷.۷۲	۳۲.۷۷	۲۴.۳۵
Gemini ۲,۵ Flash	۵.۶۴	۵۱.۲۸	۱۶.۸۴	۳۱.۷۵	۲۰.۰۶
Gemini ۲,۵ Flash-Lite	۶.۵۹	۵۲.۵۶	۱۵.۳۳	۲۹.۸۱	۲۰.۵۱
GPT-۵ Mini	۲.۰۳	۲۵.۴۸	۱۴.۰۱	۲۳.۲۹	۱۱.۱۷

جدول ۴–۸– نتایج نمایه ثابت در n=۱۶ – سناریوی چند مرجع (BDD-X + چهار بازنویسی LLM)

مدل	BLEU-٤	CIDEr-D	METEOR	ROUGE-L(F۱)	SPICE
Gemini ۲,۵ Pro	۱۶.۰۸	۴۴.۰۹	۲۳.۸۱	۴۸.۱۵	۲۶.۴۸
Gemini ۲,۵ Flash	۱۴.۱۰	۳۷.۶۶	۲۳.۱۸	۴۷.۷۰	۲۳.۵۴
Gemini ۲,۵ Flash-Lite	۱۵.۴۳	۴۰.۶۱	۲۱.۳۰	۴۵.۴۶	۲۱.۲۱
GPT-۵ Mini	۶.۳۱	۲۱.۶۰	۱۹.۸۰	۳۶.۵۳	۱۴.۸۷

تذکر: نتایج بالا در سطح corpus-level و با aac_metrics نسخه ۰.۶.۰ محاسبه شده‌اند. هیچ نمونه‌ای حذف نشده و برای LLM ها، تصاویر کانتورگذاری‌شده به‌همراه «اکشن» ADAPT مطابق ۴–۴–۲ و ۴–۴–۳ ارسال شده است.

۴–۸–۱– مدل Gemini ۲,۵ Pro

- تک مرجع: بهترین‌های هر معیار در بازه n=۱..۱۶:

BLEU-٤ در ۱۲ فریم (≈۷.۰۸)، CIDEr-D در ۱۲ فریم (≈۶۲.۲۰)، METEOR در ۱۶ فریم (≈۱۷.۷۲)، ROUGE-L در ۱۲ فریم (≈۳۲.۸۲)، و SPICE در ۱۶ فریم (≈۲۴.۳۵). روند کلی نشان می‌دهد افزودن فریم‌ها به‌ویژه برای METEOR/ROUGE-L/SPICE سودمند است.

- چند مرجع: BLEU-٤، METEOR و ROUGE-L در ۱۶ فریم بیشینه می‌شوند (به‌ترتیب ≈۱۶.۰۸، ≈۲۳.۸۱، ≈۴۸.۱۵)، SPICE در ۱۲ فریم به اوج می‌رسد (≈۲۶.۶۳)، و CIDEr-D غیر یکنواخت است و در ۱ فریم بیشینه می‌شود (≈۵۹.۲۰). این رفتار CIDEr-D با وزن‌دهی tf-idf و حساسیت به طول/واژگان قابل توجیه است.

۴–۸–۲– مدل Gemini ۲٫۵ Flash

- تک مرجع: بیشینه BLEU-٤/ROUGE-L/METEOR/SPICE در ۱۶ فریم (≈۰۶.۲۰/۸۴.۱۶/۷۵.۳۱/۶۴.۵) و CIDEr-D در ۴ فریم (≈۴۰.۵۴) به‌دست آمده است؛ با افزایش فریم‌ها، شاخص‌های معنایی (METEOR/ROUGE-L/SPICE) بهبود تدریجی دارند.
- چند مرجع: BLEU-٤ در ۴ فریم بیشینه (≈۹۰.۱۴) اما ROUGE-L و SPICE در ۱۶ فریم بالاترین‌اند (≈۴۷.۷۰ و ≈۵۴.۲۳). METEOR نیز در ۱۶ فریم به ≈۱۸.۲۳ می‌رسد. CIDEr-D در ۱ فریم بیشینه (≈۱۶.۵۳) و سپس کاهش نسبی دارد.

۴–۸–۳– مدل Gemini ۲٫۵ Flash-Lite

- تک مرجع: بهترین BLEU-٤ و CIDEr-D در ۸ فریم (≈۶.۸۸ و ≈۶۰.۳۸) رخ می‌دهد؛ اما METEOR، ROUGE-L، SPICE با افزایش فریم تا ۱۶ بهبود می‌یابند (≈۱۵.۳۳، ≈۲۹.۸۱، ≈۲۰.۵۱).

- چند مرجع: BLEU-٤ و CIDEr-D در ۸ فریم بیشینه‌اند (≈۱۸.۰۷ و ≈۵۳.۷۹)، در حالی‌که ROUGE-L و SPICE در ۱۶ فریم بالاترند (≈۴۵.۴۶ و ≈۲۱.۲۱). METEOR نیز در ۱۶ فریم ≈۲۱.۳۰ گزارش می‌شود.

۴–۸–۴– مدل GPT-٥ Mini

- تک مرجع: مقادیر در سطح پایین‌تر و با روند افزایشی ملایم با افزایش فریم‌ها: BLEU-٤ در ۱۶ فریم ≈۲.۰۳، CIDEr-D در ۱۶ ≈۲۵.۴۸، METEOR در ۱۲ ≈۱۴.۰۵ (اوج)، ROUGE-L در ۱۶ ≈۲۳.۲۹ و SPICE در ۱۲ ≈۱۱.۴۰.
- چند مرجع: BLEU-۴، ROUGE-L، METEOR به‌ترتیب در ۱۶ فریم ≈۶.۳۱، ≈۳۶.۵۳، ≈۱۹.۷۹ بهینه می‌شوند؛ SPICE در ۱ فریم ≈۱۵.۵۰ بیشینه است؛ CIDEr-D در ۱۶ فریم ≈۲۱.۶۰ گزارش شده است.

۴–۸–۵– مقایسه تجمیعی با ADAPT در هر دو سناریو

- تک مرجع: در قیاس با روش پایه (۷–۴) ADAPT ، مدل‌های Gemini (به‌ویژه Pro) در شاخص‌های معنایی/عبارتی مانندMETEOR ، ROUGE-L و SPICE با افزایش فریم‌ها به سطح رقابتی نزدیک می‌شوند، هرچند در BLEU-٤ و به‌ویژه CIDEr-D عموماً پایین‌تر از روش پایه باقی می‌مانند.
- چند مرجع: با افزودن چهار مرجع هم‌معنا، شکافLLM ها با مرجع انسانی در-BLEU ۴/METEOR/ROUGE-L به‌وضوح کاهش می‌یابد؛ به‌طور خاص، Gemini ۲٫٥ Pro در n =۱۶ به سطح بالایی در ROUGE-L و/METEOR SPICE می‌رسد و در برخی شاخص‌ها از روش پایه عبور می‌کند، در حالی‌که CIDEr-D همچنان مزیت محسوس ADAPT را نشان می‌دهد. این الگو مؤید آن است کهLLM ها در بازنویسی‌های معنایی هم‌تراز با مرجع

(چندرفرنس) توانمندترند، اما در هم‌پوشانی n-gram وزندار (CIDEr-D) به سطح روش پایه تخصصی (ADAPT) نمی‌رسند.

- تأثیر تعداد فریم‌ها: برای LLM ها، افزایش n غالباً بهبود METEOR ، ROUGE-L، SPICE را در پی دارد، اما CIDEr-D الزاماً یکنواخت افزایش نمی‌یابد و حتی گاهی در ورودی‌های کوتاه‌تر بیشینه می‌شود. نتیجه‌ای که با سازوکار tf-idf و حساسیت به طول/واژگان همخوان است.
- یادداشت درباره GPT-٥: نتایج نهایی در این نسخه ارائه نشده است؛ بر اساس مشاهده اولیه نگارنده، GPT-٥ تنها اندکی از GPT-٥ Mini بهتر بوده و انتظار تغییر کیفی بزرگ در مقایسه بالا نمی‌رود.

پیش از بررسی نتایج FT در جمع‌بندی، Gemini ٢,٥ Pro در سناریوی چندرفرنس و با n =١۶ فریم بهترین توازن را میان معیارها ارائه می‌کند؛ Flash و Flash-Lite نیز نزدیک‌اند. روش پایه ADAPT در CIDEr-D (و گاهی BLEU-٤) برتری دارد، اما LLM ها در METEOR، ROUGE-L و SPICE رقابتی یا برترند. یافته‌ای که در ۹–۴ و ۱۰–۴ با مثال‌های بصری و تحلیل خطا تکمیل می‌شود.

۴–۹– مطالعه حذف مؤلفه‌ها و تحلیل حساسیت

در این بخش، اثر مؤلفه‌های اصلی خط لوله و نیز حساسیت نتایج نسبت به تغییرات ورودی/پیکربندی بررسی می‌شود. گزارش بر مجموعه آزمون و با همان پروتکل ۶–۴ انجام شده است؛ اعداد از فایل‌های نتایج ارسالی (اکسل‌های تک‌مرجع و چندمرجع) استخراج گردیده‌اند.

۴–۹–۱– اثر تعداد فریم‌های ورودی (n-fractions) بر معیارها

- گرایش کلی: با افزایش تعداد فریم‌های ورودی به LLM (از ۱ تا ۱۶ فریم نخست کلیپ)، معیارهای مبتنی بر هم‌پوشانی معنایی/عبارتی مانند METEOR و ROUGE-L عموماً افزایشی

هستند و در n =۱۶ به سقف خود می‌رسند؛ برای نمونه در Gemini ۲٫۵ Pro (چندمرجع) ROUGE-Lاز ۴۵.۳۰ در n=۱ به ۴۸.۱۵ در n =۱۶ و METEOR از ۲۲.۱۷ به ۲۳.۸۱ افزایش می‌یابد. در حالت تک‌مرجع نیز METEOR از ۱۴.۸۳ به ۱۷.۷۲ و ROUGE-L از ۲۷.۵۸ به ۳۲.۷۷ (۱۶ فریم) می‌رسد.

- رفتار نایکنواخت CIDEr-D/BLEU-٤ در برخی مدل‌ها: در چندمرجع، CIDEr-D گاهی در ورودی‌های کوتاه‌تر بیشینه می‌شود (مثلاً Gemini Pro در n =۱: ۵۹.۲۰، درحالی‌که در n =۱۶ : ۴۴.۰۹). این پدیده با وزن‌دهی tf-idf و حساسیت CIDEr به طول/واژگان سازگار است. BLEU-٤ نیز ممکن است با افزایش فریم‌ها همواره یکنواخت رشد نکند (نمونه‌های موضعی در Flash/Flash-Lite).
- روش پایه ADAPT: برای سنجش هم‌ترازی، ADAPT در ۳۲ فریم (تنظیم بومی) بهترین عملکرد خود را دارد؛ در چندمرجع ROUGE-L=۴۷.۰۵، METEOR=۲۱.۲۰ و BLEU-٤=۱۸.۴۵. برای n-fraction های کاهش‌یافته نیز روند کلی افزایشی تا ۱۶ فریم مشاهده می‌شود (با محدودیت محاسبه در ۱۲ فریم که پیش‌تر ذکر شد).

نتیجه این بخش: اطلاعات زمانی غنی‌تر (فریم‌های بیشتر) غالباً به توضیح‌های زبانی نزدیک‌تر به مرجع انسانی منجر می‌شود؛ بااین‌حال، به‌ویژه در CIDEr-D، بالاترین امتیاز لزوماً در بیشترین فریم حاصل نمی‌شود.

۴–۹–۲– اثر افزوده‌شدن مراجع LLM (چندمرجع در برابر تک‌مرجع)

افزودن چهار بازنویسی هم‌معنا به مرجع انسانی (چندمرجع) هم‌پوشانی سطح واژگانی/عبارتی را افزایش می‌دهد و معمولاً باعث رشد محسوس BLEU-٤، METEOR و ROUGE-L می‌شود؛ برای نمونه در Gemini ۲٫۵ Pro در n =۱۶ (مقایسه چندمرجع با تک‌مرجع):

- BLEU-٤: از ۱۶.۰۸ ← ۷.۰۱ (Δ=+۹٫۰۷)

- METEOR: از ۲۳.۸۱ ⟵ ۱۷.۷۲(Δ=+٦،۰۹)
- ROUGE-L: از ۴۸.۱۵ ⟵ ۳۲.۷۷ (Δ=+۱۵،۳۸)
- SPICE: از ۲۶.۴۸ ⟵ ۲۴.۳۵ (Δ=+۲،۱۳)

در مقابل، CIDEr-D می‌تواند کاهش یابد (برای همان مثال: ۴۴.۰۹ ⟵ ۵۸.۴۷، Δ =−۱۴.۳۸، که با ماهیت tf-idf و تفاوت توزیع واژگان در بازنویسی‌های LLM قابل توضیح است. همین الگو در ADAPT نیز دیده می‌شود: در ۳۲ فریم، BLEU-٤ از ۱۸.۴۵ ⟵ ۱۱.۴۰(Δ=+۷،۰۵)، METEOR از ۲۱.۲۰ ⟵ ۱۵.۲۰ (Δ=+٦،۰۰) و ROUGE-L از ۴۷.۰۵ ⟵ ۲۱.۰۰ (Δ=+۱۵،۰۵) افزایش یافته، درحالی‌که CIDEr-D از ۹۵.۱۳ ⟵ ۱۰۲.۶۰ (Δ=−۷،٤۷) و SPICE از ۱۸.۲۲ ⟵ ۲۵.۶۰ (Δ=−۷،۳۸) کاهش نشان می‌دهد.

نتیجه این بخش: چندمرجع، همبستگی با بیان‌های معتبرِ معادل را تقویت می‌کند و در اکثر شاخص‌ها (جز CIDEr-D) سودمند است؛ تفسیر نتایج بر اساس مجموعه معیارها و نه یک معیار منفرد ضروری است.

۴–۹–۳– اثر Fine-Tuning

قواعد مقایسه: هر مدلِ FT–شده با بهترین نمره نسخه پایه همان مدل در میان تمام n–ها (بر اساس جدول ۴–۱۰ و جدول ۴–۱۱) مقایسه می‌شود؛ نه با نسخه ۱۲ فریم. اعداد FT مربوط به پیکربندی ۱۲ فریم هستند، اما «مبنای مقایسه» بهترینِ پایه در کل n–هاست.

- Gemini ۲،۵ Pro - FT در ۱۲ فریم، مقایسه با بهترینِ پایه همان مدل (میان همه-n ها):
  - تک‌مرجع:

    BLEU-٤: ۴.۷۵+ (۱۱.۸۳ در برابر بهترین پایه ۷.۰۸)

    CIDEr-D: ۵۲.۵۳+ (۱۱۴.۷۳ در برابر ۶۲.۲۰)

    METEOR: ۱.۷۳+ (۱۹.۴۵ در برابر ۱۷.۷۲)

ROUGE-L: ۷.۷۵+ (۴۰.۵۷ در برابر ۳۲.۸۲)

SPICE: ۴.۱۱+ (۲۸.۴۷ در برابر ۲۴.۳۵)

- چندمرجع:

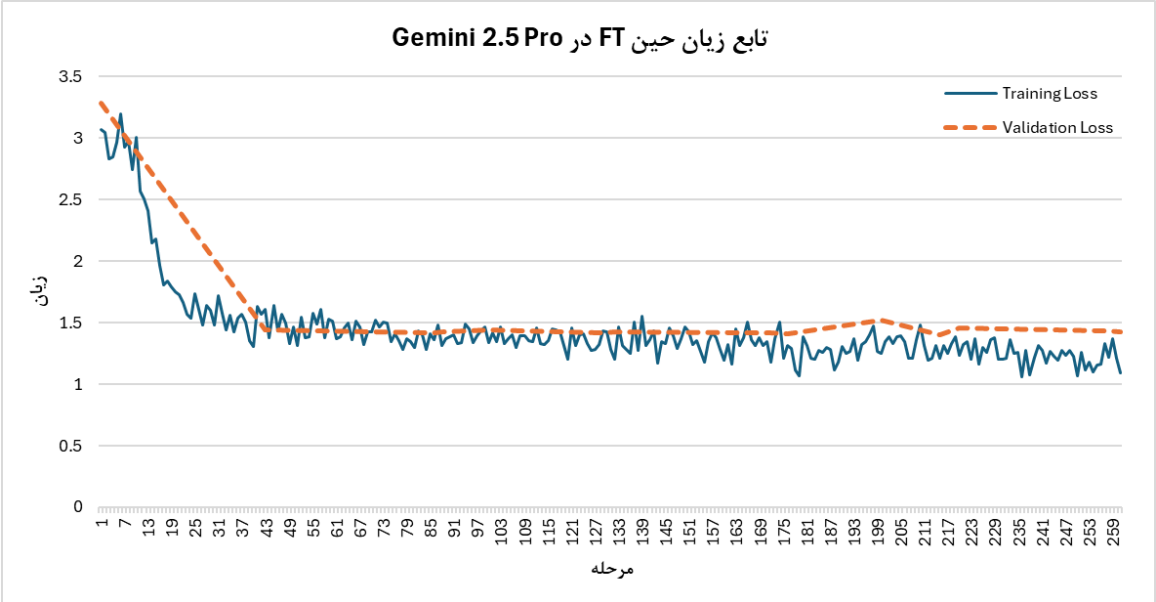
BLEU-٤: ۷.۸۲+ (۲۳.۹۰ در برابر بهترین پایه ۱۶.۰۸)

CIDEr-D: ۴۱.۵۵+ (۱۰۰.۷۵ در برابر ۵۹.۲۰)

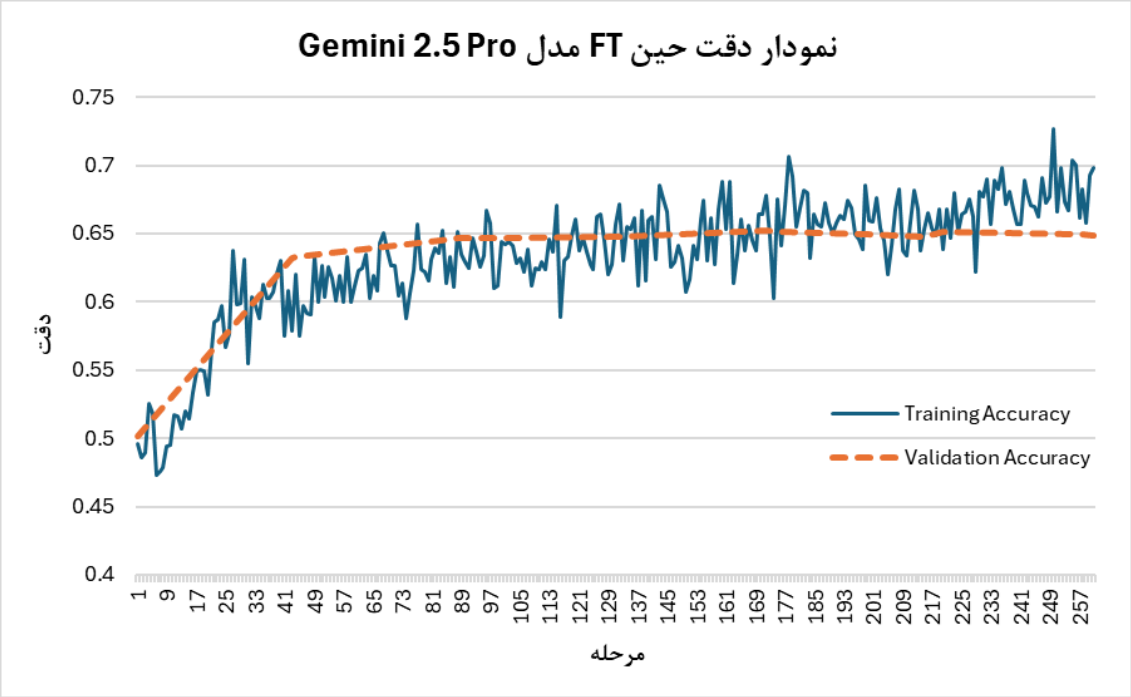
METEOR: ۰.۷۷+ (۲۴.۵۸ در برابر ۲۳.۸۱)

ROUGE-L: ۹.۴۴+ (۵۷.۵۹ در برابر ۴۸.۱۵)

SPICE: ۰.۸۲+ (۲۷.۴۶ در برابر ۲۶.۶۳)



تصویر ۴–۷– همگرایی تابع زیان در حین Fine Tuning مدل Gemini ۲٫۵ Pro



تصویر ۴–۸– نمودار دقت مدل در حین Fine Tuning در مدل Gemini ۲٫۵ Pro

فرآیند FT برای مدل Gemini ۲٫۵ Pro به دقت پایش شد تا از انتخاب بهینه مدل اطمینان حاصل شود. نمودارهای ۴–۷ و ۴–۸ روند همگرایی تابع زیان و دقت را در طول مراحل آموزش و اعتبارسنجی نشان می‌دهند. همان‌طور که مشاهده می‌شود، در حالی که معیارهای آموزش (زیان و دقت) به طور پیوسته بهبود می‌یابند، معیارهای اعتبارسنجی پس از یک بهبود اولیه، به یک سطح ثابت[۱۱۴] می‌رسند و بهبود معناداری از خود نشان نمی‌دهند.

این رفتار، که نشان‌دهنده توقف یادگیری ویژگی‌های قابل تعمیم است، ضرورت به‌کارگیری راهبرد توقف زودهنگام[۱۱۵] را برای جلوگیری از بیش‌برازش احتمالی و انتخاب بهترین نقطه عملکرد، آشکار می‌سازد. با تحلیل نمودارها، بهترین عملکرد مدل بر روی داده‌های

---

[۱۱۴] plateau

[۱۱۵] Early Stopping

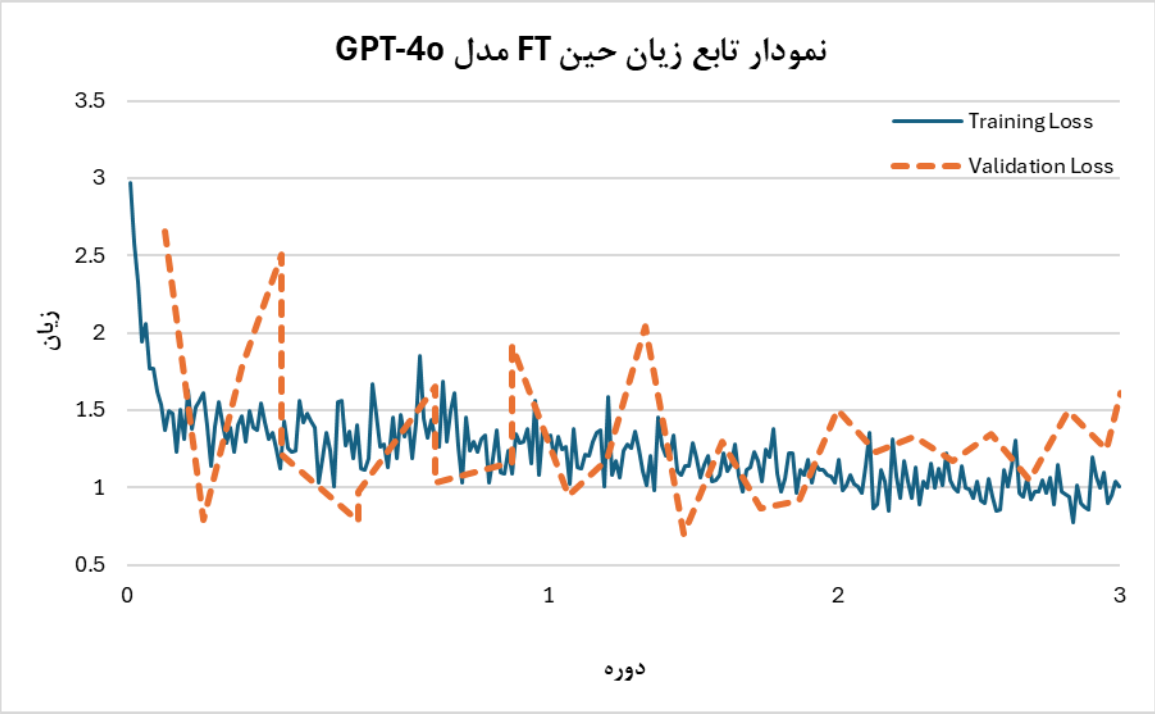
اعتبارسنجی در حوالی مرحله ۱۵۷ مشاهده شد، جایی که دقت اعتبارسنجی به پایداری رسیده و زیان اعتبارسنجی در یکی از پایین‌ترین سطوح خود قرار دارد.

بر همین اساس، مدلی که برای ارزیابی‌های نهایی در جداول بعدی به کار گرفته شده، از این نقطه بهینه ذخیره گردیده است تا اطمینان حاصل شود که مدل نهایی دارای بهترین قابلیت تعمیم‌پذیری ممکن است.

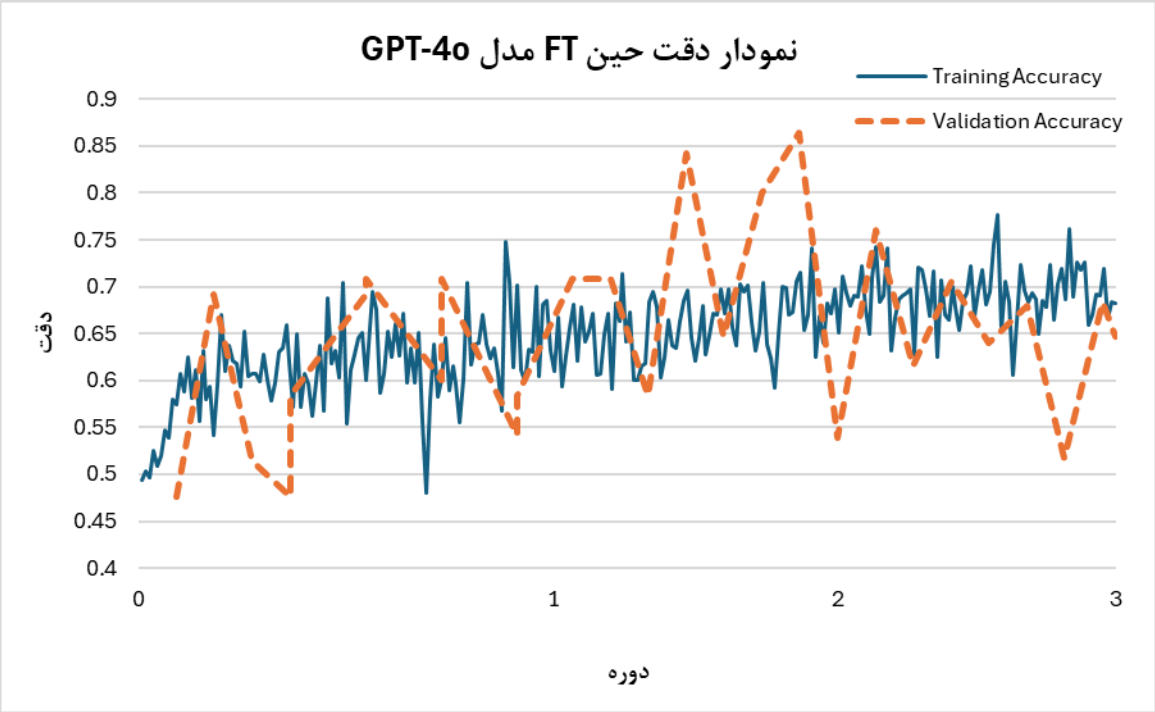
- GPT-٤o - FT فریم (گزارش مطلق):

مطابق محدودیت‌های پلتفرم OpenAI، تنها مدلِ پشتیبانِ Fine-Tuning چندوجهی (متنی–تصویری) GPT-٤o است؛ لذا نتایج FT مربوط به GPT-٤o گزارش می‌شود و برای این مدل، ارقام «پایه» (بدون FT) در فایل‌های ارسالی موجود نبود تا مقایسه تفاضلی انجام شود. بنابراین فقط مقادیر مطلق زیر (در ۱۲ فریم) ارائه می‌گردد:

  - تک‌مرجع: BLEU-٤ = ۱۱٫۰۹، CIDEr-D = ۱۰۳٫۰۰، METEOR = ۱۵٫۲٤، ROUGE-L = ۳۳٫۸٤، SPICE = ۲٦٫۲۵
  - چندمرجع: BLEU-٤ = ۲۳٫٦۵، CIDEr-D = ۹۵٫۳۳، METEOR = ۲۲٫۲۰، ROUGE-L = ۵٤٫۱٦، SPICE = ۲۲٫۷۵



تصویر ۴–۹– همگرایی تابع زیان در حین Fine Tuning مدل GPT-٤o



تصویر ۴–۱۰– نمودار دقت مدل در حین Fine Tuning در مدل GPT-٤o

در مقابل، فرآیند آموزش تکمیلی برای مدل GPT-٤o رفتاری کاملاً متفاوت و ناپایدار از خود نشان داد. نمودارهای ۴—۹ و ۴—۱۰ که روند زیان و دقت این مدل را به تصویر می‌کشند، با نوسانات شدید، به‌ویژه در معیارهای اعتبارسنجی، مشخص می‌شوند.

همان‌طور که در نمودارها مشهود است، منحنی‌های اعتبارسنجی دارای قله‌های تیز و دره‌های عمیق هستند. برای مثال، دقت اعتبارسنجی (نمودار ۴—۱۰) به صورت متناوب بین مقادیر بالا (حدود ۰.۸۵) و پایین (حدود ۰.۵) در نوسان است. این رفتار، به جای یک روند همگرایی پایدار یا شروع تدریجی بیش‌برازش، نشان‌دهنده ناپایداری در قابلیت تعمیم[۱۱۶] مدل است. به عبارت دیگر، مدل در مواجهه با دسته‌های مختلفی از داده‌های نادیده، عملکردی بسیار متغیر از خود به نمایش می‌گذارد.

این ناپایداری به احتمال زیاد ناشی از ابرپارامترهای فرآیند آموزش، مانند نرخ یادگیری است. لازم به ذکر است که این مقادیر به صورت خودکار توسط پلتفرم آموزش OpenAI انتخاب شده‌اند و به دلیل هزینه‌های محاسباتی بالا، امکان جستجوی دستی و بهینه‌سازی ابرپارامترها برای دستیابی به همگرایی پایدارتر، وجود نداشته است. در چنین شرایطی، انتخاب یک نقطه بهینه برای مدل نهایی با چالش روبرو است، زیرا یک قله عملکرد در منحنی اعتبارسنجی لزوماً نماینده یک مدل با قابلیت تعمیم پایدار نیست و ممکن است نتیجه‌ای تصادفی بر روی یک دسته خاص از داده‌ها باشد.

جمع‌بندی: با معیار مقایسه اصلاح‌شده (در برابر «بهترین پایه همان مدل»)، Fine-Tuning چندوجهی در Gemini ۲٫۵ Pro بهبود پایدار در همه معیارها (تک‌مرجع و چندمرجع) نشان می‌دهد. برای GPT-٤o، به‌سبب نبودِ نتایج پایه، تنها سطوح مطلق FT گزارش شد و قضاوت تفاضلی به‌صورت رسمی ممکن نیست.

---

[۱۱۶] instability in generalization

جدول ۴–۹– بهترین نتایج هر مدل در هر معیار (تک مرجع)، اعداد از ۷-۴، ۸-۴ و ۳-۹-۴ استخراج شده اند.

	BLEU-٤		CIDEr-D		METEOR		ROUGE-L		SPICE	
	امتیاز	n-best	امتیاز	n-best	امتیاز	n-best	امتیاز	n-best	امتیاز	n-best
ADAPT	۱۱.۴	۳۲	۱۰۲.۶	۳۲	۱۵.۲	۳۲	۳۲	۳۲و۱۶	۲۵.۶	۳۲
DriveGPT٤	۱۰.۰۲	۸	۹۸.۷۱	۸			۳۱.۵۲	۸		
Gemini ۲,۵ Pro	۷.۰۸	۱۲	۶۲.۲	۱۲	۱۷.۷۲	۱۶	۳۲.۸۲	۱۲	۲۴.۳۵	۱۶
Gemini ۲,۵ Flash	۵.۶۴	۱۶	۵۴.۴	۴	۱۶.۸۴	۱۶	۳۱.۷۵	۱۶	۲۰.۰۶	۱۶
Gemini ۲,۵ Flash Lite	۶.۸۸	۸	۶۰.۳۸	۱۶	۱۵.۳۳	۱۶	۲۹.۸۱	۸	۲۰.۵۱	۱۶
GPT-٥ mini	۲.۰۳	۱۶	۲۵.۴۸	۱۶	۱۴.۰۵	۱۲	۲۳.۲۹	۱۶	۱۱.۴	۱۶
FT-Gemini ۲.۵ □□□	**۱۱.۸۳**	۱۲	**۱۱۴.۷۳**	۱۲	**۱۹.۴۵**	۱۲	**۴۰.۵۷**	۱۲	**۲۸.۴۷**	۱۲
FT-GPT-٤o	۱۱.۰۹	۱۲	۱۰۳	۱۲	۱۵.۲۴	۱۲	۳۳.۸۴	۱۲	۲۶.۲۵	۱۲

جدول ۴–۱۰– بهترین نتایج هر مدل در هر معیار (چند مرجع)، اعداد از ۷-۴، ۸-۴ و ۳-۹-۴ استخراج شده اند.

	BLEU-٤		CIDEr-D		METEOR		ROUGE-L		SPICE	
	امتیاز	n-best	امتیاز	n-best	امتیاز	n-best	امتیاز	n-best	امتیاز	n-best
ADAPT	۱۸.۴۵	۳۲	۹۵.۱۳	۳۲	۲۱.۰۲	۳۲	۴۷.۰۸	۳۲و۱۶	۱۸.۲۲	۳۲
Gemini ۲.۵ □□□	۱۶.۰۸	۱۶	۵۹.۰۲	۱۶	۲۳.۸۱	۱۶	۴۸.۱۵	۱	۲۶.۶۳	۱۲
Gemini ۲.۵ Flash	۱۴.۰۹	۴	۵۳.۱۶	۱۶	۲۳.۱۸	۱۶	۴۷.۷	۱	۲۳.۵۴	۱۶
Gemini ۲.۵ Flash Lite	۱۸.۰۷	۸	۵۳.۷۹	۱۶	۲۱.۳	۱۶	۴۵.۴۶	۸	۲۱.۲۱	۱۶
GPT-٥ mini	۶.۳۱	۱۶	۲۱.۶	۱۶	۱۹.۸	۱۶	۳۶.۵۳	۱۶	۱۵.۵	۱

FT-Gemini ۲.۵ □□□	**۲۳.۹**	۱۲	**۱۰۰.۷۵**	۱۲	**۲۴.۵۸**	۱۲	**۵۷.۵۹**	۱۲	**۲۷.۴۶**	۱۲
FT-GPT-٤o	۲۳.۶۵	۱۲	۹۵.۳۳	۱۲	۲۲.۲	۱۲	۵۴.۱۶	۱۲	۲۲.۷۵	۱۲

۴–۱۰– منابع محاسباتی، زمان اجرا و هزینه

این بخش تصویر روشنی از منابع محاسباتی، زمان‌های اجرا و محرک‌های هزینه‌ای ارائه می‌کند و در پایان، محدودیت‌ها و مسیرهای مقیاس‌پذیری را جمع‌بندی می‌نماید.

۴–۱۰–۱– جمع‌بندی کمّی منابع و زمان

- آموزش روش پایه ADAPT: با پیکربندی ۴–۴–۱ روی Google Colab انجام شده است؛ GPU غالب A۱۰۰ (۴۰ GB) و در برخی اجراها L٤ به‌کار رفته است (آخرین اجرا با L٤ پردازش شده است). زمان آموزش به‌طور میانگین ≈۱۰۳ دقیقه برای هر دوره و در مجموع ≈۶۲۱ دقیقه (تک‌کارت) بوده است. لاگ‌ها به‌طور منظم ذخیره و با TensorBoard پایش شده‌اند.
- Fine-Tuning چندوجهی مدل‌های زبانی:

جدول ۴–۱۱– منابع و زمان Fine-Tuning

مدل	پلتفرم	زمان کل FT	تعداد دوره	توضیح زیرساخت
Gemini ۲,۵ Pro	Vertex AI (Google)	۱ ساعت و ۳۵ دقیقه	۶	نوع GPU/ منابع توسط پلتفرم افشا نمی‌شود
GPT-٤o (OpenAI)	OpenAI FT	۲ ساعت و ۳۸ دقیقه	۳	نوع GPU/ منابع توسط پلتفرم افشا نمی‌شود

- استنتاج LLM ها: برای سطوح ورودی $n \in \{1,4,8,12,16\}$ و برای هر مدل یک پاسخ تولید کرده است؛ زمان پاسخ‌گویی در پلتفرم‌های ابری برون‌سامانه مدیریت می‌شود و عدد دقیقی گزارش‌پذیر نیست.

۴–۱۰–۲– محرک‌های هزینه و راهبردهای کنترل هزینه

۱. خط‌مشی محاسبه هزینه API:

مطابق مستندات هر دو پلتفرم، فراخوانی‌هایی که با خطا خاتمه می‌یابند مشمول هزینه نمی‌شوند؛ در این پژوهش نیز سیاست «تا ۴ تلاش در صورت خطا» به‌کار رفته و موارد شکست، هزینه‌ای ایجاد نکرده‌اند.

۲. کاهش بار محاسباتی/حجمی در آموزش و FT:

- در آموزش ADAPT، تمامی فریم‌ها پس از برش مربعی به ۲۲۴×۲۲۴ کاهش یافته‌اند؛ این کاهش ابعاد، هزینه محاسباتی (GPU-hour) را به‌طور معناداری پایین آورده است.
- در OpenAI (فقط در فرایند Fine-Tuning)، تصاویر آموزشی به ۵۱۲×۵۱۲ کاهش اندازه داده شده‌اند. (در استنتاج معمول، کاهش اجباری به ۵۱۲×۵۱۲ اعمال نشده است.)

۳. انتخاب مسیرهای ارزان‌تر فراخوانی در OpenAI:

- برای استنتاج عادی از API نوع Flex استفاده شده است؛ Flex با هزینه کمتر و زمان پاسخ بیش‌تر همراه است، اما دقت مدل ثابت می‌ماند (برای سناریوهای این پژوهش کفایت داشته است).

- برای مدل FT شده GPT-٤o نوع Flex در دسترس نیست لذا از پردازش دسته‌ای[۱۱۷] در پلتفرم OpenAI (آپلود دسته درخواست‌ها و دریافت خروجی) بهره گرفته شده است؛ ساختار هزینه این روش معادل Flex و از نظر دقت هم‌ارز است.

٤. بهره‌گیری از سهمیه‌ها و اعتبارات رایگان:

- OpenAI: با پذیرش سیاست اشتراک‌گذاری داده، روزانه ۲.۵ میلیون توکن رایگان برای مدل GPT-٥ Mini و ۲۵۰ هزار توکن رایگان برای GPT-٥ تخصیص می‌یابد؛ این سهمیه‌ها طی مدت پژوهش روی دو حساب کاربری استفاده شده‌اند.
- Gemini: سهمیه روزانه رایگان شامل ۱۰۰۰ درخواست یا ۲۵۰ هزار توکن برای Gemini ۲٫٥ Flash-Lite، ۲۵۰ درخواست یا ۲۵۰ هزار توکن برای Gemini ۲٫٥ Flash و ۱۰۰ درخواست یا ۲۵۰ هزار توکن برای Gemini ۲٫٥ Pro است؛ این سهمیه‌ها روی سه حساب کاربری مصرف شده‌اند. افزون‌براین، با احراز هویت، اعتبار ۳۰۰ دلاری رایگان دریافت و برای Fine-Tuning و استنتاج به‌کار گرفته شده است.

٥. جمع‌بندی عملی:

هزینه کل پروژه عمدتاً از GPU-hour آموزش ADAPT و تعداد فراخوانی‌های LLM ناشی می‌شود. با کوچک‌سازی تصویر در آموزش (۲۲۴×۲۲۴)، استفاده از OpenAI Flex یا Batch Jobs (برای GPT-٤o FT)، و مصرف هدفمند سهمیه‌ها/اعتبارات رایگان، هزینه‌ها در محدوده قابل‌قبول نگه داشته شده‌اند بدون افت دقت در سناریوهای ارزیابی گزارش‌شده.

۴–۱۰–۳– محدودیت‌های مرتبط با منابع

در این پژوهش سه محدودیت عملی وجود داشته است:

۱. نوسان بستر Colab:

---

[۱۱۷] Batch Jobs

در اجرای نهایی، همه آموزش‌ها بر روی GPU نوع NVIDIA L٤ انجام شد؛ بنابراین نوسان ناشی از تغییر نوع GPU یا محدودیت/تمدید نشست‌ها در نتایج دخیل نیست. بااین‌حال، زمان محاسبه در Colab ثابت نیست و بسته به ساعات شبانه‌روز با ثابت‌بودن کد، داده و پیکربندی زمان هر گام/ایپاک می‌تواند به‌طور محسوسی افزایش یابد. این ناپایداری در زمان پاسخ LLM نیز مشاهده شد، هرچند دامنه آن به‌مراتب کمتر است (افزایش تأخیر حداکثر تا حدود ۴۰٪) و اثر معناداری بر نتایج کمّی گزارش‌شده نداشته است. برای کاهش اثر، اجرای آزمایش‌ها تا حد امکان در بازه‌های زمانی مشابه زمان‌بندی شد و نتایج در سطح corpus-level گزارش گردید.

۲. نامشهود بودن جزییات منابع و زمان در فرآیند Fine-Tuning روی پلتفرم‌های مدیریت‌شده (صرفاً زمان کل و شمار دوره گزارش‌پذیر بوده است)؛ و
۳. محدودیت‌های نرخ و خطاهای گذرا در APIها که با سیاست تلاشِ تکراری (تا ۴ بار) جبران شده و-طبق مستندات-فراخوانی‌های ناموفق هزینه‌ای ایجاد نکرده‌اند. این موارد بر صحت نتایج کمی اثر نگذاشته و صرفاً به‌عنوان قیود اجرایی مستندسازی می‌شوند.

۴–۱۰–۴– ملاحظات مقیاس‌پذیری

برای مقیاس‌های بزرگ‌تر، سه راهبرد عملی توصیه می‌شود:

۱. بهینه‌سازی مصرف GPU در آموزش با Mixed-Precision، Gradient Accumulation و در صورت امکان ZeRO مرحله بالاتر؛
۲. کاهش محاسبه تکراری با Cache کردن خروجی‌های پیش‌پردازش (فریم‌های ۲۲۴×۲۲۴، نقشه/کانتور توجه) و ذخیره نتایج استنتاج LLM بر اساس هش ورودی؛
۳. مدیریت کارای فراخوانی‌ها با صف‌بندی/تجمیع درخواست‌ها و استفاده از مسیرهای کم‌هزینه پلتفرم‌ها (نظیر Flex API یا Batch Jobs)، بدون تغییر در پروتکل ارزیابی. این اقدامات ماهیت روش را تغییر نمی‌دهند و صرفاً کارایی و هزینه اجرا را بهبود می‌بخشند.

۴–۱۱– محدودیت‌ها و تهدیدهای اعتبار نتایج

این بخش به‌صورت نظام‌مند محدودیت‌های پژوهش و تهدیدهای اعتبار نتایج را بیان می‌کند. چارچوب ارائه‌شده چهار بُعد اصلی اعتبار سازه‌ای، اعتبار درونی، اعتبار بیرونی (تعمیم‌پذیری) و اعتبار استنباطی را پوشش می‌دهد و در هر مورد (در صورت امکان) راهکارهای کاهش اثر یا پیشنهاد کارهای بعدی ذکر می‌شود.

۴–۱۱–۱– اعتبار سازه‌ای[۱۱۸]

- معیارهای زیرنویس‌محور به‌عنوان جانشینِ «کیفیت توضیح»: سنجه‌های BLEU-٤ ، METEOR، ROUGE-L، CIDEr-D و SPICE شاخص‌های استانداردِ شباهت متنی‌اند، نه لزوماً کیفیت «توضیح علّی»؛ بنابراین همبستگی آن‌ها با «قابلیت اقناع/تبیین» کامل نیست.
- کاهش اثر: استفاده از چندمرجع (افزودن ۴ بازنویسی هم‌معنا) برای کاهش حساسیت لغوی انجام شد؛ پیشنهاد می‌شود در کارهای بعدی ارزیابی انسانیِ کنترل‌شده و/یا سنجه‌های اختصاصیِ XAI اضافه شود.
- فرض درستیِ تصمیم ADAPT در دستور: دستور LLM به‌صراحت تصمیم ADAPT را «درست» فرض می‌کند؛ این امر کیفیت «توضیح» را به‌سمت توجیه سوق می‌دهد.

  پیشنهاد: طراحی سناریوهایی که در آن‌ها «توضیح» بدون پیش‌فرضِ صحت تصمیم نیز سنجیده شود.
- استفاده از کانتور توجه به‌جای نقشه پیوسته: تبدیل نقشه توجه به کانتور (آستانه=τ ۰٫۷، ضخامت ۳ px) ماهیت بازنمایی را تغییر می‌دهد و ممکن است بخشی از نشانه‌ها حذف/ساده‌سازی شود.

پیشنهاد: مطالعه حذف مؤلفه‌ها برای ارزیابی «بدون کانتور» و ابلیشنِ آستانه/ضخامت در آینده.

---

[۱۱۸] Construct Validity

۴–۱۱–۲– اعتبار درونی[۱۱۹]

- نوسان زمان محاسبه پلتفرم اجرا. هرچند اجرای نهایی بر GPU نوع L٤ انجام شده و تغییر نوع GPU دخیل نیست، اما زمان محاسبه در Colab ثابت نیست و بسته به ساعات شبانه‌روز با ثابت‌بودن کد/داده/پیکربندی می‌تواند افزایش یابد؛ تأثیر این نوسان بر استنتاج LLM کمتر و در حد افزایش تأخیر تا ≈٪۴۰ بوده است.

  کاهش اثر: زمان‌بندی اجراها در بازه‌های مشابه و گزارش نتایج در سطح corpus-level.
- غیرقطعی‌بودن بخشی از عملیات GPU و .API برخی عملیات (به‌ویژه در GPU و سرویس‌های ابری) غیرقطعی‌اند؛ نتایج single-run ممکن است تحت تأثیر نوسان‌های جزیی قرار گیرد.

  کاهش اثر: تنظیمSeed ها و ثبات پیکربندی رعایت شده؛ تکرار چندباره آزمون‌های حساس به‌عنوان کار تکمیلی پیشنهاد می‌شود.

۴–۱۱–۳– اعتبار بیرونی[۱۲۰] و تعمیم‌پذیری

- وابستگی به BDD-X و توزیع داده: نتایج بر یک مجموعه داده مشخص (BDD-X) و سناریوهای دید از جلو گزارش شده‌اند؛ تعمیم به سایر محیط‌ها/شرایط جوی/سنسورها یا توزیع‌های زبانی متفاوت نیازمند ارزیابی مجدد است.

  پیشنهاد: آزمون بر مجموعه‌داده‌های متنوع‌تر و سناریوهای چندسنسوری.
- پیکربندی ورودی LLM (۱ تا ۱۶ فریم نخست): ارزیابی کیفی/کمی با پنجره زمانی محدود انجام شده است؛ رویدادهای دیرترِ کلیپ ممکن است دیده نشوند.

  پیشنهاد: بررسی پنجره‌های لغزان و ورودی‌های طولانی‌تر (در چارچوب هزینه).

---

[۱۱۹] Internal Validity

[۱۲۰] External Validity

۴–۱۱–۴– اعتبار استنباطی[۱۲۱]

- single-run و فقدان آزمون معناداری آماری: به‌دلیل مقیاس محاسباتی، نتایج هر پیکربندی یک‌بار اجرا شده و آزمون‌های آماری (bootstrap/randomization) اعمال نشده‌اند؛ بنابراین برتری‌های نزدیک باید با احتیاط تفسیر شوند.

  پیشنهاد: اجرای چندrun و آزمون معناداری در کارهای بعدی.
- ناهم‌سانی رفتار معیارها: به‌ویژه CIDEr-D ممکن است در ورودی‌های کوتاه‌تر بیشینه شود (به‌واسطه وزن‌دهیtf-idf)؛ نتیجه‌گیری باید بر مجموعه معیارها استوار باشد، نه یک شاخص منفرد.
- Fine-Tuning پلتفرم‌محور با جزییات نامشهود: در FT (Gemini ۲٫۵ Pro، GPT-٤o) زمان کل و شمار دوره ثبت شده است، اما نوع/تعداد دقیق GPU سمت پلتفرم افشا نمی‌شود؛ این امر دقت مقایسه هزینه/منابع را محدود می‌کند.

۴–۱۱–۵– سایر ملاحظات

- کاهش ابعاد تصویر: در آموزش ADAPT کاهش به ۲۲۴×۲۲۴ ضروری و مؤثر بر هزینه بوده است، اما ممکن است جزئیات ظریف صحنه را حذف کند؛ اثر خالص در نتایج نهایی به‌صورت غیرمستقیم دیده می‌شود.
- چندمرجع‌های تولیدشده با LLM: هرچند چندمرجع، ارزیابی را مقاوم‌تر می‌کند، اما بازنویسی‌های LLM می‌توانند توزیع واژگانی را به‌گونه‌ای تغییر دهند که برخی معیارها (مثلاً CIDEr-D) کاهش یابند؛ این موضوع در تفسیر نتایج لحاظ شده است.

---

[۱۲۱] Conclusion Validity

- اخلاق و حریم خصوصی: تصاویر صرفاً نمای جلو و عاری از اطلاعات شناسایی مستقیم استفاده شده‌اند؛ هرگونه انتشار بیرونی تصاویر باید با ملاحظات محوسازی/ناشناس‌سازی همراه باشد.

جمع‌بندی: با وجود محدودیت‌های فوق، طراحی پروتکل-شامل کنترل یکسان‌سازی ورودی‌ها، استفاده از چندمرجع، هم‌ترازی زمانیِ فریم‌ها با اکشن ADAPT، و گزارش نتایج در سطح corpus تا حد زیادی تهدیدهای اعتبار را کاهش داده است. بااین‌حال، برای تقویت بیشتر استنباط‌ها، پیشنهاد می‌شود در کارهای بعدی: (۱) آزمون‌های آماری اجرا شود، (۲) سناریوی «بدون کانتور توجه» و ابعاد جایگزینِ ورودی بررسی گردد، (۳) ارزیابی انسانیِ کنترل‌شده به سنجش اضافه شود، و (۴) تعمیم‌پذیری بر مجموعه‌داده‌ها/محیط‌های دیگر آزموده شود.

۴–۱۲– جمع‌بندی فصل چهارم و گذار به فصل پنجم

در این فصل، چارچوب ارزیابی و نتایج تجربی روش پیشنهادی به‌صورت کمّی و کیفی گزارش شد. روش پایه ADAPT در ورودیِ بومیِ ۳۲ فریم بهترین کارایی را نشان داد و افزودن چندمرجع (در کنار مرجع انسانی) موجب بهبود معنادار در شاخص‌های زبانیِ BLEU-٤ ، METEOR و ROUGE-L شد؛ هرچند CIDEr-D همواره یکنواخت افزایش نیافت. مدل‌های زبانیِ چندوجهی نیز، به‌ویژه پس از Fine-Tuning (برای Gemini ۲٫۵ Pro و GPT-٤o)، در معیارهای زبانیِ مذکور به سطوح رقابتی یا برتر نسبت به روش پایه نزدیک شدند و در مطالعه حساسیتِ تعداد فریم‌ها روندی عمدتاً افزایشی (تا ۱۶ فریم) نشان دادند. در تحلیل کیفی، نمونه‌های موفق نشان داد که متن تولیدی با کانتورِ توجه ADAPT و مرجع انسانی هم‌خوان است؛ در مقابل، ضدنمونه‌ها عمدتاً ناشی از ابهام بصری، کمبود زمینه زمانی یا ناهمخوانی اشاره متنی با ناحیه توجه بودند. مسیر جایگزینِ طبقه‌بندی سرعت/جهت-به‌رغم تلاش‌های اجرایی-به سقف دقتی حوالی ٪۶۰ محدود شد و به‌سبب هم‌سویی کمتر با هدف تبیینی کنار گذاشته شد. ملاحظات منابع، زمان اجرا و هزینه نیز مستندسازی گردید و

راهبردهای کنترل هزینه (کاهش ابعاد، استفاده از مسیرهای ارزان‌تر فراخوانی و سهمیه‌های رایگان) بدون افت دقت به‌کار گرفته شد.

با اتکا به این یافته‌ها، فصل پنجم به نتیجه‌گیری کلّی و پاسخ صریح به اهداف/پرسش‌های پژوهش می‌پردازد، دلالت‌های علمی و کاربردی کار را جمع‌بندی می‌کند و با نگاه انتقادی، پیشنهادهای بهبود و مسیرهای آینده-از جمله ارزیابی انسانی کنترل‌شده، سنجش سناریوی «بدون کانتور توجه» و آزمون بر مجموعه‌داده‌های متنوع‌تر-را صورت‌بندی خواهد کرد. این گذار، پیوند منسجم بین شواهد تجربی فصل حاضر و برداشت‌های نهایی فصل بعد را فراهم می‌سازد.

فصل پنجم

نتیجه گیری و کار های آتی

۵–۱– کلیات فصل

این فصل به‌منزله جمع‌بندی نهایی پژوهش، تفسیر نتایج و ترسیم افق‌های آتی تنظیم شده است. در امتداد فصل چهارم-که در آن چارچوب ارزیابی، یافته‌های کمّی/کیفی، و ملاحظات اجرایی به‌تفصیل گزارش شد-اینجا تمرکز از ارائه جزئیات اجرایی به تبیین و نتیجه‌گیری منتقل می‌شود: نخست، مهم‌ترین دستاوردهای پژوهش در پیوند «مدل تصمیم‌گیر مبتنی بر ADAPT» و «ماژول توضیح‌گر مبتنی بر LLM» با تکیه بر معیارهای زبانی و شواهد بصری-متنی خلاصه می‌گردد؛ سپس میزان تحقق اهداف/پرسش‌های تحقیق و دلالت‌های نظری و کاربردی کار روشن می‌شود.

سازمان فصل به این ترتیب است: در ۲–۵، یافته‌های کلیدی به‌صورت فشرده جمع‌بندی می‌گردد. ۳–۵، پاسخ صریح به اهداف و پرسش‌های پژوهش را عرضه می‌کند. در ۴–۵ و ۵–۵، به‌ترتیب دلالت‌های علمی (نظری/روش‌شناختی) و دلالت‌های کاربردی برای استقرار در سامانه‌های خودران تشریح می‌شود. ۶–۵، حدود اعتبار و تعمیم‌پذیری را—بدون تکرار جزئیات—مرزبندی می‌کند و ۷–۵، پیشنهادهای پژوهشی آینده را طرح می‌نماید. نهایتاً ۷–۵ با یک جمع‌بندی نهایی جهت‌دهنده، جایگاه

کار را در ادبیات و مسیرهای بعدی تثبیت می‌کند. بدین‌سان، فصل حاضر حلقه اتصال نتایج تجربی فصل چهارم با نتیجه‌گیری‌های کلان و توصیه‌های آتی است.

۵–۲– جمع‌بندی یافته‌های کلیدی پژوهش

۵–۲–۱– روش پایه ADAPT و رفتار معیارها

ارزیابی‌های فصل چهارم نشان داد که روش پایه ADAPT در پیکربندی بومیِ ۳۲ فریم، بالاترین هم‌خوانی متنی را با مراجع انسانی به دست می‌آورد (۷–۴). افزودن چندمرجع (چهار بازنویسی هم‌معنا در کنار مرجع انسانی) به‌طور پایدار موجب بهبود معیارهای زبانیِ حساس به هم‌معنایی و ساخت جمله (به‌ویژه METEOR و ROUGE-L) می‌شود، در حالی‌که CIDEr-D همواره روندی یکنواخت ندارد و گاه در ورودی‌های کوتاه‌تر مقادیر بیشتری نشان می‌دهد (۶–۴، ۷–۴). همچنین افزایش تعداد فریم‌های ورودی تا مرزهای معقول (n=۱۶ برای مقایسه با LLM ها) عموماً به ارتقای هم‌معنایی منجر می‌شود، هرچند رفتار همه معیارها کاملاً یکنواخت نیست (۱–۹–۴). در جمع‌بندی، ADAPT مرجع سنجش معتبری برای مقایسه متنِ توضیح محسوب می‌شود و نقش چندمرجع در پایدارسازی ارزیابی و کاهش حساسیت واژگانی تأیید شد.

۵–۲–۲– نقش LLM ها در تولید توضیح متنی (پیش و پس از Fine-Tuning)

نتایج نشان دادند که LLMهای چندوجهی پیش از آموزش تکمیلی در تولید توضیح‌های موجز و هم‌معنا با مرجع انسانی رقابتی ظاهر می‌شوند، و با افزایش تعداد فریم‌های ورودی، هم‌خوانی بصری-متنی (با اتکا به کانتور توجه ADAPT) بهبود می‌یابد (۸–۴،۱۰–۴). پس از Fine-Tuning، این هم‌خوانی و کیفیت زبانی به‌طور معناداری ارتقا می‌یابد؛ به‌ویژه برای Gemini ۲٫۵ Pro، که نسبت به نسخه پایه همان مدل (در بهترین حالتش) بهبود همزمان در معیارهای اصلی مشاهده شد (۳–۹–۴). برای GPT-٤o نیز سطوح مطلقِ مطلوبی گزارش شد، هرچند به‌دلیل نبود اعداد پایه متناظر، قضاوت

تفاضلی محدود است (۳–۹–۴). در ارزیابی کیفی، نمونه‌های موفق نشان دادند که متن تولیدی عموماً به همان شواهدی ارجاع می‌دهد که در ناحیه توجه برجسته شده‌اند، و ضدنمونه‌ها عمدتاً به ابهام بصری، کمبود زمینه زمانی، یا ناسازگاری اشاره متنی با ناحیه توجه بازمی‌گردند (۱۰–۴). به‌طور کلی، ترکیب «تصمیم ADAPT + کانتور توجه + دستور موجز» بستر مؤثری برای توضیح‌پذیری زبانی فراهم کرده است و چندمرجع ارزیابی را منصفانه‌تر و پایدارتر کرده است.

۵–۳– پاسخ به اهداف/سؤالات پژوهش و میزان تحقق آن‌ها

در این بند، نتایج فصل چهارم و مستندات فصول پیشین با اهداف تعریف‌شده در فصل‌های اوّل و دوم هم‌تراز می‌شوند و برای هر هدف، میزان تحقق به‌صورت روشن بیان می‌گردد.

۱. طراحی و پیاده‌سازی سامانه ترکیبی «تصمیم‌گیر–توضیح‌گر».

هدف، ساخت خط لوله‌ای بود که تصمیم را با ADAPT و توضیح متنی را با LLM تولید کند. این خط لوله با مؤلفه‌های استخراج فریم، اجرای ADAPT، استخراج توجه و تزریق آن به LLM (به‌همراه اکشن) طبق ۴–۳ و ۵–۳ پیاده‌سازی و در فصل چهارم ارزیابی شد.

نتیجه: هدف به‌طور کامل محقق شد؛ رفتار سامانه در ارزیابی کمی و کیفی (۷–۴ تا ۴–۹ و **پیوست ب**) مستند است.

۲. بهره‌گیری از نقشه توجه ADAPT در ماژول توضیح‌گر.

هدف، سنجش این بود که آیا هدایت LLM با نشانه مکانیِ توجه (کانتور قرمز بر اساس نقشه توجه) به هم‌خوانی بصری–متنی بهتر می‌انجامد. تبدیل نقشه توجه به کانتور و الحاق آن به ورودی LLM در ۱–۵–۳ و ۲–۵–۳ تشریح و در تحلیل کیفی **پیوست ب** ارزیابی شد.

نتیجه: هدف تحقق یافته و شواهد کیفی نشان می‌دهد متنِ تولیدی غالباً به شواهد داخل کانتور ارجاع می‌دهد؛ سناریوی «بدون کانتور» به‌عنوان کار تکمیلی پیشنهاد شده است (۴–۱۱).

۳. طراحی پروتکل ارزیابی مبتنی بر معیارهای زیرنویس و چندمرجع.

هدف، سنجش توضیح‌های متنی با BLEU-٤ ، METEOR، ROUGE-L، CIDEr-Dو SPICEدر دو رژیم تک‌مرجع/چندمرجع و سطح corpus بود. این پروتکل در ۶–۴ صورت‌بندی و اجرا شد و نتایج ADAPT وLLM ها بر همین مبنا گزارش گردید(۷–۴ و ۸–۴).

نتیجه: هدف کامل محقق شد؛ چندمرجع، پایداری و منصفانه‌بودن ارزیابی را افزایش داده است.

٤. مطالعه رفتار معیارها و حساسیت به اطلاعات زمانی (n-fraction).

هدف، فهم اثر تعداد فریم‌های ورودی بر کیفیت توضیح بود. در ۱–۹–۴، روندهای معیارها بر حسب $n \in \{1,4,8,12,16\}$ تحلیل شد.

نتیجه: هدف تحقق یافته؛ شاخص‌های معنایی (به‌ویژه METEOR/ROUGE-L) عموماً با افزایش n بهبود پیدا می‌کنند، هرچند CIDEr-D رفتاری یکنواخت ندارد.

٥. نقش LLM ها پیش و پس از Fine-Tuning.

هدف، مقایسه نسخه‌های پایه LLM با نسخه‌های Fine-Tuned بود. نتایج نشان داد Gemini ۲٫٥ Pro (FT) نسبت به بهترین حالت پایه همان مدل در اکثر معیارها بهبود دارد (–۹–۴ ۳). برای GPT-٤o (FT) اعداد مطلق گزارش شد و به‌سبب فقدان نتایج پایه متناظر، داوری تفاضلی محدود است.

نتیجه: هدف تا حدّ زیاد محقق شد؛ برای GPT-٤o، تنها مقایسه مطلق ممکن بود.

٦. مرجع‌گذاری و مقایسه با روش پایه ADAPT.

هدف، تثبیت ADAPT به‌عنوان خط‌کش مقایسه برای متون توضیح بود. نتایج ۷–۴ نشان دادند ADAPT در ۳۲ فریم عملکرد مرجع معتبری ارائه می‌کند و مبنای مناسبی برای سنجش LLM هاست.

نتیجه: هدف کامل محقق شد.

۷. بازتولیدپذیری و شفاف‌سازی منابع/هزینه.

هدف، ثبت دقیق پیکربندی‌ها، مسیرهای کُد، داده، و زمان/هزینه اجرا بود. در ۲–۴، ۳–۴ و ۴–۱۰، جزئیات محیط، پیش‌پردازش، زمان‌های کلیدی و راهبردهای مهار هزینه مستند شد.

نتیجه: هدف تحقق یافته؛ قیود پلتفرم‌های FT (نامشهود بودن جزئیات منابع) نیز شفاف‌سازی شد (۴–۱۰).

اهداف اصلی پژوهش-از طراحی و ارزیابی سامانه ترکیبی تا سنجش نقش توجه و چندمرجع و بررسی اثر Fine-Tuningبه‌طور عمده محقق شده‌اند و شواهد کمّی/کیفی فصل چهارم از آن‌ها پشتیبانی می‌کند. مواردی چون سناریوی «بدون کانتور» و مقایسه تفاضلی کامل برای GPT-٤o، به‌عنوان کارهای تکمیلی در فصل چهارم و ۴–۱۱ مشخص شده‌اند.

۵–۴– دلالت‌های علمی (نظری/روش‌شناختی)

این پژوهش در سه محورِ به‌هم‌پیوسته سهم‌هایی را عرضه می‌کند که هم جنبه نظری دارند و هم برای طراحی و ارزیابی سامانه‌های «توضیح‌پذیر» قابل استفاده‌اند.

۵–۴–۱– پیوند «توجهِ ADAPT + توضیح زبانیِ LLM» به‌مثابه سازوکار هم‌ترازی بصری–متنی

تعبیه نقشه توجهِ مدل تصمیم‌گیر (ADAPT) در ورودیِ ماژول زبانی (به‌صورت نمایش کانتوریِ نواحی پرتوجه) از حیث نظری نقش «قیودِ گراندینگ» را ایفا می‌کند: LLM به‌جای تولید دلیلِ آزاد، به ناحیه‌هایی مقید می‌شود که واقعاً در فرایند تصمیم‌گیری مؤثر بوده‌اند. این سازوکار، از دید تبیینی، دو پیامد دارد: (۱) کاهش گرایش به توصیف‌های کلی/هذیانی و حرکت به‌سوی «دلیلِ مبتنی بر شواهدِ موضعی»، و (۲) فراهم‌کردنِ بستری برای سنجش هم‌خوانی بصری–متنی (در ۴–۱۰ نشان داده شد که ارجاعات متنی عموماً با ناحیه‌های پرتوجه هم‌راستاست). از منظر روش‌شناسی XAI، این پیوند را

می‌توان به‌منزله «شرطِ پیشینِ گراندینگ» تلقی کرد که میان مولفه ادراکی[۱۲۲] و مولفه زبانی[۱۲۳] پلی شفاف ایجاد می‌کند و قابلیت ممیزیِ دلایل را بالا می‌برد.

۵–۴–۲– نقش چندمرجع در پایدارسازی و منصفانه‌سازیِ ارزیابی متنی

به‌کارگیری چهار بازنویسیِ هم‌معنا در کنار مرجع انسانی، ارزیابی را از حساسیت افراطی به انتخاب‌های واژگانی و نحویِ منفرد دور می‌کند و سازگاری بیشتری با ماهیت «چندجوابه»ی توضیح‌های معتبر دارد. دلالت نظریِ این مشاهده آن است که برای سنجش «کیفیت توضیح» (که ذاتاً طیفی و چندبیانی است)، استفاده از چندمرجع نه یک گزینه تجملی بلکه یک اقتضای روش‌شناختی است. پیام عملیِ آن نیز روشن است: گزارش نتایج در دو رژیم تک‌مرجع/چندمرجع، به‌همراه تصریح سطح محاسبه (corpus-level) و پیش‌پردازش متن، ارزیابی را پایدارتر و مقایسه‌ها را منصفانه‌تر می‌سازد.

۵–۴–۳– ناهمسان‌رفتاری معیارها و پیامدهای آن برای سنجش توضیح

مشاهده شد که معیارها رفتار همگن ندارند: شاخص‌های نزدیک به هم‌معنایی و ساخت جمله (مانند METEOR و ROUGE-L) با افزایش اطلاعات زمانی و چندمرجع عموماً بهبود می‌یابند، درحالی‌که CIDEr-D (به‌سبب وزن‌دهی tf–idf و حساسیت به توزیع واژگان) می‌تواند در ورودی‌های کوتاه‌تر یا در حضور بازنویسی‌های زبانیِ متنوع افت‌وخیز نشان دهد. این ناهمسانی یک نتیجه روش‌شناختی مهم دارد: هیچ معیار منفردی نباید نماینده «کیفیت توضیح» تلقی شود. ازاین‌رو، (۱) گزارش چندمعیاره و تفسیر «بر مبنای مجموعه معیارها» ضروری است؛ (۲) در کنار سنجه‌های مبتنی بر n-gram و هم‌معنایی، استفاده از ارزیابی‌های کیفیِ کنترل‌شده و/یا سنجه‌های ساختاری (مانند سنجش

---

[۱۲۲] Vision

[۱۲۳] Language

اشاره به اشیاءِ واقعاً حاضر در ناحیه توجه) به‌عنوان مکمل توصیه می‌شود؛ و (۳) شفاف‌سازیِ رژیم ارزیابی (تک‌مرجع/چندمرجع) بخشی از گزارش استاندارد کیفیت است.

در سطح نظری، کار حاضر نشان می‌دهد که «توجهِ مدل تصمیم‌گیر» می‌تواند به‌صورت قیود گراندینگ، تولید دلیلِ زبانی را هدایت کند و فاصله بین «آنچه مدل می‌بیند» و «آنچه مدل می‌گوید» را کاهش دهد. در سطح روش‌شناسی، به‌کارگیری چندمرجع و قرائت چندمعیاره نتایج، چارچوبی منصفانه‌تر و پایدارتر برای سنجش توضیح فراهم می‌آورد و از تکیه بیش‌ازحد بر یک شاخص خاص (به‌ویژه در حضور بازنویسی‌های زبانی) پرهیز می‌دهد. این دلالت‌ها، هم برای بازطراحیِ سامانه‌های توضیح‌پذیر آینده و هم برای استانداردسازی گزارش‌های تجربی در حوزه XAI، قابل تعمیم و استفاده‌اند.

۵–۴–۴– مقایسه روش پیشنهادی با رویکردهای مبتنی بر RAG در توضیح رانندگی خودران

RAG-Driver[۷۱] به‌واسطه استفاده از بازیابی نمونه‌های توضیحی، عملکرد بدون نیاز به آموزش مجدد در محیط‌های جدید را نشان می‌دهد؛ اما این روش به مدل‌های نسبتاً کوچک (۷ میلیارد پارامتری) اتکا دارد که منجر به بروز خطاهایی از جمله تولید محتوای نامعتبر (هالوسیناسیون) شده است [۷۱]. در مقابل، روش ما با تأکید بر آموزش بازتنظیمی و محدودیت‌های صریح دیداری بر پایه نقشه‌های توجه، بر ارائه توضیحاتی تفسیرپذیر و هم‌راستا با شواهد تأکید دارد. مطالعات مختلف نشان داده‌اند که Fine-Tuning در بسیاری از موارد دقت و استحکام بالاتری نسبت به روش‌های صرفاً مبتنی بر RAG ارائه می‌کند [۷۹]. در مجموع، در حالی که RAG-Driver قابلیت تعمیم سریع به موقعیت‌های ناشناخته را دارد، روش پیشنهادی ما بر شفافیت و انسجام بیشتر توضیحات برای فهم انسانی تمرکز می‌کند.

۵–۵– دلالت‌های کاربردی و توصیه‌ها برای استقرار در سامانه‌های خودران

این بخش، نتیجه عملی‌سازی چارچوب «ADAPT (تصمیم‌گیر) + LLM (توضیح‌گر)» را به‌صورت توصیه‌های اجرایی خلاصه می‌کند تا ادغام آن در یک سامانه خودران واقعی با کمترین سربار محاسباتی و بیشترین شفافیت انجام شود.

- تنظیم ورودی و محدودسازی طول فریم‌ها

۱. پنجره زمانی ورودی به LLM (n): برای حالت برخط، مقدار $n \in [۴,۱۶]$ توصیه می‌شود؛ افزایش n معمولاً به بهبود هم‌معنایی (METEOR/ROUGE-L) می‌انجامد، اما باید با قیود تأخیر مبادله شود. راهبرد عملی:
   - سناریوهای سرعت‌بالا/بلادرنگ: $n = ۴$ یا $n = ۸$؛
   - سناریوهای تحلیل‌محور/پس از واقعه: $n = ۱۲$ یا $n = ۱۶$.

۲. استانداردسازی تصویر: برش به مربع، کاهش اندازه به ۲۲۴×۲۲۴ و نرمال‌سازی (مطابق فصل ۳) را ثابت نگه دارید تا هزینه و تأخیر کنترل شود.

۳. هم‌ترازی زمانی: فریم‌های ارسال‌شده به LLM دقیقاً باید با بازه تصمیم ADAPT هم‌پوشانی زمانی داشته باشند (همان پنجره تصمیم).

- استفاده از نمایش کانتوریِ توجه به‌عنوان قید گراندینگ

۱. کانتور به‌جای نقشه رنگی: نمایش خطوط قرمز دور نواحی پرتوجه (با آستانه τ و ضخامت مشخص) برای LLM خواناتر و از نظر حافظه/پهنا کارآمدتر است.

۲. پارامترهای پایدار: آستانه ثابتی (مثلاً $\tau \approx 0.7$) و ضخامت ۲–۳px کفایت دارد؛ تغییر پیاپی این پارامترها می‌تواند رفتار زبانی را ناپایدار کند.

۳. حفظ هم‌خوانی بصری-متنی: کانتور را قبل از فشرده‌سازی روی تصویر رندر کنید تا در اندازه‌های کوچک نیز واضح بماند؛ هرگونه بازنمونه‌گیریِ بعدی باید همان نسبت‌ها را حفظ کند.

- سیاست تولید متن: ایجاز، قطعیت، و پرهیز از هذیان

۱. قیود خروجی: تولید ۱یا۲ جمله کوتاه، بدون ارجاع به «متن ورودی/تصویر» و بدون درخواست اطلاعات اضافی؛ این سیاست هم ثبات سبک را تضمین می‌کند و هم قابل‌خواندن بودن را برای راننده/اپراتور افزایش می‌دهد.

۲. قطعیّت کنترل‌شده: دمای تولید را پایین نگه دارید (نزدیک به ۰/۵) و از یک خروجی در هر درخواست استفاده کنید تا هم‌سانی متن‌ها حفظ شود.

۳. عدم مداخله در کنترل: متنِ توضیح نباید دستور کنترلی القا کند (مانند «بپیچ» یا «بایست»). نقش آن صرفاً شرح علت تصمیم است؛ تصمیم مرجع، ماژول ADAPT است.

- یکپارچه‌سازی سامانه‌ای و مسیر داده

۱. جریان ایمن داده: (توضیح → افزودن کانتور → توجه/اکشن → افزودن کانتور → ADAPT → تصویر) باید بی‌دررفت و قابل‌ردیابی باشد. برای هر تصمیم، شناسه کلیپ/زمان، اکشن ADAPT، پارامترهای کانتور و دستور فعلی را ثبت کنید.

۲. ماژولاریتی: ماژول توضیح‌گر باید در صورت اختلال سرویس LLM بی‌اثر شود (fail-open)؛ تصمیم ADAPT ادامه می‌یابد و تنها توضیح غیرفعال می‌شود.

۳. پایش و گزارش‌پذیری: لاگ‌های خروجی (توضیح، امتیازهای ارزیابی داخلی، کُد نسخه/دستور) را برای ممیزی و بهبود آتی ذخیره کنید.

- کارایی، تأخیر و هزینه

۱. بودجه تأخیر: برای برخط، بودجه زمان را به دو بخش تقسیم کنید: «زمان ADAPT» و «زمان LLM» با انتخاب n کوچک‌تر و تصاویر ۲۲۴×۲۲۴، زمان LLM قابل‌کنترل می‌ماند.

۲. کَش نتایج تکراری: درخواست‌هایی با ورودی یکسان (فریم‌ها/اکشن/دستور) را هش کنید و نتایج را کَش نمایید تا هزینه و تأخیر کاهش یابد.

۳. مسیرهای فراخوانی: بسته به پلتفرم، استفاده از مسیرهای کم‌هزینه برای استنتاج انبوه (نظیر پردازش دسته‌ای) یا API با هزینه کمتر، بدون تغییر کیفیت خروجی، توصیه می‌شود.

- نمایش و تجربه کاربری[۱۲۴] برای توضیح‌پذیری

۱. هم‌نمایی تصویر+متن: در رابط انسانی[۱۲۵]، تصویر کانتورگذاری‌شده در کنار توضیح ۱یا۲ جمله‌ای نمایش داده شود؛ از نمادهای رنگی ساده برای نوع علت (مانع، علامت راهنمایی، وسیله نقلیه جلویی، عابر) استفاده کنید.

۲. پرهیز از بار شناختی: توضیح طولانی، اصطلاحات فنی یا عدم قطعیت زائد[۱۲۶] را حذف کنید؛ هدف «فهم سریع» است.

۳. زبان و بومی‌سازی: زبان خروجی را با زبان کاربر/بازار مقصد هماهنگ کنید؛ سبک یکسان و واژگان محدود، پایایی ادراک را بالا می‌برد.

- ایمنی، اخلاق و حریم خصوصی

۱. اولویت ایمنی: توضیح هرگز جایگزین منطق کنترل یا قید ایمنی نمی‌شود؛ در تضادِ احتمالیِ متن با حسگرها/قواعد، قواعد ایمنی ارجح‌اند.

۲. حریم خصوصی: در ذخیره/نمایش، چهره‌ها و پلاک‌ها را محو کنید؛ داده تصویری را طبق خط‌مشی سازمان نگه‌داری کنید.

۳. بازخورد انسانی: سازوکاری برای اخذ بازخورد اپراتور درباره کیفیت توضیح (مفید/نامرتبط/نادرست) پیش‌بینی کنید تا داده لازم برای بهبود آینده فراهم شود.

- راهبرد بهبود و نگهداری

---

[۱۲۴] UX

[۱۲۵] HMI

[۱۲۶] Hedging

۱. نسخه‌بندی دستور/مدل: هر تغییر در دستور، وزن‌ها یا پارامترهای کانتور باید شماره نسخه داشته باشد تا اثرات قابل‌رهگیری باشند.
۲. پایش کیفیت در میدان: به‌صورت دوره‌ای، نمونه‌های واقعی را با چندمرجع داخلی ارزیابی و انحراف از کیفیت هدف را ثبت کنید.
۳. مسیرهای آینده: ارزیابی انسانی کنترل‌شده در سناریوهای بحرانی، آزمون «بدون کانتور» به‌عنوان روش پایه مقایسه‌ای، و بررسی پنجره‌های زمانی لغزان برای غنای زمینه‌ای پیشنهاد می‌شود.

- جمع‌بندی

با محدودسازی ورودی (فریم/ابعاد)، اعمال قید گراندینگ از طریق کانتور توجه، رعایت سیاست ایجاز در تولید، و پیاده‌سازیِ یکپارچه با پایش و ثبت کافی، ماژول توضیح‌گر متنی می‌تواند بدون خدشه به مسیر تصمیم‌گیریِ ADAPT، شفافیت ادراکی سامانه را افزایش دهد و مبنایی عملی برای ممیزی و اعتماد فراهم سازد.

۵–۶– مرزبندی و حدود اعتبار نتایج

یافته‌های این پژوهش در چهارچوبی مشخص به‌دست آمده‌اند و دامنه اعتبار آن‌ها به همان چارچوب مقید است. نخست، نتایج بر داده BDD-X (نمای جلوی خودرو، سناریوهای خیابانی) و پیکربندی‌های تعیین‌شده در فصل‌های سوم و چهارم بنا شده‌اند؛ ورودی زبانی/تصویریِ LLM ها به $n \in \{1,4,8,12,16\}$ فریم نخست هر کلیپ محدود بوده و نشانه بصریِ «توجه» به‌صورت کانتور به مدل زبانی تزریق شده است. بنابراین، تعمیم به سناریوهای حسگریِ دیگر (چندسنسوری، نماهای جانبی/عقبی، شرایط نادیده) یا پنجره‌های زمانی بلندتر/لغزان نیازمند ارزیابی مجدد است. همچنین سنجه‌های ارزیابی عمدتاً متنی‌اند (BLEU-٤، METEOR، ROUGE-L، CIDEr-D، SPICE) و

کیفیت «توضیح» را از منظر هم‌پوشانی زبانی می‌سنجند؛ در نتیجه، برداشت از «توضیح‌پذیری» باید چندمعیاره و مکمل با ارزیابی‌های کیفی/انسانی تفسیر شود (نگاه کنید به **پیوست ب** و ۴–۱۱).

دوم، در پروتکل دستور، تصمیم ADAPT به‌عنوان درست مفروض شده و نقش LLM «توجیه/تبیین» آن تصمیم است؛ لذا نتایج مستقیماً درباره «درستی تصمیم» داوری نمی‌کنند، بلکه درباره هم‌خوانی بصری–متنیِ دلیلِ تولیدشده اظهار نظر دارند. افزون بر این، اجرای آزمایش‌ها بر بسترهای ابری انجام شده و بخشی از منابع FT و زمان پاسخِ سرویس‌ها نامشهود است؛ نوسان زمانی گزارش‌شده (کنترل‌شده در سطح corpus) در ۴–۱۱ مستندسازی شده و احتمال اثرات جزئیِ اجرای تکی (single-run) باقی است. در خصوص معیارها نیز ناهمسان‌رفتاری مشاهده شد (مثلاً CIDEr-D)، ازاین‌رو نتیجه‌گیری‌ها بر مجموعه سنجه‌ها استوار است نه یک شاخص منفرد.

در جمع‌بندی، نتایج این فصل برای سامانه ترکیبی «ADAPT + LLM» با ورودی‌ها و رژیم ارزیابیِ مشخصِ این پژوهش معتبَر است و برای تعمیم به محیط‌ها/سخت‌افزارها/پروتکل‌های ارزیابی متفاوت باید بازارزیابی تجربی انجام شود. مرزهای یادشده به‌تفصیل در ۴–۱۱ (محدودیت‌ها و تهدیدهای اعتبار) آمده و پیشنهادهای لازم برای تقویت اعتبار بیرونی (گسترش داده، ارزیابی انسانی کنترل‌شده، سناریوی «بدون کانتور توجه»، پنجره زمانی لغزان) در همان فصل ارائه شده است.

۵–۷– پیشنهادهای پژوهشی آینده

در ادامه، مسیرهایی پیشنهاد می‌شود که می‌تواند اعتبار بیرونی، تبیینی و کارایی سامانه «ADAPT + LLM» را ارتقا دهد و تصویر کامل‌تری از کیفیت توضیح فراهم آورد:

۱. ارزیابی انسانیِ کنترل‌شده: طراحی یک مطالعه کاربر با داوران خبره/نیمه‌خبره و سنجه‌های شفاف (درستی، کفایت، اختصار، اتکا به شواهد تصویری) به‌صورت Likert یا مقایسه دوتایی؛ گزارش توافق بین‌داوران (مثلاً κ) و همبستگی با سنجه‌های خودکار برای سنجش روایی.

۲. مطالعه حذف مؤلفه‌ها برای کانتور توجه: اجرای سناریوهای «بدون کانتور»، «کانتور با آستانه/ضخامت متفاوت»، و «نمایش پیوسته نقشه توجه» و مقایسه نظام‌مندِ نتایج در هر دو رژیم تک‌مرجع/چندمرجع؛ هدف، کمی‌سازی دقیق سهم قیود گراندینگ در کیفیت توضیح.

۳. گسترش داده و سناریو: ارزیابی بر داده‌های متنوع‌تر (شرایط جوی/نوری گوناگون، بافت‌های ترافیکی متفاوت، ویدئوهای طولانی‌تر، نماهای غیرِ جلو) و (در صورت امکان) حسگرهای افزوده (چنددوربینه/چندسنسوری) برای تقویت تعمیم‌پذیری.

٤. پنجره‌های زمانی طولانی‌تر/لغزان. بررسی پنجره‌های لغزان[۱۲۷] با طول و گام‌های متفاوت و تحلیل موازنه «اطلاعات زمانی غنی‌تر در برابر تأخیر و هزینه»؛ گزارش اثر بر هم‌خوانی بصری–متنی و سنجه‌های متنی.

٥. سنجه‌های مکمل برای توضیح‌سنجی: افزودن سنجه‌های هم‌معنایی/استلزام و گراندینگ (مثلاً مقایسه اشاره‌های متنی با نواحی علامت‌گذاری‌شده، یا محاسبه هم‌پوشانی اشاره متنی با نقشه/کانتور توجه) در کنارBLEU-٤/METEOR/ROUGE/CIDEr-D/SPICE ، تا تصویری چندبعدی‌تر از «کیفیت توضیح» به دست آید.

٦. پایداری و رُباسـت[۱۲۸]: سنجش حساسیت خروجی‌ها نسبت به تغییرات کوچک در دستور، تعداد فریم‌ها و نویز تصویری؛ ارزیابی یکنواختی سبک و معنا در بازتکرارها و تخمین عدم‌قطعیتِ متنی[۱۲۹] برای استفاده ایمن‌تر.

۷. بهبود ادغام سامانه‌ای: طراحی الگوهای کَش و صف‌بندی در استنتاج انبوه، ثبت دقیق نسخه‌های دستور/مدل/پارامترهای کانتور، و پایش میدانی کیفیت به‌همراه سازوکار بازخورد انسانی جهت بهبود پیوسته.

---

[۱۲۷] Sliding

[۱۲۸] Robustness

[۱۲۹] Confidence

۸. بهبود روش‌های یادگیری: بررسی اتصال‌های سبک‌وزن میان نمایش‌های بصری ADAPT و LLM (adapter/connector) یا آموزش مشترک محدود (بدون تغییر ماهیت تصمیم‌گیر) با قید ایمنی و جداسازی نقش‌ها؛ هدف، افزایش وفاداری دلیل به شواهد.

۹. اخلاق، حریم خصوصی و جانبداری: توسعه رویه‌های ناشناس‌سازی پایدار، پایش سوگیری‌های احتمالی در توضیح‌ها (مثلاً سوگیری نسبت به اشیاء/صحنه‌های خاص) و تعریف خط‌مشی‌های انتشار/استفاده داده مطابق استانداردهای صنعت.

۱۰. کارایی و هزینه: تدوین بودجه تأخیر/هزینه و بررسی راهبردهای کم‌هزینه استنتاج (پردازش دسته‌ای، مسیرهای ارزان پلتفرم، فشرده‌سازی ورودی) همراه با سنجش اثر آن‌ها بر کیفیت توضیح؛ هدف، دستیابی به نقطه کار بهینه کیفیت–هزینه برای استقرار عملی.

این مسیرها—با حفظ هسته روش حاضر (تصمیم ADAPT، قید گراندینگ از طریق توجه، و تولید موجز LLM)—می‌توانند به استانداردسازی ارزیابی، افزایش اعتمادپذیری و تسهیل استقرار صنعتیِ توضیح‌های متنی در سامانه‌های خودران کمک کنند.

۵–۸– جمع‌بندی نهایی

این پایان‌نامه با صورت‌بندی و ارزیابیِ یک سامانه ترکیبی برای تبیین تصمیمات رانندگی—متشکل از مدل تصمیم‌گیر ADAPT و ماژول توضیح‌گر متنی مبتنی بر—LLM نشان داد که می‌توان با اتکا به نشانه‌گذاری توجه (نمایش کانتوری نواحی پرتوجه) شکاف میان «آنچه سامانه می‌بیند» و «آنچه توضیح می‌دهد» را کاهش داد و هم‌خوانی بصری–متنیِ توضیح‌ها را به‌طور معناداری ارتقا بخشید. طراحی دستور موجز و کنترل‌شده، انتخاب پنجره زمانی سازگار با قیود تأخیر، و ارزیابی چندمعیاره در دو رژیم تک‌مرجع/چندمرجع، چارچوبی منسجم برای سنجش «کیفیت توضیح» فراهم کرد و نشان داد که LLM ها—به‌ویژه پس از آموزش تکمیلی—در تولید توضیح‌های موجز، هم‌معنا با مرجع انسانی و هم‌راستا با شواهد تصویری توانمندند؛ در عین حال، رفتار متفاوت معیارها (مثلاً CIDEr-D در

برابر METEOR/ROUGE-L) ضرورت قرائت چندمعیاره و پرهیز از اتکای تک‌معیاری را برجسته ساخت.

از منظر سهم علمی، پیوند «توضیح زبانی LLM + توجه ADAPT» به‌مثابه قید گراندینگ، سازوکاری عملی و قابل‌ممیزی برای افزایش شفافیت فراهم آورد و مسیر کاربردیِ تلفیق تبیین با سامانه‌های ادراکی را روشن کرد. از منظر مهندسی، چارچوب اجرا—با استانداردسازی پیش‌پردازش، ثبت نسخه‌ها، کنترل هزینه/تأخیر و گزارش‌پذیری نشان داد که می‌توان ماژول توضیح‌گر را بدون مداخله در مسیر کنترل و با حداقل سربار عملیاتی کنار مدل تصمیم‌گیر مستقر کرد. ثبت مسیر ناموفقِ طبقه‌بندی سرعت/جهت نیز، ضمن شفاف‌سازی حدود مسئله، بر این نکته تأکید کرد که «برچسب‌گذاری گسسته» الزاماً جایگزین مناسبی برای «توضیح علّی» نیست.

در کنار این دستاوردها، محدودیت‌های روش—از جمله وابستگی به دادهBDD-X ، پنجره زمانی محدود، و نامشهودبودن برخی جزئیات در پلتفرم‌های آموزش تکمیلی—صادقانه مستندسازی شد و برای تقویت اعتبار بیرونی و تبیینیِ نتایج، مسیرهای آینده‌ای چون ارزیابی انسانی کنترل‌شده، سنجش سناریوی بدون کانتور توجه، گسترش داده/حسگر، پنجره‌های زمانی لغزان و معیارهای مکمل گراندینگ پیشنهاد گردید.

به این ترتیب، فصل پنجم پرونده استدلالی پایان‌نامه را می‌بندد: روش پیشنهادی، جایگاه خود را هم در ادبیات هوش مصنوعی توضیح‌پذیر برای سامانه‌های خودران و هم در مسیر استقرار صنعتیِ ماژول‌های توضیح‌گر تثبیت می‌کند. امید است این چارچوب—با توسعه‌های آتی—به استانداردسازی سنجش تبیین، افزایش اعتمادپذیری و ارتقای مسئولیت‌پذیری سامانه‌های خودران یاری رساند.

فهرست منابع

[۱] A. B. Arrieta *et al*., "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information fusion,* vol. ۵۸, pp. ۸۲–۱۱۵, .۲۰۲۰

[۲] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems,* vol. ۳۰, .۲۰۱۷

[۳] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization ",in *Proceedings of the IEEE international conference on computer vision*, ۲۰۱۷, pp. ۶۱۸– .۶۲۶

[۴] M. T. Ribeiro, S. Singh, and C. Guestrin, "" Why should i trust you?" Explaining the predictions of any classifier," in *Proceedings of the ۲۲nd ACM SIGKDD international conference on knowledge discovery and data mining*, ۲۰۱۶, pp. ۱۱۳۵– .۱۱۴۴

[۵] P. Mavrepis, G. Makridis, G. Fatouros, V. Koukos, M. M. Separdani, and D. Kyriazis, "XAI for All: Can Large Language Models Simplify Explainable AI?," *arXiv preprint arXiv*:۲۴۰۱.۱۳۱۱۰, .۲۰۲۴

[۶] F. Siddiqui, " ۱۷fatalities, ۷۳۶crashes: The shocking toll of Tesla's Autopilot," in *The Washington Post*, ed, .۲۰۲۳

[۷] T. Partners. "Can Tesla be Held Liable For a Car Accident Caused by the Autopilot?" Tekedia. https://www.tekedia.com/can-tesla-be-held-liable-for-a-car-accident-caused-by-the-autopilot/?srsltid=AfmBOoposHNUcxfa۳sKFo_۳jvO۲۳J_۰RfBvZ۱a۵qtJ۷YxBYVvqZPooZ#۷:~:text=Public۲۰%perception۲۰%and۲۰%trust۲۰%in,te

chnology٪E٩٩٪٨٠٪٢s٢٠٪readiness٢٠٪for٢٠٪widespread٢٠٪use
(accessed.

[٨] R. Mitchell, "Your Tesla could explain why it crashed. But good luck getting its Autopilot data," in *Los Angeles Times*, ed, .٢٠٢٠

[٩] B. Jin *et al*., "Adapt: Action-aware driving caption transformer," in ٢٠٢٣*IEEE International Conference on Robotics and Automation (ICRA)*, ٢٠٢٣: IEEE, pp. ٧٥٥٤– .٧٥٦١

[١٠] A. Bhattacharjee, R. Moraffah, J. Garland, and H. Liu, "LLMs as Counterfactual Explanation Modules: Can ChatGPT Explain Black-box Text Classifiers?," *arXiv preprint arXiv*:٢٣٠٩.١٣٣٤٠, .٢٠٢٣

[١١] X. Wu *et al*., "Usable xai: ١٠strategies towards exploiting explainability in the llm era," *arXiv preprint arXiv*:٢٤٠٣.٠٨٩٤٦, .٢٠٢٤

[١٢] J. Gu *et al*., "A systematic survey of prompt engineering on vision-language foundation models," *arXiv preprint arXiv*:٢٠٢٣ ,٢٣٠٧.١٢٩٨٠

[١٣] O. Rubin, J. Herzig, and J. Berant, "Learning to retrieve prompts for in-context learning," *arXiv preprint arXiv*:٢١١٢.٠٨٦٣٣, .٢٠٢١

[١٤] J. Wei *et al*., "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. ٣٥, pp. ٢٤٨٢٤–٢٤٨٣٧, .٢٠٢٢

[١٥] Q. Dong *et al*., "A survey on in-context learning," *arXiv preprint arXiv*:٢٣٠١.٠٠٢٣٤, .٢٠٢٢

[١٦] *Regulation (EU)* ١٦٨٩/٢٠٢٤*of the European Parliament and of the Council of* ١٣*June* ٢٠٢٤ *laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*, ١٦٨٩/٢٠٢٤, .٢٠٢٤

[١٧] Z. Dai and J. Callan, "Deeper text understanding for IR with contextual neural language modeling," in *Proceedings of the* ٤٢*nd international ACM SIGIR conference on research and development in information retrieval*, ٢٠١٩, pp. ٩٨٥– .٩٨٨

[۱۸] C. Zhao, W. Qian, Y. Shi, M. Huai, and N. Liu, "Automated Natural Language Explanation of Deep Visual Neurons with Large Models," *arXiv preprint arXiv*:۲۰۲۳ ,۲۳۱۰.۱۰۷۰۸

[۱۹] P. H. Winston, "Learning and reasoning by analogy," *Communications of the ACM,* vol. ۲۳, no. ۱۲, pp. ۶۸۹–۷۰۳, .۱۹۸۰

[۲۰] L. Chen *et al*., "Driving with llms: Fusing object-level vector modality for explainable autonomous driving," *arXiv preprint arXiv*:۲۳۱۰.۰۱۹۵۷*,* .۲۰۲۳

[۲۱] M. Bojarski, "End to end learning for self-driving cars," *arXiv preprint arXiv*:۱۶۰۴.۰۷۳۱۶*,* .۲۰۱۶

[۲۲] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "Spice: Semantic propositional image caption evaluation," in *European conference on computer vision*, ۲۰۱۶: Springer, pp. ۳۸۲– .۳۹۸

[۲۳] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, ۲۰۱۵, pp. ۴۵۶۶– .۴۵۷۵

[۲۴] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the ۴۰th annual meeting of the Association for Computational Linguistics*, ۲۰۰۲, pp. ۳۱۱– .۳۱۸

[۲۵] C.-Y. Lin and F. J. Och, "Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics," in *Proceedings of the ۴۲nd annual meeting of the association for computational linguistics (ACL-۰۴)*, ۲۰۰۴, pp .۶۱۲–۶۰۵ .

[۲۶] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation*

measures for machine translation and/or summarization, ۲۰۰۵, pp. ۶۵– .۷۲

[۲۷] A. W. Lee *et al*., "Semantic Integrity Constraints: Declarative Guardrails for AI-Augmented Data Processing Systems," *arXiv preprint arXiv:*۲۵۰۳.۰۰۶۰۰, .۲۰۲۵

[۲۸] C. Rivera *et al*., "Conceptagent: Llm-driven precondition grounding and tree search for robust task planning and execution," in ۲۰۲۵*IEEE International Conference on Robotics and Automation (ICRA)*, ۲۰۲۵: IEEE, pp. ۸۹۸۸– .۸۹۹۵

[۲۹] C. Patrício, J. C. Neves, and L. F. Teixeira, "Explainable deep learning methods in medical image classification: A survey," *ACM Computing Surveys,* vol. ۵۶, no. ۴, pp. ۱–۴۱, .۲۰۲۳

[۳۰] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in ۲۰۱۸*IEEE ۵th International Conference on data science and advanced analytics (DSAA)*, ۲۰۱۸: IEEE, pp. ۸۰– .۸۹

[۳۱] S. Atakishiyev, M. Salameh, H. Yao, and R. Goebel, "Explainable artificial intelligence for autonomous driving: A comprehensive overview and field guide for future research directions," *IEEE Access,* .۲۰۲۴

[۳۲] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:*۱۷۰۲.۰۸۶۰۸, .۲۰۱۷

[۳۳] A. Bilal, D. Ebert, and B. Lin, "Llms for explainable ai: A comprehensive survey," *arXiv preprint arXiv:*۲۵۰۴.۰۰۱۲۵, .۲۰۲۵

[۳۴] S. Han, M. Wang, J. Zhang, D. Li, and J. Duan, "A Review of Large Language Models: Fundamental Architectures, Key Technological Evolutions, Interdisciplinary Technologies Integration, Optimization and Compression Techniques, Applications, and Challenges,"

Electronics, vol. ۱۳, no. ۲۴, p. ۵۰۴۰, ۲۰۲۴. [Online]. Available: https://www.mdpi.com/.۵۰۴۰/۲۴/۱۳/۹۲۹۲-۲۰۷۹

[۳۵] A. Vaswani *et al*., "Attention is all you need," *Advances in neural information processing systems,* vol. ۳۰, .۲۰۱۷

[۳۶] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI blog,* vol. ۱, no. ۸, p. ۹, .۲۰۱۹

[۳۷] T. Brown *et al*., "Language models are few-shot learners," *Advances in neural information processing systems,* vol. ۳۳, pp. ۱۸۷۷–۱۹۰۱, .۲۰۲۰

[۳۸] C. Raffel *et al*., "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of machine learning research,* vol . ,۲۱no. ۱۴۰, pp. ۱–۶۷, .۲۰۲۰

[۳۹] J. Wei *et al*., "Finetuned language models are zero-shot learners. arXiv ۲۰۲۱," *arXiv preprint arXiv*:۲۱۰۹.۰۱۶۵۲, .۲۰۲۳

[۴۰] Y. Sun *et al*., "Ernie ۳.۰: Large-scale knowledge enhanced pre-training for language understanding and generation," *arXiv preprint arXiv*:۲۱۰۷.۰۲۱۳۷, .۲۰۲۱

[۴۱] A. Chowdhery *et al*., "Palm: Scaling language modeling with pathways," *Journal of Machine Learning Research,* vol. ۲۴, no. ۲۴۰, pp. ۱–۱۱۳, .۲۰۲۳

[۴۲] J. Achiam *et al*., "Gpt- ۴technical report ",*arXiv preprint arXiv*:۲۳۰۳.۰۸۷۷۴, .۲۰۲۳

[۴۳] G. Team *et al*., "Gemini: a family of highly capable multimodal models," *arXiv preprint arXiv*:۲۳۱۲.۱۱۸۰۵, .۲۰۲۳

[۴۴] Z. Xu *et al*., "Drivegpt۴: Interpretable end-to-end autonomous driving via large language model ",*IEEE Robotics and Automation Letters,* .۲۰۲۴

[۴۵] X. Yang and X. Wang, "Language model as visual explainer," *Advances in Neural Information Processing Systems*, vol. ۳۷, pp. ۱۳۵۰۹۴–۱۳۵۱۲۸, ۲۰۲۴.

[۴۶] M. Mesinovic, P. Watkinson, and T. Zhu, "Explainability in the age of large language models for healthcare," *Communications Engineering*, vol. ۴, no. ۱, p. ۱۲۸, ۲۰۲۵.

[۴۷] "GPT-۴." https://openai.com/index/gpt-۴-research/ (accessed.

[۴۸] B. Chen, Z. Zhang, N. Langrené, and S. Zhu, "Unleashing the potential of prompt engineering in large language models: a comprehensive review," *arXiv preprint arXiv:۲۳۱۰.۱۴۷۳۵*, ۲۰۲۳.

[۴۹] G. Team. "Prompt engineering: overview and guide." https://cloud.google.com/discover/what-is-prompt-engineering (accessed.

[۵۰] Y. Mao, J. He ,and C. Chen, "From prompts to templates: A systematic prompt template analysis for real-world LLMapps," in *Proceedings of the ۳۳rd ACM International Conference on the Foundations of Software Engineering*, ۲۰۲۵, pp. ۷۵– ۸۶.

[۵۱] M. Zheng, J. Pei, L. Logeswaran, M. Lee, and D. Jurgens, "When" a helpful assistant" is not really helpful: Personas in system prompts do not improve performances of large language models," in *Findings of the Association for Computational Linguistics: EMNLP* ۲۰۲۴, ۲۰۲۴, pp. ۱۵۱۲۶– ۱۵۱۵۴.

[۵۲] X. Amatriain, "Prompt design and engineering: Introduction and advanced methods," *arXiv preprint arXiv:۲۴۰۱.۱۴۴۲۳*, ۲۰۲۴.

[۵۳] K. Chang *et al.*, "Efficient prompting methods for large language models: A survey," *arXiv preprint arXiv:۲۴۰۴.۰۱۰۷۷*, ۲۰۲۴.

[۵۴] Z. Yu, X. Ouyang, Z. Shao, M. Wang, and J. Yu, "Prophet: Prompting large language models with complementary answer heuristics for

knowledge-based visual question answering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, ۲۰۲۵.

[۵۵] Nexla. "Prompt Engineering vs. Fine-Tuning—Key Considerations and Best Practices." https://nexla.com/ai-infrastructure/prompt-engineering-vs-fine-tuning (accessed.

[۵۶] J. Kim and J. Canny, "Interpretable learning for self-driving cars by visualizing causal attention," in *Proceedings of the IEEE international conference on computer vision*, ۲۰۱۷, pp. ۲۹۴۲–۲۹۵۰.

[۵۷] J. Kim, A. Rohrbach, T. Darrell, J. Canny, and Z. Akata, "Textual explanations for self-driving vehicles," in *Proceedings of the European conference on computer vision (ECCV)*, ۲۰۱۸, pp. ۵۶۳–۵۷۸.

[۵۸] K. Ding *et al.*, "Hint-ad: Holistically aligned interpretability in end-to-end autonomous driving," *arXiv preprint arXiv*:۲۴۰۹.۰۶۷۰۲, ۲۰۲۴.

[۵۹] ن. حسین ابراهیم, "بازشناسی اتوماتیک علایم راهنمایی و رانندگی در تصویر از روی ویژگی خروجی فیلتر های گابور با استفاده از شبکه عصبی احتمالاتی," presented at the پانزدهمین کنفرانس بین المللی سالانه انجمن کامپیوتر ایران, ۱۳۸۸. [Online]. Available: https://civilica.com/doc/۷۹۱۱۷.

[۶۰] ا. سیدعلی رضوی, "بررسی و ارزیابی روشهای شناسایی علائم راهنمایی و رانندگی مبتنی بر پردازش تصویر," presented at the دومین کنفرانس ملی تصادفات جاده ای، سوانح ریلی و هوایی, ۱۳۹۰. [Online]. Available: https://civilica.com/doc/۱۳۲۶۱۰.

[۶۱] علیرضا خدایاری, "ارائه یک الگوریتم پردازش تصویر هوشمند جدید برای تشخیص و شناسایی علائم راهنمایی و رانندگی مبتنی بر منطق فازی," مهندسی مکانیک امیرکبیر, vol. سال پنجاهم, pp. ۲۰۷–۲۱۸, ۱۳۹۷.

[۶۲] م. ح. پور, "شبکه سریع و کم وزن برای شناسایی خطوط جاده با استفاده از معماری موبایل نت و توابع هزینه متفاوت," (in Farsi), نشریه مهندسی برق و مهندسی کامپیوتر ایران, vol. ۲۱, no. ۳, pp. ۲۰۲–۲۱۰, ۲۰۲۴. [Online]. Available: https://www.magiran.com/paper/۲۶۹۱۱۷۱.

[۶۳] ع. غفاری, "شبیه سازی ماشین های خودران با and ع. هویدافرد, س. ف. م. نیا, ب. گلچین, in Fa استفاده از روش های یادگیری ماشین," فصلنامه مهندسی حمل و نقل, (vol. ۱۵, no. ۲, pp. ۳۴۸۳–۳۵۰۷, ۲۰۲۳. [Online]. Available: https://www.magiran.com/paper/۲۶۹۲۱۷۳.

[۶۴] د. محمد. "هوش مصنوعی توضیح پذیر: گامی به سوی شفافیت و اعتماد در هوش مصنوعی." https://civilica.com/doc/۲۱۵۵۷۶۶/ (accessed.

[۶۵] م. فریدون فر, "یک روش هوش مصنوعی قابل توضیح در معماری کلان داده," ۱۴۰۴.

[۶۶] ح. اخوان تفتی, "تاثیر هوش مصنوعی توضیح پذیر بر تمایل به and س. ح. حسینی طاهری اشتراک گذاری دانش در سیستم های مدیریت ارتباط با مشتری تلفیق یافته با هوش مصنوعی," presented at the نوزدهمین کنفرانس ملی پژوهش های کاربردی در علوم برق، کامپیوتر و مهندسی پزشکی, ۱۴۰۳. [Online]. Available: https://civilica.com/doc/۲۰۲۳۳۴۵.

[۶۷] Ç .Kaymak and A. Uçar, "Semantic image segmentation for autonomous driving using fully convolutional networks," in ۲۰۱۹ *International Artificial Intelligence and Data Processing Symposium (IDAP)*, ۲۰۱۹: IEEE, pp. ۱– ۸.

[۶۸] H. Mankodiya, D. Jadav, R. Gupta ,S. Tanwar, W.-C. Hong, and R. Sharma, "OD-XAI: Explainable AI-Based Semantic Object Detection for Autonomous Vehicles," *Applied Sciences*, vol. ۱۲, no. ۱۱, p. ۵۳۱۰, ۲۰۲۲. [Online]. Available: https://www.mdpi.com/۲۰۷۶-۵۳۱۰/۱۱/۱۲/۳۴۱۷.

[۶۹] L. Cultrera, L .Seidenari, F. Becattini, P. Pala, and A. Del Bimbo, "Explaining autonomous driving by learning end-to-end visual attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, ۲۰۲۰, pp. ۳۴۰– ۳۴۱.

[۷۰] H. A. Tahir, W. Alayed, W. U. Hassan, and A. Haider, "A novel hybrid XAI solution for autonomous vehicles: Real-time interpretability through LIME–SHAP integration," *Sensors*, vol. ۲۴, no. ۲۱, p. ۶۷۷۶, ۲۰۲۴.

[۷۱] J. Yuan *et al*., "Rag-driver: Generalisable driving explanations with retrieval-augmented in-context learning in multi-modal large language model," *arXiv preprint arXiv*:۲۴۰۲.۱۰۸۲۸, .۲۰۲۴

[۷۲] É. Zablocki, H. Ben-Younes, P. Pérez, and M. Cord, "Explainability of deep vision-based autonomous driving systems :Review and challenges," *International Journal of Computer Vision,* vol. ۱۳۰, no. ۱۰, pp. ۲۴۲۵–۲۴۵۲, .۲۰۲۲

[۷۳] M. Bojarski *et al*., "Explaining how a deep neural network trained with end-to-end learning steers a car," *arXiv preprint arXiv*:۱۷۰۴.۰۷۹۱۱, .۲۰۱۷

[۷۴] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, and F. Viegas, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)," in *International conference on machine learning*, ۲۰۱۸: PMLR, pp. ۲۶۶۸– .۲۶۷۷

[۷۵] B. Ding *et al*., "Data augmentation using large language models: Data perspectives, learning paradigms and challenges," *arXiv preprint arXiv*:۲۴۰۳.۰۲۹۹۰, .۲۰۲۴

[۷۶] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition ",in *Proceedings of the IEEE conference on computer vision and pattern recognition*, ۲۰۱۶, pp. ۷۷۰– .۷۷۸

[۷۷] M. Kilickaya, A. Erdem, N. Ikizler-Cinbis, and E. Erdem, "Re-evaluating automatic metrics for image captioning," *arXiv preprint arXiv*:۱۶۱۲.۰۷۶۰۰, .۲۰۱۶

[۷۸] X. Chen *et al*., "Microsoft coco captions: Data collection and evaluation server," *arXiv preprint arXiv*:۱۵۰۴.۰۰۳۲۵, .۲۰۱۵

[۷۹] B. Pingua *et al*., "Medical LLMs: Fine-Tuning vs. Retrieval-Augmented Generation," *Bioengineering,* vol. ۱۲, no. ۷, p. .۲۰۲۵ ,۶۸۷

پیوست ها

(پیوست الف)

دستور استفاده شده برای تولید زیرنویس های افزوده:

You are a caption generation system for driving scenarios.

Your task is to take two captions: one describing an action and the other providing a justification for that action. From each caption, generate four additional captions that convey the same meaning but use different wording.

Your input will be a JSON object containing the original 'action' caption and the 'justification' caption. The action caption describes what the driving system is doing, while the justification caption explains why that decision was made.

Captions must be simple and brief, expressed in only one or two short sentences, similar to the original captions. Avoid including extra explanations, requests for more information, or any context.

For example:

- Original caption: 'The car makes a left turn at the intersection.'

- Generated caption: 'The car turns left.'

This format is correct because it retains the meaning without adding incorrect details. If the original 'action' or 'justification' is null, you will return null for that part as well.

Here is an example of how the input and output will look:

- Input: {'action': 'The car is moving at a steady speed', 'justification': 'because traffic is clear.'} - Output: {'action': ['The car moves forward', 'The car continues to drive', 'The car is driving down the road', 'The car is driving at normal speed'], 'justification': ['because

the street is clear', 'because the road is clear', 'as the road ahead is clear', 'since the roadway is clear.']}

(پیوست ب)

- **تحلیل کیفی و مطالعات موردی**

این بخش با تکیه بر رژیم چندمرجع (BDD-X + چهار بازنویسی هم‌معنا) و بر پایه نتایج فصل حاضر، رفتار سامانه را به‌صورت کیفی ارزیابی می‌کند. تمرکز بر خروجی مدل‌های FT (Gemini ۲٫۵ Pro و GPT-٤o) در کنار روش پایه ADAPT است تا هم توان توضیح‌گری و هم هم‌خوانی بصری–متنی روشن شود. موارد انتخابی از میان نمونه‌هایی برگزیده شده‌اند که امتیازهای ROUGE-L/METEOR چندمرجع در آن‌ها مرتفع (موفق) یا منخفض (ضدنمونه/خطا) بوده‌اند. در هر مورد، تصویر(های) کانتورگذاری‌شده، «اکشنِ ADAPT»، خروجی مدل و مرجع انسانی کنار هم تحلیل شده‌اند.

- **نمونه‌های موفق**

بر اساس بررسی های کیفی، الگوی مشترک در موارد موفق چنین است:

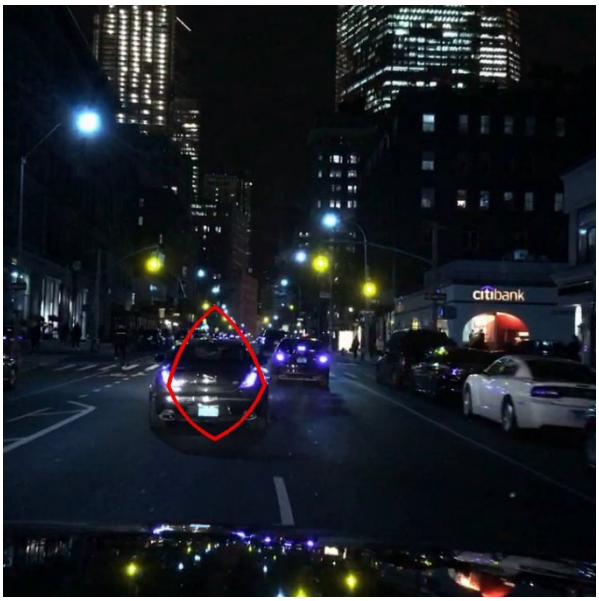
۱. زمینهٔ بصریِ روشن: شیء یا قرینه علّی (چراغ راهنمایی، خودروی جلویی آهسته، عابر، انسداد مسیر) در داخل کانتور قرمز به‌صورت واضح دیده می‌شود.
۲. هم‌خوانی معناییِ کوتاه‌گو: متن تولیدی به همان قرینه داخل کانتور ارجاع ضمنی می‌دهد و با مرجع انسانی هم‌معنا است (در ۱ یا ۲ جمله‌ی موجز، بدون حاشیه).
۳. پایداری با فریم‌های بیشتر: در مدل‌هایFT ، افزایش فریم (مطابق بخش‌های کمی) غالباً به غنای قرائن و بهبود ROUGE-L/METEOR منجر شده و در نمونه‌های موفق نیز این روند دیده می‌شود.

نمونه A (چراغ قرمز/توقف): کانتور روی چراغ قرمز و ناحیه تقاطع متمرکز است؛ اکشن ADAPT«کاهش سرعت/توقف».

تصویر ۱–۱– مثال نمونه A - چراغ قرمز / توقف (موفق)

- خروجی FT-Gemini Pro: «چون چراغ قرمز است.»
- مرجع انسانی: «به‌دلیل چراغ قرمز در تقاطع. »
- تحلیل: هم‌جانبگی دقیق میان متن و ناحیه توجه؛ طول و واژگان متن کوتاه و هم‌ارزش با مرجع.

نمونه B (پیروی از خودروی جلویی): کانتور روی خودروی جلویی و فاصله طولی.



تصویر ۱–۲– مثال نمونه B - پیروی از خودروی جلوییِ کند (موفق)

- خروجی FT-GPT-٤o : «به‌دلیل خودروی جلویی که آهسته حرکت می‌کند.»
- مرجع انسانی: «چون خودروی جلویی سرعت کمی دارد.»
- تحلیل: اشاره مستقیم به عامل علّی در کانتور (خودروی جلویی)، سازگاری معنایی با مرجع؛ عدم هذیان.

نمونه C (عابر/اولویت عبور): کانتور روی عابر در گذرگاه.

تصویر ۱–۳– مثال نمونه C - اولویت عبور عابر پیاده (موفق)

- خروجی FT-Gemini Pro : «برای عبور عابر پیاده.»
- مرجع انسانی: «به‌دلیل حضور عابر در گذرگاه.»
- تحلیل: متن دقیق و کوتاه؛ هم‌خوانی بصری–متنی کامل؛ رفتاری ایمن و قابل‌توجیه.

- **ضدنمونه‌ها/خطاها و تحلیل علل**

طبقات خطا که در موارد با امتیاز پایین مشاهده شد، عمدتاً به یکی از دسته‌های زیر تعلق دارند:

۱. عدم تطابق بصری-متنی[۱۳۰]: متن به عاملی اشاره می‌کند که در داخل کانتور دیده نمی‌شود (مثلاً اشاره به «عابر» وقتی کانتور روی خودروی جلویی است).

۲. ابهام صحنه و شرایط تصویری: شب/نور کم، باران/بازتاب‌ها، یا پوشیدگی[۱۳۱] باعث می‌شود قرینه علّی نامشخص یا کم‌کنتراست باشد؛ متن به حدس‌های کلی گرایش پیدا می‌کند.

---

[۱۳۰] Mis-grounding

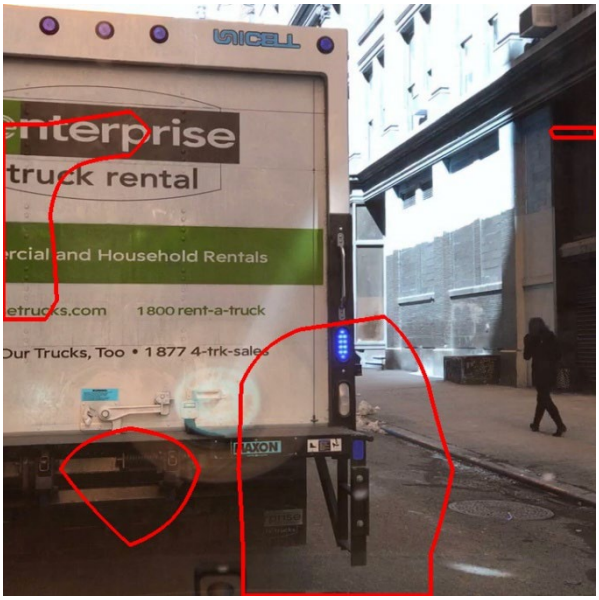
[۱۳۱] occlusion

۳. ناهم‌زمانی/کمبود زمینه زمانی: در برخی نمونه‌ها، سرنخ اصلی در فریم‌های غیرارسالی رخ می‌دهد؛ متن بر اساس نشانه‌های ناکافی شکل می‌گیرد (این امر با کاهش n شدیدتر است).

٤. نویز یا اختصار مرجع انسانی: مرجع BDD-X گاه بسیار کوتاه/کلی است؛ در نتیجه ممکن است متن مدل از نظر زبانی «بهتر» به نظر برسد اما به‌سبب دوری از عبارت مرجع، امتیاز کمی بگیرد (به‌ویژه در BLEU-٤/CIDEr-D).

٥. هذیان جزئی[۱۳۲]: افزودن جزئیاتی که در تصویر نیست (مثلاً «چراغ سبز» در حالی‌که چراغ دیده نمی‌شود). این موارد در نسخه‌های FT کاهش یافته اما صفر نشده‌اند.

نمونه D (عدم تطابق بصری-متنی): کانتور بر خودروی پارک‌شده؛ اکشن «کاهش سرعت».



تصویر ۱–۴– مثال ضدنمونه D - عدم تطابق بصری-متنی

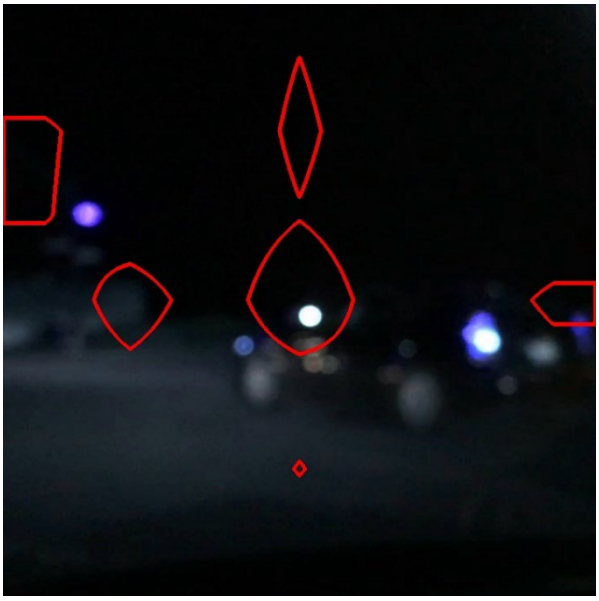
- خروجی مدل: «برای عبور عابر.»
- مرجع انسانی: «به‌دلیل خودروی جلویی که مسیر را تنگ کرده.»

---

[۱۳۲] Hallucination

- تحلیل: ارجاع متنی به «عابر» با نشانه بصری موجود در کانتور سازگار نیست؛ نمونه کلاسیک عدم تطابق بصری-متنی.

نمونه E (ابهام نوری/شب): کانتور پراکنده و کم‌کنتراست؛ نور چراغ‌ها غالب.



تصویر ۱–۵– مثال ضدنمونه E - ابهام نوری/شب و کلی‌گویی متن

- خروجی مدل: «چون ترافیک سنگین است.»
- مرجع انسانی: «به‌دلیل توقف خودروی جلویی. »
- تحلیل: متن به کلی‌گویی متمایل شده؛ قرینه دقیق در تصویر به‌سختی قابل اتکا است.

- **هم‌خوانی متن تولیدی با ناحیه‌های توجه (مطابقت بصری-متنی)**

چارچوب ارزیابی کیفی: برای هر مورد، سه پرسش به‌عنوان چک‌لیست بررسی شده است:

۱. محل‌یابی علت در تصویر: آیا «شیء/قرینه ذکرشده» در متن، در داخل کانتور قرمز دیده می‌شود؟

۲. هم‌معنایی با مرجع انسانی: آیا متن مدل از نظر معنا با Justification انسانی هم‌ارزش است (ولو با واژگان متفاوت)؟
۳. کفایت و اختصار: آیا متن کوتاه، مستقیم و بی‌حاشیه است (مطابق قیود دستور) و از افزودن اطلاعات ناموجود پرهیز می‌کند؟

یافته کلی: در موارد موفق، پاسخ هر سه پرسش مثبت است؛ به‌ویژه FT-Gemini Pro بیشترین نرخ هم‌خوانی بصری-متنی را نشان می‌دهد و FT-GPT-٤o نیز عملکرد رقابتی دارد. در ضدنمونه‌ها، بیشترین تخطی از بند (۱) رخ می‌دهد. (Mis-grounding) هم‌چنین در صحنه‌های مبهم/کم‌نور یا با nکوچک، بند (۲) آسیب‌پذیرتر می‌شود (افت هم‌معنایی با مرجع انسانی). مطابق نتایج کمی فصل، افزایش فریم‌های ورودی و استفاده از چندمرجع، در مجموع به بهبود هم‌خوانی انجامیده و بروز خطاهای هذیانی/کلی‌گویی را کاهش داده است.

- **الگوهای خطای پرتکرار**

بر اساس نمونه‌های خروجی و ماهیت برچسب‌گذاری، الگوهای خطای غالب چنین‌اند:

- ابهام‌های معنایی نزدیک: SLIGHTLY_LEFT با STRAIGHT_FORWARD و SLIGHTLY_RIGHT؛ SLOWLY_STOP با STOP و INCHES_FORWARD .
- کلاس‌های کم‌نمونه/درازدُم: U_TURN ، K_TURN، PULL_OVER و تغییر خط (GO_TO_LEFT/RIGHT_LANE) با بسامد کمتر و نشانه‌های بصریِ ضعیف‌تر.
- وابستگی زمانیِ بلندمدت: تمایز برخی کلاس‌ها به نشانه‌های دیرتر در کلیپ وابسته است؛ با کاهش تعداد فریم‌های ورودی، مدل به حدس‌های محافظه‌کارانه متمایل می‌شود.
- نوفه‌ی برچسب‌های مشتق از متن: نگاشت "کلاس → متنیِ زیرنویس" با LLM گاهی چندمعنایی یا ناسازگار با زمینه بصری است و به مدل سیگنال مبهم می‌دهد.

- از دست رفتن دقت پیوسته: تبدیل مسئله ذاتاً پیوسته (سرعت/زاویه) به طبقه‌بندی گسسته باعث از دست رفتن اطلاعات جزئی و ایجاد مرزهای مصنوعی می‌شود.

(پیوست پ)

- **تلاش‌های ناموفق: بازطراحی مدل بر اساس کلاس‌بندی تصمیم‌ها**

در این مرحله تلاش شد تا به‌جای تولید توضیحات متنی آزاد، مدل به‌صورت یک طبقه‌بند برای تصمیم‌های رانندگی بازطراحی شود. بدین منظور، اقدام‌های پیوسته رانندگی به دو دسته مجزای کلاس‌های سرعت و کلاس‌های جهت/مسیر تبدیل گردید. هر تصمیم رانندگی که پیش‌تر به‌صورت جمله توصیف می‌شد، اکنون به یک کلاس در بُعد سرعت و یک کلاس در بُعد جهت نگاشت می‌یابد. ایده اصلی این بود که با ساده‌سازی فضای خروجی به مجموعه‌ای از برچسب‌های محدود، مدل بتواند تصمیم‌ها را با دقت بهتری طبقه‌بندی کند. با اینکه این رویکرد پیچیدگی تولید زبان طبیعی را حذف می‌کرد، در نهایت دقت حاصل‌شده رضایت‌بخش نبود و این تلاش‌ها ناموفق ارزیابی شدند.

- **تبدیل اکشن‌ها به کلاس‌های سرعت و جهت**

برای طبقه‌بندی تصمیم‌ها، ابتدا مجموعه‌ای از کلاس‌های از پیش‌تعریف‌شده برای سرعت و جهت حرکت خودرو تعیین شد. کلاس‌های سرعت انواع حالت‌های شتاب و ترمز را شامل می‌شوند؛ از جمله:

- افزایش سرعت،
- حرکت با سرعت معمول،
- کاهش سرعت،
- حرکت خزشی به جلو،
- توقف آهسته،
- توقف کامل،
- حرکت آهسته به عقب،
- دنده عقب رفتن و

- دنبال کردن خودروی جلویی.

همچنین کلاس‌های جهت یا اقدام انواع مانورهای حرکتی خودرو را در بر می‌گیرند؛ مانند:

- حرکت مستقیم به جلو،
- پیچیدن به راست/چپ،
- تغییر به لاین راست/چپ،
- انحراف جزئی به راست/چپ،
- کنار زدن و توقف در حاشیه،
- دور زدن کامل[۱۳۳] و
- دور زدن سه‌نقطه‌ای[۱۳۴]

سایر موارد که حاوی زیرنویس های نادرست یا غیر قابل دسته بندی بودند نیز در یک کلاس خنثی قرار گرفتند تا در آموزش مورد استفاده قرار نگیرند. این کلاس‌ها به گونه‌ای انتخاب شدند که طیف کاملی از رفتارهای رانندگی تحت پوشش قرار گیرند. به عنوان نمونه، شرح متنیِ «خودرو سرعتش را کم می‌کند و برای چراغ قرمز می‌ایستد» به دو برچسب «توقف آهسته» در بُعد سرعت و «حرکت مستقیم به جلو» در بُعد جهت تبدیل می‌شود.

در این کار نیز ما توضیحات متنی رانندگی را به دو مؤلفه مجزای عددی (سرعت و جهت) تبدیل کردیم تا مدل بتواند آنها را یاد بگیرد. بدین ترتیب هر توضیح انسانی به یک جفت کلاس speed, action نگاشت شد و مسئله‌ی یادگیری مدل به شکل یک مسئله طبقه‌بندی دودوییِ چندخروجی[۱۳۵] در آمد که در آن دو برچسب همزمان پیش‌بینی می‌شوند (یکی برای سرعت و یکی برای جهت تصمیم رانندگی).

---

[۱۳۳] U-turn

[۱۳۴] K-turn

[۱۳۵] multi-output classification

- **پیاده‌سازی و ارزیابی مدل کیم و همکاران و ADAPT (به‌عنوان طبقه‌بند)**

برای ارزیابی این ایده، دو مدل شناخته‌شده در زمینه توضیح‌پذیری رانندگی را به عنوان طبقه‌بند تصمیم به کار گرفتیم: مدل Kim et al (۲۰۱۷)[۵۶] و مدل ADAPT (۲۰۲۳)[۹]. هریک از این مدل‌ها در اصل برای تولید توضیحات زبانی طراحی شده بودند، اما ما با اعمال تغییراتی، خروجی آن‌ها را به کلاس‌های از پیش‌تعریف‌شده سرعت و جهت محدود کردیم تا به‌عنوان یک طبقه‌بند عمل کنند.

- **مدل کیم و همکاران ۲۰۱۷**

شمای کلی این مدل در تصویر ۲–۱ نشان داده شده است. این معماری یک خط لوله ادراکی-کنترلیِ تفسیرپذیر برای رانندگی ارائه می‌دهد که در آن یک backbone دیداری، توصیف‌های مکانی–زمانی از فریم‌های ویدیو استخراج می‌کند، سپس یک LSTM با توجه علّی[۱۳۶] توالی را خلاصه می‌نماید و در خروجی زاویه فرمان را به‌صورت رگرسیون پیوسته پیش‌بینی می‌کند [۵۶]. ما برای سازگار کردن این مدل با ایده «طبقه‌بندی تصمیم»، شاخه خروجی رگرسیونی را بازطراحی کردیم: هد خروجیِ پیوسته حذف و به‌جای آن دو هد طبقه‌بندی قرار داده شد. یکی برای کلاس‌های سرعت و دیگری برای کلاس‌های جهت/مسیر (یادگیری چندوظیفه‌ای). تابع هزینه هر هد Cross-Entropy تعریف شد و در آموزش، مجموع/میانگین دو زیان به‌عنوان loss نهایی به‌کار رفت. امید ما این بود که حضور LSTM و توجه زمانی-مکانیِ مدل ۲۰۱۷ به بهبود دقت طبقه‌بندی کمک کند؛ بااین‌حال، علی‌رغم پایداری آموزش، دقت نهایی حدود ۶۰٪ باقی ماند.

- **مدل ADAPT ۲۰۲۳**

معماری ADAPT که در سال ۲۰۲۳ معرفی شده است، یک مدل مبتنی بر ترنسفورمر ویدیو برای توضیح‌پذیری خودران‌ها است. این مدل نیز به صورت مشترک دو وظیفه را انجام می‌دهد: تولید توضیحات متنیِ رفتار رانندگی و پیش‌بینی سیگنال‌های کنترلی وسیله نقلیه [۹]. در واقع ADAPT یک

---

[۱۳۶] causal attention

رمزگذار ویدیویی (مبتنی برSwin Transformer) را به کار می‌گیرد تا ویژگی‌های ویدیو را استخراج کند و سپس دو هد جداگانه بر پایه این ویژگی‌ها خروجی می‌دهند که یکی جمله‌ی توضیحی اقدام رانندگی را تولید می‌کند و دیگری دنباله کنترل‌های خودرو (مانند سرعت/شتاب) را پیش‌بینی می‌کند [۹]. نتایج گزارش‌شده برای ADAPT روی مجموعه‌داده BDD-X حاکی از بهبود عملکرد در هر دو بخش توصیف متن و پیش‌بینی کنترل بوده است [۹].

برای سازگار کردن این مدل با مسئله طبقه‌بندی تصمیم‌ها، رویکرد مشابهی با مدل قبلی اتخاذ شد؛ بدین ترتیب که هد متن‌ساز در مرحله استنتاج غیرفعال گردید و تنها هد کنترل مورد استفاده قرار گرفت. لایه‌های خروجی هد کنترلِ ADAPT نیز مطابق نیاز به دو طبقه‌بند (یکی برای سرعت و یکی برای جهت) تغییر یافتند و با تابع هزینه Cross-Entropy آموزش داده شدند. به خاطر پیچیدگی بیشتر ADAPTو ماهیت ترنسفورمری آن، انتظار می‌رفت که این مدل شاید در طبقه‌بندی تصمیم‌ها بهتر عمل کند. همچنین، به دلیل اشتراک ویژگی بین دو وظیفه در معماری اصلیADAPT، حذف یکی از وظایف (تولید متن) ممکن است به افت جزئی کارایی منجر شود؛ با این حال، ما فرض را بر این گذاشتیم که شبکه ویژگیِ ADAPT برای خروجی کنترل به اندازه کافی توانمند است که طبقه‌بندی دقیق‌تری ارائه دهد.

- **تهیه داده‌های آموزشی**

یک چالش در این بازطراحی، تهیه برچسب‌های کلاس سرعت و جهت برای هر نمونه از داده‌ها بود. توضیحات متنی موجود در مجموعه‌داده BDD-X به صورت جملات زبان طبیعی بیان شده‌اند و مستقیماً حاوی برچسب کلاس از پیش تعریف‌شده نیستند. برای تبدیل این توضیحات به برچسب‌های تعریف‌شده خود، از مدل‌های زبانی بزرگ بهره گرفتیم. به این صورت که برای هر جمله توضیحی، یک مدل زبانی (از طریق API) محتوای جمله را تحلیل کرده و نزدیک‌ترین کلاس سرعت و کلاس جهت متناسب با آن را انتخاب می‌کرد. این فرایند به شکل خودکار برای تمام زیرنویس‌های مجموعه‌داده

انجام شد تا یک مجموعه داده برچسب‌دار برای آموزش طبقه‌بند به‌دست آید. هرچند ممکن است برچسب‌گذاری خودکار کاملاً عاری از خطا نباشد، اما به دلیل هزینه‌بر بودن برچسب‌گذاری دستیِ هزاران نمونه، استفاده از LLM راهکار مناسبی برای تسریع این فرایند بود.

- **نتایج مسیر ناموفقِ طبقه‌بندی سرعت/جهت**

این بخش خلاصه‌ی نتایج مسیر جایگزینی را گزارش می‌کند که در آن، به‌جای تولید زیرنویس، تصمیم رانندگی به‌صورت دو مسئله طبقه‌بندی هم‌زمان مدل‌سازی شد:

الف) جهت/اقدام[۱۳۷] و ب) وضعیت سرعت[۱۳۸]. این مسیر را هم با مدل Kim et al., ۲۰۱۷ (شبکه بصری + LSTM برای دنباله فریم‌ها) و هم با تغییر ADAPT به طبقه‌بند (استفاده از سرِ Signal Prediction و جایگزینی لایه نهایی رگرسیون با دو سرِ classification) ارزیابی کردیم. تابع زیان برای هر سر Cross-Entropy و آموزش به‌صورت چندوظیفه‌ای (دو سرِ هم‌زمان) انجام شد؛ تنظیمات اجرایی مطابق فصل ۴ (بستر Colab، GPU های A۱۰۰/L٤، PyTorch ۲،٥،۱، CUDA ۱۲،٤) است.

- **تنظیمات دقیق و دقت کلی**

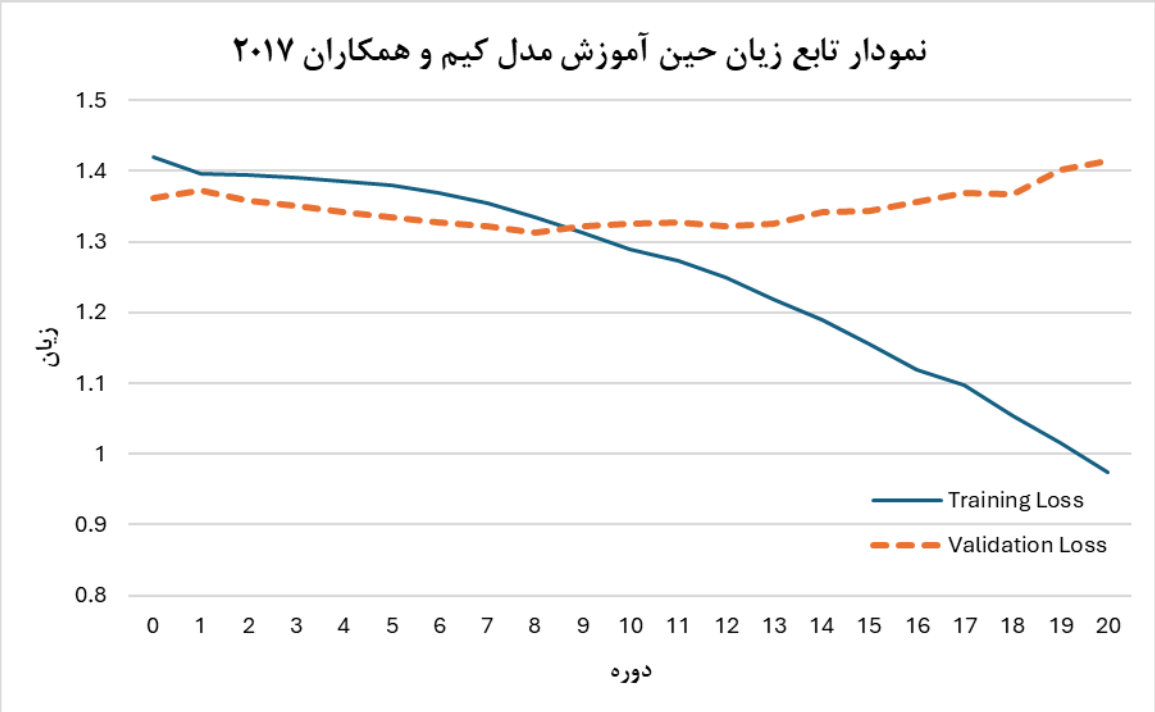
- نگاشت برچسب‌ها: زیرنویس‌های انسانیِ BDD-X به دو واژگان جهت و سرعت نگاشت شدند (به‌کمک LLM). این نگاشت، با وجود کنترل‌ها، دارای ابهام‌های معنایی و ناهماهنگی‌های موردی بود (مثلاً تمایز SLOWLY_STOP در برابر STOP یا SLIGHTLY_LEFT در برابر STRAIGHT_FORWARD).
- مدل Kim ۲۰۱۷[٥٦]: ورودی توالی فریم‌های پیش‌پردازش‌شده (مطابق۳-۴) و خروجی دو سرِ طبقه‌بندی بود.

---

[۱۳۷] Action

[۱۳۸] Speed

- دقت اعتبارسنجی[۱۳۹]: بر اساس نمودارهای ثبت‌شده، دقتِ سرِ Action در حوالی ۶۰–۶۴٪ و دقتِ سرِ Speed در حوالی ۴۶–۵۲٪ نوسان داشته است؛ دقت کل (میانگین دو سر) حدود ۵۴–۵۸٪. روند کاهش Loss/Action در آموزش دیده می‌شود، اما Loss/Action و Loss/Speed در اعتبارسنجی از یک نقطه به بعد افزایش/نوسان دارد. نشانه‌ای از بیش‌برازش و/یا نوفه‌ی برچسب‌ها[۱۴۰].

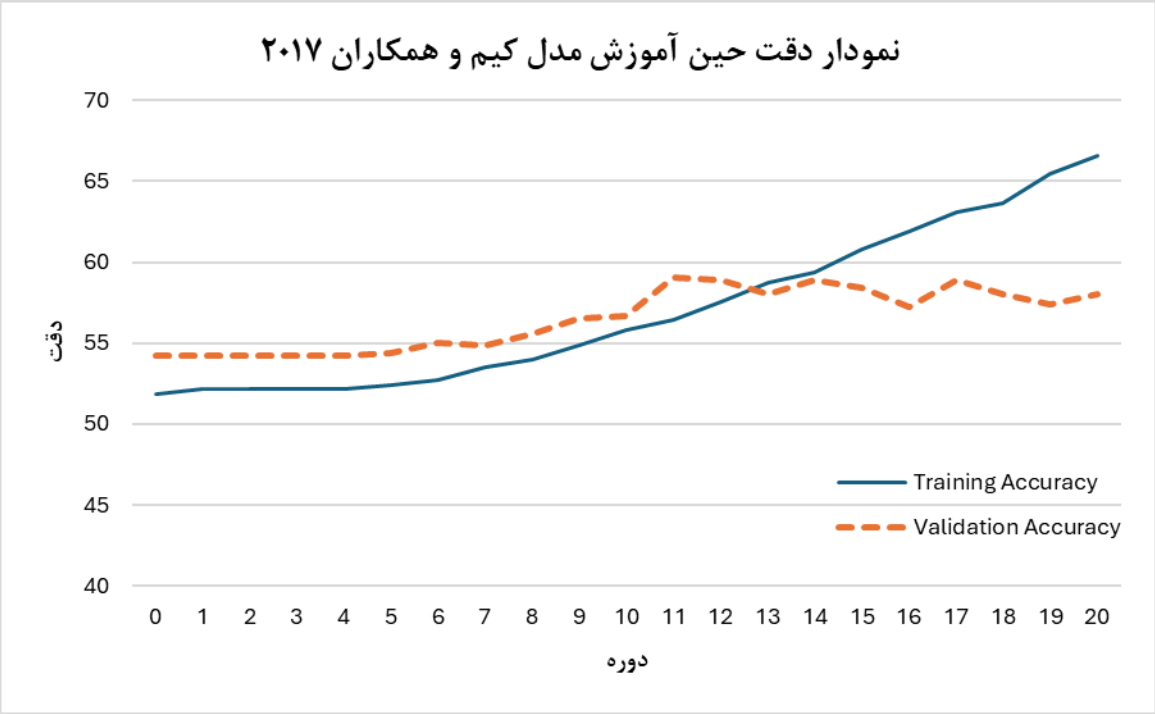


تصویر ۱–۶– همگرایی تابع زیان در مدل تغییر یافته کیم و همکاران ۲۰۱۷ حین آموزش

---

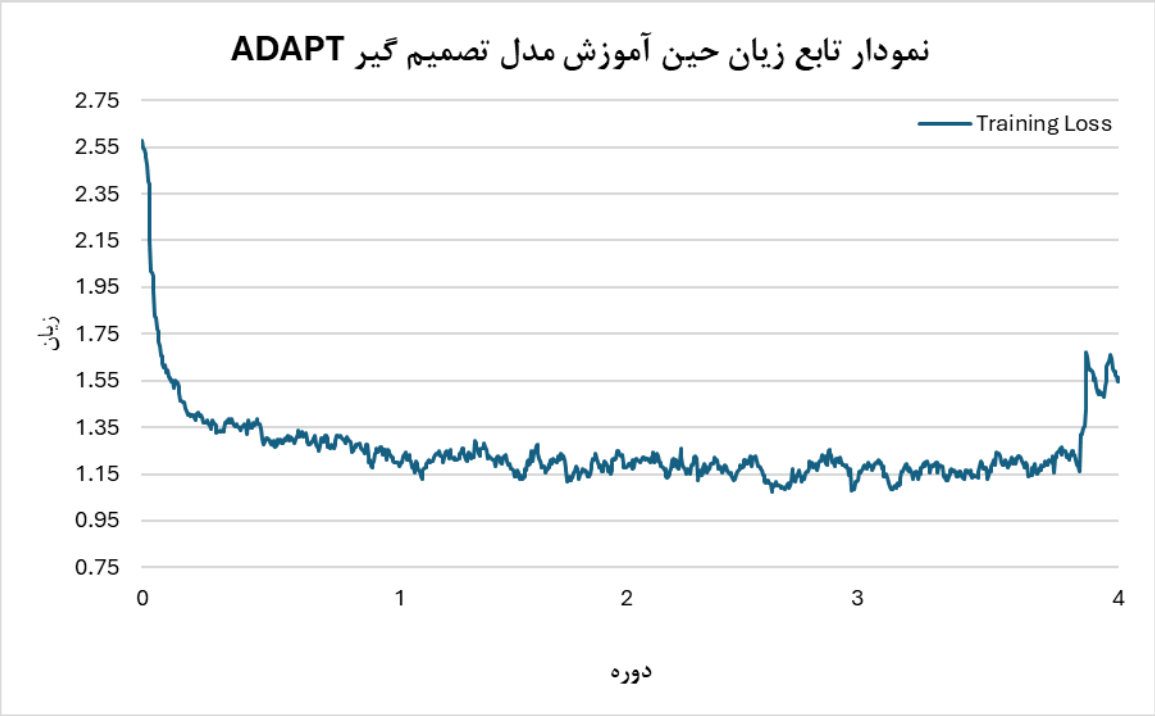
[۱۳۹] Validation Accuracy

[۱۴۰] label noise



تصویر پ–۲– نمودار دقت مدل تغییر یافته کیم و همکاران (۲۰۱۷) در حین آموزش

- ADAPT به‌صورت طبقه‌بند: در این مسیر، سرِ رگرسیونِ «سیگنال» ADAPT به دو سرِ طبقه‌بندیِ جهت/سرعت تبدیل شد.
  - نمودارهای آموزش نشان می‌دهد Loss در ابتدا کاهش می‌یابد، اما در ادامه (در حوالی گام‌های پایانی) افزایش پیدا می‌کند و دقت اعتبارسنجی به سقفی نزدیک به ۶۰٪ می‌رسد. این الگو نیز حاکی از کف عملکرد پایین و پایداری ناکافی است.



تصویر پ–۳– همگرایی تابع خطا در مدل تغییر یافته ADAPT حین آموزش

پس از آموزش مدل‌ها (هر دو مدل Kim و ADAPT) بر روی داده‌های برچسب‌خورده، عملکرد آن‌ها بر روی مجموعه‌آزمون ارزیابی شد. معیار ارزیابی، دقت طبقه‌بندی کلی بود که نشان‌دهنده درصد مواردی است که مدل توانسته هر دو برچسب سرعت و جهت را درست پیش‌بینی کند. دقت به‌دست‌آمده برای هر دو مدل در حدود ٪۶۰ بود. به بیان دیگر، مدل‌های بازطراحی‌شده تنها در ۶۰ درصد حالات قادر به تولید تصمیم کاملاً منطبق با توضیح انسانی بودند که رقم نسبتاً پایینی محسوب می‌شود. این سطح دقت بیانگر آن است که ساده‌سازی مسئله به طبقه‌بندی دودویی نیز نتوانست به پیش‌بینی کاملاً قابل اعتمادی منجر شود. در مجموع، با وجود به‌کارگیری مدل‌های پیشرفته پیشین و طراحی مجموعه‌ای جامع از کلاس‌های تصمیم، رویکرد طبقه‌بندی تصمیم‌ها موفقیت چشم‌گیری حاصل نکرد و دقت مدل در حدی نبود که بتوان آن را به عنوان راه‌حل نهایی انتخاب نمود. این یافته ما را به این نتیجه رساند که مسئله توضیح‌پذیری تصمیم‌های رانندگی نیازمند روش‌های غنی‌تری

(احتمالاً ترکیب اطلاعات متنی و عددی به‌صورت همزمان) است و رویکرد صرفاً مبتنی بر طبقه‌بندی ممکن است برای دستیابی به عملکرد مطلوب کافی نباشد.

- **جمع‌بندی چرایی کنارگذاری این مسیر**

۱. سقف عملکرد پایین و ناپایداری عمومی: هر دو پیاده‌سازی (Kim ۲۰۱۷ و نسخه طبقه‌بندی‌شده ADAPT) در بهترین حالت به ≈٪۶۰ دقت کل می‌رسند؛ روند Loss/val غیر یکنواخت و شواهد بیش‌برازش دیده می‌شود.

۲. هم‌سویی ضعیف با هدف پژوهش: هدف پایان‌نامه تولید توضیح علّیِ قابل‌فهم است. طبقه‌بندیِ دوگانه (سرعت/جهت) حتی در صورت بهبود عددی، اطلاعات علّی/زمینه‌ای موردنیاز برای توضیح را فراهم نمی‌کند و در بهترین حالت، صرفاً «برچسب» می‌دهد.

۳. ریسک‌های برچسب‌گذاری و عدم قطعیت مفهومی: اتکا به نگاشت کلاس ← متن ، ناگزیر بخشی از تناقضات زبانی/بصری را به مدل تزریق می‌کند و مرزهای گسسته (کلاس‌ها) با ماهیت پیوسته‌ی رفتار رانندگی سازگار نیست.

٤. راهکار جایگزین موفق‌تر: خط لوله اصلی ADAPTبرای تصمیم و LLM برای توضیح با نشانه‌گذاری کانتور توجه در ارزیابی‌های کمی (METEOR/ROUGE-L/SPICE) و کیفی (هم‌خوانی بصری-متنی) سازگارتر با هدف تبیینی بوده و بنابراین مسیر طبقه‌بندی کنار گذاشته شد.

نتیجه: مسیر طبقه‌بندی سرعت/جهت، با وجود تلاش برای بهینه‌سازی ابرپارامترها، به کارایی پایدار و رضایت‌بخش نرسید و از نظر تبیینی نیز ارزش افزوده معنادار ایجاد نکرد؛ لذا در ادامه پایان‌نامه به‌عنوان «مسیر ناموفق» ثبت و کنار گذاشته شد و تمرکز بر خط لوله ADAPT + LLM قرار گرفت.

Abstract

Surname: Mosalmani	**Name :** Omid
Title : Design and Implementation of an AI explanation system for self-driving vehicle decisions	
Supervisor/s: Dr. Mohammad Javad Rashti	
Advisor/s: Dr. Seyed Enayat Alavi	
Degree: Master's	
University: Shahid Chamran University of Ahvaz	
Faculty: Engineering	**Department :** Computer Engineering
Keywords: Explainable Artificial Intelligence (XAI), Prompt Engineering, Large Language Models (LLMs), Autonomous Vehicles, Textual Explanation, Attention Map	

Abstract: With the rapid advancement of autonomous driving systems, the transparency and explainability of their decision-making processes have become a critical necessity for gaining user trust, enhancing safety, and ensuring fair performance evaluation. This study proposes a hybrid framework in which the ADAPT decision-making model is employed to predict driving actions, while its attention maps serve as a bridge between the visual input and the explanation module. Subsequently, large language models (LLMs), including the Gemini and GPT families, are used to generate concise and human-understandable textual explanations by receiving the final decision, attention map, and a carefully designed prompt. The main novelty of this approach lies in leveraging the broad knowledge of LLMs that goes beyond the limitations of dataset-bound explanations, thereby enabling the generation of more accurate and diverse justifications.

Experiments were conducted on the BDD-X dataset using standard text generation metrics such as BLEU-٤ and CIDEr-D. Quantitative results demonstrated that the proposed LLM-based fine-tuned approach significantly outperforms the ADAPT baseline; for instance, in the single-reference setting, the CIDEr-D score improved from ١٠٢,٦ to ١١٤,٧٣, representing an ١١٪ increase. Improvements were also observed in other metrics, confirming that the generated explanations are not only linguistically more fluent but also better aligned with the underlying reasons for decisions. Furthermore, complementary experiments verified that employing contour-based representations of attention maps enhances both the visual–textual alignment and the stability of results.

In summary, this research introduces an effective mechanism for bridging visual attention with the linguistic capabilities of LLMs, thereby contributing to the development of more explainable and trustworthy autonomous driving systems.

Shahid Chamran University of Ahvaz

Faculty of Engineering

Department of Computer Engineering

Title

Design and Implementation of an AI explanation system for self-driving vehicle decisions

A thesis submitted In partial fulfillment of the requirement for the Ph.D. degree

By

Omid Mosalmani

Supervised by

Dr. Mohammad Javad Rashti

Advised by

Dr. Seyed Enayat Alavi

October ۲۰۲۵