دانشکده مهندسی کامپیوتر

مطالعه مدلهای زبانی-بینایی برای برنامه‌ریزی وظایف چندمرحله‌ای در رباتیک

گزارش سمینار کارشناسی ارشد

رشته مهندسی کامپیوتر – گرایش مهندسی نرم‌افزار

نام دانشجو:

علیرضا نظری

استاد راهنما:

دکتر بهروز مینایی بیدگلی

آبان ماه ۱۴۰۴


بسم الله الرحمن الرحيم

چکیده

ایجاد عامل‌های رباتیک[۱] هوشمند که قادر به درک دستورات زبان طبیعی[۲] سطح بالا و اجرای آن‌ها در محیط‌های فیزیکی پیچیده باشند، یکی از چالش‌های دیرینه در هوش مصنوعی است. به طور سنتی، این امر مستلزم برنامه‌ریزی‌های صریح و پرهزینه یا آموزش‌های گسترده و خاص دامنه بود. ظهور مدل‌های زبانی-بینایی[۳](VLM [4]) در مقیاس بزرگ، که بر روی داده‌های عظیم اینترنتی آموزش دیده‌اند، الگوی جدیدی را معرفی کرده است. این مدل‌ها، دانش عقل سلیم و قابلیت‌های استدلال معنایی بی‌سابقه‌ای را ارائه می‌دهند که پتانسیل تحول در حوزه برنامه‌ریزی وظایف رباتیک را دارد.

با این حال، انتقال این دانش مفهومی از درک منفعلانه[۵] به عمل تجسم‌یافته و هدفمند[۶] – یعنی گذار از VLM به مدل‌های بینایی-زبانی-عمل[۷] (VLA [8]) یک چالش تحقیقاتی باز و اساسی است. مسئله اصلی در «زمینه‌سازی[۹]» مفاهیم انتزاعی زبانی در تعاملات فیزیکی دقیق و ممکن در دنیای واقعی نهفته است.

این مطالعه، به بررسی عمیق و تحلیل انتقادی معماری‌ها و استراتژی‌های محاسباتی می‌پردازد که با تمرکز ویژه بر برنامه‌ریزی وظایف[۱۰] رباتیک، برای پر کردن این شکاف طراحی شده‌اند. ما ابتدا مفاهیم بنیادی هوش مصنوعی، یادگیری عمیق، پردازش زبان طبیعی[۱۱] و بینایی کامپیوتر[۱۲] را که زیربنای این مدل‌ها هستند، مرور می‌کنیم. سپس، هسته اصلی این پژوهش را با کالبدشکافی و دسته‌بندی معماری‌های پیشرفته مدل‌های VLA ارائه می‌دهیم. این دسته‌بندی شامل مدل‌های یکپارچه[۱۳]، رویکردهای سلسله‌مراتبی[۱۴] مبتنی

---

[۱] robotic

[۲] natural language

[۳] vision language models

[4] vision language model

[۵] Passive Perception

[۶] Embodied action

[۷] vision language action model

[8] vision language action model

[۹] Grounding

[۱۰] task planning

[۱۱] natural language processing

[۱۲] computer vision

[۱۳] monolith

بر تجزیه وظیفه، مدل‌های مبتنی بر قابلیت‌دهی[۱] و مدل‌های جهان[۲] است. با تحلیل انتقادی مدل‌های معرف در هر، چالش‌های کلیدی نظیر تعمیم‌پذیری[۳]، کارایی داده، استدلال چندمرحله‌ای[٤] و اجرای بلادرنگ[٥] را شناسایی می‌کنیم. این تحلیل، چارچوبی جامع برای درک چشم‌انداز فعلی پژوهش و ترسیم مسیرهای آتی برای توسعه عامل‌های رباتیک همه‌منظوره[٦] و هوشمند فراهم می‌آورد.

واژه‌های کلیدی: برنامه‌ریزی وظایف، مدل‌های VLM، مدل‌های VLA، هوش مصنوعی تجسم‌یافته[۷].

---

[١٤] Hierarchical

[١] Affordance-based

[٢] World Models

[٣] Generalizability

[٤] multi stage

[٥] Real-time

[٦] General-purpose Robotic Agents

[٧] embodied

فهرست مطالب

عنوان **صفحه**

فهرست شکل‌ها

عنوان **صفحه**

فصل ۱:

مقدمه

۱-۱- زمینه و انگیزه: به سوی ربات‌های همه‌منظوره

یکی از اهداف غایی هوش مصنوعی، ایجاد عامل‌های هوشمندی است که بتوانند به‌طور یکپارچه با انسان‌ها تعامل کرده و وظایف پیچیده‌ای را در محیط‌های پویا[۱] و نادیده به انجام رسانند. حوزه رباتیک تجسم‌یافته مستقیماً این چالش را هدف قرار می‌دهد. به طور سنتی، برنامه‌ریزی رباتیک به شدت به الگوریتم‌های کنترل کلاسیک، مدل‌سازی دقیق محیط و مهارت‌های از پیش تعریف‌شده متکی بوده است. این رویکردها، اگرچه در محیط‌های صنعتی کنترل‌شده موفق بوده‌اند، اما در مواجهه با عدم قطعیت دنیای واقعی و نیاز به درک دستورات مبهم زبان طبیعی انسان، به شدت شکننده و فاقد انعطاف‌پذیری هستند. آن‌ها فاقد «عقل سلیم» لازم برای درک این هستند که «یک میان‌وعده مناسب برای فرد خسته» چیست یا چگونه باید «میز را پس از مهمانی تمیز کرد» [۱].

انقلاب اخیر در مدل‌های پایه[۲]، به ویژه مدل‌های زبانی بزرگ[۳] و به دنبال آن، مدل‌های VLM، چشم‌انداز را به طور کامل دگرگون کرده است. این مدل‌ها که بر روی حجم عظیمی از داده‌های متنی و تصویری آموزش دیده‌اند، قابلیت‌های شگفت‌انگیزی در استدلال، درک مفاهیم و تعمیم‌پذیری بدون-داده[۴] از خود نشان داده‌اند. این پیشرفت، این پرسش اساسی را مطرح کرده است: چگونه می‌توان از این «مغز» دیجیتال که دانش گسترده‌ای از جهان دارد، برای کنترل یک «بدن» فیزیکی ربات استفاده کرد؟[۳،۴].

۱-۲- چالش اصلی: از درک تا عمل در برنامه‌ریزی وظایف

چالش اصلی این حوزه، گذار از درک چندوجهی منفعل[۵] به برنامه‌ریزی و اجرای وظیفه تجسم‌یافته[۶] است. یک مدل VLM استاندارد مانند CLIP [۵،۶] می‌تواند تشخیص دهد که یک تصویر حاوی «سیب» و «کاسه» است، اما نمی‌داند «چگونه» سیب را بردارد و در کاسه بگذارد. این شکاف میان بازنمایی‌های معنایی و اجرای فیزیکی، به «چالش زمینه‌سازی» معروف است [۱،۷،۸].

---

[۱] dynamic environments

[۲] Foundation Models

[۳] large language models

[۴] Zero-shot

[۵] Passive Multimodal Understanding

[۶] Embodied Task Planning and Execution

موضوع این مطالعه، یعنی برنامه‌ریزی وظایف، دقیقاً در قلب این چالش قرار دارد [۹،۱۰]. برنامه‌ریزی وظایف فراتر از کنترل واکنشی[۱] است؛ این فرآیند مستلزم آن است که ربات یک دستورالعمل سطح بالا و بالقوه مبهم (مانند «برایم قهوه درست کن») را دریافت کند، آن را به توالی‌ای منطقی از زیروظایف قابل اجرا (مانند: ۱. فنجان را پیدا کن، ۲. آن را بردار، ۳. زیر دستگاه قهوه‌ساز بگذار) تجزیه کند [۱۱،۱۳] و در عین حال، این برنامه را بر اساس مشاهدات بصری فعلی از محیط، اعتبارسنجی[۲] و اجرا نماید [۱۴].

۱-۳- اهداف و رویکرد پژوهش

هدف اصلی این مطالعه، ارائه یک تحلیل جامع، ساختاریافته[۳] و انتقادی از معماری‌های محاسباتی است که برای توانمندسازی مدل‌های VLM در برنامه‌ریزی وظایف رباتیک توسعه یافته‌اند [۱۵،۱۷]. این پژوهش به دنبال پاسخ به این پرسش کلیدی است: «استراتژی‌های معماری غالب برای تبدیل مدل‌های VLM به مدل‌های بینایی-زبانی-عمل کارآمد برای برنامه‌ریزی وظایف کدامند و نقاط قوت و ضعف هر یک چیست؟»

برای دستیابی به این هدف، ما رویکردی نظام‌مند را اتخاذ می‌کنیم. ابتدا، مفاهیم اساسی و پایه‌ای این حوزه، از هوش مصنوعی و یادگیری عمیق گرفته تا اجزای کلیدی پردازش زبان و بینایی را معرفی می‌کنیم. سپس، در هسته اصلی این مطالعه، به کالبدشکافی عمیق معماری‌های بینایی-زبانی-عمل می‌پردازیم. ما این معماری‌ها را به چند دسته اصلی تقسیم می‌کنیم:

۱. مدل‌های حسگر-موتور[۴]: رویکردهای یکپارچه‌ای که مشاهدات را مستقیماً به عمل نگاشت می‌دهند (مانند RT-2 [۴] و VIMA [۱۸]).

۲. معماری‌های سلسله‌مراتبی[۵]: رویکردهایی که برنامه‌ریزی سطح بالا (تولید زیروظیفه) را از اجرای سطح پایین (موتور) جدا می‌کنند (مانند SayCan [۱] و PaLM-E [۳]).

۳. مدل‌های مبتنی بر قابلیت‌دهی: مدل‌هایی که ابتدا پیش‌بینی می‌کنند چه تعاملاتی با اشیاء ممکن است (مانند CLIPort [۵]).

۴. رویکردهای نوین دیگر: شامل مدل‌های جهان [۱۹] و استراتژی‌های مبتنی بر تولید کد [۲۰].

---

[۱] Reactive Control

[۲] evaluation

[۳] structured

[۴] Sensorimotor Models

[۵] Hierarchical Architectures

با بررسی دقیق مدل‌های معرف در هر دسته، ما به ارزیابی چگونگی مواجهه آن‌ها با چالش‌های کلیدی مانند استدلال چندمرحله‌ای، کارایی داده [۲۱،۲۲]، و تعمیم‌پذیری به وظایف و محیط‌های نادیده می‌پردازیم [۲۳،۲۴].

۱-۴- ساختار گزارش

این مطالعه، به شرح زیر سازماندهی شده است:

- فصل دوم (تعاریف و مفاهیم مبنایی): به مرور ادبیات و تعریف مفاهیم پایه در هوش مصنوعی، یادگیری عمیق، پردازش زبان طبیعی، بینایی کامپیوتر و مدل‌های چندوجهی می‌پردازد که برای درک معماری‌های پیشرفته در فصول بعدی ضروری هستند.
- فصل سوم (مطالعه کارهای پیشین): این فصل، که هسته تحلیلی این رساله را تشکیل می‌دهد، به بررسی عمیق و دسته‌بندی معماری‌های مختلف مدل‌های بینایی-زبانی با تمرکز بر قابلیت‌های برنامه‌ریزی وظایف آن‌ها می‌پردازد.
- فصل چهارم (نتیجه‌گیری و چشم‌انداز آینده): در این فصل، یافته‌های کلیدی این مطالعه را خلاصه کرده، به محدودیت‌های پژوهش حاضر اشاره نموده و مسیرهای تحقیقاتی آتی در جهت دستیابی به ربات‌های هوشمند همه‌منظوره را ترسیم می‌کنیم.

فصل ۲:

تعاریف و مفاهیم مبنایی

۲-۱- مقدمه

مدل‌های چندوجهی، در نقطه تلاقی دو حوزه پردازش زبان طبیعی و بینایی کامپیوتر قرار دارند که هر دو از شاخه‌های هوش مصنوعی به شمار می‌روند. بنابراین پیش از بررسی کارهای مرتبط، باید با مفاهیم اساسی این دو حوزه آشنا شویم. در ادامه این فصل، به طرح مفاهیم و تعاریف پایه پرداخته می‌شود.

۲-۲- هوش مصنوعی

هوش مصنوعی یک حوزه گسترده است که شامل یادگیری ماشین، شبکه‌های عصبی[۱] و سایر تکنیک‌های محاسباتی می‌شود که به ماشین‌ها امکان انجام وظایفی را می‌دهد که معمولاً به هوش انسانی نیاز دارند.

هوش مصنوعی در حوزه‌های علمی مختلف، از جمله تشخیص الگو[۲]، کاربرد دارد و به تحلیل داده‌ها[۳]، شبیه‌سازی ناهنجاری‌ها[٤] و استنتاج سیستمی[٥] کمک می‌کند .

تعریف هوش مصنوعی به‌طور عام، توانایی ماشین‌ها برای انجام وظایف هوشمندانه است. ازطرفی، مفهوم هوش در حوزه‌های مختلفی مانند روانشناسی، علم اعصاب و فلسفه مورد بحث قرار گرفته است.

برخی از پژوهشگران تلاش می‌کنند تا هوش مصنوعی را به‌طور رسمی تعریف کنند تا آن را از الگوهای سنتی متمایز سازند، تأکید آنها بر قابلیت یادگیری و تطبیق‌پذیری[٦] این فناوری است.

۲-۳- یادگیری عمیق

یادگیری عمیق، شاخه‌ای از یادگیری ماشین است که از شبکه‌های عصبی مصنوعی با چندین لایه برای یادگیری بازنمایی‌ها مستقیماً از داده‌ها استفاده می‌کند. این روش در حوزه‌های مختلف از جمله بینایی کامپیوتر، پردازش زبان طبیعی و رباتیک به‌طور گسترده‌ای به کار گرفته شده است.

مبانی نظری یادگیری عمیق شامل درک نحوه انتشار سیگنال‌ها در شبکه‌های عصبی و یادگیری بازنمایی‌ها از طریق تبدیلات غیرخطی است. تحقیقات نشان داده‌اند که شبکه‌های عمیق رفتار تقریباً

---

[۱] neural networks

[۲] pattern recognition

[۳] data analysis

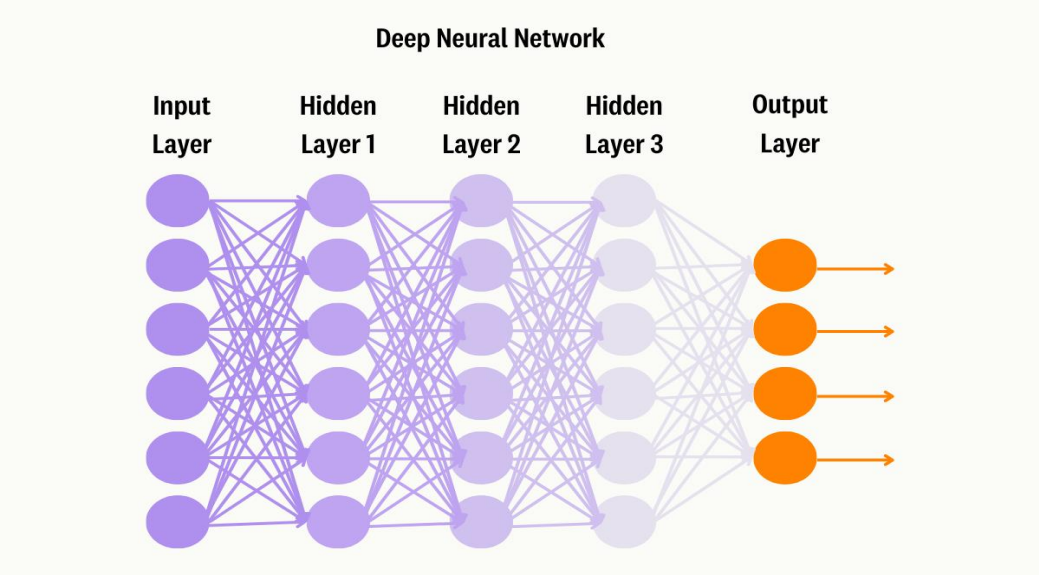
[٤] Anomaly Simulation

[٥] Systemic Inference

[٦] Adaptability

گاؤسی[۱] دارند و نسبت عمیق به عرض آن‌ها بر پیچیدگی و قابلیت تعمیم مدل تأثیر می‌گذارد.

علاوه بر این، مدل‌های یادگیری عمیق به‌طور گسترده‌ای برای تخمین علم قطعی در کاربردهای پرخطر مانند رانندگی خودران[۲] و تشخیص پزشکی مطالعه شده‌اند. یادگیری عمیق مبتنی بر شواهد چارچوبی را برای تخمین قابلیت اطمینان عدم قطعیت با سربار محاسباتی حداقلی ارائه می‌دهد



شکل (۲-۱) یک نمونه از ساختار شبکه‌های عصبی عمیق.

۲-۴- پردازش زبان طبیعی (NLP)

پردازش زبان طبیعی، شاخه‌ای از هوش مصنوعی و زبان‌شناسی[۳] است که بر تواناسازی رایانه‌ها برای درک، تفسیر و تولید زبان انسانی تمرکز دارد. این حوزه تحولی چشمگیر داشته و کاربردهای آن شامل ترجمه ماشینی[۴]، تشخیص اسم اشخاص، استخراج اطلاعات[۵]، خلاصه‌سازی[۶]، تشخیص‌های پزشکی و پاسخ‌گویی به سؤالات می‌شود.

این رشته به‌طور کلی به دو بخش تقسیم می‌شود:

---

[۱] gussian

[۲] self driving car

[۳] Linguistics

[۴] machine translation

[۵] data extraction

[۶] summerization

۱) درک زبان طبیعی

۲) تولید زبان طبیعی که به‌ترتیب شامل فهم و تولید زبان انسانی هستند. پردازش زبان طبیعی اجزای مختلف زبان‌شناسی مانند آواشناسی، ساختار کلمه، ساختار جمله، معناشناسی و کاربردشناسی را دربر می‌گیرد.

تحقیقات اخیر نشان داده‌اند که NLP[۱] با سرعت زیادی در حال پیشرفت و گسترش است و نقش مهمی در پردازش و تولید مفاهیم زبان طبیعی دارد . این حوزه به‌مرور قادر می‌شود چالش‌هایی مانند ابهام‌های معنایی، پیچیدگی نحوی و نیاز به مدل‌های زبانی پیشرفته‌تر را بررسی و حل کند.

۲-۵- بینایی کامپیوتر

دانشمندان، مدل‌های پایه مانند Florence [۲۵] قابلیت‌های بینایی کامپیوتر را با تلفیق داده‌های چندرسانه‌ای[۲] از جمله تصاویر، متن و اطلاعات عمیق گسترش داده‌اند، که باعث افزایش تصمیم‌پذیری آدم‌ها در وظایف مختلف شده است. بینایی کامپیوتر در صنایعی مانند بهداشت، رانندگی خودکار، امنیت و سرگرمی به‌طور گسترده‌ای مورد استفاده قرار می‌گیرد. برای مثال، تصویربرداری پزشکی از بینایی کامپیوتر برای تشخیص بیماری‌ها بهره می‌برد، در حالی که اتومبیل‌های خودران از آن برای تشخیص اشیاء[۳] و هدایت در دنیای واقعی استفاده می‌کنند. در ادامه، مفاهیم کلیدی مرتبط با این موضوع مورد بررسی قرار می‌گیرند:

- تشخیص اشیاء: تشخیص اشیاء شامل شناسایی و تعیین موقعیت اشیاء در یک تصویر یا ویدیو است. روش‌های سنتی از ویژگی‌های دستی استفاده می‌کردند، در حالی که تکنیک‌های مدرن با بهره‌گیری از یادگیری عمیق، به‌ویژه شبکه‌های عصبی پیچشی[۴] و مدل‌های مبتنی بر مبدل[۵]، دقت و کارایی این فرایند را بهبود داده‌اند.
- طبقه‌بندی تصاویر[۶]: طبقه‌بندی تصاویر فرآیند اختصاص دادن برچسب به تصاویر بر اساس محتوای آن‌هاست. روش‌های اولیه به استخراج ویژگی‌ها متکی بودند، در حالی که یادگیری عمیق، به‌ویژه

---

[۱] natural language processing

[۲] multi-media

[۳] object detection

[۴] convolutional neural network

[۵] transformers

[۶] image classification

شبکه‌های عصبی پیچشی، عملکرد این حوزه را به شکل قابل توجهی بهبود بخشیده است. پیشرفت‌های اخیر شامل مبدل‌های بصری و مدل‌های مولد برای افزایش دقت طبقه‌بندی هستند.

- شناسایی چهره[۱]: شناسایی چهره شامل تشخیص هویت افراد بر اساس ویژگی‌های چهره آن‌هاست و در کاربردهایی مانند امنیت، احراز هویت و شبکه‌های اجتماعی مورد استفاده قرار می‌گیرد. مدل‌های یادگیری عمیق مانند مبدل‌های مبتنی بر توجه، دقت شناسایی چهره را به میزان قابل توجهی بهبود بخشیده‌اند.
- بازسازی صحنه: بازسازی صحنه شامل تولید نمایش سه‌بعدی از محیط بر اساس تصاویر یا دنباله‌های ویدیویی دوبعدی است. این فناوری در حوزه‌هایی مانند رباتیک، واقعیت افزوده و رانندگی خودکار کاربرد دارد. تکنیک‌های پیشرفته مبتنی‌بر میدان‌های تابشی عصبی[2] و پخش‌کردن گوسی[۳] برای بازسازی بادقت صحنه‌های خبری مورد استفاده قرار می‌گیرند.

۲-۶- شبکه‌های عصبی پیچشی

شبکه‌های عصبی پیچشی نوعی از مدل‌های یادگیری عمیق هستند که برای پردازش داده‌های ساختارمند مانند تصاویر طراحی شده‌اند. این مدل‌ها با الهام از نقش شبکیه بیولوژیکی، انقلاب بزرگی در وظایف بینایی کامپیوتر ایجاد کرده‌اند و فرایند بصورت خودکار سلسله‌مراتب فضایی ویژگی‌ها را از داده‌های خام استخراج می‌کنند.

شبکه‌های عصبی پیچشی از لایه‌های متعددی مانند لایه‌های پیچش، لایه‌های نمونه‌برداری، و لایه‌های کاملاً متصل تشکیل شده‌اند. لایه‌های پیچشی مسئول استخراج ویژگی‌ها از طریق اعمال فیلترهای قابل یادگیری هستند، در حالی که لایه‌های نمونه‌برداری بعد فضایی را کاهش داده و کارایی محاسبات را افزایش می‌دهند. سپس لایه‌های کاملاً متصل[٤] این ویژگی‌ها را برای وظایفی مانند طبقه‌بندی یا بازگویی تفسیر می‌کنند.

یکی از مزایای کلیدی شبکه‌های عصبی پیچشی، توانایی آن‌ها در یادگیری نمایش‌های سلسله‌مراتبی است، که باعث می‌شود این مدل‌ها برای شناسایی تصاویر، تشخیص اشیاء و بخش‌بندی تصاویر بسیار مناسب

---

[۱] face detection

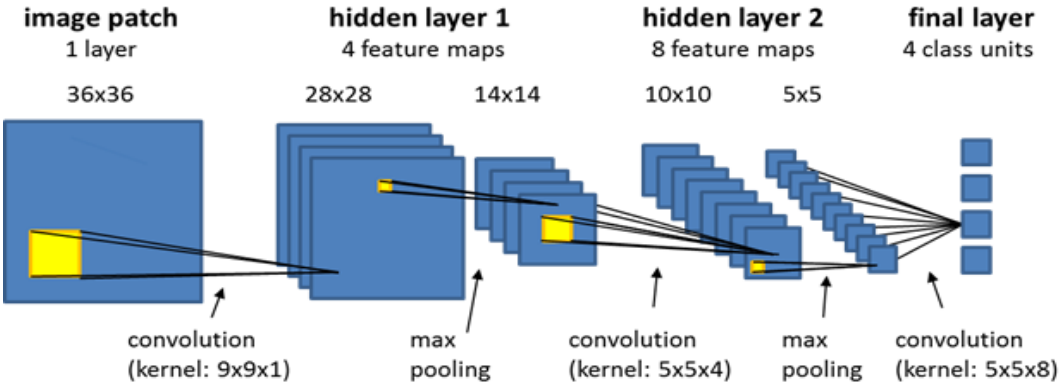
[2] Neural Radiance Fields

[۳] Gaussian Splatting

[٤] dense

باشند.

موفقیت آن‌ها منجر به پذیرش گسترده در حوزه‌هایی مانند تصویربرداری پزشکی، رانندگی خودکار، و پردازش زبان طبیعی شده است.



شکل (۲-۲) یک نمونه از ساختار شبکه‌های عصبی پیچشی.

۲-۷- مکانیزم توجه

مکانیزم توجه[۱]، یکی از اجزای کلیدی در مدل‌های یادگیری عمیق است که به‌ویژه در پردازش زبان طبیعی و بینایی رایانه‌ای استفاده می‌شود. این مکانیزم به مدل‌ها اجازه می‌دهد تا هنگام پردازش داده‌ها بر بخش‌های مهم‌تر تمرکز کنند، که باعث افزایش دقت و کارایی آن‌ها می‌شود. مکانیزم توجه به خودی[۲] که در مدل مبدل [۲۶] معرفی شده است، به هر تکواژ در یک توالی اجازه می‌دهد تا اهمیت سایر تکواژ را ارزیابی کند.

مکانیزم توجه متقاطع این مفهوم را گسترش می‌دهد و امکان تعامل بین چندین توالی[۳] یا مدل‌داده را فراهم می‌کند. به‌جای تمرکز بر عناصر داخل یک توالی، توجه متقاطع به یک توالی دیگر اجازه می‌دهد تا از اطلاعات یک منبع خارجی تمرکز کند. این مکانیزم در یادگیری چندمداله و ترکیب تصاویر بسیار کاربردی است. به‌عنوان مثال، در شناسایی احساسات چند رسانه‌ای، توجه متقاطع توانسته اطلاعات را از منابع مختلف همچون صوت و تصویر به‌طور مؤثر ترکیب کند. همچنین، در ترکیب تصاویر، مکانیزم توجه متقاطع برای استخراج اطلاعات مکمل و کاهش ویژگی‌های زاید[۴] پیشنهاد شده است.

---

[۱] attention mechanism

[۲] self attention

[۳] sequence

[۴] redundant

۲-۸- مدل‌های چندوجهی

مدل‌های چندوجهی، سیستم‌های هوش مصنوعی هستند که قادر به پردازش و ترکیب چندین نوع داده (مانند متن، تصویر، صوت و ویدئو) در یک چارچوب یکپارچه هستند. این مدل‌ها از ویژگی‌های گوناگون داده‌ها برای افزایش قابلیت‌های درک و استدلال خود استفاده می‌کنند و تصمیم‌گیری‌های دقیق‌تر و آگاهانه‌تری را امکان‌پذیر می‌سازند.

رویکردهای معماری مدل‌های چندوجهی را به چهار دسته اصلی تقسیم می‌کنند بر اساس شیوه‌ی ترکیب داده‌های مختلف:

۱) نوع اول: از مکانیزم‌های توجه متقابل استاندارد برای ترکیب ورودی‌های چندوجهی در لایه‌های داخلی مدل استفاده می‌کند.

۲) نوع دوم: لایه‌های سفارشی برای ترکیب ویژگی‌ها طراحی می‌کنند، که تعامل عمیق‌تری بین انواع داده‌ها را فراهم می‌کند.

۳) نوع سوم: شامل رمزگذارهای[۱] اختصاصی برای هر نوع داده است، که ابتدا هر ورودی به‌طور مستقل پردازش می‌شود و سپس یکپارچه‌سازی صورت می‌گیرد.

۴) نوع چهارم: از رمزگذارهای ورودی برای پردازش انواع داده‌ها در مرحله اولیه استفاده می‌کند، که ترکیب زودهنگام[۲] را امکان‌پذیر می‌سازد [۲۷].

مدل‌های چندوجهی پیشرفته، با بهره‌گیری از یادگیری عمیق و شبکه‌های عصبی، توانسته‌اند در ترکیب چندوجهی به نتایج بسیار چشم‌گیری برسند. به‌ویژه این مدل‌ها در حوزه رباتیک اهمیت زیادی دارند، زیرا به سامانه‌های خودمختار این امکان را میدهد که با استفاده از ورودی‌های چندوجهی، بتوانند با محیط در ارتباط باشند .

۲-۹- تنظیم دقیق

تنظیم دقیق[۳] فرآیندی در یادگیری ماشین است که طی آن یک مدل از پیش آموزش‌دیده برای یک وظیفه خاص تنظیم و تطبیق داده می‌شود. این روش با استفاده از یک مجموعه داده کوچک‌تر و تخصصی باعث

---

[۱] encoder

[۲] early fusion

[۳] fine tunning

افزایش عملکرد مدل در یک دامنه خاص می‌شود.

رویکردهای تنظیم دقیق شامل تنظیم دقیق نظارتی[۱] (که از داده‌های برچسب‌دار استفاده می‌کند)، تنظیم دقیق غیرنظارتی[۲] (که بدون برچسب کار می‌کند)، و تنظیم دقیق مبتنی بر دستورالعمل (که از راهنمایی‌های خاص برای هدایت مدل بهره می‌برد) هستند.

یک مراحل استاندارد تنظیم دقیق شامل پیش‌پردازش داده‌ها[۳]، تنظیم اولیه مدل، تنظیم ابرپارامترها[۴]، آموزش، ارزیابی و استقرار مدل است. بهینه‌سازی این فرآیند به کاهش هزینه‌های محاسباتی و افزایش قابلیت تعمیم کمک می‌کند.

تنظیم دقیق نیز یکی از حوزه‌های نوظهور است که مدل‌ها را بر اساس بازخورد پیوسته تنظیم می‌کند، به جای اینکه فقط از داده‌های ایستا بیاموزند. این روش در مدل‌های چندوجهی موفق بوده است [۲۸]، و باعث بهبود استدلال و دقت در زمینه‌های خاص می‌شود.

چالش‌های آینده تنظیم دقیق شامل بهبود کارایی، حفظ حریم خصوصی، و ملاحظات اخلاقی است، زیرا تنظیم مدل‌های بزرگ نیازمند رویکردهای مؤثرتر است .

۲-۱۰- مدل‌های مبدل

مدل‌های مبدل یک کلاس از معماری‌های شبکه‌ی عصبی هستند که در سال ۲۰۱۷ [۲۶] معرفی شدند و انقلابی در حوزه‌ی پردازش زبان طبیعی و به دنبال آن، در سایر حوزه‌های هوش مصنوعی ایجاد کردند.

برخلاف معماری‌های قدیمی‌تر مانند شبکه‌های بازگشتی که داده‌ها را به صورت ترتیبی[۵] پردازش می‌کردند، مدل‌های مبدل این قابلیت را دارند که تمام بخش‌های ورودی را به صورت موازی پردازش کنند. این ویژگی نه تنها سرعت آموزش را به شدت افزایش داد، بلکه امکان درک روابط پیچیده بین کلمات در یک متن طولانی را فراهم کرد.

هسته‌ی اصلی و نوآوری کلیدی مدل مبدل، مکانیسمی به نام توجه-خودی است.

مکانیسم کلیدی: توجه-خودی

---

[۱] supervised

[۲] unsupervised

[۳] preprocessing

[۴] hyper-parameter tunning

[۵] Sequential

به جای پردازش کلمه‌ی «it» در یک جمله، بدون دانستن اینکه به چه چیزی اشاره دارد، مکانیسم توجه-خودی به مدل اجازه می‌دهد تا به تمام کلمات دیگر در همان جمله «نگاه» کند و به هر کلمه یک «امتیاز اهمیت[۱]» اختصاص دهد.

این قابلیت، درک عمیق متنی[۲] را امکان‌پذیر می‌سازد. معماری مبدل از اجزای دیگری نیز تشکیل شده است:

- توجه چند-سری[۳]: اجرای همزمان چندین مکانیسم توجه-خودی به صورت موازی، که به مدل اجازه می‌دهد انواع مختلف روابط (مثلاً روابط دستورزبانی، روابط معنایی و...) را به طور همزمان یاد بگیرد.
- رمزگذاری موقعیتی[۴]: از آنجایی که مدل، کلمات را به صورت همزمان و موازی می‌بینند (و نه ترتیبی)، اطلاعات مربوط به «موقعیت» یا «ترتیب» کلمات از طریق این بردارها به ورودی اضافه می‌شود تا مدل بداند کدام کلمه اول آمده و کدام آخر.
- معماری رمزگذار-رمزگشا[۵]: معماری اصلی مبدل (که برای ترجمه‌ی ماشین طراحی شده بود) شامل دو بخش است:

۱. رمزگذار: ورودی (مثلاً جمله‌ی آلمانی) را می‌خواند و آن را به یک بازنمایی عددی غنی[۶] تبدیل می‌کند.

۲. رمزگشا: آن بازنمایی را دریافت می‌کند و خروجی (مثلاً جمله‌ی انگلیسی) را کلمه به کلمه تولید می‌کند.

۲-۱۰-۱- خانواده‌ی مدل‌های مبدل

این معماری پایه، منجر به ایجاد سه خانواده‌ی اصلی از مدل‌ها شد که امروزه اساس اکثر سیستم‌های هوش مصنوعی را تشکیل میدهد:

---

[۱] Attention Score

[۲] Contextual Understanding

[۳] Multi-Head Attention

[۴] Positional Encoding

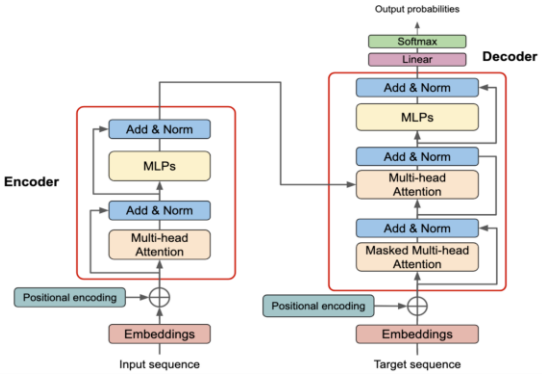
[۵] Encoder-Decoder

[۶] Contextual Representation

۱. مدل‌های مبتنی بر رمزگذار[۱]: مانند BERT [۲۹]. این مدل‌ها کل متن را به صورت دوطرفه می‌بینند و برای وظایف «درک مطلب» مانند تحلیل احساسات[۲]، پاسخ به پرسش و دسته‌بندی متن عالی هستند.

۲. مدل‌های مبتنی بر رمزگشا[۳]: مانند سری مدل‌های GPT (از جمله ChatGPT [۳۰]). این مدل‌ها در «تولید» متن تخصص دارند. آن‌ها بر اساس متن قبلی، کلمه‌ی بعدی را پیش‌بینی می‌کنند و برای نوشتن خلاقانه و خلاصه‌سازی استفاده می‌شوند.

۳. مدل‌های رمزگذار-رمزگشا: مانند T5 [۳۱] و BART [۳۲]. این مدل‌ها برای وظایف «تبدیل دنباله-به-دنباله[٤]» مانند ترجمه‌ی ماشین، خلاصه‌سازی (که هم نیاز به درک و هم تولید دارد) و پاسخ به پرسش بهینه‌سازی شده‌اند.



شکل (۲-۳) ساختار مدل‌های تبدیل‌گر.

۲-۱۱- مدل‌های بینایی-زبانی (VLM)

در سال‌های اخیر، مدل‌های یادگیری عمیق با بهره‌گیری از ترکیب اطلاعات متنی و تصویری، تحولات بزرگی در حوزه‌ی بینایی ماشین ایجاد کرده‌اند. یکی از مهمترین رویکردهای نوظهور در این زمینه، استفاده‌ی همزمان از داده‌های زبانی و تصویری برای آموزش مدل‌های چندوجهی است. این رویکرد منجر به پیدایش خانواده‌ای از مدل‌ها تحت عنوان مدل‌های بینایی-زبانی شده است [۲،۱۶].

در این مدل‌ها، برخلاف شبکه‌های بینایی‌ای که تنها از داده‌های تصویری استفاده می‌کنند، از جفت‌های

---

[۱] Encoder-Only

[۲] sentiment analysis

[۳] Decoder-Only

[٤] Seq2Seq

تصویر-متن برای آموزش بهره گرفته می‌شود. هدف اصلی این مدل‌ها، نگاشت تصاویر و متون به فضای ویژگی مشترکی است که در آن، توصیف متنی و محتوای تصویری از نظر معنایی هم‌راستا باشند. این هم‌ترازی منجر به یادگیری بازنمایی‌های[۱] چندوجهی غنی می‌شود که درک عمیق‌تری از مفاهیم موجود در تصویر و متن را ممکن می‌سازد. مدل‌های معروفی مانند:

- CLIP [۶]
- LEO [۳۶]
- UNIPI [۳۷]
- ECOT [۳۸]

توانسته‌اند در طیف گسترده‌ای از وظایف گوناگون مانند طبقه‌بندی تصویر، تشخیص اشیاء، بازشناسی بدون نمونه و یادگیری با نمونه کم[۲] عملکرد چشمگیری از خود نشان دهند.

این مدل‌ها عموماً با استفاده از روش‌های خودنظارتی[۳] آموزش دیده‌اند. برخلاف روش‌های یادگیری نظارت‌شده که نیاز به برچسب‌گذاری گسترده دارند، در اینجا از داده‌های عظیم جفت‌شده‌ی تصویر-متن استخراج‌شده استفاده می‌شود. برای مثال:

- مدل CLIP توسط شرکت OpenAI با بیش از ۴۰۰ میلیون جفت تصویر-متن [۶] آموزش دیده است.
- مدل ALIGN نیز با بهره‌گیری از بیش از ۱ میلیارد جفت تصویر-متن [۳۳] توسط Google طراحی شده است.

با وجود این پیشرفت‌ها، یکی از چالش‌های اساسی در استفاده از این مدل‌ها، مسئله‌ی تطبیق مؤثر آن‌ها با وظایف خاص موردنظر است. به عبارت دیگر، چگونه می‌توان مدل‌هایی مانند CLIP را که به صورت عمومی آموزش دیده‌اند، برای مسائل خاصی نظیر بازشناسی چهره، مسائل طبقه‌بندی، یا تشخیص فعالیت به گونه‌ای تطبیق داد که عملکرد آن‌ها بهینه شود.

روش‌های متداولی برای تطبیق مدل‌های بینایی-زبانی با وظایف موردنظر وجود دارد که دو مورد اصلی آن عبارتند از:

تنظیم دقیق کامل: در این روش، تمامی پارامترهای مدل آموزش‌دیده دوباره به‌روزرسانی می‌شوند. با این

---

[۱] representation

[۲] few-shot learning

[۳] self-supervised

که این روش درصورت وجود مجموعه داده‌ی غنی می‌تواند عملکرد مدل را در وظیفه‌ی هدف بهبود دهد، اما اغلب منجر تخریب نمایش مشترک یادگرفته‌شده‌ی بینایی-زبانی قبلی می‌شود.

کاوشگری خطی[۱]: در این روش، تنها یک لایه طبقه‌بند ساده روی ویژگی‌های استخراج‌شده‌ی مدل آموزش داده می‌شود، بدون اینکه وزن‌های اصلی مدل تغییر کنند. با این حال، این روش از امکان استفاده مدل‌های بینایی-زبانی برای تشخیص‌های بدون نمونه را محدود می‌کند.

۲-۱۲- مدل‌های بینایی-زبانی-عمل (VLA)

در امتداد تکامل مدل‌های چندوجهی، گام بعدی فراتر از درک منفعلانه و به سوی تعامل فعال با محیط است. این امر منجر به پیدایش مدل‌های بینایی-زبانی-عمل شده است که ستون فقرات نسل جدید هوش مصنوعی تجسم‌یافته و رباتیک را تشکیل می‌دهند [۱۵،۱۷].

برخلاف مدل‌های بینایی-زبانی که هدفشان هم‌تراز کردن بازنمایی‌های متن و تصویر است، هدف اصلی مدل‌های VLM-عمل، ایجاد یک خط‌مشی[۲] است. این مدل‌ها یک ورودی چندوجهی، شامل بینایی از دوربین ربات و زبانی از دستور کاربر را دریافت می‌کنند و یک توالی از عمل[۳] یا فرامین حرکتی را به عنوان خروجی تولید می‌کنند.

ایده‌ی کلیدی در این مدل‌ها، استفاده از دانش معنایی گسترده‌ی آموخته‌شده توسط مدل‌های پایه‌ی بینایی-زبانی چندوجهی برای تعمیم‌بخشی به وظایف رباتیک است. نوآوری اصلی در این حوزه، روش برخورد با «عمل» به عنوان یک مُدالیته‌ی دیگر است. در این روش‌ها:

۱. تکواژ کردن عمل[۴]: فرامین کنترلی ربات (مانند مختصات x, y, z برای بازوی ربات، یا میزان فشار گریپر) به تکواژ گسسته‌ای تبدیل می‌شوند، درست مانند کلمات در یک متن.

۲. آموزش ترتیبی: مدل به صورت یک وظیفه‌ی «دنباله-به-دنباله» آموزش می‌بیند. ورودی شامل تکواژهای تصویر و تکواژهای دستور متنی است و خروجی، پیش‌بینی تکواژهای عمل است [۱۵،۱۷].

این رویکرد به ربات اجازه می‌دهد تا دستورات پیچیده و مفهومی مانند «میوه‌ای که برای یک فرد خسته مناسب است را به من بده» را درک کند، از طریق بخش بینایی خود محیط را تحلیل کند مثلاً یک

---

[۱] linear exploration

[۲] Policy

[۳] Action

[۴] Action Tokenization

نوشیدنی انرژی‌زا یا یک سیب را تشخیص دهد و سپس توالی اقدامات صحیح برای برداشتن آن را اجرا کند.

مدل‌های معروفی در این حوزه عبارتند از:

- RT-2[1] [۴]: مدلی از گوگل که نشان داد با تنظیم دقیق یک مدل VLM قدرتمند (مانند PaLM-E [۳]) روی داده‌های رباتیک، می‌توان دانش وب-مقیاس مدل را مستقیماً به کنترل ربات منتقل کرد.

OpenVLA [۳۴]: یک چارچوب متن-باز که رویکردهای مشابه RT-2 را پیاده‌سازی می‌کند.

۲-۱۳- برنامه‌ریزی وظایف در مدل‌های LLM و VLM

فراتر از درک منفعلانه (مانند دسته‌بندی تصاویر) و یا حتی اجرای عمل مستقیم ، یکی از مرزهای پژوهشی پیشرفته، استفاده از این مدل‌ها برای برنامه‌ریزی وظایف است [۹،۱۰]. در این سناریو، مدل به جای تولید یک «عمل» واحد، یک «برنامه» یا توالی‌ای از اقدامات منطقی را برای رسیدن به یک هدف پیچیده تولید می‌کند.

هدف اصلی در اینجا، تجزیه‌ی یک دستورالعمل سطح بالا و انتزاعی مانند «برایم یک فنجان قهوه درست کن» را به مجموعه‌ای از گام‌های اجرایی سطح پایین و مشخص است مانند: ۱. به سمت آشپزخانه برو، ۲. فنجان را بردار، ۳. زیر دستگاه قهوه‌ساز بگذار، و الی آخر [۱۱،۱۳].

برنامه‌ریزی در مدل‌های زبانی

مدل‌های زبانی بزرگ به دلیل دانش گسترده‌ای که از متن‌های موجود به دست آورده‌اند، ذاتاً قابلیت برنامه‌ریزی معنایی[2] را دارند.

- نحوه عملکرد: این مدل‌ها با استفاده از قابلیت زنجیره‌ی تفکر[3] [۳۵]عمل می‌کنند. هنگامی که یک هدف به مدل داده می‌شود، مدل می‌تواند با تولید متن به صورت "مرحله-به-مرحله"، یک برنامه‌ی منطقی را تدوین کند.
- محدودیت: این نوع برنامه‌ریزی کاملاً انتزاعی[4] و جدا از محیط[5] است. مدل نمی‌داند که آیا اجرای این گام‌ها در دنیای واقعی «ممکن» است یا خیر.

---

[1] Robotic Transformer

[2] Semantic Planning

[3] Chain-of-Thought

[4] Abstract

[5] Disembodied

برنامه‌ریزی در مدل‌های بینایی-زبانی: برنامه‌ریزی تجسم‌یافته

اینجاست که این مدل‌ها وارد می‌شوند تا بزرگترین ضعف مدل‌های زبانی را برطرف کنند: آن‌ها برنامه‌ریزی را با واقعیت "زمین‌پایه" [۷،۸] می‌کنند.

در برنامه‌ریزی تجسم‌یافته، مدل نه تنها باید بداند چه کاری باید انجام دهد، بلکه باید بر اساس آنچه می‌بیند، تشخیص دهد که آیا می‌تواند آن کار را انجام دهد یا خیر.

این فرآیند یک حلقه‌ی بازخورد ایجاد می‌کند:

۱. مشاهده بینایی-زبانی: وضعیت فعلی محیط را از طریق ورودی بصری درک می‌کند.
۲. برنامه‌ریزی: مدل زبانی (یا بخش زبانی مدل بینایی-زبانی) یک گام یا یک برنامه را پیشنهاد می‌دهد.
۳. ارزیابی بینایی-زبانی: بررسی می‌کند که آیا گام پیشنهادی با توجه به تصویر فعلی، «عملی[۱]» است.
۴. عمل[۲]: گام تایید شده که می‌تواند یک فرمان حرکتی یا یک زیر-هدف زبانی باشد، اجرا می‌شود.
۵. حلقه به گام ۱ بازمی‌گردد [۱،۱۲].

[۱] Feasible

[۲] Act

فصل ۳:

مروری بر کارهای پیشین

۳-۱- مقدمه

مدل‌های VLM طیف گسترده‌ای از طرح‌های معماری را در بر می‌گیرند ، که منعکس‌کننده استراتژی‌های متنوعی برای یکپارچه‌سازی ادراک، دستورالعمل، و کنترل هستند [۱۵-۱۷]. یک رویکرد بسیار پذیرفته‌شده، مدل یکپارچه است، که به طور مشترک نمایش‌های بصری، زبانی، و عمل را می‌آموزد. این مدل‌ها تصاویر و زبان را به عنوان ورودی دریافت کرده و مستقیماً اعمال را خروجی می‌دهند [۱۵،۱۶]، و می‌توانند ساختار مسطح[۱] یا سلسله مراتبی را با معماری‌های ستون فقرات[۲] متفاوت اتخاذ کنند [۳۹]. در حالی که مدل‌های یکپارچه یک کلاس بنیادی از سیستم‌های VLA را تشکیل می‌دهند، چندین معماری جایگزین نیز پیشنهاد شده است. مدل‌های جهان تکامل آینده مُدالیته‌های حسی، معمولاً بصری را پیش‌بینی می‌کنند، که مشروط به ورودی زبان است، و از این پیش‌بینی‌ها برای هدایت تولید عمل استفاده می‌کنند [۱۹،۴۰،۴۱]. مدل‌های مبتنی بر قابلیت‌دهی یک نوع دیگر هستند که قابلیت‌دهی‌های بصری مرتبط با عمل را بر اساس زبان پیش‌بینی می‌کنند و سپس بر اساس آن عمل‌ها را تولید می‌کنند [۱،۵].

۳-۲- معماری‌ها و اجزای سازنده

۳-۲-۱- مدل یکپارچه

در حال حاضر هفت نوع معماری از مدل‌های یکپارچه وجود دارد.

- **مبدل+ تکواژ عمل گسسته**

این معماری هم تصاویر و هم زبان را به عنوان تکواژ نمایش می‌دهد، که برای پیش‌بینی عمل بعدی به یک مبدل ارسال می‌شوند، که معمولاً به شکل تکواژهای گسسته است [۱۵-۱۷]. این دسته همچنین شامل مدل‌هایی است که از تکواژهای CLS[3] استفاده می‌کنند و عمل‌های پیوسته را از طریق یک شبکه عصبی چندلایه[۴] (MLP) تولید می‌کنند، مانند RT-2 [۴]. نمونه معرف دیگر شامل VIMA است [۱۸].

- یکی از مدل‌های پیشگام مبتنی بر مبدل، توسط آقای جیانگ و همکاران در مقاله‌ی مدل VIMA [۱۸] آورده شده است که الگوی «یادگیری مبتنی بر دستورات متنی» را با معرفی دستورات متنی

---

[۱] flat

[۲] Backbone

[3] classification token

[۴] multi-layer perceptron

چندوجهی[۱] به حوزه رباتیک گسترش می‌دهد. این مدل قادر است طیف گسترده‌ای از وظایف دستکاری رباتیک، از جمله وظایف مبتنی بر هدف بصری، تقلید از نمایش تک‌نمونه[۲]، درک مفاهیم جدید، ارضای محدودیت‌های بصری و استدلال بصری را تنها با استفاده از یک معماری واحد، از طریق دستورات متنی که تکواژهای متنی و بصری (تصاویر یا فریم‌های ویدئویی) را در هم می‌آمیزند، فرموله کند. به عنوان یک مدل VLM مدل VIMA از یک رمزگذار T5 [۳۱] از پیش‌آموزش‌دیده و منجمد[۳] برای پردازش دستورات متنی چندوجهی استفاده می‌کند. نوآوری کلیدی آن در بازنمایی شیء-محور[۴] ورودی‌های بصری نهفته است؛ به جای استفاده از پنجره‌های تصویر خام یا تکواژهای فشرده‌شده تصویر، VIMA ابتدا اشیاء موجود در تصاویر دستورات متنی یا مشاهدات ربات را با استفاده از یک آشکارساز اشیاء[۵] (مانند Mask R-CNN [۴۲]) شناسایی کرده و هر شیء را به یک تکواژ متشکل از ویژگی‌های استخراج‌شده از تصویر برش‌خورده آن و کادر مرزی[۶] آن (با استفاده از MLP) تبدیل می‌کند [۱۸].

برای برنامه‌ریزی وظایف VIMA از یک معماری مبدل رمزگذار-رمزگشا [۲۶،۳۱] بهره می‌برد. رمزگذار T5 [۳۱] دستورات متنی چندوجهی را پردازش می‌کند و رمزگشا که یک مبدل سببی[۷] است [۳۰]، تکواژهای عمل گسسته[۸] را به صورت خودبازگشتی[۹] تولید می‌کند. تکنیک اصلی برای مرتبط ساختن دستورات متنی با اقدامات ربات، استفاده از لایه‌های توجه متقابل[۱۰] [۲۶] در رمزگشا است که به صورت متناوب با لایه‌های توجه به خود [۲۶] قرار گرفته‌اند. رمزگشا به تاریخچه تعاملات (دنباله‌ای از تکواژهای مشاهدات شیء-محور و تکواژهای عمل قبلی) از طریق لایه‌های توجه به خود و به تکواژهای رمزگذاری‌شده دستورات متنی از طریق لایه‌های توجه متقابل توجه می‌کند تا تکواژ عمل بعدی را پیش‌بینی کند [۱۵-

---

[۱] Multimodal Prompts

[۲] One-shot Demonstration

[۳] Freeze

[۴] Object-centric

[۵] object detector

[۶] Bounding Box

[۷] Causal Transformer

[۸] Discretized Action Tokens

[۹] Autoregressively

[۱۰] Cross-Attention

۱۸]. تکواژهای عمل گسسته نمایانگر پارامترهای حرکات پایه سطح بالا مانند «برداشتن و گذاشتن» یا «هل دادن» هستند. مزیت اصلی VIMA در یکپارچه‌سازی وظایف متنوع رباتیک و بهره‌وری داده بالا است که با ۱۰ برابر داده کمتر، عملکرد بهتری نسبت به رقبا نشان می‌دهد [۱۸]. معماری مبتنی بر توجه متقابل نیز مقیاس‌پذیری خوبی با افزایش اندازه مدل نشان داده و به‌ویژه در وظایف تعمیم سخت‌تر عملکرد قوی‌تری دارد. با این حال، VIMA دارای محدودیت‌هایی نیز هست؛ اتکای آن به یک آشکارساز اشیاء خارجی می‌تواند خطاها را به سیستم منتقل کند، اگرچه نویسندگان با استفاده از افزودن داده[۱] سعی در کاهش این مشکل داشته‌اند [۱۸]. علاوه بر این، ارزیابی این مدل تنها در محیط شبیه‌سازی انجام شده و عملکرد آن در دنیای واقعی نامشخص است و از اعمال سطح بالا به جای کنترل سطح پایین استفاده می‌کند که ممکن است دقت لازم برای برخی وظایف پیچیده را نداشته باشد.

◘ مبدل + سَرِ کنش انتشار[۲]

این معماری یک سیاست انتشار[۳] [۴۳] را به عنوان سَرِ عمل[٤] پس از مبدل گنجانده است. در حالی که تکواژهای عمل گسسته اغلب فاقد پاسخگویی و صافی[٥] بلادرنگ هستند، این مدل‌ها با استفاده از مدل‌های انتشار به خروجی‌های عمل پیوسته و پایدار دست می‌یابند. نمونه معرف شامل کارآقای گوش و همکاران در مقاله مدل Octo است [۲۴].

مدل Octo یک سیاست ربات همه‌منظوره[٦] متن-باز و مبتنی بر مبدل است که بر یادگیری یک سیاست حسگر-حرکتی[۷] مستقیم و همه‌منظوره تمرکز دارد [۲۴]. به عنوان یک VLM، این مدل از یک رویکرد «مبدل -محور» استفاده می‌کند. Octo ورودی‌های وظیلفه (دستورات زبانی یا تصاویر هدف) و مشاهدات (تصاویر دوربین‌های متعدد و داده‌های حسگرهای ربات) را با استفاده از تکواژ کننده‌های سبک مانند T5 [۳۱] برای زبان و CNN کم‌عمق برای تصاویر را به تکواژهای مجزا تبدیل می‌کند. این تکواژها سپس در یک

---

[۱] Augmentation

[۲] Diffusion Action Head

[۳] diffusion policy

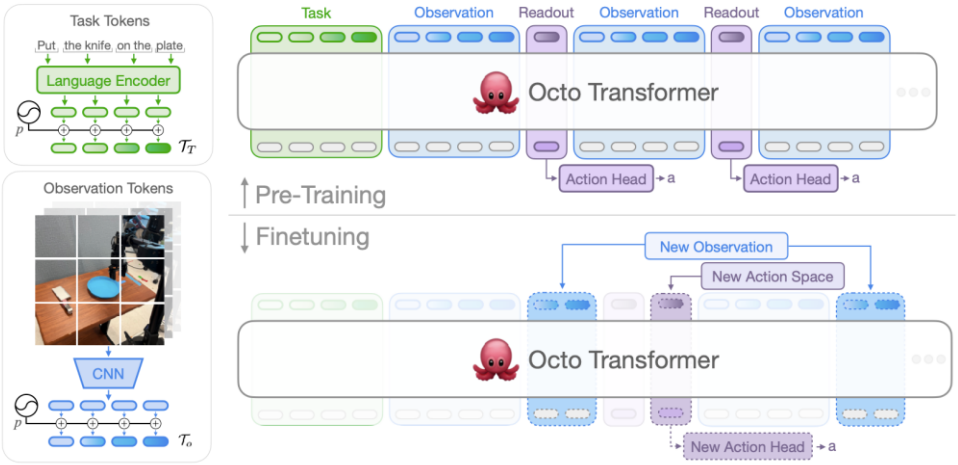
[٤] action head

[٥] smoothness

[٦] GRP

[۷] Sensorimotor

توالی واحد به بدنه اصلی مبدل خورانده می‌شوند. تکنیک کلیدی این مدل برای تولید عمل، استفاده از تکواژهای بازخوانی[۱] قابل یادگیری است که به زمینه قبلی توجه کرده و بازنمایی فشرده‌ای از حالت فعلی و هدف تولید می‌کنند [۲۴]. سپس، بردار خروجی این تکواژهای بازخوانی، به عنوان ورودی شرطی به یک بخش تولید عمل سبک و مجزا داده می‌شود. این بخش، یک سیاست انتشار است [۴۳،۴۴] که قادر است یک «قطعه» از اقدامات پیوسته آتی را به‌صورت موازی تولید کند. نوآوری اصلی Octo در معماری چندتکه[۲] و انعطاف‌پذیر آن برای تنظیم دقیق کارآمد نهفته است؛ این معماری به کاربران اجازه می‌دهد تا مدل عظیم از پیش‌آموزش‌دیده را به سادگی برای ربات‌های جدید یا حسگرهای جدید تطبیق دهند [۲۴]. با این حال، به عنوان یک مدل مبتنی بر یادگیری تقلیدی، عملکرد آن به شدت به توزیع داده‌های پیش‌آموزش وابسته است [۱۸،۲۴] و به دلیل کمبود داده‌های زبانی، عملکرد مدل در حالت شرطی‌سازی با زبان ضعیف‌تر از حالت شرطی‌سازی با تصویر هدف است.



شکل (۳-۱) معماری مدل Octo.

◘ مبدل مبتنی بر مدل انتشار[۳] (DiT[۴])

مدل مبدل مبتنی بر انتشار [۴۵]، مبدل و سَرِ کنش انتشار را یکپارچه می‌کند و فرآیند انتشار را مستقیماً درون مبدل اجرا می‌نماید. این امر مدل را قادر می‌سازد تا فرآیند انتشار را مستقیماً مشروط به تکواژهای تصویر و زبان انجام دهد. به عنوان مثال، مدل RDT-1B [۴۶] که حاصل پژوهش آقای لیو و همکارانشان

---

[۱] Readout Tokens

[۲] modular

[۳] diffusion transformer

[4] diffusion transformer

است، بر پایه این معماری ساخته شده است، دنباله‌ای از تکواژهای عمل را از طریق توجه متقاطع [۲۶] با یک دستور بینایی و زبان تولید می‌کند که متعاقباً از طریق یک MLP به عمل‌های ربات قابل اجرا نگاشت می‌شوند.

درمدل RDT-1B [۴۶] وظایف پیچیده‌ای مانند دستکاری دو بازویی[۱]، چالش اساسی، چندوجهی بودن فضای عمل است. مدل‌های خطی[۲] تمایل دارند این مُدها را "میانگین‌گیری" کنند و مدل‌های گسسته فاقد دقت لازم هستند، در حالی که مدل‌های مبتنی بر تکواژ کردن گسسته (مانند RT-2 [۴]) فاقد دقت لازم برای کنترل پیوسته و ظریف مورد نیاز در هماهنگی دو بازو هستند [۱۸،۴۶]. برای حل این مشکل، پژوهش آقای لیو بهمراه همکارانشان در مدل RDT-1B به عنوان یک مدل پایه عظیم با ۱.۲ میلیارد پارامتر، معماری مبدل مبتنی بر انتشار [۴۵] را به حوزه رباتیک معرفی می‌کند. این مدل، به جای پیش‌بینی مستقیم عمل، توزیع شرطی پیوسته را مدل‌سازی می‌کند.

زخلدر این معماری، فرآیند مدل انتشار مستقیماً در ستون فقرات مبدل ادغام شده است. شبکه نویززدا[۳] همان بدنه مبدل می‌باشد [۴۵،۴۶]. در زمان استنتاج، یک «قطعه عمل» نویزی گاوسی به مدل داده می‌شود و مبدل به صورت تکرارشونده این نویز را، با مشروط شدن بر تکواژهای زبان و تصویر، می‌زداید تا یک توالی عمل پیوسته تولید کند [۴۶].

مدل RDT-1B [۴۶] در ابتدا برای مقابله با کمبود داده‌های دو بازویی، یک فضای عمل یکپارچه با تفسیر فیزیکی[٤] معرفی می‌کند که امکان پیش‌آموزش بر روی مجموعه داده‌های عظیم و از ربات‌های مختلف (تک‌بازو، دو بازو، چرخ‌دار) را با نگاشت ابعاد عمل آن‌ها به یک بردار ۱۲۸ بعدی مشترک بر اساس معنای فیزیکی (مانند مفاصل بازوی راست، مفاصل بازوی چپ) فراهم می‌سازد. دوم، برای جلوگیری از غلبه اطلاعات بصری بر دستورات زبانی، از تکنیک تزریق شرطی متناوب (ACI[٥]) استفاده می‌کند که در آن تکواژهای تصویر و زبان به صورت متناوب به لایه‌های توجه متقابل تزریق می‌شوند. سوم، با جایگزینی LayerNorm با RMSNorm، پایداری آموزش را در مواجهه با داده‌های رباتیک پرنوسان و غیرخطی افزایش می‌دهد. مزیت اصلی این رویکرد، توانایی بی‌نظیر آن در مدل‌سازی توزیع‌های عمل پیچیده و چندوجهی و دستیابی به تعمیم‌پذیری بدون نمونه قوی به اشیاء و صحنه‌های نادیده است. با این حال، محدودیت اصلی

---

[۱] Bimanual Manipulation

[۲] regression

[۳] Denoising Network

[٤] Physically Interpretable Unified Action Space

[٥] alternating conditional injection

آن این است که علی‌رغم پیش‌آموزش گسترده بر روی داده‌های عمدتاً تک‌بازویی، همچنان برای پر کردن شکاف تجسم[۱] و دستیابی به عملکرد بهینه در ربات دو بازویی هدف، نیازمند یک مجموعه داده تنظیم قابل توجه (شش هزار مسیر دو بازویی) است [۴۶].

تلاش دیگر در این حوزه پژوهش آقای روس و مدل MDT [۴۴] است که به چالش مقیاس‌پذیری یادگیری تقلیدی می‌پردازد که همان اتکا به داده‌های کاملاً برچسب‌گذاری‌شده است، در حالی که بسیاری از مجموعه داده‌های رباتیک در مقیاس بزرگ، مانند داده‌های «یادگیری از بازی» (LfP[۲]) ، دارای برچسب‌های زبانی بسیار پراکنده هستند. مدل مبدل-انتشار چندوجهی MDT [۴۴] به طور خاص برای یادگیری رفتارهای همه‌منظوره از اهداف چندوجهی ، حتی زمانی که حاشیه‌نویسی‌های زبانی کمیاب هستند طراحی شده است.

این مدل، به عنوان یک VLM، از معماری مبدل رمزگذار-رمزگشا [۲۶] استفاده می‌کند. رمزگذار، حالت فعلی و یک هدف چندوجهی (دستور زبان یا تصویر هدف، که هر دو با رمزگذار CLIP منجمد

[۶] پردازش می‌شوند) را دریافت کرده و آن‌ها را به یک بازنمایی نهفته[۳] غنی از اطلاعات حالت، نگاشت می‌دهد. سپس، این بازنمایی نهفته به عنوان شرط به رمزگشا، که یک سیاست انتشار مبتنی بر مدل GPT [۳۰،۴۳،۴۴] است، داده می‌شود. تکنیک برنامه‌ریزی وظیفه در MDT متکی بر مدل‌سازی توزیع پیوسته عمل است؛ رمزگشا از طریق توجه متقابل [۲۶] به بازنمایی نهفته حالت-هدف توجه کرده و با استفاده از یک فرآیند نویززدایی تکراری، یک قطعه عمل پیوسته (مثلاً ده گام زمانی آینده) را از نویز خالص تولید می‌کند.

نوآوری کلیدی MDT برای یادگیری مؤثر از داده‌های با برچسب پراکنده، معرفی دو هدف کمکی خودنظارتی[۴] است که فضای نهفته را مجبور به هم‌ترازی می‌کنند. اولین هدف، هم‌ترازی نهفته متضاد، از یک تابع ضرر InfoNCE [۶] استفاده می‌کند تا بازنمایی‌های نهفته حالتی که تحت مودالیته‌های هدف مختلف تولید شده‌اند را به یکدیگر نزدیک کند و یک فضای معنایی مشترک برای اهداف، مستقل از مودالیته ورودی، ایجاد نماید. دومین هدف، آینده‌نگری مولد ماسک‌دار[۵]، این بازنمایی نهفته را وادار می‌سازد تا اطلاعات کافی برای پیش‌بینی آینده را در خود رمزگذاری کند؛ این کار از طریق یک رمزگشای تصویر مجزا

---

[۱] Embodiment Gap

[۲] learn from playing

[۳] Latent Representation

[۴] Self-Supervised Auxiliary Objectives

[۵] masked generative foresight

انجام می‌شود که وظیفه دارد پنجره‌های ماسک‌دار یک فریم در آینده را تنها بر اساس بازنمایی نهفته حالت-هدف فعلی بازسازی کند [۴۴]. مزیت اصلی این رویکرد، دستیابی بـه عملکـرد پیشـرفته در بنچمارک‌هـای چالشی مانند CALVIN [۴۷] و LIBERO [۴۸]، حتی با پارامترهایی ۱۰ برابر کمتر از رقبا و بدون نیاز به پیش‌آموزش در مقیاس بزرگ است [۴۴]. با این حال، MDT به عنوان یک سیاسـت مبتنـی بـر انتشـار، در زمان استنتاج به دلیل نیاز به چندین مرحله نـویززدایی، کنـدتر از روش‌هـای تک‌مرحلـه‌ای عمـل می‌کنـد. همچنین، نویسندگان اشاره می‌کنند که تأثیر مثبت اهداف کمکی در تمام وظایف یکسان نبوده و به‌ویژه در وظایف طولانی‌مدت که فاقد زیرهدف‌های مشخص هستند، بهبود قابل توجهی مشاهده نشده است [۴۴].

◘ VLM + تکواژ عمل گسسته

مدل‌های VLM + تکواژ عمل گسسته، با جایگزینی مبدل با یک مدل VLM که بر روی داده‌های اینترنتـی بزرگ‌مقیاس پیش‌آموزش دیده است تعمیم‌پذیری را بهبود می‌بخشـند [۴،۳۴]. اسـتفاده از VLM بـه ایـن مدل‌ها اجازه می‌دهد تا دانش عقل سلیم انسانی را در خود جای دهند و از قابلیت‌های یادگیری درون‌متنی[۱] [۳۰] بهره‌مند شوند.

مدل RT-2 که نتیجه پژوهش آقای بروهان و همکارانشان است به جای طراحی یـک معمـاری جدیـد، یـک دستورالعمل[۲] آموزشی برای انتقال مستقیم دانـش مفهـومی از مـدل‌های VLM عظـیمِ از پیش‌آموختـه بـه سیاست‌های کنترلی ربات ارائه می‌دهد [۴]. این مدل از ستون فقرات VLMهای پیشرفته ماننـد PaLM-E (با حداکثر ۵۵ میلیارد پارامتر) [۳] استفاده می‌کند که بر روی داده‌های اینترنتی آموزش دیده‌انـد. تکنیـک محوری RT-2 یکپارچه‌سازی کامل اقدام در مُدالیته زبان است. در این رویکرد، اقدامات پیوسته ربات را ابتدا به سطل‌های[۳] گسسته (۲۵۶ سطل برای هر بُعد) تقسیم‌بندی می‌شوند. سپس، ایـن سـطل‌های گسسـته بـه تکواژهای متنی در واژگان موجود VLM نگاشت داده می‌شوند [۴] برخلاف تنظـیم دقیـق سـاده، RT-2 از هم-تنظیم دقیق[۴] استفاده می‌کند؛ یعنی مدل به طور همزمان هم بر روی داده‌های رباتیـک و هـم بـر روی داده‌های زبان-بینایی اصلی خود آموزش می‌بیند. این فرآیند از «فراموشی فاجعـه‌بار» دانـش وب جلـوگیری می‌کند. در زمان استنتاج، مدل به صورت خودبازگشتی، توالی نشانه‌های متنی را تولید می‌کند که مستقیماً

---

[۱] in-context learning

[۲] recipe

[۳] bin

[۴] Co-fine-tuning

به دستورات عمل ربات رمزگشایی می‌شوند [۴].

نوآوری اصلی و مزیت بزرگ RT-2 ، بروز قابلیت‌های نوظهور[۱] است به این معنا که مدل می‌تواند دانش معنایی و استدلال آموخته‌شده از اینترنت را به مهارت‌های فیزیکی آموخته‌شده از ربات اعمال کند و وظایفی را انجام دهد که هرگز در داده‌های رباتیک ندیده است؛ مانند درک مفاهیم انتزاعی (مثلاً «کوچکترین شیء»)، نمادها (مانند «عدد ۳») یا حتی انجام استدلال زنجیره-فکر (CoT) [۳۵] برای حل مسائل چندمرحله‌ای (مانند استنتاج اینکه «نوشیدنی انرژی‌زا» برای فرد «خسته» مناسب است) [۴،۳۵]. با این حال، محدودیت اصلی آن این است که این انتقال دانش، حرکات فیزیکی جدیدی ایجاد نمی‌کند و مهارت‌های حرکتی همچنان به داده‌های نمایشی ربات محدود هستند. علاوه بر این، اندازه عظیم مدل، اجرای حلقه-بسته[۲] بلادرنگ را به یک چالش محاسباتی جدی تبدیل می‌کند که نیازمند زیرساخت ابری است [۴]. علاوه بر این، مدل‌های دیگری نیز این معماری را پذیرفته‌اند که در ادامه آورده شده‌اند.

OpenVLA توسط آقای کیم و همکاران [۳۴] یک نقطه عطف کلیدی در گذار به مدل‌های پایه رباتیکِ شفاف و در دسترس است که معماری مدل‌های بسته مانند RT-2 [۴] را بازتولید می‌کند. این مدل به عنوان یک مدل VLM عمل می‌کند که بر پایه یک ستون فقرات یک VLM از پیش‌آموخته (Prismatic [۴۹] مبتنی بر Llama-2 [۵۰]) ساخته شده است [۳۴]. نوآوری اصلی آن در رمزگذار بصری دوگانه نهفته است که ویژگی‌های معنایی سطح بالا را با ویژگی‌های فضایی دقیق و ریز-دانه از DINOv2 [۵۱] ترکیب می‌کند تا درک جامع‌تری از صحنه ارائه دهد. برای برنامه‌ریزی وظایف مدل OpenVLA از یک راهبرد یکپارچه حسگر-حرکتی استفاده می‌کند که در آن، استدلال و اجرا به یک مسئله واحد پیش‌بینی تکواژ[۳] تبدیل می‌شوند. تکنیک اصلی تولید عمل، گسسته‌سازی مستقیم فضای عمل[۴] است. در این روش، بردار عمل پیوسته ربات به سطل‌های مجزا (معمولاً ۲۵۶ سطل برای هر بُعد) تقسیم می‌شود [۴،۳۴]. سپس، این سطل‌ها مستقیماً به ۲۵۶ تکواژ کم‌استفاده در واژگان[۵] موجود Llama [۵۰] نگاشت داده می‌شوند و مدل

---

[۱] Emergent Capabilities

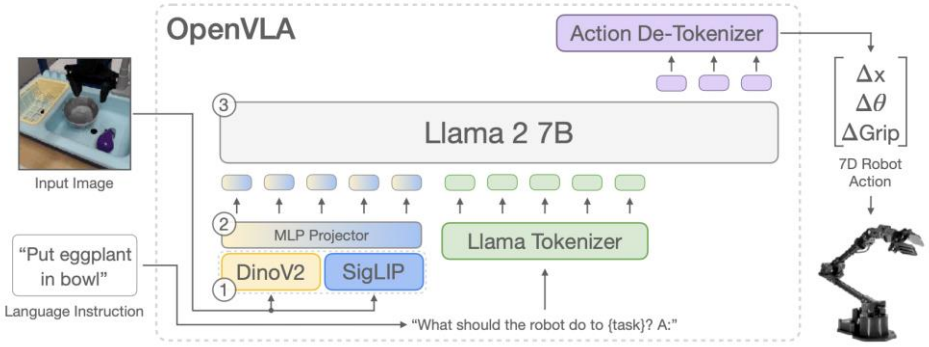
[۲] closed-loop

[۳] Token Prediction

[۴] direct action discretization

[۵] vocabulary

زبانی با استفاده از تابع زیان متقاطع[۱] استاندارد، برای تولید این «تکواژهای عمل» خاص به صورت خودبازگشتی تنظیم دقیق می‌شود [۳۴].



شکل (۳-۲) معماری مدل OpenVLA.

مزیت اصلی OpenVLA ماهیت متن-باز[۲] آن است که تحقیقات رباتیک را همگانی می‌کند. همچنین، این مدل با وجود داشتن پارامترهایی ۷ برابر کمتر، به دلیل بهره‌گیری از داده‌های آموزشی متنوع‌تر و معماری بصری برتر، در بسیاری از معیارها از مدل بسته RT-2 عملکرد بهتری داشته است [۳۴]. علاوه بر این، این پژوهش برای اولین بار، امکان‌پذیری تنظیم دقیق پارامتربهینه و تقسیم بندی را برای VLAها بررسی کرد که امکان آموزش و استنتاج بر روی سخت‌افزارهای همگانی[۳] را فراهم می‌کند. محدودیت‌های اصلی این مدل شامل اتکای آن به مشاهدات تک-تصویری (عدم وجود حافظه زمانی) و سرعت استنتاج نسبتاً پایین (حدود ۶ هرتز) است که آن را برای کنترل حلقه-بسته با فرکانس بالا نامناسب می‌سازد [۳۴].

برخلاف مدل‌هایی که صرفاً بر داده‌های تصویری دوبعدی اتکا می‌کنند، پژوهش آقای هوانگ وهمکارانشان در مدل LEO [۵۲] به عنوان یک عامل فراگیر[۴] تجسم‌یافته در دنیای سه‌بعدی طراحی شده است که هدف آن درک، استدلال، برنامه‌ریزی و عمل در محیط‌های سه‌بعدی است. به عنوان یک VLM، معماری LEO ورودی‌های چندوجهی شامل تصاویر خودمحوری[۵] دوبعدی و به‌طور ویژه، ابر نقاط سه‌بعدی را می‌پذیرد. نوآوری کلیدی این مدل در بازنمایی شیء-محور سه‌بعدی آن نهفته است؛ مدل ابتدا با استفاده از PointNet

---

[۱] cross-entropy

[۲] Open-Source

[۳] consumer GPUs

[۴] Generalist Agent

[۵] ego centeric

[۵۳] ویژگی‌های هر شیء را در ابر نقاط استخراج کرده و سپس با بهره‌گیری از یک مبدل فضایی[۱]، روابط فضایی میان اشیاء را مدل می‌کند. این تکواژهای بصری دوبعدی و سه‌بعدی، همراه با تکواژهای متنی، به یک ستون فقرات LLM (مدل Vicuna-7B [۵۴]) که با LoRA [۵۵] سازگار شده است، خورانده می‌شوند [۵۲].

استراتژی LEO برای برنامه‌ریزی وظایف و تولید عمل، بر یک چارچوب پیش‌بینی توالی خودبازگشتی یکپارچه استوار است. در این چارچوب، تمامی وظایف، از جمله درک صحنه سه‌بعدی (مانند پرسش و پاسخ یا تولید کپشن)، برنامه‌ریزی سطح بالا (مانند تولید لیست گام‌های وظیفه) و اقدامات تجسم‌یافته سطح پایین (ناوبری)، همگی به عنوان یک مسئله واحد تولید تکواژ در نظر گرفته می‌شوند [۵۲]. تکنیک اصلی برای تولید عمل، گسسته‌سازی فضای عمل است [۴،۳۴،۵۲]؛ اقدامات پیوسته (مانند مختصات کنترل کننده نهایی در ناوبری) و اقدامات گسسته (مانند «حرکت به جلو» در ناوبری) همگی به سطل‌های مجزا تقسیم شده و به تکواژهای رزرو شده در واژگان متنی موجود LLM نگاشت داده می‌شوند. بدین ترتیب، LLM به طور مستقیم هم پاسخ‌های متنی (استدلال و برنامه‌ریزی) و هم تکواژهای عمل اجرایی را تولید می‌کند [۵۲].

مزیت اصلی LEO [۵۲]، توانایی آن در زمینه‌سازی عمیق سه‌بعدی است که با استفاده از یک خط لوله نوآورانه تولید داده مبتنی بر LLM با استفاده از «گراف‌های صحنه» و «زنجیره-فکر شیء-محور[۲]» [۵۲] پشتیبانی می‌شود. با این حال، نویسندگان اشاره می‌کنند که یک «شکاف قابل توجه» بین یادگیری VL (درک بصری-زبانی) و یادگیری VLA (کنترل تجسم‌یافته) وجود دارد [۱۵،۱۶] و آموزش مشترک این دو می‌تواند به قابلیت‌های استدلال VL آسیب بزند. همچنین، عدم استفاده از مکانیزم بازگشتی[۳] در سیاست‌گذاری، توانایی آن را در وظایفf ناوبری طولانی‌مدت محدود می‌سازد [۵۲].

مدل‌های VLA استاندارد (مانند OpenVLA [۳۴]) معمولاً به صورت واکنشی[۴] عمل کرده و مشاهدات حسی را مستقیماً به تکواژهای عمل نگاشت می‌دهند. این رویکرد، اگرچه در کارهای تکراری موفق است، اما در تعمیم به وظایف جدید یا صحنه‌های نادیده، به دلیل فقدان استدلال میانی، دچار چالش می‌شود. برای

---

[۱] Spatial Transformer

[۲] (O-CoT)

[۳] Recurrence

[۴] Reactive

حل این مشکل، آقای زاوالسکی و همکارانشان مدل ECOT[۱] [۳۸] را پیشنهاد دادند که یک استراتژی آموزشی نوین را بر روی بدنه‌های VLA موجود (مانند OpenVLA [۳۴]) پیاده‌سازی می‌کند تا قابلیت تفکر گام‌به‌گام یا زنجیره-فکر را به آن‌ها بیافزاید. برخلاف CoT صرفاً معنایی در LLMها [۳۵]، تکنیک ECOT یک استدلال زنجیره‌ای تجسم‌یافته است [۳۸] که مدل را وادار می‌سازد تا قبل از تولید تکواژهای عمل گسسته، به صورت خودبازگشتی مجموعه‌ای از تکواژهای استدلال میانی را تولید کند. این زنجیره استدلال به دو بخش تقسیم می‌شود:

۱) استدلال زبانی سطح بالا، شامل بازنویسی وظیفه، تولید برنامه گام‌به‌گام و شناسایی زیروظیفه فعلی[۲]

۲) استدلال تجسم‌یافته سطح پایین که حیاتی‌ترین بخش نوآوری است و مدل را مجبور به «نگاه دقیق» می‌کند. این بخش شامل پیش‌بینی ویژگی‌های بصری-فضایی صریح مانند موقعیت پیکسل دست ربات[۳] و کادرهای مرزی اشیاء شناسایی‌شده در صحنه به عنوان خروجی متنی است [۳۸]. از آنجایی که داده‌های آموزشی برای این استدلال‌های میانی وجود ندارد، نویسندگان یک خط لوله مقیاس‌پذیر برای تولید داده مصنوعی ارائه می‌دهند که با استفاده از مدل‌های پایه دیگر (مانند Gemini [۵۶] Grounding DINO [۵۸،۵۹]OWL+SAM) مجموعه داده‌های رباتیک موجود را به صورت آفلاین با این زنجیره‌های استدلال برچسب‌گذاری می‌کند.

مزیت اصلی ECOT [۳۸] افزایش چشمگیر تعمیم‌پذیری (تا ۲۸٪ بهبود مطلق نسبت به OpenVLA [۳۴]) بدون نیاز به داده رباتیک جدید است. علاوه بر این، این مدل ذاتاً قابل تفسیر[۴] شده و امکان اصلاح تعاملی[۵] را فراهم می‌آورد؛ اگر مدل در استدلال خود (مثلاً شناسایی شیء) خطا کند، انسان می‌تواند با بازخورد زبان طبیعی، زنجیره-فکر متنی را اصلاح کرده و بدین ترتیب به صورت آنی، عمل خروجی را تصحیح نماید [۳۸]. محدودیت اصلی این روش، سرعت پایین استدلال است، زیرا تولید زنجیره استدلال طولانی (مثلاً ۳۵۰ تکواژ) قبل از هر اقدام (مثلاً هفت تکواژ عمل)، فرکانس کنترل را به شدت کاهش

---

[۱] Embodied Chain-of-Thought

[۲] SUBTASK

[۳] GRIPPER POS

[۴] Interpretable

[۵] Interactive Correction

می‌دهد، هرچند نویسندگان راهکارهایی مانند اجرای ناهمزمان[۱] را برای کاهش این مشکل پیشنهاد می‌کنند [۳۸].

مدل GR-1 که ثمره پژوهش آقای لو و همکارانشان است [۶۰]، استدلال می‌کند که تعمیم‌پذیری ربات به طور قابل توجهی از پیش‌آموزش مولد ویدئویی در مقیاس بزرگ بهره می‌برد، زیرا پیش‌بینی فریم‌های آینده، درکی قوی از پویایی‌های فیزیکی و پیش‌آمدهای یک عمل را فراهم می‌کند. مدل GR [۶۰] از یک معماری ساده به سبک GPT (مبدل فقط-رمزگشا) [۳۰] استفاده می‌کند که ورودی‌های چندوجهی شامل دستور زبان (رمزگذار CLIP منجمد) [۶]، تاریخچه تصاویر و حالات ربات (رمزگذاری شده با MLP) را دریافت می‌کند. برخلاف مدل‌هایی که صرفاً عمل را پیش‌بینی می‌کنند، تکنیک GR-۱ برای برنامه‌ریزی وظایف، یادگیری مشترک دو-هدفه[۲] در یک چارچوب خودبازگشتی واحد است:

مدل به طور همزمان هم تصویر آینده و هم عمل ربات را پیش‌بینی می‌کند. این کار از طریق تکواژهای ویژه‌ی قابل یادگیری انجام می‌شود: تکواژ [OBS] که خروجی آن به یک رمزگشای بصری برای بازسازی پیکسلی فریم آتی هدایت می‌شود، و تکواژ [ACT] که خروجی آن به یک هد MLP مجزا برای رگرسیون مستقیم اعمال پیوسته و وضعیت باینری گریپر ارسال می‌گردد [۶۰].

نوآوری اصلی این مدل، فرآیند آموزش دو مرحله‌ای آن است: ابتدا، مدل به طور کامل بر روی یک مجموعه داده عظیم غیر-رباتیک (Ego4D) [۶۱] فقط برای وظیفه «پیش‌بینی ویدئوی مشروط به زبان» پیش‌آموزش داده می‌شود. سپس، این مدل که اکنون یک «مدل جهان ضمنی[۳]» را فراگرفته، بر روی داده‌های رباتیک برای یادگیری مشترک عمل و پیش‌بینی ویدئو تنظیم دقیق می‌شود [۶۰].

مدل‌های VLA مبتنی بر تصویر دوبعدی، علی‌رغم توانایی درک معنایی، در استدلال روابط فضایی سه‌بعدی دچار چالش هستند، در حالی که مدل‌های بومی سه‌بعدی به دلیل کمبود داده‌های آموزشی مقیاس‌پذیر نیستند. پژوهش آقای لی و همکاران مدل 3D-VLA [۶۲] را برای پل زدن بر این شکاف طراحی کرده است و یک VLM دوبعدی از پیش‌آموخته (مبتنی بر LLaMA [۵۰] و CLIP [۶]) را با آگاهی فضایی سه‌بعدی جامع تجهیز می‌کند. نوآوری اصلی این مدل در مشاهده فضایی سه‌بعدی نهفته است که به رمزگذار بصری دوبعدی اجازه می‌دهد تا همزمان تصاویر دوبعدی و ابر نقاط سه‌بعدی را با استفاده از یک انکودر مشترک

---

[۱] Asynchronous Execution

[۲] Dual-Task

[۳] Implicit World Model

پردازش کند. در حالی که تصاویر دوبعدی به صورت استاندارد تکواژ می‌شوند، ابر نقاط ابتدا توسط یک تکواژ کننده سه‌بعدی غیرپارامتری (مبتنی بر FPS و kNN) [۵۳] به تکواژهای سه‌بعدی تبدیل می‌گردند [۶۲].

سپس، تکنیک کلیدی هم‌ترازی موقعیتی دوبعدی-به-سه‌بعدی[۱] به کار گرفته می‌شود که به جای استفاده از بردارهای موقعیتی سه‌بعدی جدید، هر تکواژ سه‌بعدی با استفاده از پارامترهای دوربین به صفحه تصویر دوبعدی بازتابانده می‌شود و امبدینگ موقعیتی دوبعدی از پیش‌آموخته‌ی متناظر با آن پنجره تصویری، به آن تکواژ سه‌بعدی تخصیص می‌یابد. این هم‌ترازی هندسی به VLM اجازه می‌دهد تا دانش فضایی دوبعدیِ از پیش‌آموخته‌ی خود را مستقیماً برای استدلال سه‌بعدی به کار گیرد [۶۲]. تکنیک دوم این مدل برای برنامه‌ریزی وظایف، فراتر رفتن از الگوی ادراک-به-عمل[۲] از طریق قیود فضایی سه‌بعدی است. این قیود، که به صورت نقاط کلیدی[۳] سه‌بعدی متوالی تعریف می‌شوند، با استفاده از مدل‌های خارجی (مانند Grounded SAM) استخراج می‌گردند. نکته حیاتی این است که این قیود به جای استفاده به عنوان مختصات هدف، به بردار متنی[۴] تبدیل شده و به عنوان بخشی از دستورالعمل به LLM خورانده می‌شوند. این کار، مدل را وادار می‌سازد تا به صراحت در مورد روابط فضایی-زمانی استدلال کند. در نهایت، مدل به صورت خودبازگشتی، تکواژهای عمل گسسته را برای پیش‌بینی ژست[۵] ربات تولید می‌کند. مزیت اصلی این روش، بهره‌گیری از دانش VLMهای دوبعدی برای داده‌های سه‌بعدی بدون سربار محاسباتی اضافه و بهبود استدلال فضایی-زمانی است، اما محدودیت آن، وابستگی به اطلاعات عمق دوربین و اتکا به مدل‌های خارجی برای استخراج نقاط کلیدی است.

مدل آقای چن و همکاران با عنوان WorldVLA [۶۳]، یک چارچوب نوآورانه را به عنوان یک «مدل جهان-عمل خودبازگشتی[۶]» ارائه می‌دهد که دو الگوی مجزا را در یک معماری واحد یکپارچه می‌سازد: یک مدل زبان-بینایی-عمل (VLA) برای تولید کنش، و یک مدل جهان برای پیش‌بینی حالت آتی. به عنوان یک VLM، این مدل از تکواژسازهای مجزا برای تصویر، متن و عمل (گسسته‌سازی با سطل‌بندی [۴،۳۴]) استفاده می‌کند و تمامی این‌ها را به یک فضای واژگان مشترک نگاشت می‌دهد. تکنیک اصلی برنامه‌ریزی

---

[۱] 2D-to-3D Positional Alignment

[۲] perception-to-action

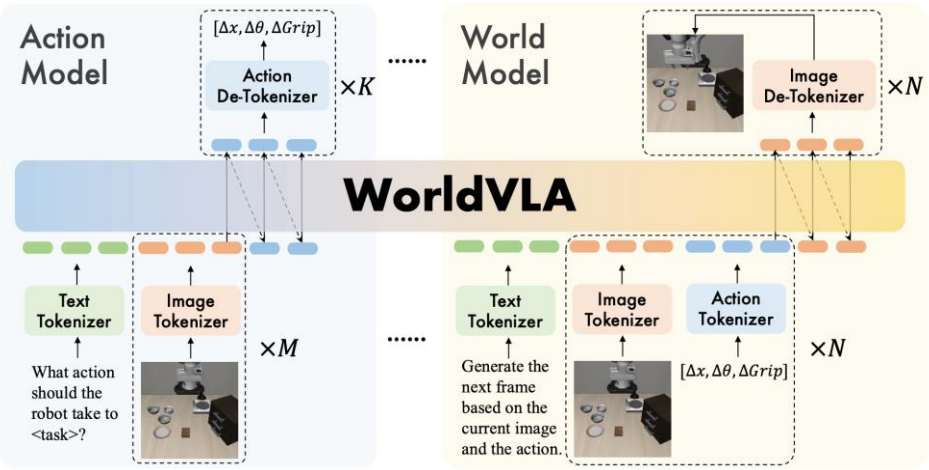
[۳] keypoints

[۴] text-based formulation

[۵] pose

[۶] Autoregressive Action World Model

وظیفه در WorldVLA [۶۳] مبتنی بر یک یادگیری دو-هدفه همزمان است که در آن، این دو قابلیت، یکدیگر را به صورت متقابل تقویت می‌کنند[۱]. مدل هم به عنوان مدل عمل آموزش می‌بیند (تولید نشانه‌های عمل بر اساس تاریخچه تصویر و دستور زبان و هم به عنوان مدل جهان (تولید نشانه‌های تصویر آتی بر اساس تاریخچه تصویر و نشانه‌های عمل ورودی). این یادگیری مشترک، مدل را مجبور به درک «دینامیک‌های فیزیکی زیربنایی» محیط می‌کند که به نوبه خود، تصمیم‌گیری برای تولید عمل را بهبود می‌بخشد.



شکل (۳-۳) مرور کلی WorldVLA.

نوآوری کلیدی این مدل، ارائه راهکاری برای چالش انتشار خطا در تولید قطعه عمل[۲] است. نویسندگان دریافتند که تولید خودبازگشتی دنباله‌ای از اعمال، به دلیل ظرفیت تعمیم ضعیف بر روی نشانه‌های عمل، منجر به انتشار خطای اعمال قبلی به اعمال بعدی می‌شود. برای حل این مشکل، WorldVLA [۶۳] یک استراتژی ماسک‌گذاری توجه عمل[۳] را معرفی می‌کند. این ماسک، هنگام تولید یک نشانه عمل، از توجه کردن به نشانه‌های عمل قبلی در همان قطعه جلوگیری می‌کند و آن را مجبور می‌سازد تا فقط بر ورودی‌های متنی و بصری اتکا کند، که این امر عملاً تولید موازی‌گونه را ممکن ساخته و از تجمع خطا جلوگیری می‌کند. مزیت اصلی این چارچوب، یادگیری هم‌افزای بین درک و تولید عمل و تصویر است؛ با این حال، محدودیت اصلی آن، «بیانگری ادراکی محدود»[۴] در نشانه‌ساز تصویر گسسته در مقایسه با رمزگذارهای

---

[۱] Mutual Enhancemen

[۲] Action Chunking

[۳] Action Attention Masking

[۴] limited perceptual expressiveness

پیوسته است.

◘ VLM + سَرِ کنش انتشار

مدل‌های VLM + سَرِ کنش انتشار مبدل را با یک VLM جایگزین می‌کنند. این معماری VLMها (که تعمیم‌پذیری بهتر را ممکن می‌سازند) را با مدل‌های انتشار (که دستورات عمل ربات پیوسته و روان را تولید می‌کنند) ترکیب می‌کند. این مدل‌ها هم تکواژهای گسسته را به صورت خودرگرسیوی تولید می‌کنند و هم از یک سَرِ عمل انتشار برای تولید عمل‌های پیوسته در یک مدل واحد استفاده می‌نمایند.

مدل‌های VLA استاندارد معمولاً هنگام تنظیم دقیق بر روی داده‌های رباتیک، دچار پدیده «فراموشی فاجعه‌بار»[۱] [۶۵] می‌شوند و توانایی‌های استدلال زبانی و درک بصری عمومی خود را که از پیش‌آموزش اینترنتی کسب کرده‌اند، از دست می‌دهند. آن‌ها اغلب یا در استدلال خوب عمل می‌کنند یا در اجرا، اما نه در هر دو. مدل ChatVLA به عنوان نتیجه پژوهش آقای ژایو و همکاران [۶۶] برای حل این تعارض و ایجاد یک مدل واحد که هم در درک چندوجهی (مانند گفتگو) و هم در کنترل تجسم‌یافته توانمند باشد، معرفی شده است. این مدل که در دسته VLM + سر کنش انتشار قرار می‌گیرد، از یک ستون فقرات Qwen-VL [۶۷] و یک سربرگ سیاست مبتنی بر انتشار [۴۳] برای تولید اعمال پیوسته استفاده می‌کند.

نوآوری کلیدی ChatVLA برای برنامه‌ریزی و اجرای وظایف، بهره‌گیری از یک معماری ترکیبی از متخصصین[۲] (MoE) است. در این چارچوب، لایه‌های مدل به چندین «متخصص» موازی (که MLP هستند) تقسیم می‌شوند. تکنیک اصلی این مدل، آموزش یک مسیریاب پویا است که تکواژهای ورودی را بر اساس نوع وظیفه به متخصصین مجزا هدایت می‌کند: تکواژهای مرتبط با زبان به «متخصصین VLM» و تکواژهای مرتبط با عمل (از داده‌های ربات) به «متخصصین VLA» فرستاده می‌شوند [۶۶]. این جداسازی از طریق یک استراتژی آموزشی دو مرحله‌ای اجرا می‌شود: در مرحله اول، تنها متخصصین VLM بر روی داده‌های عظیم زبان-بینایی آموزش می‌بینند تا قابلیت استدلال را کسب کنند. در مرحله دوم یا تنظیم دقیق رباتیک، ستون فقرات VLM و متخصصین VLM منجمد شده و تنها متخصصین VLA و سربرگ سیاست انتشار بر روی داده‌های مانور رباتیک تنظیم دقیق می‌شوند. مزیت اصلی این رویکرد جداسازی پارامتریک است؛ با تفکیک وزن‌های استدلال VLM از وزن‌های عمل VLA، مدل می‌تواند مهارت‌های رباتیک را بدون بازنویسی یا تخریب دانش عقل سلیم و توانایی‌های زبانی خود فرا بگیرد، که این امر منجر به

---

[۱] Catastrophic Forgetting

[۲] Mixture of Experts

یک عامل واحد با قابلیت‌های دوگانه و تنظیم دقیق بسیار بهینه (از جهت پارامتر) می‌شود [۶۶].

مدل‌های VLA که به صورت یکپارچه تنظیم دقیق می‌شوند، اغلب با چالش «قطع ارتباط استدلال-عمل[۱]» مواجه هستند و دانش عقل سلیم پیش‌آموخته‌ی خود را در طی تنظیم دقیق رباتیک از دست می‌دهند که به «فراموشی فاجعه‌بار» [۶۵] معروف است. این امر منجر به مدل‌هایی می‌شود که یا در استدلال پیچیده یا در کنترل دقیق موفق عمل می‌کنند، اما نه در هر دو. مدل آقای ون و همکارانشان Diffusion-VLA [۶۸] برای حل این تعارض، یک الگوی «تولید عمل مشروط به استدلال[۲]» را با استفاده از یک ستون فقرات VLM از پیش‌آموخته (مانند Qwen-VL) [۶۷] معرفی می‌کند.

تکنیک اصلی برنامه‌ریزی وظیفه در این مدل، «استدلال خود-تولید[۳]» (SGR) است. برخلاف نگاشت مستقیم ادراک-به-عمل، Diffusion-VLA [۶۸] ابتدا ستون فقرات VLM خود را فراخوانی می‌کند تا یک زنجیره استدلال متنی (یک برنامه) را به صورت خودبازگشتی تولید کند. این زنجیره، وظیفه را به زیروظیفه‌ها تجزیه کرده و روابط اشیاء را مشخص می‌سازد. سپس، این خروجی متنی استدلالی، به عنوان یک شرط اضافی، به یک ماژول اجرایی مجزا داده می‌شود. این ماژول، یک «متخصص انتشار قابل اتصال[۴]» سبک است که به رمزگذار بصری مدل متصل شده و وظیفه تولید اعمال پیوسته را بر عهده دارد [۶۸].

نوآوری کلیدی در استراتژی آموزش نهفته است:

ستون فقرات VLM (مسئول استدلال) در طول تنظیم دقیق رباتیک کاملاً منجمد باقی می‌ماند و تنها «متخصص انتشار» (مسئول عمل) با استفاده از روش‌های بهینه (PEFT مانند LoRA [۵۵]) آموزش می‌بیند. این جداسازی کامل پارامترها، همزمان قابلیت‌های استدلال عمومی VLM را حفظ کرده و هم امکان تنظیم دقیق بسیار کارآمد (تنها ۰.۷ میلیون پارامتر) برای یادگیری مهارت‌های حرکتی را فراهم می‌آورد. این رویکرد منجر به مدلی قابل تفسیر با تعمیم‌پذیری بالا در وظایف پیچیده می‌شود. محدودیت اصلی این معماری دو مرحله‌ای، تأخیر ذاتی در زمان استنتاج است، زیرا مدل باید ابتدا زنجیره استدلال را تولید و سپس بر اساس آن، عمل را تولید کند [۶۸].

---

[۱] reasoning-action disconnection

[۲] reasoning-conditioned action generation

[۳] Self-Generated Reasoning

[۴] plug-in diffusion expert

◘ VLM + سَرِ عمل تطبیق جریان

مدل‌های VLM + سَرِ عمل تطبیق جریان،[۱] مدل انتشار را با یک سَرِ عمل تطبیق جریان [۶۹] جایگزین می‌کنند و پاسخگویی بلادرنگ را بهبود می‌بخشند در حالی که کنترل پیوسته و روان را حفظ می‌نمایند. یک نمونه معرف pi-0 [۷۰] است که بر پایه PaliGemma [۷۱] بنا شده و به نرخ کنترل تا ۵۰ هرتز دست می‌یابد.

مدل‌های قبلی مانند Octo [۲۴] و RDT-1B [۴۶] از سیاست‌های انتظار برای تولید عمل استفاده می‌کنند. این روش در مدل‌سازی توزیع‌های پیچیده و پیوسته عمل (مثلاً برای حرکات دقیق دست) عالی است، اما یک نقطه ضعف اساسی دارد: سرعت استنتاج پایین.

برای تولید یک عمل، مدل انتشار باید یک فرآیند نویززدایی تکرارشونده را طی کند [۴۳،۴۶]. این فرآیند نیازمند چندین مرحله محاسباتی است تا به تدریج از یک نویز تصادفی به یک بردار عمل دقیق برسد. این ماهیت چندمرحله‌ای، دستیابی به فرکانس‌های کنترل بالا (مانند ۳۰ تا ۵۰ هرتز) که برای کنترل ربات به صورت بلادرنگ و روان ضروری است را تقریباً غیرممکن می‌سازد.

در اینجا مدل pi-0 [۷۰] مستقیماً برای حل این مشکل «تأخیر محاسباتی» طراحی شده است. این مدل، سیاست دیفیوژن را به طور کامل با یک تکنیک جدیدتر به نام تطبیق جریان جایگزین می‌کند [۶۹].

معماری pi-0 [۷۰] به صورت هوشمندانه‌ای اجزای ادراکی و اجرایی را ترکیب می‌کند. مدل از PaliGemma [۷۱] (یک VLM متن-باز قدرتمند) به عنوان ستون فقرات اصلی خود استفاده می‌کند. این بخش وظیفه درک صحنه بصری (تصویر ورودی) و درک دستورالعمل چندوجهی (متن یا تصویر هدف) را بر عهده دارد. ستون فقرات PaliGemma [۷۱] پس از پردازش ورودی‌ها، یک بازنمایی فشرده و غنی از اطلاعات را در قالب یک «نشانه بازخوانی» خاص خروجی می‌دهد. این نشانه، عصاره‌ی "درک" مدل از وضعیت فعلی و هدف وظیفه است.

اینجاست که تطبیق جریان وارد می‌شود [۶۹]. «متخصص عمل» یک تکه مجزا است که با تطبیق جریان آموزش دیده است. برای تولید عمل نهایی، دو ورودی به «متخصص عمل» داده می‌شود:

۱. نشانه بازخوانی (از VLM، که می‌گوید «چه کاری» باید انجام شود).

۲. اطلاعات حالت ربات[۲] (مانند وضعیت فعلی مفاصل یا کنترل‌کننده نهایی، که می‌گوید ربات

---

[۱] Flow Matching Action Head

[۲] Proprioception

«کجاست»).

این متخصص عمل سپس با استفاده از نگاشت تک‌مرحله‌ای تطبیق جریان، به سرعت یک «قطعه عمل» پیوسته و روان را تولید می‌کند [۷۰].

مدل‌های زبان-بینایی-عمل (VLA) همه‌منظوره، مانند OpenVLA [۳۴]، در درک دستورات سطح بالا و برنامه‌ریزی توالی وظایف (مثلاً «سیب را بردار و در کاسه بگذار») پیشرفت‌های چشمگیری داشته‌اند. با این حال، مدل‌های VLA همه‌منظوره اغلب در عمل فیزیکیِ «گرفتن[۱]» ضعیف عمل می‌کنند، زیرا فاقد درک عمیق فیزیکی هستند.

مشکلات اصلی این مدل‌ها در سه بخش دسته‌بندی می‌شود:

۱. کمبود دانش فیزیکی: مدل‌های همه‌منظوره که بر روی داده‌های اینترنتی آموزش دیده‌اند، درک عمیقی از فیزیک سه‌بعدی و تعامل را ندارند. آنها ممکن است «چکش» را تشخیص دهند، اما دانش ذاتی در مورد اینکه بهترین نقطه برای گرفتن چکش «دسته» آن است و نه «سر» آن، ندارند.

۲. محدودیت داده‌های آموزشی ربات: مجموعه داده‌های واقعی ربات (مانند Open-X Embodiment [۷۴]) اگرچه بزرگ هستند، اما در مقایسه با میلیاردها تصویر اینترنتی، هنوز «محدود» محسوب می‌شوند. این داده‌ها نمی‌توانند تمام روش‌های ممکن برای گرفتن تمام اشیاء ممکن در تمام زوایای ممکن را پوشش دهند. این کمبود داده، آموزش یک سیاست گرفتن دقیق را بسیار دشوار می‌کند.

۳. چالش‌های معماری (سرعت و دقت): همانطور که اشاره شد، مدل‌های VLA برای تولید عمل یا از نشانه‌های (تکواژهای) گسسته استفاده می‌کنند (که سریع هستند اما دقت کافی برای ژست‌های پیوسته را ندارند [۴،۳۴]) و یا از مدل‌های انتشار (که پیوسته و دقیق هستند اما به دلیل نیاز به چندین مرحله نویززدایی، بسیار کند هستند و برای کنترل بلادرنگ مناسب نیستند) [۲۴،۴۶].

به طور خاص به عنوان یک «مدل پایه تخصصی برای گرفتن» طراحی شده است. این مدل، مشکلات فوق را از طریق دو تکنیک اصلی و مکمل حل می‌کند:

۱. پیش‌آموزش در مقیاس میلیارد بر روی داده‌های مصنوعی (حل مشکل داده)

GraspVLA [۷۵] توسط آقای دنگ و همکاران به جای اتکای صِرف به داده‌های محدود واقعی، از یک استراتژی دو مرحله‌ای «مصنوعی-سپس-واقعی[۲]» استفاده می‌کند:

---

[۱] Grasping

[۲] Synthetic-then-Real

- مرحله پیش‌آموزش: مدل ابتدا بر روی یک مجموعه داده عظیم و اختصاصی به نام Grasp-Anything [۷۶] آموزش می‌بیند. این یک مجموعه داده مصنوعی در مقیاس میلیارد است که به طور خاص برای آموزش «تعامل گرفتن» ساخته شده است. این داده‌ها به مدل یاد می‌دهند که چگونه ویژگی‌های بصری (از VLM) را به ژست‌های گرفتنِ با کیفیت بالا مرتبط سازد. این مرحله، دانش فیزیکی و هندسی عمیقی را که در داده‌های واقعی کمیاب است، به مدل تزریق می‌کند.
- مرحله تنظیم دقیق: پس از اینکه مدل با داده‌های مصنوعی به یک «متخصص گرفتن» تبدیل شد، سپس بر روی داده‌های واقعی ربات (مانند Open-X [۷۴]) تنظیم دقیق می‌شود. این مرحله به مدل اجازه می‌دهد تا دانش نظری خود را با فیزیک دنیای واقعی و خصوصیات ربات خاص تطبیق دهد.

۲. معماری VLM + تطبیق جریان برای حل مشکل معماری [۶۹،۷۵]

GraspVLA [۷۵] برای اجرای این دانش، از یک معماری کارآمد VLA استفاده می‌کند که سرعت و دقت را همزمان دارد:

- ستون فقرات VLM: مدل از یک VLM قدرتمند (InternLM2-VL [۷۷]) به عنوان ستون فقرات ادراکی خود استفاده می‌کند. این بخش وظیفه دارد ورودی تصویر و دستورالعمل متنی (مثلاً «بطری را بردار») را پردازش کرده و درک کند.
- رمزگشای عمل با تطبیق جریان: این بخش، تکنیک کلیدی مشترک با pi-0 [۷۰] است. به جای استفاده از مدل انتشار کُند [۴۳،۴۶]، GraspVLA [۷۵] از یک «متخصص عمل» مبتنی بر تطبیق جریان استفاده می‌کند [۶۹]. این ماژول، بازنمایی فشرده VLM را دریافت کرده و آن را مستقیماً به یک عمل پیوسته نگاشت می‌دهد.

چرا تطبیق جریان مهم است؟ زیرا این تکنیک (برخلاف مدل‌های انتشار) می‌تواند در یک گام محاسباتی واحد، یک ژست گرفتن دقیق و پیوسته تولید کند [۶۹]. این ویژگی GraspVLA [۷۵] را قادر می‌سازد تا به فرکانس‌های کنترل بالا مورد نیاز برای مانورهای بلادرنگ دست یابد، در حالی که دقت بالای مدل‌سازی پیوسته را نیز حفظ می‌کند.

- تفاوت نگرش GraspVLA

در حالی که مدل‌هایی مانند OpenVLA [۳۴] سعی می‌کنند «همه‌چیزدان» باشند (هم برنامه‌ریزی سطح بالا و هم اجرای سطح پایین)، GraspVLA [۷۵] یک نگرش متفاوت دارد. این مدل می‌پذیرد که «گرفتن» یک مهارت بنیادی، بسیار پیچیده و حیاتی است که نیازمند یک مدل پایه تخصصی است [۷۵].

بنابراین، GraspVLA یک مدل همه‌منظوره برای برنامه‌ریزی وظایف پیچیده نیست؛ بلکه یک مدل پایه تخصصی برای عمل گرفتن است. این مدل با ترکیب پیش‌آموزش گسترده بر روی داده‌های مصنوعی (برای یادگیری «چه چیزی» و «کجا» باید گرفته شود) [۷۶] و یک رمزگشای تطبیق جریان [۶۹] (برای اجرای «چگونه» گرفتن به صورت سریع و دقیق)، به طور قابل توجهی قوی‌ترین مدل در این حوزه تخصصی محسوب می‌شود [۷۵].

◘ VLM + مبدل انتشار

مدل‌های VLM + مبدل انتشار، یک VLM را با یک مبدل انتشار [۴۵] ترکیب می‌کنند. VLM معمولاً به عنوان یک سیاست سطح بالا (سیستم ۲) عمل می‌کند، در حالی که مبدل انتشار به عنوان یک سیاست سطح پایین (سیستم ۱) عمل می‌نماید. مبدل انتشار [۴۳] ممکن است با استفاده از انتشار یا تطبیق جریان [۶۹] پیاده‌سازی شود. یک مدل معرف GR00T N1 از شرکت انویدیا [۱]است [۷۸] که توجه متقاطع [۲۶] را از مبدل انتشار به تکواژهای VLM اعمال می‌کند و عمل‌های پیوسته را از طریق تطبیق جریان [۶۹] تولید می‌نماید. این طراحی در CogACT [۷۹] نیز استفاده می‌شود.

پژوهش آقای لی و همکاران که به مدل CogAct [۷۹] ختم شد به عنوان یک راه‌حل مستقیم برای چالش «فراموشی فاجعه‌بار» [۶۵] در مدل‌های VLA یکپارچه مانندOpenVLA ارائه شده است؛ مشکلی که در آن، تنظیم دقیق مدل برای وظایف کنترلی ربات، منجر به تخریب قابلیت‌های استدلال عقل سلیم و درک زبانی VLM پیش‌آموخته می‌شود. CogACT [۷۹] این تعارض شناختی-عملی[۲] را با پیاده‌سازی یک معماری دو-سیستمی[۳] صریح حل می‌کند که استدلال سطح بالا (سیستم ۲) را از اجرای عمل سطح پایین (سیستم ۱) جدا می‌سازد. به عنوان یک VLM، این مدل از یک OpenVLA (مبتنی بر LLaMA [۵۰]) به عنوان «مدل شناختی» خود استفاده می‌کند، اما برخلاف معماری اصلی، این ستون فقرات در طول آموزش رباتیک کاملاً منجمد (frozen) باقی می‌ماند [۷۹].

تکنیک برنامه‌ریزی وظیفه در CogACT به این صورت عمل می‌کند که VLM منجمد، دستور زبان و مشاهدات بصری را پردازش کرده و به جای تولید مستقیم تکواژهای عمل، یک «راهنمایی نهفته» یا بازنمایی شناختی سطح بالا تولید می‌کند. این بازنمایی سپس از طریق توجه متقابل [۲۶] به یک «مدل

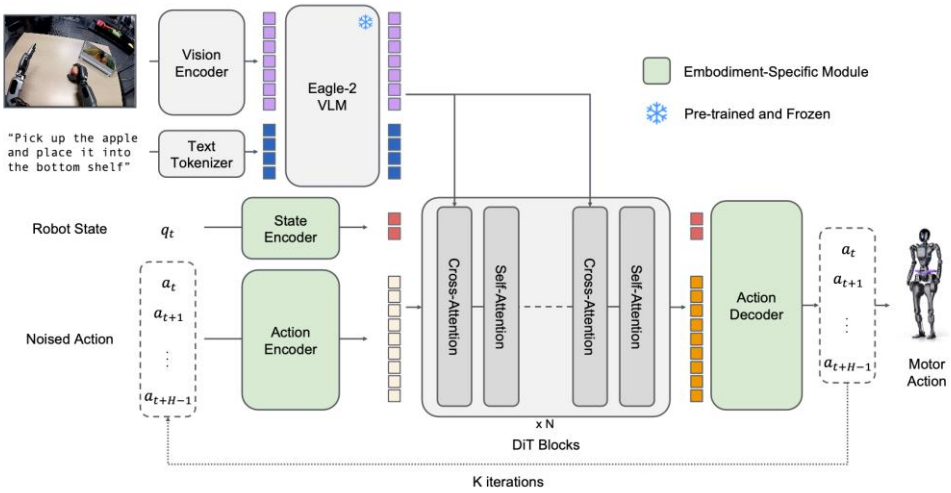
---

[۱] Nvidia

[۲] cognition-action

[۳] Dual-System

عمل» مجزا و تخصصی خورانده می‌شود. این مدل عمل، یک مبدل انتشار (DiT) [۴۵] است که با الهام از pi-0 [۷۰] از تطبیق جریان [۶۹] برای تولید اَعمال پیوسته به صورت بسیار سریع و کارآمد استفاده می‌کند. نوآوری اصلی این مدل، جداسازی پارامتری کامل بین ماژول شناختی (که دانش وب را حفظ می‌کند) و ماژول عمل (که مهارت‌های حرکتی را یاد می‌گیرد) [۷۹] است؛ این کار امکان هم‌افزایی بین استدلال پیچیده و کنترل دقیق بلادرنگ را بدون تخریب هیچ‌یک از قابلیت‌ها فراهم می‌آورد.



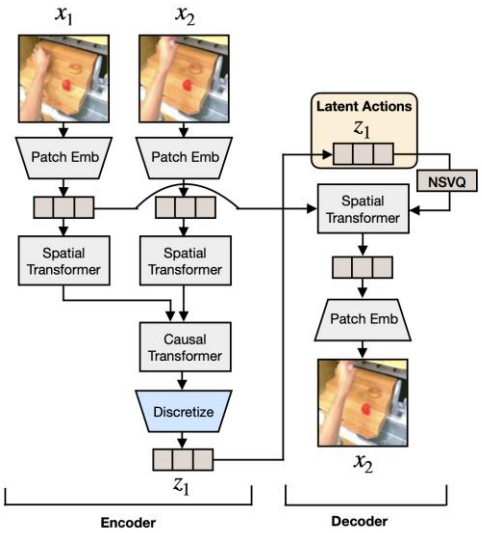
شکل (۳-۴) معماری مدل GR00T N1 بر روی مجموعه‌ای متنوع از تجسم‌ها.

مدل GR00T N1 [۷۸] یک مدل پایه متن-باز است که به طور خاص برای ربات‌های انسان‌نما[۱] طراحی شده و در دسته معماری‌های دو-سیستمی [۷۸،۷۹] قرار می‌گیرد. به عنوان یک VLM، این مدل از یک ماژول زبان به عنوان «سیستم ۲» (ادراک و استدلال) استفاده می‌کند تا محیط و دستورات زبانی را تفسیر کند. تکنیک برنامه‌ریزی وظیفه و اجرای آن در GR00T N1 [۷۸] مبتنی بر جداسازی صریح شناخت از عمل است. VLM (سیستم ۲) وظیفه استدلال سطح بالا را بر عهده دارد و خروجی آن، یک ماژول «سیستم ۱» مجزا را هدایت می‌کند. این سیستم ۱، یک مبدل انتشار [۴۵] است که با بهره‌گیری از تطبیق جریان [۶۹]، اعمال حرکتی پیوسته و روان را در زمان واقعی تولید می‌کند. نوآوری کلیدی این مدل در آموزش مشترک و سرتاسری این دو ماژول بر روی یک مجموعه داده ترکیبی ناهمگن[۲] است. GR00T N1 [۷۸] نه‌تنها از مسیرهای واقعی ربات و داده‌های شبیه‌سازی‌شده استفاده می‌کند، بلکه با بهره‌گیری از تکنیک‌هایی مانند LAPA [۸۰]، قادر است «اعمال نهفته» را مستقیماً از ویدیوهای انسانی بدون برچسب بیاموزد. این

[۱] Humanoid

[۲] heterogeneous mixture

مزیت به مدل اجازه می‌دهد تا شکاف تجسم را پر کرده و مهارت‌های پیچیده دستکاری دو بازویی را از طریق مشاهده انسان‌ها فرا گیرد، که این امر آن را برای ربات‌های انسان‌نما بسیار مقیاس‌پذیر می‌سازد [۷۸].



شکل (۳-۵) معماری مدل کنش پنهان.

۳-۲-۲- معماری‌های سلسله‌مراتبی

پیشرفت‌های اخیر جهت‌گیری به سمت معماری‌های سلسله‌مراتبی را برجسته می‌سازد. این رویکرد، بازنمایی‌های میانی ساختاریافته‌ای[۱] را بین دستورالعمل‌های زبانی و اقدامات سطح پایین معرفی می‌کند که امکان برنامه‌ریزی، تجزیه[۲] و استدلال قوی‌تری را فراهم می‌آورند.

◘ تجزیه مبتنی بر وظیفه فرعی[۳]

بنیادی‌ترین رویکرد، استفاده از VLMهای موجود به عنوان سیاست‌های سطح بالا برای تجزیه دستورالعمل‌های وظیفه به زیروظیفه‌ها[۴] است [۱،۳]. سپس این توصیفات زیروظیفه به یک VLA (سیاست سطح پایین) منتقل می‌شوند.

مدل آقای آن وهمکاران با نام SayCan [۱] به عنوان یکی از کارهای بنیادی در الگوی سلسله‌مراتبی،

[۱] structured intermediate representations

[۲] decomposition

[۳] Subtask-based Decomposition

[۴] subtasks

مشکل اساسی «عدم اتصال به واقعیت[۱]» در مدل‌های زبانی بزرگ (LLM) را هدف قرار می‌دهد؛ LLMها دانش معنایی گسترده‌ای در مورد «چه کاری باید انجام شود» دارند، اما فاقد درک تجسم‌یافته از «چه کاری می‌توان انجام داد» در یک محیط فیزیکی خاص هستند. در این معماری، یک LLM عظیم (مانند PaLM [۳] یا FLAN [۸۱]) به عنوان یک برنامه‌ریز وظیفه سطح بالا عمل می‌کند، اما برخلاف مدل‌های یکپارچه، مستقیماً دستورات حرکتی تولید نمی‌کند.

تکنیک برنامه‌ریزی وظیفه در SayCan [۱] یک فرآیند دو مرحله‌ای پویا است: ابتدا، در مرحله « Say»، LLM دستورالعمل سطح بالای کاربر (مثلاً «من نوشیدنی‌ام را ریختم، کمکم کن») را دریافت کرده و لیستی از وظایف فرعی متنیِ محتمل (مانند «یک اسفنج پیدا کن»، «نزدیک‌ترین نوشابه را بردار») را به عنوان «نمایش میانی[۲]» پیشنهاد می‌دهد. سپس، در مرحله «Can»، این پیشنهادات معنایی باید در دنیای فیزیکی «زمینه‌مند» شوند. این کار از طریق مجموعه‌ای از سیاست‌های سطح پایین یا «توابع ارزش[۳]» از پیش‌آموخته انجام می‌شود که هر کدام مختص یک مهارت اتمی هستند. این توابع، «کنش‌پذیری» یا همان امکان‌پذیریِ[۴] اجرای هر یک از آن وظایف فرعی پیشنهادی را با توجه به مشاهده بصری فعلی ربات، تخمین می‌زنند.

در نهایت، SayCan [۱] با ضرب کردن امتیاز معنایی LLM (آنچه از نظر زبانی مفید است) در امتیاز کنش‌پذیری (آنچه از نظر فیزیکی ممکن است)، بهترین اقدام بعدی را انتخاب کرده و سیاست سطح پایین مربوطه را فراخوانی می‌کند. نوآوری اصلی SayCan [۱] در استفاده از دانش عقل سلیم LLMها برای برنامه‌ریزی وظایف پیچیده و طولانی‌مدت و در عین حال، جلوگیری از «هذیان‌گویی[۵]» از طریق فیلتر کردن گزینه‌های غیرممکن فیزیکی است [۱]. محدودیت اصلی این رویکرد، اتکای آن به یک کتابخانه مهارت ثابت و پیش‌تعریف‌شده است؛ مدل قادر به یادگیری مهارت‌های حرکتی جدید نیست، بلکه تنها مهارت‌های موجود را به شیوه‌ای هوشمندانه توالی‌بندی می‌کند.

پژوهش آقای دریس و همکاران با نام مدل PaLM-E [۳] یک مدل زبان چندوجهی در مقیاس بزرگ، نشان‌دهنده یک پیشرفت کلیدی در معماری‌های سلسله‌مراتبی است. به عنوان یک VLM، این مدل یک

---

[۱] un-groundedness

[۲] intermediate representation

[۳] value functions

[۴] probability of success

[۵] hallucination

مدل زبانی بزرگ (LLM) از پیش‌آموخته (مانند PaLM) را با ورودی‌های پیوسته و تجسم‌یافته (مانند تصاویر و حالات ربات) ادغام می‌کند. نوآوری اصلی PaLM-E [۳] در تکنیک «تزریق مُدالیته[۱]» آن نهفته است؛ به جای استفاده از تکواژهای بصری انتزاعی، این مدل مشاهدات حسی (مانند تصاویر) را از طریق یک رمزگذار بصری پردازش کرده و بردار حاصل را مستقیماً به فضای امبدینگ LLM «تزریق» می‌کند. این امبدینگ‌های بصری سپس دقیقاً مانند تکواژهای متنی، بخشی از «جملات چندوجهی» ورودی (مانند «<تصویر> چیپس را برایم بیاور») می‌شوند [۳].

برای برنامه‌ریزی وظایف، PaLM-E [۳] به عنوان یک برنامه‌ریز سطح بالا عمل می‌کند و مستقیماً دستورات حرکتی سطح پایین را تولید نمی‌کند. در عوض، مدل با دریافت یک دستورالعمل و مشاهدات بصریِ تزریق‌شده، به صورت خودبازگشتی یک دنباله از دستورات متنی وظایف فرعی را به عنوان نمایش میانی یا همان برنامه تولید می‌کند (مانند: «۱. به سمت کشو برو»، «۲. کشوی بالا را باز کن»، «۳. چیپس سبز را بردار»). این برنامه‌های متنی سپس برای اجرا به مجموعه‌ای مجزا از سیاست‌های سطح پایین از پیش‌آموخته ارسال می‌شوند. مزیت اصلی این رویکرد، توانایی مدل در انتقال دانش[۲] از پیش‌آموزش عظیم وب (مانند درک مفاهیم) به وظایف برنامه‌ریزی رباتیکِ زمینه‌مند و پیچیده است. با این حال، محدودیت اساسی آن این است که قابلیت‌های اجرایی مدل به طور کامل به کتابخانه مهارت سطح پایینِ از پیش‌آموخته محدود می‌شود [۱،۳] و خود مدل قادر به تولید حرکات جدید نیست.

مدل RT-H یا (Robotics Transformer with Hierarchies) توسط آقای بلکال و همکاران [۳۹]، یک چارچوب سلسله‌مراتبی ارائه می‌دهد که شکاف بین مدل‌های کاملاً تجزیه‌شده[۳] مانند SayCan [۱] و مدل‌های یکپارچه مانند RT-2 [۴] را پر می‌کند. به عنوان یک VLM، این مدل مستقیماً بر پایه معماری RT-2 [۴] (و در نتیجه ستون فقرات VLM آن مانند PaLI-X [۴،۳۹]) ساخته شده است. برخلاف مدل‌هایی که برنامه‌ریز و اجراکننده کاملاً مجزا دارند، RT-H [۳۹] یک مدل واحد و یکپارچه است که می‌تواند به صورت پویا بین دو سطح انتزاع عمل کند. تکنیک برنامه‌ریزی وظیفه در این مدل، معرفی یک نمایش میانی جدید به نام «حرکت زبانی[۴]» است. این «حرکت زبانی» یک توصیف متنی از یک مهارت اتمی

---

[۱] modality injection

[۲] knowledge transfer

[۳] decoupled

[۴] Language Motion

و سطح پایین است (مانند «بازو را به جلو حرکت بده»، «بازو را به راست بچرخان»، یا «دست را ببند») [۳۹].

فرآیند برنامه‌ریزی و اجرا در RT-H [۳۹] به این صورت عمل می‌کند:

۱) VLM ابتدا دستورالعمل سطح بالا (مثلاً «در شیشه پسته را ببند») و تصویر فعلی را دریافت می‌کند و با دستورات متنی شدن برای «حرکت زبانی»، یک زیروظیفه متنی (مثلاً «بازو را به راست بچرخان») را به صورت خودبازگشتی پیش‌بینی می‌کند.

۲) سپس، این «حرکت زبانیِ» تولید شده، به عنوان یک شرط جدید، به دستورات متنی ورودی مدل اضافه می‌شود.

۳) در مرحله بعد، همان VLM، این بار با دستورات متنی حاوی «حرکت زبانی»، فراخوانی می‌شود تا نشانه‌های (تکواژهای) عمل گسسته سطح پایین و اجرایی (مشابه RT-2 [۴]) را تولید کند که متناظر با آن زیروظیفه خاص هستند. این حلقه تکرار می‌شود و مدل به طور متناوب بین تولید «حرکت زبانی» (برنامه‌ریزی) و تولید «عمل ربات» (اجرا) جابجا می‌شود.

نوآوری اصلی RT-H [۳۹] در این است که یک مدل واحد می‌تواند هم به عنوان برنامه‌ریز سطح بالا (تولیدکننده زبان) و هم به عنوان سیاست سطح پایین (تولیدکننده تکواژ عمل) عمل کند و این جابجایی تنها از طریق تغییر دستورات متنی[۱] صورت می‌گیرد. مزیت بزرگ این رویکرد، قابلیت تفسیرپذیری و مداخله[۲] است؛ از آنجایی که نمایش میانی (زیروظیفه) به زبان طبیعی است، یک انسان می‌تواند اجرای ناموفق را با ارائه یک «حرکت زبانی» اصلاحی، مستقیماً هدایت کند و ربات از آن نقطه به بعد به صورت یکپارچه ادامه دهد. این مدل همچنین در وظایف پیچیده و طولانی‌مدت عملکرد بهتری از خود نشان می‌دهد. با این حال، محدودیت اصلی آن این است که همچنان به فضای عمل گسسته و تکواژشده‌ی RT-2 متکی است و دقت مدل‌های با فضای عمل پیوسته را ندارد [۴].

مدل آقای شی و همکاران با نام HiRobot [۷۷] یک چارچوب سلسله‌مراتبی پیشرفته است که برای اجرای دستورالعمل‌های طولانی‌مدت[۳] و پایان-باز طراحی شده است. برخلاف رویکردهایی که از VLMهای از پیش‌آموخته‌ی موجود به عنوان برنامه‌ریز سطح بالا استفاده می‌کنند (مانند SayCan) [۱] یک سیاست

---

[۱] prompt switching

[۲] Interpretability and Intervention

[۳] long-horizon

سطح بالای سفارشی را آموزش می‌دهد. به عنوان یک VLM، این مدل از یک ستون فقرات مانند PaliGemma [۷۱] برای درک صحنه و دستورالعمل کاربر استفاده می‌کند. تکنیک برنامه‌ریزی وظیفه در این مدل، یک تجزیه سلسله‌مراتبی صریح است: VLM سطح بالا دستورالعمل کلی کاربر (مثلاً «میز را تمیز کن») را دریافت کرده و آن را به دنباله‌ای از دستورات اتمی یا زیروظایف قابل تفسیر (مانند «اسفنج را بردار»، «به سمت لکه برو»، «لکه را پاک کن») تجزیه می‌کند. نکته کلیدی اینجاست که این زیروظیفه‌های متنی، به عنوان «نمایش میانی»، مستقیماً به یک سیاست سطح پایین مجزا ارسال می‌شوند. این سیاست سطح پایین، خود یک VLA قدرتمند مانند Pi-0 [۷۰] است که با استفاده از تطبیق جریان [۶۹] آموزش دیده است تا دستورات اتمی و زمینه‌مند را به اَعمال پیوسته و روان تبدیل کند. مزیت اصلی این رویکرد دوگانه، ترکیب توانایی استدلال سطح بالای VLM (برای برنامه‌ریزی) با کارایی و سرعت بالای سیاست سطح پایین مبتنی بر تطبیق جریان است. با دریافت دستورات اتمی و تمیز، سیاست سطح پایین می‌تواند اَعمال را با اطمینان بیشتری نسبت به زمانی که با دستورالعمل‌های پیچیده و مبهم مواجه است، اجرا کند [۶۹،۷۰].

◘ مدل‌های مبتنی بر برنامه‌نویسی

کار آقای لیانگ و همکاران با عنوان CaP: Code as Policies[۱۱] یک چارچوب پیشگام در الگوی سلسله‌مراتبی است که از توانایی‌های نوظهور مدل‌های زبانی بزرگِ آموزش‌دیده بر روی کد (مانند Codex [۷۸]) به عنوان سیاست اصلی ربات بهره می‌برد. تکنیک برنامه‌ریزی وظیفه در این مدل، تولید برنامه است. به جای تولید مستقیم تکواژهای عمل یا زیروظیفه‌های متنی، LLM از طریق دستورات متنی چند-نمونه[۱] آموزش می‌بیند تا یک برنامه کامل پایتون را در پاسخ به دستورالعمل زبان طبیعی تولید کند. این برنامه تولید شده، که به آن «برنامه مدل زبانی» گفته می‌شود، به عنوان نمایش میانی عمل کرده و به صورت سلسله‌مراتبی، APIهای سطح پایین ادراک و کنترل را فراخوانی می‌کند. نوآوری اصلی CaP [۱۱] در این است که با بهره‌گیری از کتابخانه‌های خارجی (مانند NumPy) و ساختارهای منطقی (مانند حلقه‌ها و شرط‌ها)، قادر به انجام استدلال فضایی-هندسی پیچیده (مثلاً «حرکت به نقطه میانی بین A و B») و درک «عقل سلیم رفتاری» است؛ قابلیت‌هایی که در مدل‌های یکپارچه به سادگی قابل دستیابی نیستند. با این حال، این رویکرد به شدت به در دسترس بودن و کیفیت یک کتابخانه API از پیش‌تعریف‌شده وابسته است

[۱] few-shot prompting

.[۱۱]

مطالعه آقای سینگ و همکاران با عنوان PROGPROMPT [۱۳] نیز در الگوی سلسله‌مراتبی مبتنی بر کد قرار می‌گیرد و چالش تولید برنامه‌های وظیفه‌ای که هم «زمینه‌مند» و هم منطبق بر قابلیت‌های ربات باشند را هدف قرار می‌دهد. برخلاف مدل‌هایی که صرفاً دستورات متنی تولید می‌کنند، PROGPROMPT [۱۳] از یک LLM (مانند GPT-3) [۳۰] برای تولید مستقیم کد پایتون به عنوان برنامه اجرایی استفاده می‌کند. تکنیک کلیدی این مدل در مهندسی دستورات متنی برنامه‌مانند[۱] آن نهفته است. به جای دستورات متنی زبان طبیعی ساده، LLM با یک ساختار متنی تغذیه می‌شود که شامل سه بخش است:

۱) تعاریف APIهای موجود (مهارت‌های ربات)

۲) لیستی از اشیاء موجود در صحنه به همراه وضعیت آن‌ها (درک صحنه)

۳) نمونه‌های چند-نمونه از برنامه‌های کامل. این ساختار غنی، LLM را مجبور می‌کند تا برنامه‌هایی تولید کند که به طور همزمان به دستورالعمل، وضعیت فعلی محیط و قابلیت‌های خاص ربات وفادار باشند. مزیت اصلی این روش، تولید برنامه‌هایی است که به طور پیش‌فرض زمینه‌مند هستند و خطاهای ناشی از هذیان‌گویی مدل زبانی برای پیشنهاد اقدامات غیرممکن را کاهش می‌دهند.

مدل Instruct2Actl توسط آقای هوانگ و همکاران [۷۹] نیز یک چارچوب سلسله‌مراتبی مبتنی بر تولید کد است که بر ادغام مدل‌های پایه بصری به عنوان APIهای ادراکی تأکید دارد. در این مدل، یک LLM (مانند GPT-3 [۳۰]) وظیفه تولید یک برنامه کامل پایتون را بر عهده دارد که حلقه ادراک، برنامه‌ریزی و عمل را پیاده‌سازی می‌کند. تکنیک برنامه‌ریزی وظیفه در اینجا، یک فرآیند کدنویسی ساختاریافته است: برنامه تولید شده ابتدا APIهای ادراکی قدرتمندی مانند Segment Anything Model (SAM) [۵۸] را برای مکان‌یابی دقیق و بخش‌بندی[۲] اشیاء کاندید فراخوانی می‌کند؛ سپس از CLIP [۶] برای طبقه‌بندی معنایی آن ماسک‌ها و تطبیق آن‌ها با دستورالعمل ورودی استفاده می‌کند. پس از این مرحله ادراکِ دقیق و زمینه‌مند، LLM APIهای عملِ از پیش‌تعریف‌شده (مانند PickPlace) را با پارامترهای صحیح فراخوانی می‌کند. نوآوری کلیدی Instruct2Act [۷۹] در بهره‌گیری مستقیم از قابلیت‌های تعمیم صفر-نمونه مدل‌های پایه ادراکی مدرن مانند SAM [۵۸] است که به آن اجازه می‌دهد دستورات پیچیده چندوجهی را به کدهای اجرایی دقیق تبدیل کند. این مدل نیز مانند سایر رویکردهای این دسته، در اجرای حرکات جدید

[۱] program-like prompt structure

[۲] segmentation

به کتابخانه APIهای حرکتی خود محدود است.

◘ مدل مبتنی بر قابلیت‌دهی[۱]

قابلیت‌دهی به امکانات عملیاتی اشاره دارد که یک محیط برای یک عامل ارائه می‌دهد. در رباتیک، این مفهوم به ویژگی‌های قابل اجرا اشیا یا صحنه‌ها (چه اقداماتی ممکن است) تطبیق داده می‌شود. VLAهای مبتنی بر پیش‌بینی قابلیت‌دهی در سه نوع دسته‌بندی می‌شوند:

(۱) پیش‌بینی قابلیت‌دهی و تولید عمل با استفاده از VLMها VLMهای از پیش‌آموزش‌دیده اغلب برای تخمین قابلیت‌دهی‌ها و تولید اعمال مربوطه استفاده می‌شوند.

مدل VoxPoser که توسط آقای هیوانگ و همکارانشان [۸۰] عرضه شد یک معماری سلسله‌مراتبی نوآورانه است که LLMها را به مولدین ترکیبیِ[۲] توابع هزینه سه‌بعدی تبدیل می‌کند. یک LLM سطح بالا (مانند GPT-4) [۸۱] به عنوان برنامه‌ریز وظیفه عمل می‌کند. تکنیک کلیدی، تولید کد پایتون است که به صورت پویا، ماژول‌های ادراک مبتنی بر VLM (مانند CLIP [۶]) را فراخوانی و ترکیب می‌کند. این کد تولید شده، صحنه سه‌بعدی را جستجو کرده و مجموعه‌ای از «نقشه‌های ارزش وُکسلی[۳]» را می‌سازد. این نقشه‌های سه‌بعدی، دانش معنایی LLM را به صورت فضایی زمینه‌مند می‌کنند و «کنش‌پذیری‌های» مرتبط با وظیفه و «محدودیت‌های» (نواحی ممنوعه) را کدگذاری می‌کنند. این نقشه‌های ارزش ترکیبی سپس به عنوان تابع هزینه نهایی به یک برنامه‌ریز حرکت مبتنی بر بهینه‌سازی[۴] (مانند MPPI [۸۲]) داده می‌شوند تا مسیر ربات را به صورت صفر-نمونه مشخص کند. مزیت اصلی، تعمیم‌پذیری فوق‌العاده به اشیاء و دستورات جدید است، اما محدودیت آن، وابستگی به بازنمایی‌های سه‌بعدی دقیق و هزینه محاسباتی بالای برنامه‌ریزی حرکت است [۸۰].

(۲) استخراج قابلیت‌دهی از مجموعه‌های داده انسانی این خط کاری بر استخراج قابلیت‌دهی‌ها از ویدئوهای حرکت انسانی، اغلب بدون حاشیه‌نویسی، تمرکز دارد. VRB [۸۳] نقاط تماس و مسیرهای دست را از ویدئوهای EPIC-KITCHENS [۸۴] با استفاده از آشکارساز دست-شیء (HOD) می‌آموزد. HRP [۸۵] برچسب‌های قابلیت‌دهی را از Ego4D [۶۱] استخراج می‌کند. VidBot [۸۶] نمایش‌های قابلیت‌دهی

---

[۱] AFFORDANCE-BASED MODEL

[۲] composable generators

[۳] Voxel-based Value Maps

[۴] optimization-based motion planner

دوبعدی را به سه‌بعدی گسترش می‌دهد.

(۳) یکپارچه‌سازی مدل‌های یکپارچه و مدل‌های مبتنی بر قابلیت‌دهی این رویکرد پیش‌بینی قابلیت‌دهی را در VLA ادغام می‌کند.

مدل CLIPort از آقای اشنایدر و همکاران [۵] به عنوان یک چارچوب بنیادی و تأثیرگذار، نشان داد که چگونه می‌توان دانش معنایی غنیِ VLM را برای کارهای مانور رباتیک که نیازمند دقت فضایی هستند، «زمینه‌مند» کرد. این مدل از یک VLM از پیش‌آموخته (مشخصاً CLIP [۶]) به عنوان یک استخراج‌کننده ویژگی دو-مسیره[۱] استفاده می‌کند. تکنیک برنامه‌ریزی وظیفه در CLIPort [۵]، یک برنامه‌ریزی فضایی است، نه زمانی؛ این مدل به جای تجزیه وظیفه به گام‌های متوالی، آن را به دو پرسش اساسی «چه چیزی» و «کجا» تفکیک می‌کند. برای این منظور، مدل دستورالعمل زبانی (مانند «بلوک قرمز را در کاسه آبی بگذار») را با رمزگذار متنی CLIP [۶] و تصویر صحنه را به صورت تماماً کانولوشنی[۲] با رمزگذار بصری CLIP [۶] پردازش می‌کند تا یک نقشه ویژگی پیکسلی متراکم به دست آورد. سپس، در «مسیر توجه[۳]»، این دو بازنمایی با هم ترکیب می‌شوند تا یک «نقشه کُنش‌پذیری برداشتن» پیکسلی را به عنوان نمایش میانی تولید کنند. این نقشه حرارتی[۴]، احتمال موفقیت برداشتن را در هر پیکسل از تصویر نشان می‌دهد. به طور همزمان، «مسیر انتقال[۵]» با در نظر گرفتن تصویر محصول از نقطه برداشتِ پیش‌بینی‌شده، یک «نقشه کُنش‌پذیری قرار دادن» تولید می‌کند. ربات سپس به سادگی با انتخاب نقاطی که دارای بالاترین امتیاز در این دو نقشه حرارتی هستند، عمل برداشتن و قرار دادن را اجرا می‌کند. نوآوری اصلی CLIPort [۵] در استفاده از دانش صفر-نمونه برای تعمیم‌پذیری به اشیای نادیده و دستورات زبانی جدید بود، اما محدودیت عمده آن، عدم توانایی در برنامه‌ریزی وظایف چندمرحله‌ایِ فراتر از یک عمل برداشتن-و-گذاشتن[۶] واحد است.

مدل RoboPoint آقای یوآن و همکاران [۸۷] نیز چالش دقت فضایی در VLMها را با رویکردی متفاوت حل می‌کند. این مدل یک VLM (مانند GPT-4o) [۸۸] را نه به عنوان یک برنامه‌ریز سلسله‌مراتبی، بلکه به

---

[۱] two-stream

[۲] fully convolutional

[۳] Attention pathway

[۴] heatmap

[۵] Transport pathway

[۶] pick-and-place

عنوان یک مولد مستقیم کُنِش‌پذیری فضایی بازآموزی می‌کند. برخلاف CLIPort [۵] که نقشه‌های حرارتی متراکم تولید می‌کند، تکنیک برنامه‌ریزی وظیفه در RoboPoint [۸۷]، پیش‌بینی مستقیم نقاط کلیدی است. برای دستیابی به این هدف، نویسندگان ابتدا یک خط لوله تولید داده مصنوعی در مقیاس بزرگ ایجاد کردند که دستورالعمل‌های زبانی پیچیده (شامل روابط فضایی مانند «سمت راستِ»، «بین» و «زیر») را با مختصات پیکسلی دقیق نقاط کلیدی مرتبط، جفت می‌کند. سپس، VLM پایه بر روی این داده‌ها «تنظیم دقیق دستوری» می‌شود تا یاد بگیرد در پاسخ به یک دستورات متنی چندوجهی (تصویر و متن)، مستقیماً لیستی از مختصات پیکسلی را به عنوان نمایش میانی خروجی دهد. این نقاط کلیدی دوبعدی پیش‌بینی‌شده، سپس با استفاده از اطلاعات عمق، به مختصات سه‌بعدی در دنیای واقعی نگاشت داده شده و به عنوان اهداف (مثلاً نقاط گرفتن یا رها کردن) برای کنترلرهای سطح پایین ربات عمل می‌کنند. نوآوری اصلی RoboPoint در مقیاس‌پذیری آن نهفته است، زیرا با حذف نیاز به داده‌های واقعی ربات و آموزش VLM برای درک زبان فضایی پیچیده، به تعمیم‌پذیری بالایی دست می‌یابد. با این حال، این مدل نیز عمدتاً بر روی زمینه‌مندی یک مرحله‌ای تمرکز دارد و فاقد قابلیت برنامه‌ریزی توالی‌های طولانی و پیچیده است [۲۱،۲۲].

مدل HAMSTER از آقای لی و همکاران [۸۳] که در سال ۲۰۲۵ منتشر شد یک معماری سلسله‌مراتبی صریح [۱،۳،۳۹] را برای ترکیب نقاط قوت VLMها در استدلال سطح بالا و کنترلرهای دقیق سطح پایین ارائه می‌دهد. در این چارچوب، یک VLM قدرتمند (مانند VILA) [۸۴] به عنوان سیاست سطح بالا عمل می‌کند. تکنیک برنامه‌ریزی وظیفه این مدل، تنظیم دقیق VLM برای تولید یک مسیر دوبعدی به عنوان نمایش میانی است. VLM با دریافت دستورالعمل متنی و تصویر RGB ورودی، توالی‌ای از نقاط کلیدی دوبعدی را مستقیماً بر روی صفحه تصویر پیش‌بینی می‌کند که مسیر مطلوب کنترل‌کننده نهایی را مشخص می‌سازد [۸۳]. این مسیر دوبعدی سپس به عنوان «راهنما[۱]» به یک سیاست کنترل سطح پایینِ آگاه از سه‌بعدی و کاملاً مجزا ارسال می‌شود. این سیاست سطح پایین (که می‌تواند یک مدل انتشار سه‌بعدی باشد [۴۳])، ورودی‌های سه‌بعدی دقیق را پردازش کرده و حرکات فیزیکی ظریف و پیوسته لازم برای دنبال کردن مسیر دوبعدی پیشنهادی را اجرا می‌کند. نوآوری کلیدی HAMSTER [۸۳] در این است که با تنزل دادن وظیفه VLM به تولید مسیرهای دوبعدی، آن را از یادگیری کنترل دقیق سه‌بعدی بی‌نیاز می‌کند. این

[۱] guidance

جداسازی به مدل اجازه می‌دهد تا از داده‌های ارزان‌تر و متنوع‌تر «خارج-از-دامنه[۱]» مانند ویدیوهای انسانی برای آموزش برنامه‌ریزی سطح بالا بهره ببرد [۶۱،۸۰].

۳-۲-۳- مدل جهان

مدل‌های جهان قابلیت پیش‌بینی مشاهدات آینده یا نمایش‌های نهفته را بر اساس ورودی‌های فعلی دارند. قابلیت‌های پیش‌بینی رو به جلوی آن‌ها باعث شده است که به طور فزاینده‌ای در سیستم‌های VLA محوریت یابند. این رویکردها به سه نوع گروه‌بندی می‌شوند:

(۱) تولید عمل در مدل‌های جهان برخلاف مدل‌هایی که مستقیماً عمل‌ها را تولید می‌کنند، مدل‌های جهان مشاهدات بصری آینده، مانند تصاویر یا دنباله‌های ویدئویی، را تولید می‌کنند که سپس برای هدایت تولید عمل استفاده می‌شوند [۴۰،۶۰].

- مدل UniPi (Learning Universal Policies via Text-Guided Video Generation) [۳۷] که توسط آقای دو و همکاران منتشر شده یک رویکرد منحصر به فرد مبتنی بر مدل جهان است که برنامه‌ریزی را مستقیماً در فضای بصری انجام می‌دهد. این مدل از یک ستون فقرات سفارشی استفاده می‌کند و بر VLMهای عظیم از پیش‌آموخته متکی نیست. تکنیک برنامه‌ریزی وظیفه در UniPi دو مرحله دارد:
- ۱) برنامه‌ریزی بصری: یک مدل انتشار ویدئویی [۳۷،۴۳]، با دریافت مشاهده فعلی و هدف (متنی یا تصویری)، دنباله‌ای از مشاهدات بصری آینده را تولید می‌کند - اساساً یک «ویدئوی خیالی» از چگونگی تکمیل وظیفه. این دنباله تصاویر، «برنامه» مدل را تشکیل می‌دهد.

۲) استنتاج عمل: سپس، یک مدل پویایی معکوس[۲] این مسیر بصریِ تولید شده را دریافت کرده و دنباله اَعمالی را که برای گذار بین آن فریم‌های بصری متوالی لازم است، استنتاج می‌کند. بنابراین، مدل ابتدا «می‌بیند» چه اتفاقی باید بیفتد و سپس محاسبه می‌کند که «چگونه» آن را انجام دهد. جنبه جهان باز UniPi [۳۷] در تمرکز آن بر یادگیری و پیش‌بینی پویایی‌های بصری نهفته است. با برنامه‌ریزی مستقیم در فضای پیکسل، مدل یاد می‌گیرد که چگونه ظاهر محیط در پاسخ به اَعمال تغییر می‌کند. این درک از پویایی‌های سطح-پیکسل می‌تواند به طور بالقوه نسبت به تغییرات ظاهری (مانند بافت‌ها، نورپردازی) مقاوم‌تر باشد و به مدل اجازه دهد تا در سناریوهای نادیده که در آن‌ها پویایی‌های بصری اصلی حفظ

---

[۱] off-domain

[۲] Inverse Dynamics Model - IDM

می‌شوند، بهتر تعمیم یابد، حتی اگر اشیاء یا پیکربندی‌های خاص جدید باشند. این رویکرد، یادگیری را مستقیماً به مشاهدات خام گره می‌زند.

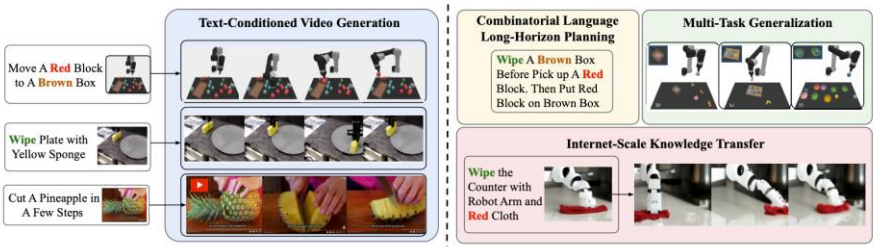
(۲) تولید عمل نهفته از طریق مدل‌های جهان این دسته از VLAها از مدل‌های جهان برای یادگیری نمایش‌های عمل نهفته از نمایش‌های انسانی بهره می‌برند.

مدل LAPA (Latent Action Pre-training from Videos) [۸۰] از آقای یی و همکاران مستقیماً یک سیاست VLA نیست، بلکه یک تکنیک پیش‌آموزش قدرتمند است که از الگوی مدل جهان برای یادگیری بازنمایی‌های عمل قابل انتقال استفاده می‌کند. این تکنیک از یک مدل جهان بهره می‌برد. تکنیک اصلی در LAPA [۸۰]، یادگیری «اَعمال نهفته از ویدئوهای بدون برچسب است. مدل جهان آموزش می‌بینند تا فریم‌های آینده ویدئو را پیش‌بینی کند. در طی این فرآیند پیش‌بینی، مدل به صورت ضمنی[۱] یک بازنمایی فشرده و معنادار از «عملی» که بین فریم‌های مشاهده‌شده رخ داده است را در فضای نهفته خود یاد می‌گیرد. این بازنمایی عمل نهفته، که بدون نیاز به هیچ‌گونه برچسب عمل انسانی یا رباتیک استخراج شده، سپس به عنوان یک ویژگی ورودی قدرتمند یا دانش پیشین برای آموزش مدل‌های VLAی پایین‌دستی استفاده می‌شود. جنبه جهان باز LAPA [۸۰] در توانایی آن برای استخراج دانش عملیاتی از مجموعه داده‌های عظیم و نامحدود ویدئوهای انسانی موجود در اینترنت نهفته است. این رویکرد به طور کامل از نیاز به جمع‌آوری داده‌های گران‌قیمت رباتیک یا حاشیه‌نویسی دقیق اَعمال اجتناب می‌کند. با یادگیری از طیف گسترده‌ای از فعالیت‌های انسانی «در طبیعت»، LAPA [۸۰] یک درک بنیادی و قابل تعمیم از تعاملات فیزیکی را کسب می‌کند که می‌تواند به طور قابل توجهی تعمیم‌پذیری جهان باز و کارایی داده سیاست‌های VLAی آموزش‌دیده با این بازنمایی‌ها را افزایش دهد.

(۳) مدل‌های یکپارچه با مدل‌های جهان ضمنی این دسته به VLAهایی اشاره دارد که به طور مشترک هم عمل‌ها و هم پیش‌بینی‌های مشاهدات آینده را برای بهبود عملکرد خروجی می‌دهند [۶۰،۶۳].

---

[۱] implicitly

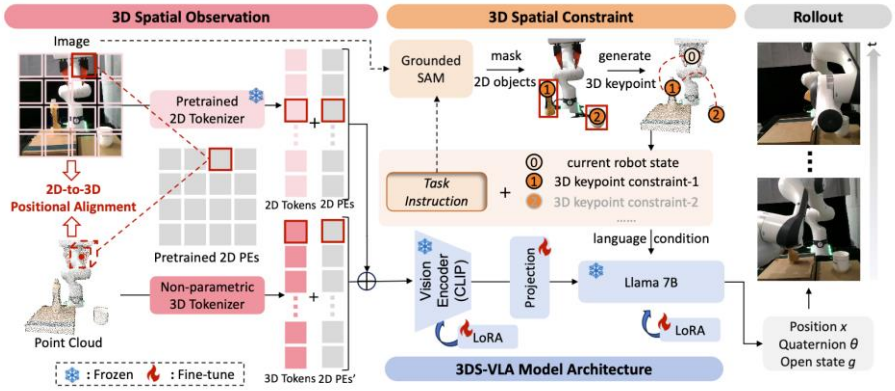


شکل (۳-۶) تولید ویدئوی مشروط به متن به عنوان سیاست‌های جهانی.

مدل WorldVLA آقای چن [۶۳] که قبل‌تر به آن اشاره شد به عنوان یک «مدل جهان-عمل خودبازگشتی معرفی شده و تلاش می‌کند تا به طور همزمان هم توانایی تولید عمل (مانند یک VLA استاندارد) و هم توانایی پیش‌بینی آینده بصری (مانند یک مدل جهان) را در یک چارچوب واحد بیاموزد. این مدل از یک VLM پایه استفاده می‌کند که هم تکواژهای بصری گسسته [۴،۳۴] (از VQ-GAN [۸۵]) و هم تکواژهای عمل گسسته را پردازش می‌کند. تکنیک برنامه‌ریزی وظیفه در این مدل مبتنی بر یادگیری مشترک و هم‌افزا است. مدل به صورت خودبازگشتی آموزش می‌بیند تا هم نشانه‌های عمل بعدی را بر اساس تاریخچه بصری و دستورالعمل پیش‌بینی کند (مانند مدل عمل) و هم نشانه‌های بصری فریم بعدی را بر اساس تاریخچه بصری و نشانه‌های عمل ورودی پیش‌بینی نماید (مانند مدل جهان) [۶۳]. جنبه جهان باز این مدل دقیقاً در همین یادگیری مشترک نهفته است. با وادار کردن مدل به پیش‌بینی پیامدهای بصری اَعمالش، WorldVLA [۶۳] مجبور می‌شود پویایی‌های فیزیکی زیربنایی محیط را درک کند، نه اینکه صرفاً توالی‌های عمل خاصی را حفظ نماید. این درک عمیق‌تر از فیزیک به مدل کمک می‌کند تا در موقعیت‌های نادیده که پویایی‌های مشابهی حاکم است، بهتر عمل کند و تعمیم یابد. برای مقابله با انتشار خطا در تولید قطعات عمل، مدل از ماسک‌گذاری توجه عمل استفاده می‌کند تا تولید هر عمل در یک قطعه، بیشتر به ورودی‌های بصری و متنی متکی باشد تا اَعمال قبلی [۶۳].

مدل 3D-VLA [۶۲] به عنوان یک مدل جهان مولد عمل می‌کند که قابلیت پیش‌بینی آینده بصری را به فضای سه‌بعدی گسترش می‌دهد. این مدل از یک VLM پایه مبتنی بر سه‌بعدی استفاده می‌کند که می‌تواند هم تصاویر دوبعدی آینده و هم ابر نقاط سه‌بعدی آینده را پیش‌بینی کند. تکنیک برنامه‌ریزی وظیفه در این مدل، استفاده ضمنی از این پیش‌بینی‌های سه‌بعدی برای هدایت تولید عمل است. مدل با پیش‌بینی چگونگی تغییر ساختار سه‌بعدی محیط در پاسخ به اَعمال، درک عمیق‌تری از پویایی‌ها و تعاملات فضایی به دست می‌آورد. اَعمال نیز به صورت تکواژهای عمل گسسته تولید می‌شوند [۶۲]. جنبه جهان باز این مدل در توانایی آن برای یادگیری و پیش‌بینی پویایی‌های ساختاری سه‌بعدی نهفته است. برخلاف مدل‌هایی که فقط تصاویر دوبعدی آینده را پیش‌بینی می‌کنند، 3D-VLA [۶۲] با پیش‌بینی ابر نقاط، درک

غنی‌تری از هندسه، روابط فضایی و انسداد[1] در محیط کسب می‌کند. این درک سه‌بعدی عمیق‌تر، پایه و اساس قوی‌تری برای تعمیم‌پذیری به اشیاء، چیدمان‌ها و تعاملات فیزیکی نادیده در دنیای واقعی فراهم می‌کند، جایی که پویایی‌ها ذاتاً سه‌بعدی هستند.

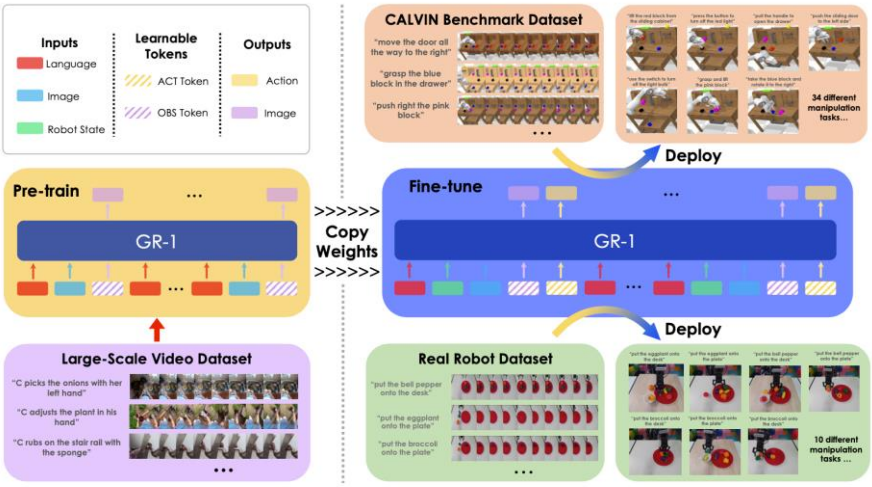


شکل (۳-۷) معماری مدل 3D-VLA.

همچنین کار آقای وو و همکارن با نام مدل GR-1 [۶۰] با هدف بهبود تعمیم‌پذیری ربات، بر اهمیت پیش‌آموزش مولد ویدئویی در مقیاس بزرگ تأکید می‌کند. این مدل از یک معماری ساده به سبک GPT استفاده می‌کند که به طور همزمان مشاهده تصویری و عمل بعدی را پیش‌بینی می‌نماید. VLM در این مدل به صورت ضمنی و از ابتدا ساخته می‌شود و از ستون فقرات VLMهای عظیم از پیش‌آموخته استفاده نمی‌کند. تکنیک برنامه‌ریزی وظیفه، یادگیری مشترک این دو هدف با استفاده از نشانه‌های ویژه [OBS] و [ACT] است. پیش‌بینی مشاهده بعدی، مدل را مجبور به یادگیری یک مدل جهان ضمنی می‌کند. اَعمال به صورت پیوسته با یک هد رگرسیون تولید می‌شوند [۶۰]. جنبه جهان باز و نوآوری اصلی GR-1 [۶۰] در استراتژی آموزش دو مرحله‌ای آن نهفته است:

مدل ابتدا منحصراً بر روی مجموعه داده عظیم و بسیار متنوع غیر-رباتیک Ego4D [۶۱]، فقط برای وظیفه «پیش‌بینی ویدئو مشروط به زبان» پیش‌آموزش داده می‌شود. این پیش‌آموزش گسترده بر روی ویدئوهای انسانی «در طبیعت» و بدون برچسب عمل، به مدل اجازه می‌دهد تا پویایی‌های فیزیکی جهان را به صورت عمومی و مستقل از ربات بیاموزد [۶۰،۶۱]. سپس، این مدلِ جهانِ ضمنیِ از پیش‌آموخته، بر روی داده‌های رباتیک برای یادگیری مشترک عمل و پیش‌بینی ویدئو تنظیم دقیق می‌شود. این دانش پیشین از پویایی‌های جهان، منجر به تعمیم‌پذیری صفر-شات فوق‌العاده به سناریوهای رباتیک نادیده و کارایی داده بسیار بالا می‌شود، زیرا مدل از قبل «قوانین فیزیک» را آموخته است [۶۰].

[1] occlusion



شکل (۳-۸) نمای کلی GR-1.

فصل ۴:

نتیجه‌گیری و کارهای آینده

۴-۱- نتیجه‌گیری

این پژوهش، به بررسی چالش اساسی گذار از «درک منفعلانه» مدل‌های زبانی-بینایی (VLM) به «عمل تجسم‌یافته» و هدفمند در رباتیک پرداخت .مسئله اصلی، زمینه‌سازی مفاهیم انتزاعی زبان در تعاملات فیزیکی دقیق و ممکن در دنیای واقعی است .در حالی که VLM ها، دانش عقل سلیم و قابلیت‌های استدلال معنایی بی‌سابقه‌ای را ارائه می‌دهند ، تبدیل این دانش مفهومی به برنامه‌های وظایف چندمرحله‌ای و قابل اجرا، یک چالش تحقیقاتی باز و اساسی باقی مانده است .هدف اصلی این مطالعه، ارائه یک تحلیل جامع، ساختاریافته و انتقادی از معماری‌های محاسباتی بود که برای پر کردن این شکاف، یعنی گذار از VLM به مدل‌های بینایی-زبانی-عمل (VLA)، طراحی شده‌اند.

در این راستا، هسته اصلی این پژوهش به کالبدشکافی و دسته‌بندی معماری‌های پیشرفته VLA اختصاص یافت. رویکردهای موجود به چند دسته اصلی تقسیم‌بندی شدند: مدل‌های یکپارچه، معماری‌های سلسله‌مراتبی مبتنی بر تجزیه وظیفه ، مدل‌های مبتنی بر قابلیت‌دهی و مدل‌های جهان . با تحلیل انتقادی چالش‌های کلیدی نظیر تعمیم‌پذیری، کارایی داده، استدلال چندمرحله‌ای و اجرای بلادرنگ در هر دسته ، این گزارش چارچوبی جامع برای درک چشم‌انداز فعلی پژوهش فراهم آورد .این تحلیل، ضمن روشن ساختن مسائل باز، مسیرهای آتی برای توسعه عامل‌های رباتیک همه‌منظوره را ترسیم می‌کند و پایه‌ای برای تعریف موضوع مطالعه در این حوزه محسوب می‌شود.

۴-۲- مسایل باز و کارهای قابل انجام

۴-۲-۱- موضوع اول: توسعه معماری‌های هیبریدی و آینده‌نگری مبتنی بر مدل جهان

یکی از شکاف‌های پژوهشی اساسی در وضعیت فعلی، وجود یک مصالحه[1] بین معماری‌های یکپارچه و سلسله‌مراتبی است. مدل‌های یکپارچه در تعمیم‌پذیری به وظایف جدید برتری دارند اما به داده‌های عظیم نیاز داشته و دقت هندسی پایینی دارند. در مقابل، مدل‌های سلسله‌مراتبی دقت بالاتری را ممکن می‌سازند اما تعمیم‌پذیری آن‌ها محدود است. یک مسیر تحقیقاتی باز و مهم، طراحی معماری‌های هیبریدی است که بتوانند به صورت پویا نقاط قوت این دو پارادایم را ترکیب کنند. چنین سیستمی می‌تواند بر اساس ماهیت وظیفه، به طور هوشمند بین یک سیاست یکپارچه برای وظایف با پیچیدگی معنایی بالا و یک برنامه‌ریز چند

[1] trade-off

مرحله‌ای برای وظایف نیازمند دقت هندسی تغییر کند.

علاوه بر این، آینده‌نگری VLM های کنونی عمدتاً به امکان‌سنجی هندسی فوری محدود است و فاقد استدلال علمی و زمانی عمیق است. تحقیقات آینده می‌تواند بر ادغامVLM ها با مدل‌های جهان تمرکز کند . در این الگو، VLMیک برنامه سطح بالا تولید می‌کند و مدل جهان می‌تواند "تصور کند" که حالت‌های آینده ناشی از این برنامه چگونه خواهند بود. این به VLM اجازه می‌دهد نه تنها امکان‌سنجی هر مرحله، بلکه مطلوبیت کل مسیر حاصل را ارزیابی کرده و از خطاهایی که ناشی از پیامدهای بلندمدت یک عمل هستند، جلوگیری کند.

۴-۲-۲- موضوع دوم: اتصال آگاه از فیزیک و مبتنی بر تعامل

چالش بنیادین ترجمه مفاهیم انتزاعی به اقدامات فیزیکی، عمدتاً بر امکان‌سنجی هندسی تمرکز دارد و یک نقطه کور مهم دارد:

ویژگی‌های دینامیکی و فیزیکی اشیاء. دستوراتی مانند "لیوان شیشه‌ای شکننده را به آرامی بگذار" نیازمند درک شهودی از مفاهیمی مانند شکنندگی، جرم و اصطکاک است کهVLM های فعلی فاقد آن هستند .این موضوع یک شکاف پژوهشی حیاتی را آشکار می‌کند توسعه مدل‌های اتصال با آگاهی فیزیکی است. تحقیقات در این حوزه باید فراتر از تولید صرف موقعیت‌های هدف حرکت کرده و مدل‌هایی را توسعه دهند که قادر باشند "پارامترهای عمل" مانند نیروی هدف، سرعت یا پروفایل‌های امپدانس را بر اساس ویژگی‌های فیزیکی استنتاج‌شده از شیء، تنظیم کنند.

همچنین، فرآیند اتصال اغلب به عنوان یک ترجمه ایستا و یک‌باره در نظر گرفته می‌شود، در حالی که در دنیای واقعی، یک فرآیند تعاملی و پویا است. دستورات انسانی اغلب مبهم هستند ("صندلی را بیاور" در اتاقی با چند صندلی (این موضوع شکاف دیگری را نمایان می‌سازد: حرکت از "اتصال ترجمه‌ای" به "اتصال گفتگومحور و تعاملی". تحقیقات آینده می‌تواند بر روی چارچوب‌هایی تمرکز کند که در آنها ربات می‌تواند به طور فعال برای حل ابهامات مربوط به اتصال، اطلاعات جستجو کرده و سؤالات شفاف‌کننده بپرسد.

۴-۲-۳- موضوع سوم: مدل‌های بنیادی ذاتاً تجسم‌یافته[۱] و معماری‌های عصبی-نمادین[۲]

محدودیت‌های VLM ها، در استدلال سه‌بعدی، درک فیزیک شهودی و شکنندگی‌های زبانی، از یک علت ریشه‌ای عمیق نشأت می‌گیرد:
عدم تطابق داده‌های آموزشی آن‌ها (تصاویر دوبعدی و ایستا از اینترنت) با الزامات هوش تجسم‌یافته. این تحلیل یک شکاف پژوهشی حیاتی را مشخص می‌کند: نیاز به حرکت فراتر از پیش-آموزش در مقیاس اینترنت و توسعه مدل‌های بنیادی ذاتاً تجسم‌یافته .این امر مستلزم ایجاد مجموعه داده‌های عظیم از ربات‌هایی است که با جهان تعامل می‌کنند و طراحی اهداف پیش-آموزش خود-نظارتی جدید بر اساس این داده‌های تعاملی (مانند پیش‌بینی حالت‌های آینده از روی اقدامات) است.

علاوه بر این، انتظار از یک VLM یکپارچه برای تسلط همزمان بر ظرافت‌های زبانی، هندسه سه‌بعدی و فیزیک، ممکن است غیرواقعی باشد. یک مسیر تحقیقاتی امیدوارکننده، توسعه معماری‌های VLM عصبی-نمادین برای رباتیک است. در این پارادایم، VLM مسئولیت درک معنایی و برنامه‌ریزی سطح بالا را بر عهده دارد، اما می‌تواند برای وظایف خاص، متخصص را فراخوانی کند.

۴-۳- معرفی موضوع مورد نظر برای پایان‌نامه

۴-۳-۱- مقدمه و بیان مسئله

در دهه‌های اخیر، رباتیک پیشرفت‌های چشمگیری در محیط‌های صنعتی کنترل‌شده داشته است. با این حال، تعامل مؤثر ربات‌ها با محیط‌های پیچیده، پویا و نامعین انسانی (مانند خانه یا دفتر کار) همچنان یک چالش اساسی محسوب می‌شود. این چالش، نیازمند گذار از ربات‌های با برنامه‌ریزی ثابت به عامل‌های هوشمند[۳] است که بتوانند دستورات را از طریق کانال‌های متنوع انسانی درک کنند، محیط خود را به صورت بصری تفسیر کرده و کنش‌های فیزیکی دقیق و آگاهانه انجام دهند. این پژوهش در صدد طراحی یک معماری نوین برای یک عامل رباتیک است که بتواند شکاف میان درک چندوجهی[۴]، استدلال سطح بالا و

---

[۱] embodiment-native

[۲] neuro-symbolic

[۳] Intelligent Agents

[۴] Multi-modal Understanding

کنش فیزیکی را پر کند.

۴-۳-۲- اهداف و رویکرد پیشنهادی

هدف اصلی این پروژه، توسعه یک مدل بینایی-زبان-کنش است که بر روی یک بازوی رباتیک مجهز به چنگک[۱] یا دست رباتیک پنج‌انگشتی پیاده‌سازی می‌شود .رویکرد پیشنهادی بر سه ستون اصلی استوار است:

الف) درک چندوجهی دستور و صحنه: سیستم برای حداکثر انعطاف‌پذیری در تعامل انسان و ربات، از چندین مدالیته ورودی پشتیبانی می‌کند:

- فرمان‌های ورودی:
  - صوتی: فرامین صوتی کاربر (مانند «تخم‌مرغ روی میز را به من بده») توسط یک ماژول تشخیص گفتار به متن تبدیل شده و برای پردازش به LLM ارسال می‌شود.
  - متنی: کاربر می‌تواند مستقیماً از طریق یک رابط متنی (مانند ارسال پیام) دستور خود را وارد کند.
  - اشاره بصری[۲]: ربات می‌تواند دستورات غیرکلامی را نیز درک کند. کاربر می‌تواند با اشاره به یک جسم خاص، به VLM فرمان دهد که شیء مورد نظر را شناسایی کند.

تجزیه و تحلیل و اتصال[۳] :

مدل زبانی بزرگ، به عنوان هسته مرکزی دانش و تصمیم‌گیری عمل می‌کند. وظیفه آن تجزیه و تحلیل معنایی دستور، صرف نظر از مدالیته ورودی آن، است (تشخیص «تخم‌مرغ» به عنوان شیء و «روی میز» به عنوان مکان.

مدل بینایی-زبانی (VLM) به عنوان چشم سیستم، وظیفه «اتصال بصری[۴]» این مفاهیم انتزاعی به دنیای واقعی را دارد. VLM، با اسکن محیط، وجود و موقعیت دقیق شیء مورد نظر را تأیید می‌کند.

ب) استدلال و برنامه‌ریزی کنش با درک فیزیکی :چالش اصلی در برداشتن اشیاء، صرفاً رسیدن به آن‌ها نیست، بلکه اعمال نیروی مناسب و انتخاب شیوه گرفتن[۵] صحیح است.

---

[۱] Gripper

[۲] Visual Pointing / Deixis

[۳] Decomposition and Grounding

[۴] Visual Grounding

[۵] Grasp Policy

استخراج دانش پیشین فیزیکی[١]: در این تحقیق، LLM صرفاً یک مترجم دستور نیست، بلکه به عنوان یک پایگاه دانش فیزیکی پیشین[٢] عمل می‌کند. مدل، اطلاعاتی پیش‌فرض در مورد ویژگی‌های فیزیکی اشیاء (مانند شکنندگی تخم‌مرغ یا وزن احتمالی یک بطری آب) را از LLM استخراج می‌کند.

حلقه بازخورد نیرو در لحظه[٣]: این دانش پیشین، با داده‌های آنی دریافتی از سنسورهای بازخورد نیرو (مانند پتانسیومترهای تعبیه‌شده در مفاصل چنگک (تلفیق می‌شود. این تلفیق داده‌ها (دانش انتزاعی LLM +داده‌های سنسور فیزیکی) به ربات اجازه می‌دهد تا نیروی بهینه را برای برداشتن شیء محاسبه و اعمال کند؛ به طوری که از شکستن اشیاء شکننده یا لغزیدن اشیاء سنگین جلوگیری شود.

ج) یادگیری سیاست‌های کنش[٤]: برای آنکه ربات بیاموزد هر جسم را چگونه بردارد (مثلاً گرفتن یک فنجان از دسته آن در مقابل گرفتن یک سیب)، از یک رویکرد یادگیری ترکیبی استفاده خواهد شد:

یادگیری از طریق مشاهده[٥]: مدل با استفاده از مجموعه داده‌های ویدیویی اینترنتی، به‌ویژه ویدیوهای خودمحور[٦]، آموزش می‌بینند. در این ویدیوها، انسان‌ها اشیاء مختلف با اَشکال و وزن‌های گوناگون را برمی‌دارند و مدل VLA می‌آموزد که الگوهای بهینه گرفتن هر نوع شیء را استخراج کند.

یادگیری تقلیدی از طریق تله‌رباتیک[٧]: در این روش، از یک اپراتور انسانی برای کنترل مستقیم ربات در «حالت سایه[٨]» استفاده می‌شود. اپراتور با استفاده از یک رابط مبتنی بر سیگنال‌های الکترومیوگرافی(EMG) و با بهره‌گیری از یک مدل پیش‌آموخته (مانند معماری `bilstm-transformer` توسعه‌یافته در پژوهش‌های پیشین محقق) حرکات دست خود را به چنگک ربات منتقل می‌کند. ربات با مشاهده و تقلید[٩] این حرکات متخصص‌گونه، فرآیند یادگیری برداشتن اشیاء پیچیده را تسریع می‌بخشد.

---

[١] Physics Prior Extraction

[٢] Physics Prior Knowledge Base

[٣] Real-time Force-Feedback Loop

[٤] Action Policies

[٥] Learning from Observation

[٦] Egocentric Videos

[٧] Imitation Learning via Teleoperation

[٨] Shadow Mode

[٩] Imitation Learning

۴-۳-۳- نتایج مورد انتظار[۱]

خروجی نهایی این پژوهش، یک مدل VLA بهینه‌سازی شده خواهد بود که صرفاً محدود به اشیاء دیده‌شده در طول آموزش نیست. نوآوری کلیدی این پروژه، تزریق دانش فیزیکی استنتاج‌شده از LLM به حلقه کنترل ربات[۲] و درک فرمان از طریق کانال‌های چندوجهی[۳] است.

انتظار می‌رود این مدل با قابلیت تعمیم‌پذیری[٤] بالا، بتواند در مواجهه با اشیاء جدید که پیش از این ندیده است، با تخمین هوشمندانه ویژگی‌های فیزیکی (مانند توزیع وزن و شکنندگی) از طریق دانش LLM و ادراکVLM ، کنش بهینه و ایمن را برای برداشتن آن‌ها تولید کند. این تحقیق گامی مهم در جهت توسعه ربات‌های همکار با قابلیت درک عمیق از زبان، بینایی و فیزیک محیط خواهد بود.

---

[۱] Contribution

[۲] LLM-in-the-Loop Control

[۳] Multi-modal HRI

[٤] Generalization

مراجع

[1] M. Ahn et al., “Do As I Can, Not As I Say: Grounding Language in Robotic Affordances,” Aug. 16, 2022, arXiv: arXiv:2204.01691. doi: 10.48550/arXiv:2204.01691.
[2] Y. Du, Z. Liu, J. Li, and W. X. Zhao, “A Survey of Vision-Language Pre-Trained Models,” July 16, 2022, arXiv: arXiv:2202.10936. doi: 10.48550/arXiv:2202.10936.
[3] D. Driess et al., “PaLM-E: An Embodied Multimodal Language Model,” Mar. 06, 2023, arXiv: arXiv:2303.03378. doi: 10.48550/arXiv:2303.03378.
[4] A. Brohan et al., “RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control”.
[5] M. Shridhar, L. Manuelli, and D. Fox, “CLIPort: What and Where Pathways for Robotic Manipulation,” Sept. 24, 2021, arXiv: arXiv:2109.12098. doi: 10.48550/arXiv:2109.12098.
[6] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Welinder, D. P. K. M. Sutskever, and I. Sutskever, “Learning Transferable Visual Models From Natural Language Supervision,” in *Proceedings of the 38th International Conference on Machine Learning*, Jul. 2021, pp. 8748–8763.
[7] X. Zhang et al., “Grounding Classical Task Planners via Vision-Language Models,” June 19, 2023, arXiv: arXiv:2304.08587. doi: 10.48550/arXiv:2304.08587.
[8] J. Gao et al., “Physically Grounded Vision-Language Models for Robotic Manipulation,” Mar. 03, 2024, arXiv: arXiv:2309.02561. doi: 10.48550/arXiv:2309.02561.
[9] M. Aghzal, E. Plaku, G. J. Stein, and Z. Yao, “A Survey on Large Language Models for Automated Planning,” Feb. 18, 2025, arXiv: arXiv:2502.12435. doi: 10.48550/arXiv:2502.12435.
[10] P. Cao et al., “Large Language Models for Planning: A Comprehensive and Systematic Survey,” May 26, 2025, arXiv: arXiv:2505.19683. doi: 10.48550/arXiv:2505.19683.
[11] J. Liang et al., “Code as Policies: Language Model Programs for Embodied Control,” May 25, 2023, arXiv: arXiv:2209.07753. doi: 10.48550/arXiv:2209.07753.
[12] W. Huang et al., “Inner Monologue: Embodied Reasoning through Planning with Language Models”.
[13] I. Singh et al., “ProgPrompt: Generating Situated Robot Task Plans using Large Language Models,” Sept. 22, 2022, arXiv: arXiv:2209.11302. doi: 10.48550/arXiv:2209.11302.
[14] N. Wake, A. Kanehira, K. Sasabuchi, J. Takamatsu, and K. Ikeuchi, “GPT-4V(ision) for Robotics: Multimodal Task Planning From Human Demonstration,” IEEE Robot. Autom. Lett., vol. 9, no. 11, pp. 10567–10574, Nov. 2024, doi: 10.1109/LRA.2024.3477090.
[15] Y. Ma, Z. Song, Y. Zhuang, J. Hao, and I. King, “A Survey on Vision-Language-Action Models for Embodied AI,” Aug. 31, 2025, arXiv: arXiv:2405.14093. doi: 10.48550/arXiv:2405.14093.
[16] R. Shao et al., “Large VLM-based Vision-Language-Action Models for Robotic Manipulation: A Survey,” Sept. 01, 2025, arXiv: arXiv:2508.13073. doi: 10.48550/arXiv:2508.13073.
[17] K. Kawaharazuka, J. Oh, J. Yamada, I. Posner, and Y. Zhu, “Vision-Language-Action Models for Robotics: A Review Towards Real-World Applications,” IEEE Access, vol. 13, pp. 162467–162504, 2025, doi: 10.1109/ACCESS.2025.3609980.

[18] Y. Jiang et al., “VIMA: General Robot Manipulation with a VIsual-Motor Makeover,” Oct. 06, 2022, arXiv: arXiv:2210.03094. doi: 10.48550/arXiv:2210.03094.
[19] J. Cen et al., “WorldVLA: Towards Autoregressive Action World Model,” June 26, 2025, arXiv: arXiv:2506.21539. doi: 10.48550/arXiv:2506.21539.
[20] J. Liang et al., “Code as Policies: Language Model Programs for Embodied Control,” May 25, 2023, arXiv: arXiv:2209.07753. doi: 10.48550/arXiv:2209.07753.
[21] S. Zhang et al., “LoHoRavens: A Long-Horizon Language-Conditioned Benchmark for Robotic Tabletop Manipulation,” Oct. 23, 2023, arXiv: arXiv:2310.12020. doi: 10.48550/arXiv:2310.12020.
[22] Y. Fan et al., “Long-VLA: Unleashing Long-Horizon Capability of Vision Language Action Model for Robot Manipulation”.
[23] P. Intelligence et al., “$π_{0.5}$: a Vision-Language-Action Model with Open-World Generalization,” Apr. 22, 2025, arXiv: arXiv:2504.16054. doi: 10.48550/arXiv:2504.16054.
[24] O. M. Team et al., “Octo: An Open-Source Generalist Robot Policy,” May 26, 2024, arXiv: arXiv:2405.12213. doi: 10.48550/arXiv:2405.12213.
[25] L. Yuan et al., “Florence: A New Foundation Model for Computer Vision,” 2021, arXiv: arXiv:2111.11432.
[26] A. Vaswani et al., “Attention Is All You Need,” Aug. 02, 2023, arXiv: arXiv:1706.03762. doi: 10.48550/arXiv:1706.03762.
[27] Y. Kim, D. Kim, J. Choi, J. Park, N. Oh, and D. Park, “A Survey on Integration of Large Language Models with Intelligent Robots,” Intel Serv Robotics, vol. 17, no. 5, pp. 1091–1107, Sept. 2024, doi: 10.1007/s11370-024-00550-5.
[28] Y. Zhai et al., “Fine-Tuning Large Vision-Language Models as Decision-Making Agents via Reinforcement Learning”.
[29] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” 2018, arXiv: arXiv:1810.04805.
[30] T. B. Brown et al., “Language Models are Few-Shot Learners,” 2020, arXiv: arXiv:2005.14165.
[31] C. Raffel et al., “Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer,” 2020, arXiv: arXiv:1910.10683.
[32] M. Lewis et al., “BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension,” 2019, arXiv: arXiv:1910.13461.
[33] C. Jia et al., “Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision,” in *Proceedings of the 38th International Conference on Machine Learning*, 2021, pp. 4776–4787.
[34] M. J. Kim et al., “OpenVLA: An Open-Source Vision-Language-Action Model,” Sept. 05, 2024, arXiv: arXiv:2406.09246. doi: 10.48550/arXiv:2406.09246.
[35] J. Wei et al., “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models,” 2022, arXiv: arXiv:2201.11903.
[36] M. N. Azadani, J. Riddell, S. Sedwards, and K. Czarnecki, “LEO: Boosting Mixture of Vision Encoders for Multimodal Large Language Models,” 2025, arXiv: arXiv:2501.06986.
[37] Y. Du et al., “Learning Universal Policies via Text-Guided Video Generation,” Nov. 20, 2023, arXiv: arXiv:2302.00111. doi: 10.48550/arXiv:2302.00111.
[38] M. Zawalski, W. Chen, K. Pertsch, O. Mees, C. Finn, and S. Levine, “Robotic Control via Embodied Chain-of-Thought Reasoning,” Mar. 06, 2025, arXiv:

arXiv:2407.08693. doi: 10.48550/arXiv:2407.08693.
[39] S. Belkhale et al., “RT-H: Action Hierarchies Using Language,” June 01, 2024, arXiv: arXiv:2403.01823. doi: 10.48550/arXiv:2403.01823.
[40] L. Yang et al., “RoboEnvision: A Long-Horizon Video Generation Model for Multi-Task Robot Manipulation,” June 27, 2025, arXiv: arXiv:2506.22007. doi: 10.48550/arXiv:2506.22007.
[41] H. Wu et al., “Unleashing Large-Scale Video Generative Pre-training for Visual Robot Manipulation,” Dec. 21, 2023, arXiv: arXiv:2312.13139. doi: 10.48550/arXiv:2312.13139.
[42] K. He, G. Gkioxari, P. Dollár, and R. Girshick, “Mask R-CNN,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.
[43] C. Chi, S. Feng, Y. Du, Z. Yang, T. M. L. T. P. Kaelbling, and D. D. Lee, “Diffusion Policy: Visuomotor Policy Learning via Action Diffusion,” 2023, arXiv: arXiv:2303.04137.
[44] M. Reuss, Ö. E. Yağmurlu, F. Wenzel, and R. Lioutikov, “Multimodal Diffusion Transformer: Learning Versatile Behavior from Multimodal Goals,” July 08, 2024, arXiv: arXiv:2407.05996. doi: 10.48550/arXiv:2407.05996.
[45] W. Peebles and S. Xie, “Scalable Diffusion Models with Transformers,” 2022, arXiv: arXiv:2212.09748.
[46] S. Liu et al., “RDT-1B: a Diffusion Foundation Model for Bimanual Manipulation,” Mar. 01, 2025, arXiv: arXiv:2410.07864. doi: 10.48550/arXiv:2410.07864.
[47] O. Mees et al., “CALVIN: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks,” *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 7327–7334, 2022.
[48] W. Yu et al., “LIBERO: A Benchmark for Lifelong Robotic Manipulation,” 2023, arXiv: arXiv:2310.08867.
[49] S. Sar-Shgi, A. Fuchs, and A. H. T. K. M. R. L. D. F. P. K. H. W. P. S. A. G. G. H. R. J. Al-Halah, “Prismatic VLMs: Investigating the Design Space of Visually-Conditioned Language Models,” 2024, arXiv: arXiv:2404.03264.
[50] H. Touvron et al., “Llama 2: Open foundation and fine-tuned chat models,” 2023, arXiv: arXiv:2307.09288.
[51] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, and V. K. A. E. G. F. S. P. B. P. J. M. A. L. S. G. D. P. d. L. C. T. L. A. J. L. P. B., “DINOv2: Learning robust visual features without supervision,” 2023, arXiv: arXiv:2304.07193.
[52] J. Huang et al., “An Embodied Generalist Agent in 3D World,” May 09, 2024, arXiv: arXiv:2311.12871. doi: 10.48550/arXiv:2311.12871.
[53] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, “PointNet: Deep learning on point sets for 3D classification and segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 652–660.
[54] W.-L. Chiang et al., “Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality,” *See: https://vicuna.lmsys.org*, 2023.
[55] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, “LoRA: Low-Rank Adaptation of Large Language Models,” 2021, arXiv: arXiv:2106.09685.
[56] Gemini Team et al., “Gemini: A Family of Multimodal Models,” 2023, arXiv: arXiv:2312.11805.
[57] S. Liu et al., “Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection,” 2023, arXiv: arXiv:2303.05499.

[58] A. Kirillov et al., “Segment Anything,” 2023, arXiv: arXiv:2304.02643.
[59] M. Minderer et al., “Simple Open-Vocabulary Object Detection with Vision Transformers,” 2022, arXiv: arXiv:2205.06230.
[60] H. Wu et al., “Unleashing Large-Scale Video Generative Pre-training for Visual Robot Manipulation,” Dec. 21, 2023, arXiv: arXiv:2312.13139. doi: 10.48550/arXiv:2312.13139.
[61] K. Grauman et al., “Ego4D: Around the World in 3,000 Hours of Egocentric Video,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 18995–19012.
[62] X. Li et al., “3DS-VLA: A 3D Spatial-Aware Vision Language Action Model for Robust Multi-Task Manipulation”.
[63] J. Cen et al., “WorldVLA: Towards Autoregressive Action World Model,” June 26, 2025, arXiv: arXiv:2506.21539. doi: 10.48550/arXiv:2506.21539.
[64] R. Ke et al., “Grounded SAM: Assembling Open-World Models for Segmenting Anything,” 2024, arXiv: arXiv:2403.09014.
[65] J. Kirkpatrick et al., “Overcoming catastrophic forgetting in neural networks,” *Proceedings of the national academy of sciences*, vol. 114, no. 13, pp. 3521–3526, 2017.
[66] Y. Zhai et al., “ChatVLA: Unified Multimodal Understanding and Robot Control with Mixture of Experts,” 2024, arXiv: arXiv:2407.02027.
[67] J. Bai et al., “Qwen-VL: A Frontier Large Vision-Language Model with Versatile Abilities,” 2023, arXiv: arXiv:2308.12966.
[68] J. Wen et al., “Diffusion-VLA: Generalizable and Interpretable Robot Foundation Model via Self-Generated Reasoning,” June 04, 2025, arXiv: arXiv:2412.03293. doi: 10.48550/arXiv:2412.03293.
[69] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, W. H. R. T. Q. Chen, and M. R. G. D. D. D. K. D. R. B. A. M. H. B.-H. Y. L. W. H. R. T. Q. C., “Flow Matching for Generative Modeling,” 2022, arXiv: arXiv:2210.02747.
[70] K. Black et al., “$π_0$: A Vision-Language-Action Flow Model for General Robot Control,” Nov. 13, 2024, arXiv: arXiv:2410.24164. doi: 10.48550/arXiv:2410.24164.
[71] Google Core Team, “PaliGemma: A versatile 3B VLM for transfer,” 2024.
[72] J. Zhai et al., “PaLI-3: A Vision-Language Model for Many-to-Many Multimodal Tasks,” 2023.
[73] S. Zhai et al., “SigLIP: Vision-Language Pre-training by Aligning Features with Sigmoid Loss,” 2023, arXiv: arXiv:2303.15343.
[74] A. Padalkar et al., “Open X-Embodiment: Robotic Learning Datasets and RT-X Models,” 2023, arXiv: arXiv:2310.08864.
[75] S. Deng et al., “GraspVLA: a Grasping Foundation Model Pre-trained on Billion-scale Synthetic Action Data,” 2024, arXiv: arXiv:2407.12781.
[76] A. D. Vuong, M. N. Vu, H. Le, B. Huang, B. Huynh, T. Vo, A. Kugi, and A. Nguyen, “Grasp-anything: Large-scale grasp dataset from foundation models,” in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 8838–8845.
[77] InternLM-XComposer Team, “InternLM-XComposer2: A Vision-Language Large Model for Advanced Nature Language and Grounding Capabilities,” 2024, arXiv: arXiv:2404.09413.
[78] NVIDIA et al., “GR00T N1: An Open Foundation Model for Generalist Humanoid Robots,” Mar. 27, 2025, arXiv: arXiv:2503.14734. doi: 10.48550/arXiv:2503.14734.
[79] Q. Li et al., “CogACT: A Foundational Vision-Language-Action Model for Synergizing Cognition and Action in Robotic Manipulation,” Nov. 29, 2024, arXiv:

arXiv:2411.19650. doi:10.48550/arXiv:2411.19650.
[80] S. Ye et al., “Latent Action Pretraining from Videos,” May 15, 2025, arXiv: arXiv:2410.11758. doi: 10.48550/arXiv:2410.11758.
[81] J. Wei et al., “Scaling Instruction-Finetuned Language Models,” 2022, arXiv: arXiv:2210.11416.
[82] W. Yuan et al., “RoboPoint: A Vision-Language Model for Spatial Affordance Prediction for Robotics,” June 15, 2024, arXiv: arXiv:2406.10721. doi: 10.48550/arXiv:2406.10721.
[83] Y. Li et al., “HAMSTER: Hierarchical Action Models For Open-World Robot Manipulation,” May 10, 2025, arXiv: arXiv:2502.05485. doi: 10.48550/arXiv:2502.05485.
[84] J. Lin, Y. Li, Z. Liu, L. Wang, Z. Gan, L. Wang, and Z. Liu, “VILA: On the Scaling of Vision Language Models,” 2023, arXiv: arXiv:2312.07533.
[85] P. Esser, R. Rombach, and B. Ommer, “Taming transformers for high-resolution image synthesis,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 12873–12883.

واژه نامه

بخش الف: واژه نامه فارسی به انگلیسی

اتصال بصری Visual Grounding
اجرای حلقه-بسته Closed-loop
اجرای ناهمزمان Asynchronous Execution
اجرای وظیفه تجسم‌یافته Embodied Task Planning and Execution
ادراک-به-عمل Perception-to-action
استخراج اطلاعات Data extraction
استخراج دانش پیشین فیزیکی Physics Prior Extraction
استدلال خود-تولید Self-Generated Reasoning
استنتاج سیستمی Systemic Inference
اشاره بصری Visual Pointing / Deixis
اصلاح تعاملی Interactive Correction
اطلاعات حالت Proprioception
اعتبارسنجی evaluation
افزودن Augmentation
امتیاز اهمیت Attention Score
امکان‌پذیری یا احتمال موفقیت probability of success
انتزاعی Abstract
انتقال دانش knowledge transfer
انسان‌نما Humanoid
انسداد occlusion
آشکارساز اشیاء object detector
آینده‌نگری مولد ماسک‌دار masked generative foresight
بازگشتی Recurrence
بازنمایی representation
بازنمایی عددی غنی Contextual Representation
بازنمایی نهفته Latent Representation
بازنمایی‌های میانی ساختاریافته‌ای structured intermediate representations
بخش‌بندی segmentation
بدون-داده Zero-shot
بردار متنی text-based formulation

برداشتن-و-گذاشتن ..... pick-and-place
برنامه‌ریز حرکت مبتنی بر بهینه‌سازی ..... optimization-based motion planner
برنامه‌ریزی معنایی ..... Semantic Planning
برنامه‌ریزی وظایف ..... task planning
بلادرنگ ..... Real-time
بیانگری ادراکی محدود ..... limited perceptual expressiveness
بینایی کامپیوتر ..... computer vision
پایگاه دانش فیزیکی پیشین ..... Physics Prior Knowledge Base
پخش‌کردن گوسی ..... Gaussian Splatting
پردازش زبان طبیعی ..... natural language processing
پردازش زبان‌های طبیعی ..... natural language processing
پردازنده‌های گرافیکی همگانی ..... consumer GPUs
پیش‌بینی تکواژ ..... Token Prediction
پیش‌پردازش ..... preprocessing
تجزیه ..... decomposition
تجزیه مبتنی بر وظیفه فرعی ..... Subtask-based Decomposition
تجزیه و تحلیل و اتصال ..... Decomposition and Grounding
تجزیه‌شده ..... decoupled
تجسم‌یافته ..... embodied
تحلیل داده‌ها ..... data analysis
ترتیبی ..... Sequential
ترجمه ماشینی ..... machine translation
ترکیب زودهنگام ..... early fusion
ترکیبی از متخصصین ..... Mixture of Experts
ترکیبی ناهمگن ..... heterogeneous mixture
تزریق مُدالیته ..... modality injection
تشخیص اشیاء ..... object detection
تشخیص الگو ..... pattern recognition
تطبیق‌پذیری ..... Adaptability
تعمیم‌پذیری ..... Generalizability
تعمیم‌پذیری ..... Generalization
تغییر دستورات متنی ..... prompt switching

تقویت متقابل ..... Mutual Enhancemen
تکنیک تزریق شرطی متناوب ..... alternating conditional injection
تکواژ کردن عمل ..... Action Tokenization
تکواژهای بازخوانی ..... Readout Tokens
تکواژهای عمل گسسته ..... Discretized Action Tokens
تماماً کانولوشنی ..... fully convolutional
تنظیم ابرپارامترها ..... hyper-parameter tunning
تنظیم دقیق ..... fine tunning
توابع ارزش ..... value functions
توالی ..... sequence
توجه به خودی ..... self attention
توجه چند-سری ..... Multi-Head Attention
توجه متقابل ..... Cross-Attention
تولید عمل مشروط به استدلال ..... reasoning-conditioned action generation
چندتکه ..... modular
چندرسانه‌ای ..... multi-media
چندمرحله‌ای ..... multi stage
چندوجهی ..... Multi-modal Understanding
چنگک ..... Gripper
حالت سایه ..... Shadow Mode
حرکت زبانی ..... Language Motion
حسگر-حرکتی ..... Sensorimotor
حسگر-موتور ..... Sensorimotor Models
حلقه بازخورد نیرو در لحظه ..... Real-time Force-Feedback Loop
خارج-از-دامنه ..... off-domain
خط‌مشی ..... Policy
خلاصه‌سازی ..... summerization
خودبازگشتی ..... Autoregressively
خودمحوری ..... ego centeric
خودنظارتی ..... self-supervised
درک چندوجهی منفعل ..... Passive Multimodal Understanding
درک عمیق متنی ..... Contextual Understanding

درک منفعلانه ..... Passive Perception
دستکاری دو بازویی ..... Bimanual Manipulation
دستورات چندوجهی ..... Multimodal Prompts
دستورات متنی برنامه‌مانند ..... program-like prompt structure
دستورات متنی چند-نمونه ..... few-shot prompting
دستورالعمل ..... recipe
دنباله-به-دنباله Seq2Seq
دو-سیستمی ..... Dual-System
دو-مسیره ..... two-stream
دو-هدفه ..... Dual-Task
ذاتاً تجسم‌یافته ..... embodiment-native
راهنما ..... guidance
رباتیک ..... robotic
رمزگذار ..... encoder
رمزگذار-رمزگشا ..... Encoder-Decoder
رمزگذاری موقعیتی ..... Positional Encoding
رمزگشا ..... sentiment analysis
زاید ..... redundant
زبان طبیعی ..... natural language
زبان‌شناسی ..... Linguistics
زمینه‌سازی ..... Grounding
زنجیره‌ی تفکر ..... Chain-of-Thought
زنجیره‌ی تفکر تجسم‌یافته ..... Embodied Chain-of-Thought
زیان متقاطع ..... cross-entropy
زیروظیفه فعلی ..... SUBTASK
زیروظیفه‌ها ..... subtasks
ژست ..... pose
سَرِ عمل ..... action head
سَرِ عمل تطبیق جریان ..... Flow Matching Action Head
سَرِ کنش انتشار ..... Diffusion Action Head
ساختاریافته ..... structured
ستون فقرات ..... Backbone

سطل، بازه ..... bin
سلسله‌مراتبی ..... Hierarchical
سلسله‌مراتبی ..... Hierarchical Architectures
سیاست انتشار ..... diffusion policy
سیاست‌های کنش ..... Action Policies
شبکه عصبی چندلایه ..... multi-layer perceptron
شبکه نویززدا ..... Denoising Network
شبکه‌های عصبی ..... neural networks
شبکه‌های عصبی پیچشی یا کانولوشنی ..... convolutional neural network
شبیه‌سازی ناهنجاری‌ها ..... Anomaly Simulation
شکاف تجسم ..... Embodiment Gap
شناختی-عملی ..... cognition-action
شناختی-عملی ..... Nvidia
شناسایی چهره ..... face detection
شیء-محور ..... Object-centric
شیء-محور ..... O-CoT
شیوه گرفتن ..... Grasp Policy
صافی ..... smoothness
ضمنی ..... implicitly
طبقه‌بندی تصاویر ..... image classification
طولانی‌مدت ..... long-horizon
عامل فراگیر ..... Generalist Agent
عامل‌های رباتیک همه‌منظوره ..... General-purpose Robotic Agents
عامل‌های هوشمند ..... Intelligent Agents
عدم اتصال به واقعیت ..... un-groundedness
عصبی-نمادین ..... neuro-symbolic
عمل ..... Act
عمل ..... Action
عمل تجسم‌یافته و هدفمند ..... Embodied action
عملی ..... Feasible
غیرنظارتی ..... unsupervised
فراموشی فاجعه‌بار ..... Catastrophic Forgetting

فضای عمل یکپارچه با تفسیر فیزیکی ..... Physically Interpretable Unified Action Space
فقط مبتنی بر رمزگذار ..... Decoder-Only
فقط مبتنی بر رمزگشا ..... Encoder-Only
قابل تفسیر ..... Interpretable
قابلیت تفسیرپذیری و مداخله ..... Interpretability and Intervention
قابلیت‌دهی ..... Affordance-based
قابلیت‌های نوظهور ..... Emergent Capabilities
قطع ارتباط استدلال-عمل ..... reasoning-action disconnection
قطعه عمل ..... Action Chunking
کادر مرزی ..... Bounding Box
کانال‌های چندوجهی ..... Multi-modal HRI
کاوشگری خطی ..... linear exploration
کلاس‌بندی تکواژها ..... classification token
کنترل واکنشی ..... Reactive Control
گاؤسی ..... gussian
گرفتن ..... Grasping
گسسته‌سازی مستقیم فضای عمل ..... direct action discretization
لایه‌های کاملاً متصل ..... dense
ماسک‌گذاری توجه عمل ..... Action Attention Masking
ماشین‌های خودران ..... self driving car
مبدل سببی ..... Causal Transformer
مبدل فضایی ..... Spatial Transformer
مبدل مبتنی بر مدل انتشار ..... diffusion transformer
مبدل‌ها ..... transformers
مبدل‌های رباتیک ..... Robotic Transformer
متخصص انتشار قابل اتصال ..... plug-in diffusion expert
متن-باز ..... Open-Source
محیط ..... Disembodied
محیط‌های پویا ..... dynamic environments
مدل پویایی معکوس ..... Inverse Dynamics Model - IDM
مدل جهان-عمل خودبازگشتی ..... Autoregressive Action World Model
مدل جهان ضمنی ..... Implicit World Model

مدل مبتنی بر قابلیت‌دهی ..... AFFORDANCE-BASED MODEL
مدل‌های بینایی-زبانی-عمل ..... vision language action model
مدل‌های پایه ..... Foundation Models
مدل‌های جهان ..... World Models
مدل‌های خطی ..... regression
مدل‌های زبانی-بینایی ..... vision language models
مدل‌های زبانی بزرگ ..... large language models
مسطح ..... flat
مسیر انتقال ..... Transport pathway
مسیر توجه ..... Attention pathway
مصالحه ..... trade-off
مصنوعی-سپس-واقعی ..... Synthetic-then-Real
مکانیزم توجه ..... attention mechanism
منجمد ..... Freeze
موقعیت پیکسل دست ربات ..... GRIPPER POS
مولدین ترکیبیِ ..... composable generators
میدان‌های تابشی عصبی ..... Neural Radiance Fields
نظارتی ..... supervised
نقاط کلیدی ..... keypoints
نقشه حرارتی ..... heatmap
نقشه‌های ارزش وُکسلی ..... Voxel-based Value Maps
نمایش تک‌نمونه ..... One-shot Demonstration
نمایش میانی ..... intermediate representation
هدف کمکی خودنظارتی ..... Self-Supervised Auxiliary Objectives
هذیان‌گویی ..... hallucination
هم-تنظیم دقیق ..... Co-fine-tuning
هم‌ترازی موقعیتی دوبعدی-به-سه‌بعدی ..... 2D-to-3D Positional Alignment
همه‌منظوره ..... GRP
واژگان ..... vocabulary
واکنشی ..... Reactive
ویدیوهای خودمحور ..... Egocentric Videos
یادگیری از بازی ..... learn from playing

یادگیری از طریق مشاهده ..... Learning from Observation
یادگیری با نمونه کم ..... few-shot learning
یادگیری تقلیدی ..... Imitation Learning
یادگیری تقلیدی از طریق تله‌رباتیک ..... Imitation Learning via Teleoperation
یادگیری درون‌متنی ..... in-context learning
یکپارچه ..... monolith

بخش ب: واژه نامه انگلیسی به فارسی

2D-to-3D Positional Alignment ..... هم‌ترازی موقعیتی دوبعدی-به-سه‌بعدی
Abstract ..... انتزاعی
Act ..... عمل
Action ..... عمل
Action Attention Masking ..... ماسک‌گذاری توجه عمل
Action Chunking ..... قطعه عمل
action head ..... سَرِ عمل
Action Policies ..... سیاست‌های کنش
Action Tokenization ..... تکواژ کردن عمل
Adaptability ..... تطبیق‌پذیری
Affordance-based ..... قابلیت‌دهی
AFFORDANCE-BASED MODEL ..... مدل مبتنی بر قابلیت‌دهی
alternating conditional injection ..... تکنیک تزریق شرطی متناوب
Anomaly Simulation ..... شبیه‌سازی ناهنجاری‌ها
Asynchronous Execution ..... اجرای ناهمزمان
attention mechanism ..... مکانیزم توجه
Attention pathway ..... مسیر توجه
Attention Score ..... امتیاز اهمیت
Augmentation ..... افزودن
Autoregressive Action World Model ..... مدل جهان-عمل خودبازگشتی
Autoregressively ..... خودبازگشتی
Backbone ..... ستون فقرات
Bimanual Manipulation ..... دستکاری دو بازویی
bin ..... سطل، بازه
Bounding Box ..... کادر مرزی
Catastrophic Forgetting ..... فراموشی فاجعه‌بار
Causal Transformer ..... مبدل سببی
Chain-of-Thought ..... زنجیره‌ی تفکر
classification token ..... کلاس‌بندی تکواژها
closed-loop ..... اجرای حلقه-بسته

هم-تنظیم دقیق ..... Co-fine-tuning
شناختی-عملی ..... cognition-action
مولدین ترکیبیِ ..... composable generators
بینایی کامپیوتر ..... computer vision
پردازنده‌های گرافیکی همگانی ..... consumer GPUs
بازنمایی عددی غنی ..... Contextual Representation
درک عمیق متنی ..... Contextual Understanding
شبکه‌های عصبی پیچشی یا کانولوشنی ..... convolutional neural network
توجه متقابل ..... Cross-Attention
زیان متقاطع ..... cross-entropy
تحلیل داده‌ها ..... data analysis
استخراج اطلاعات ..... data extraction
فقط مبتنی بر رمزگذار ..... Decoder-Only
تجزیه ..... decomposition
تجزیه و تحلیل و اتصال ..... Decomposition and Grounding
تجزیه‌شده ..... decoupled
شبکه نویززدا ..... Denoising Network
لایه‌های کاملاً متصل ..... dense
سَرِ کنش انتشار ..... Diffusion Action Head
سیاست انتشار ..... diffusion policy
مبدل مبتنی بر مدل انتشار ..... diffusion transformer
گسسته‌سازی مستقیم فضای عمل ..... direct action discretization
تکواژهای عمل گسسته ..... Discretized Action Tokens
محیط ..... Disembodied
دو-سیستمی ..... Dual-System
دو-هدفه ..... Dual-Task
محیط‌های پویا ..... dynamic environments
ترکیب زودهنگام ..... early fusion
خودمحوری ..... ego centeric
ویدیوهای خودمحور ..... Egocentric Videos
تجسم‌یافته ..... embodied
عمل تجسم‌یافته و هدفمند ..... Embodied action

زنجیره‌ی تفکر تجسم‌یافته ..... Embodied Chain-of-Thought
اجرای وظیفه تجسم‌یافته ..... Embodied Task Planning and Execution
شکاف تجسم ..... Embodiment Gap
ذاتاً تجسم‌یافته ..... embodiment-native
قابلیت‌های نوظهور ..... Emergent Capabilities
رمزگذار ..... encoder
رمزگذار-رمزگشا ..... Encoder-Decoder
فقط مبتنی بر رمزگشا ..... Encoder-Only
اعتبارسنجی ..... evaluation
شناسایی چهره ..... face detection
عملی ..... Feasible
یادگیری با نمونه کم ..... few-shot learning
دستورات متنی چند-نمونه ..... few-shot prompting
تنظیم دقیق ..... fine tunning
مسطح ..... flat
سَرِ عمل تطبیق جریان ..... Flow Matching Action Head
مدل‌های پایه ..... Foundation Models
منجمد ..... Freeze
تماماً کانولوشنی ..... fully convolutional
پخش‌کردن گوسی ..... Gaussian Splatting
عامل فراگیر ..... Generalist Agent
تعمیم‌پذیری ..... Generalizability
تعمیم‌پذیری ..... Generalization
عامل‌های رباتیک همه‌منظوره ..... General-purpose Robotic Agents
شیوه گرفتن ..... Grasp Policy
گرفتن ..... Grasping
چنگک ..... Gripper
موقعیت پیکسل دست ربات ..... GRIPPER POS
زمینه‌سازی ..... Grounding
همه‌منظوره ..... GRP
راهنما ..... guidance
گاؤسی ..... gussian

هذیان‌گویی ..... hallucination
نقشه حرارتی ..... heatmap
ترکیبی ناهمگن ..... heterogeneous mixture
سلسله‌مراتبی ..... Hierarchical
سلسله‌مراتبی ..... Hierarchical Architectures
انسان‌نما ..... Humanoid
تنظیم ابرپارامترها ..... hyper-parameter tunning
طبقه‌بندی تصاویر ..... image classification
یادگیری تقلیدی ..... Imitation Learning
یادگیری تقلیدی از طریق تله‌رباتیک ..... Imitation Learning via Teleoperation
مدل جهان ضمنی ..... Implicit World Model
ضمنی ..... implicitly
یادگیری درون‌متنی ..... in-context learning
عامل‌های هوشمند ..... Intelligent Agents
اصلاح تعاملی ..... Interactive Correction
نمایش میانی ..... intermediate representation
قابلیت تفسیرپذیری و مداخله ..... Interpretability and Intervention
قابل تفسیر ..... Interpretable
مدل پویایی معکوس ..... Inverse Dynamics Model
نقاط کلیدی ..... keypoints
انتقال دانش ..... knowledge transfer
حرکت زبانی ..... Language Motion
مدل‌های زبانی بزرگ ..... large language models
بازنمایی نهفته ..... Latent Representation
یادگیری از بازی ..... learn from playing
یادگیری از طریق مشاهده ..... Learning from Observation
بیانگری ادراکی محدود ..... limited perceptual expressiveness
کاوشگری خطی ..... linear exploration
زبان‌شناسی ..... Linguistics
طولانی‌مدت ..... long-horizon
ترجمه ماشینی ..... machine translation
آینده‌نگری مولد ماسک‌دار ..... masked generative foresight

ترکیبی از متخصصین ..... Mixture of Experts
تزریق مُدالیته ..... modality injection
چندتکه ..... modular
یکپارچه ..... monolith
چندمرحله‌ای ..... multi stage
توجه چند-سری ..... Multi-Head Attention
شبکه عصبی چندلایه ..... multi-layer perceptron
چندرسانه‌ای ..... multi-media
کانال‌های چندوجهی ..... Multi-modal HRI
دستورات چندوجهی ..... Multimodal Prompts
چندوجهی ..... Multi-modal Understanding
تقویت متقابل ..... Mutual Enhancemen
زبان طبیعی ..... natural language
پردازش زبان طبیعی ..... natural language processing
پردازش زبان‌های طبیعی ..... natural language processing
شبکه‌های عصبی ..... neural networks
میدان‌های تابشی عصبی ..... Neural Radiance Fields
عصبی-نمادین ..... neuro-symbolic
شناختی-عملی ..... Nvidia
تشخیص اشیاء ..... object detection
آشکارساز اشیاء ..... object detector
شیء-محور ..... Object-centric
انسداد ..... occlusion
شیء-محور ..... O-CoT
خارج-از-دامنه ..... off-domain
نمایش تک‌نمونه ..... One-shot Demonstration
متن-باز ..... Open-Source
برنامه‌ریز حرکت مبتنی بر بهینه‌سازی ..... optimization-based motion planner
درک چندوجهی منفعل ..... Passive Multimodal Understanding
درک منفعلانه ..... Passive Perception
تشخیص الگو ..... pattern recognition
ادراک-به-عمل ..... perception-to-action

فضای عمل یکپارچه با تفسیر فیزیکی ..... Physically Interpretable Unified Action Space
استخراج دانش پیشین فیزیکی ..... Physics Prior Extraction
پایگاه دانش فیزیکی پیشین ..... Physics Prior Knowledge Base
برداشتن-و-گذاشتن ..... pick-and-place
متخصص انتشار قابل اتصال ..... plug-in diffusion expert
خط‌مشی ..... Policy
ژست ..... pose
رمزگذاری موقعیتی ..... Positional Encoding
پیش‌پردازش ..... preprocessing
امکان‌پذیری یا احتمال موفقیت ..... probability of success
دستورات متنی برنامه‌مانند ..... program-like prompt structure
تغییر دستورات متنی ..... prompt switching
اطلاعات حالت ..... Proprioception
واکنشی ..... Reactive
کنترل واکنشی ..... Reactive Control
تکواژهای بازخوانی ..... Readout Tokens
بلادرنگ ..... Real-time
حلقه بازخورد نیرو در لحظه ..... Real-time Force-Feedback Loop
قطع ارتباط استدلال-عمل ..... reasoning-action disconnection
تولید عمل مشروط به استدلال ..... reasoning-conditioned action generation
دستورالعمل ..... recipe
بازگشتی ..... Recurrence
زاید ..... redundant
مدل‌های خطی ..... regression
بازنمایی ..... representation
رباتیک ..... robotic
مبدل‌های رباتیک ..... Robotic Transformer
بخش‌بندی ..... segmentation
توجه به خودی ..... self attention
ماشین‌های خودران ..... self driving car
استدلال خود-تولید ..... Self-Generated Reasoning
خودنظارتی ..... self-supervised

هدف کمکی خودنظارتی ..... Self-Supervised Auxiliary Objectives
برنامه‌ریزی معنایی ..... Semantic Planning
حسگر-حرکتی ..... Sensorimotor
حسگر-موتور ..... Sensorimotor Models
رمزگشا ..... sentiment analysis
دنباله-به-دنباله Seq2Seq
توالی ..... sequence
ترتیبی ..... Sequential
حالت سایه ..... Shadow Mode
صافی ..... smoothness
مبدل فضایی ..... Spatial Transformer
ساختاریافته ..... structured
بازنمایی‌های میانی ساختاریافته‌ای ..... structured intermediate representations
زیروظیفه فعلی ..... SUBTASK
تجزیه مبتنی بر وظیفه فرعی ..... Subtask-based Decomposition
زیروظیفه‌ها ..... subtasks
خلاصه‌سازی ..... summerization
نظارتی ..... supervised
مصنوعی-سپس-واقعی ..... Synthetic-then-Real
استنتاج سیستمی ..... Systemic Inference
برنامه‌ریزی وظایف ..... task planning
بردار متنی ..... text-based formulation
پیش‌بینی تکواژ ..... Token Prediction
مصالحه ..... trade-off
مبدل‌ها ..... transformers
مسیر انتقال ..... Transport pathway
دو-مسیره ..... two-stream
عدم اتصال به واقعیت ..... un-groundedness
غیرنظارتی ..... unsupervised
توابع ارزش ..... value functions
مدل‌های بینایی-زبانی-عمل ..... vision language action model
مدل‌های زبانی-بینایی ..... vision language models

اتصال بصری ..... Visual Grounding

اشاره بصری ..... Visual Pointing / Deixis

واژگان ..... vocabulary

نقشه‌های ارزش وُکسلی ..... Voxel-based Value Maps

مدل‌های جهان ..... World Models

بدون-داده ..... Zero-shot

Abstract

Creating intelligent robotic agents capable of understanding high-level natural language commands and executing them in complex physical environments has long been one of the fundamental challenges in artificial intelligence. Traditionally, this required explicit and costly programming or extensive domain-specific training.

The emergence of large-scale Vision-Language Models (VLMs), trained on massive internet data, has introduced a new paradigm. These models provide commonsense knowledge and unprecedented semantic reasoning capabilities, offering the potential to revolutionize robotic task planning.

However, transferring this conceptual knowledge from passive understanding to embodied and goal-directed action — that is, the transition from VLMs to Vision-Language-Action models (VLAs) — remains an open and fundamental research challenge. The core issue lies in *grounding* abstract linguistic concepts into precise, physically feasible interactions within the real world.

This study conducts an in-depth exploration and critical analysis of architectures and computational strategies designed to bridge this gap, with a special focus on robotic task planning. We first review the foundational concepts of artificial intelligence, deep learning, natural language processing, and computer vision that underpin these models. Then, the core of this research is presented through a dissection and categorization of advanced VLA model architectures.

The taxonomy includes unified models, hierarchical approaches based on task decomposition, affordance-based models, and world-model-based architectures. By critically analyzing representative models within each category, we identify key challenges such as generalization, data efficiency, multi-step reasoning, and real-time execution. This analysis provides a comprehensive framework for understanding the current research landscape and outlines future directions toward developing intelligent, general-purpose robotic agents.

Keywords: Task Planning, VLM Models, VLA Models, Embodied Artificial Intelligence.

Iran University of Science and Technology
School of Computer Engineering

Study of language-vision models for multi-step task planning in robotics

Degree of Master of Science in Computer Engineering
software engineering

By:
Alireza Nazari

Supervisor:
Dr. Minaei bidgoli

Fall 2025