비전-언어-액션 (VLA) 모델: 개념, 진전, 응용및도전과제

Ranjan Sapkota[a], Yang Cao[b], Konstantinos I. Roumeliotis[c], Manoj Karkee[a]

[a]*Cornell University, Biological & Environmental Engineering, Ithaca, New York, USA*
[b]*The Hong Kong University of Science and Technology, Department of Computer Science and Engineering, Hong Kong*
[c]*University of the Peloponnese, Department of Informatics and Telecommunications, Greece*

Abstract

비전-언어-액션 (VLA) 모델은인공지능의혁신적인진전을의미하며, 지각, 자연어이해, 그리고구현된행동을단일계산프레임워크내에서통합하는것을목표로합니다. 이기초리뷰는최근비전-언어-액션모델의발전을종합적으로제시하며, 빠르게진화하는이분야의지형을구조화하는다섯가지주제축에체계적으로조직되어있습니다. 우리는 VLA 시스템의개념적기반을확립하는것부터시작하며, 크로스모달학습아키텍처에서비전언어모델 (VLM), 행동계획자, 계층적컨트롤러를긴밀히통합하는범용에이전트로진화한과정을추적합니다. 저희방법론은지난 3 년간발표된 80 개이상의 VLA 모델을포함하는엄격한문헌검토체계를채택합니다. 주요진전분야로는아키텍처혁신, 효율적인학습전략, 실시간추론가속등이있습니다. 우리는자율주행차, 의료및산업로봇공학, 정밀농업, 휴머노이드로봇공학, 증강현실등다양한응용분야를탐구합니다. 이리뷰는실시간제어, 멀티모달액션표현, 시스템확장성, 보이지않는작업에대한일반화, 윤리적배포위험등주요과제들을추가로다룹니다. 최첨단기술을바탕으로, 에이전트 AI 적응, 신체간일반화, 통합신경-기호계획등목표해결책을제안합니다. 우리는 VLA 모델, VLM, 에이전트 AI 가만나사회적으로정렬되고적응적이며범용체화에이전트를강화하는미래지향적로드맵을제시합니다. 따라서이연구는지능형실제로봇공학과인공범용지능을발전시키는데기초참고자료가될것으로기대됩니다. 프로젝트저장소는 GitHub(Source Link)

Keywords: 비전-언어-액션, 액션토큰화, 인공지능, 로봇공학, 시각언어모델

1. 소개

비전-언어-액션 (VLA) 모델이개발되기전에는로봇공학과인공지능의발전이주로별도의영역에서이루어졌습니다: 이미지를획득, 해석, 인식할수있는시각시스템입니다 [55, 87, 176], 텍스트를이해하고생성할수있는언어체계들 [213, 177], 그리고이동을제어할수있는액션시스템 [60]. 이고립된시스템들은자체적으로는잘작동했지만, 함께작동하고새로운시나리오에일반화하거나실제문제의복잡성과예측불가능성에적응하는데어려움을겪었습니다 [57, 22].

그림 1 에나타난바와같이, 합성곱신경망 (CNN) 에기반한전통적인컴퓨터비전모델은객체탐지 [25, 157, 26, 24, 174] 나분류 [205, 89, 95, 73] 와같이범위가좁게정의된작업에맞추어설계되었으며, 환경이나목표가조금만달라져도대규모라벨데이터셋과번거로운재학습을필요로했습니다 [200, 77]. 이러한시각모델은그림 1 에나타난바와같이과수원의사과를식별하는등"볼수"는있었지만, 언어를이해하거나시각적정보를원하는행동으로변환하는능력은부족했습니다. 대형언어모델 (LLM) 을비롯한언어모델은텍스트기반이해와생성에서큰진전을이루었지만 [28], 물리적세계를지각하거나추론하는능력없이언어처리에만제한되어있었습니다 [98]. 한편, 수작업정책이나강화학습에크게의존하던로봇공학의행동기반시스템은객체조작과같은특정동작을수행할수있었지만 [158], 세심한공학적설계가필요했고사전에설계된시나리오밖으로일반화하기어려웠습니다 [153].

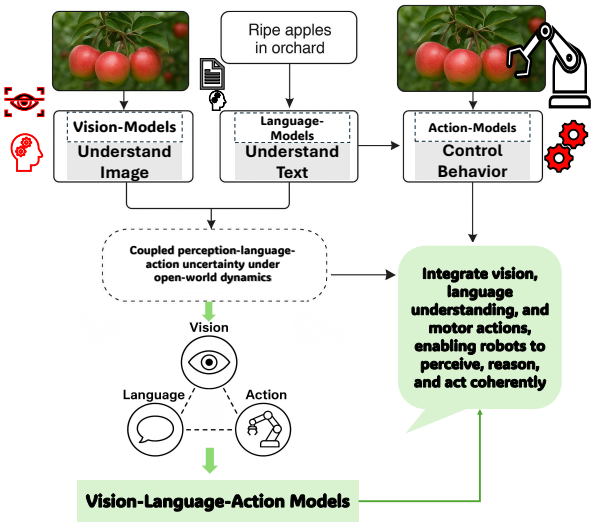

Figure 1: **고립된모달리스에서통합된비전–언어–액션모델로의진화.** 통합된지각, 언어, 행동은적응적이고일반화가능한신체적지능을가능하게합니다.

시각과언어를결합하여뛰어난멀티모달이해성능을보인 VLM 의발전에도불구하고 [190, 30, 298, 37, 10, 189], 멀티모달입력에근거하여일관된행동을생성하거나실행하는통합능력에는여전히공백이남아있었습니다 [156, 137]. 그림 1 에서 더자세히시각화된바와같이, 대부분의 AI 시스템은비전-언어, 비전-행동, 언어-행동과같이한두가지양식에특화되어있

*Ranjan Sapkota
Email address: rs2672@cornell.edu (Manoj Karkee)

었고, 세양식을하나의종단간프레임워크로완전히통합하는데어려움을겪었습니다. 따라서로봇은시각적으로"사과"를인식하거나, "사과를따라"와같은텍스트명령을이해하거나, 잡기와같은사전정의된운동동작을수행할수는있었지만, 이러한능력을유연하고적응적인행동으로통합하여실행하는데에는한계가있었습니다. 그결과새로운작업이나환경에유연하게적응하지못하는취약한파이프라인이형성되었고, 일반화가불안정하며노동집약적인엔지니어링이필요했습니다. 이러한한계는체화 AI 의중요한병목을드러냈습니다. 즉, 지각하고이해하며행동하는과정을공동으로수행할수있는시스템없이는지능형자율행동이여전히어려운목표로남아있었습니다.

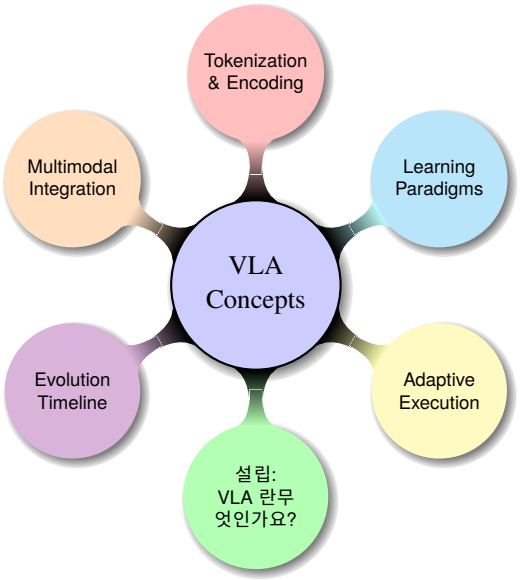

Figure 2: 핵심 VLA 개념의마인드맵. 각색상구분된분기는기본차원을강조합니다: 정의 (기초), 역사적진화, 멀티모달통합, 토큰화및인코딩, 학습패러다임, 그리고구현된환경에서적응적실행.

이러한격차를메우려는긴급한필요성이 VLA 모델의등장을촉진시켰습니다. 2021 년에서 2022 년경구상된 VLA 모델로, 구글딥마인드의 Robotics Transformer 2(RT-2) 와같은노력으로선구되었습니다 [299] 는지각, 추론, 제어를단일프레임워크내에서통합하는변혁적아키텍처를도입했습니다. 그림에제시된한계에대한해결책으로서 2, VLA 는시각입력, 언어이해, 운동제어기능을통합하여, 구현된에이전트가주변을인지하고복잡한명령을이해하며적절한행동을동적으로수행할수있게합니다. 초기 VLA 접근법은시각언어모델을확장하여로봇운동명령의수치또는기호표현동작토큰을포함시켜, 쌍으로연결된시각, 언어, 궤적데이터로부터학습할수있게했습니다 [156]. 이방법론적혁신은로봇이보이지않는물체에일반화하고, 새로운언어명령을해석하며, 구조화되지않은환경에서다단계추론을수행하는능력을획기적으로향상시켰습니다 [107].

VLA 모델은시각, 언어, 행동을별도의영역으로다루는오랜한계를극복하는통합멀티모달지능개발의변혁적단계를나타냅니다 [156]. 시각적, 언어적, 행동적정보를통합한인터넷규모의데이터셋을활용함으로써, VLA 는로봇이환경을인식하고설명할뿐만아니라맥락을따라추론하고복잡하고역동적인환경에서적절한행동을수행할수있도록지원합니다 [254]. 그림에나타난진행 2 그리고 3 고립된시각, 언어, 행동시스템에서통합 VLA 패러다임으로, 진정으로적응적이고일반화가능한체화된에이전트개발로의근본적인전환을포착합니다. 이패러다임의변혁적잠재력을고려할때, 현재문헌에대한포괄적이고비판적으로정보에기반한검토는시기적절하고필수적입니다. 첫째, 이러한검토는 VLA 를이전모델과구별하는기본개념과아키텍처원칙을명확히하기위해필요합니다. 둘째, 이분야의급속한진전과주요이정표를체계적으로설명하여연구자와실무자가알고리즘및기술발전의궤적을이해할수있도록돕습니다. 셋째, 가정용로봇공학부터산업자동화및보조기술에이르기까지 VLA 가이미변혁적잠재력을보여주고있는다양한실제응용분야를파악하기위해서는심층검토가필수적입니다. 더나아가, 데이터효율성, 안전성, 일반화, 윤리적고려사항등현재의도전과제를비판적으로검토함으로써, 광범위한배포를위해해결해야할장벽을식별합니다. 마지막으로, 이러한통찰을종합하면 AI 및로봇커뮤니티에새로운연구방향과실질적고려사항을알리고, 협력과혁신을촉진합니다.

본리뷰에서는 VLA 모델의기본원리를체계적으로분석합니다. 또한그들의개발진행상황과기술적도전과제에대해서도논의합니다. 우리의목표는 VLA 에대한현재의이해와응용을통합하고, 한계를식별하며향후진화방향을제안하는것입니다. 검토는주요개념적기초에대한상세한검토로시작됩니다 (그림참조) 2 여기에는 VLA 모델의정의, 그역사적진화, 멀티모달통합메커니즘, 그리고비전, 언어, 행동을아우르는통합토큰화및표현전략이포함됩니다. 이러한개념적설명은 VLA 가어떻게구조화되고다양한모달리즘에걸쳐작동하는지이해하는데기반을마련합니다.

이설명을바탕으로최근진행상황과훈련효율성전략에대한통합된관점을제시합니다 (그림 3). 여기에는 VLA 모델에서채택되고확장된아키텍처혁신뿐만아니라, 데이터효율적인학습프레임워크, 매개변수효율적인모델링기법, 그리고원래더넓은기계학습및로봇공학맥락에서개발된모델가속전략이포함됩니다. 이러한발전들은 VLA 시스템을실제적용에맞게확장하는데매우중요합니다.

이후 VLA 시스템이현재직면한한계에대한포괄적인논의를제시합니다 (그림참조) 3 이들중다수는구현된 AI 와로봇공학의광범위한도전을반영하지만, 시각, 언어, 행동의긴밀한통합으로인해뚜렷하고복합적인형태로나타납니다. 논의할한계에는추론병목현상, 안전문제, 높은계산요구량, 제한된일반화, 윤리적함의가포함됩니다. 우리는이러한시급한문제들을강조할뿐만아니라, 이를해결할수있는잠재적해결책에대한분석적논의도제공합니다.

이세그림은함께이리뷰에서제시된텍스트분석을뒷받침하는시각적묘사를제공합니다. 개념적배경, 최근혁신, 그리고미해결과제를개괄함으로써이연구는향후연구를안내하고보다견고하고효율적이며윤리적으로기반을둔 VLA 시스템개발을장려하는것을목표로합니다.

그림 4 이리뷰의전체구조와논리적흐름을요약하고, 원고가 VLA 연구를포괄적이고체계적으로분석하도록어떻게구성되었는지보여줍니다. 그림에서볼수있듯이, 논문은기초개념에서최근발전, 적용, 도전과제, 미래연구방향으로발전하여섹션간일관된서술을보장합니다. 이아키텍처를구축하기위해주요키워드"Vision–Language–Action" 과"Vision–Language Models", 그리고일반적으로사용되는약어"VLA"를사용하여광범위하고엄격한문헌검색이이루어졌습니다. 이키워드는 Hugging Face, arXiv, ScienceDirect, Nature, IEEE Xplore, Wiley, Springer Nature 등주요학술및기술저장소에서

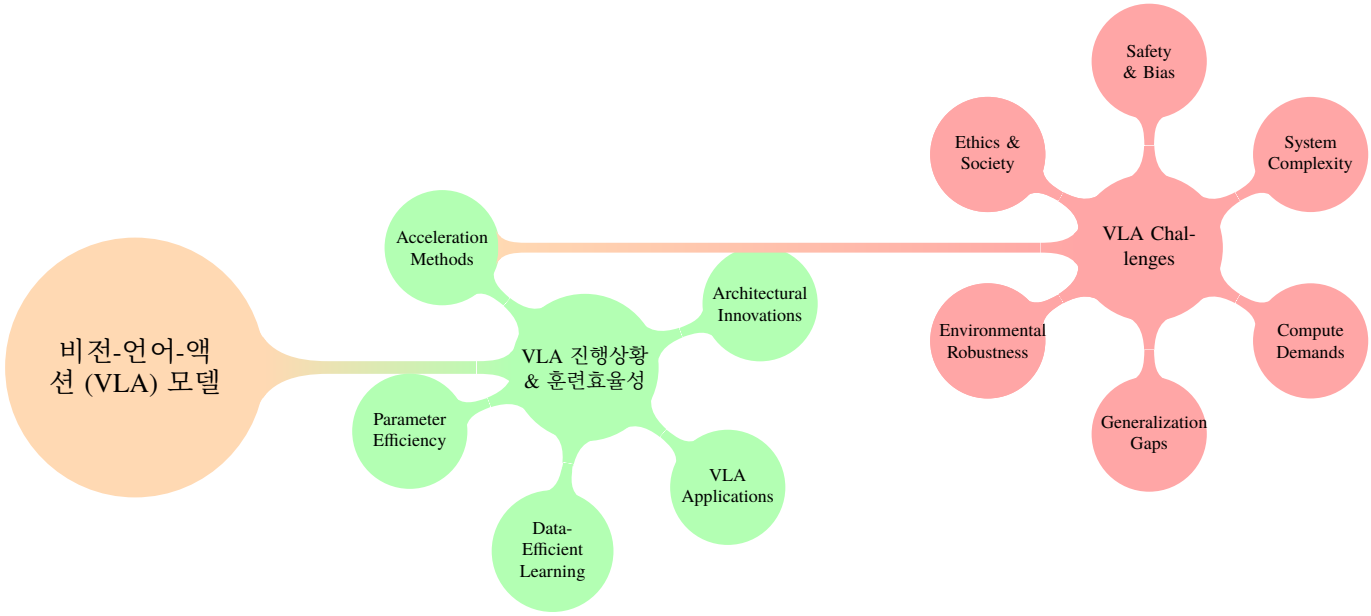

Figure 3: VLA 모델생태계를보여주는마인드맵: 학습효율성 (아키텍처혁신, 데이터/매개변수효율성, 가속) 의진전과추론제약, 멀티모달동작, 안전성, 일반화, 윤리와함께, 확장가능한실제배포를위해극복해야할주요과제들.

후보연구를검색하는데사용되었습니다. 결과말뭉치는수작업심사를통해관련성, 기술적깊이, 그리고이번검토범위와의일치를더욱정제했습니다. VLA 개념, 방법론적진전, 응용영역, 미해결과제이해에직접적으로기여한논문만이보존되었습니다. 이다단계필터링과정은기초연구와최첨단연구를균형있게포괄하면서도주변적이거나느슨하게관련된연구를피할수있게했습니다. 그결과, Figure 4 논문의주제구성을반영할뿐만아니라기저리뷰방법론을구현합니다.

본논문은구조적이고계층적인조직을따르며, VLA 모델의기초, 진화및함의를체계적으로발전시키며, 이는그림에요약되어있습니다 4. 발표는체화된지능과 VLA 의등장을동기부여하는소개로시작되며, 이어서멀티모달통합, 토큰화, 학습패러다임, 실시간제어등핵심원칙을확립하는개념섹션이이어집니다. 이러한토대를바탕으로진행섹션은아키텍처혁신, 교육및효율성향상, 그리고매개변수효율적인가속전략을살펴봅니다. 응용섹션에서는이러한발전을휴머노이드로봇, 자율주행차, 의료, 농업, 산업시스템, 인터랙티브 AR 내비게이션등실제영역에기반을두고있습니다. 이어서실시간추론, 안전성, 일반화, 시스템통합, 윤리적고려사항등주요과제에대한집중분석을진행합니다. 논문은지속적인학습, 확장성, 해석가능성, 내체적지능분야에서교차연구방향을정제한미래로드맵으로마무리됩니다.

2. 비전-언어-액션모델의개념

VLA 모델은시각적입력을공동으로처리하고, 자연어명령을해석하며, 동적환경에서작동하는물리적로봇하드웨어에구현할수있는실행가능한동작표현을생성하는지능형시스템의한군을나타냅니다. 기술적으로 VLA 는비전인코더 (예: CNN, ViT), 언어모델 (예: LLM, 트랜스포머), 정책모듈또는플래너를결합하여작업조건제어를달성합니다. 이모델은일반적으로교차주의, 연결임베딩, 토큰통합과같은시각언어모델에서확립된멀티모달융합기법을기반으로하며, 감각관찰, 언어지시, 액션표현을정렬하도록확장합니다.

전통적인가시모터파이프라인과달리, VLA 는의미론적그라운딩 (semantic grounding) 를지원합니다 [154], 문맥인식추론을가능하게합니다 [228], affordance detection [79], 그리고시간계획 [141]. 일반적인 VLA 모델은카메라또는센서데이터를통해환경을관찰하고, 언어로표현된목표 (예: ” 빨간사과를집어라”) 를해석한뒤 (그림 5 참조), 자동화된시스템이실행할수있는저수준또는고수준동작시퀀스를출력합니다. 최근발전은모방학습, 강화학습, 또는검색증강모듈을통합하여표본효율성과일반화를향상시키고있습니다. 본리뷰는 VLA 모델이기초융합아키텍처에서로봇공학, 내비게이션, 인간-AI 협업전반에걸쳐실제세계에배포할수있는범용에이전트로어떻게진화했는지살펴봅니다.

VLA 모델은시각적지각, 언어이해, 신체액션생성을하나의프레임워크로통합하는멀티모달인공지능시스템입니다. 이모델은로봇이나 AI 에이전트가감각입력 (예: 이미지, 텍스트) 을해석하고, 맥락적의미를이해하며, 실제환경에서자율적으로작업을수행할수있게합니다. 이는고립된하위시스템이아닌종단간학습과행동을통해이루어집니다. 개념적으로그림 5에나타난바와같이, VLA 모델은이전로봇및 AI 시스템의능력을제한했던시각인식, 언어이해, 운동실행간의역사적간극을연결합니다.

2.1. 발전과연대표

2022 년부터 2025 년까지 VLA 모델의급속한발전은세가지뚜렷한진화단계를보여줍니다:

1. **기초통합 (2022–2023).** 초기 VLA 는멀티모달융합구조를통해기본적인시각-운동조정을확립했습니다. CLIPort [202] 는 CLIP 임베딩과모션프리미티브를결합했고, Gato [181] 는 604 개과제에서범용역량을보였으며, RT-1 [19] 은대규모모방학습을통해높은조작성공률을달성했습니다. VIMA [112] 는트랜스포머기반계획기를통해시간적추론을도입했으며, 2023 년에는 RT-2 [299]

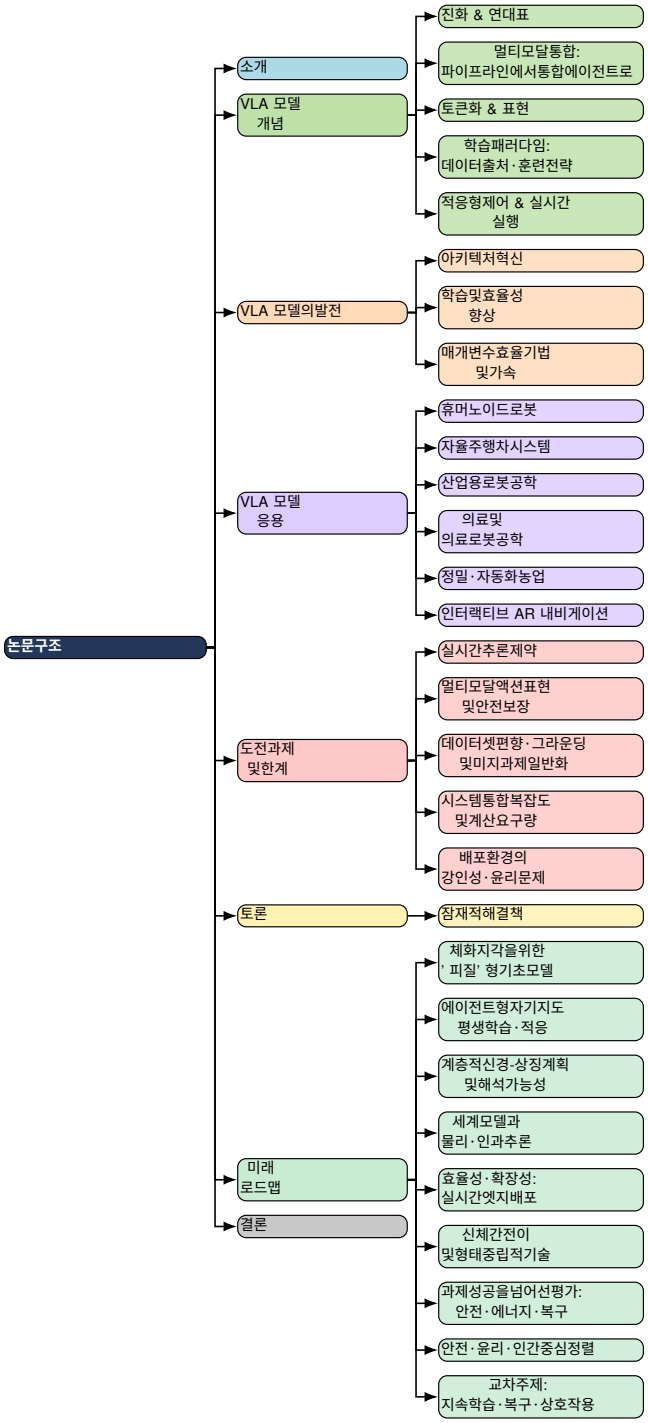

Figure 4: 이논문의전체구성과논리적흐름. 서론과 VLA 개념에서출발해최근진전, 응용, 도전과제, 논의및미래로드맵으로이어지는구조를요약한다.

가시각적사고연쇄추론을가능하게하고 Diffusion Policy [40] 가확산과정을통한고급확률적행동예측을발전시켰습니다. 이러한기초모델은저수준제어를다루었지만, 복잡한작업을재사용가능한의미기반하위행동으로분해하고새로운맥락에서재결합하는구성적추론은여전히제한적이었으며, 이는이후의유연성중심연구를촉진했습니다 [287, 100].

2. **전문화와체화된추론 (2024).** 2 세대 VLA 는도메인특화유도편향을도입했습니다. Deer-VLA [269] 는검색보조학습을통해 few-shot 적응을개선했고, Uni-NaVid [280] 는 3D 장면그래프통합을통해내비게이션을최적화했으며, ReVLA [48] 는메모리효율적인가역아키텍처를도입했습니다. Occllama [239] 는물리기반주의메커니즘으로부분관측가능성을다루었습니다. 동시에 Arai 등 [5] 은객체중심얽힘해소를통해구성적이해를향상시켰고, Zhou 등 [293] 은멀티모달센서융합을통해 VLA 의적용범위를자율주행으로확장했습니다. 이러한발전은새로운벤치마킹방법론의필요성을부각했습니다 [254].

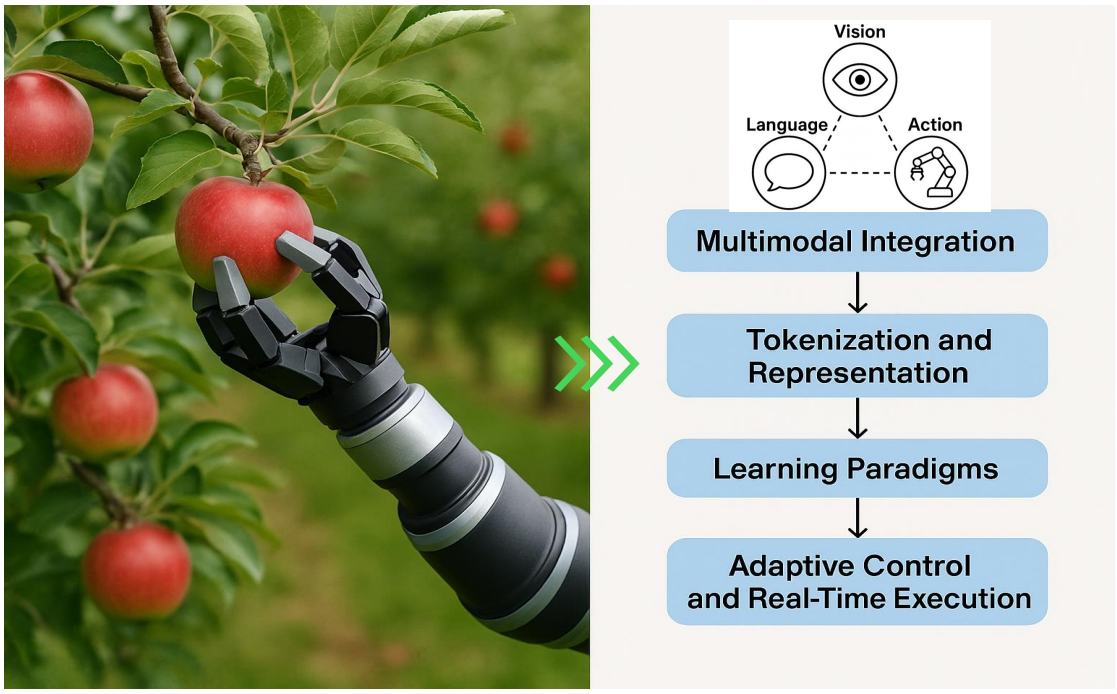

Figure 5: **VLA 모델의기초개념 (사과따기시나리오에서)** 이그림은로봇팔이과수원에서자율적으로익은사과를따는모습을보여주며, VLA 모델의안내를받는다. 오른쪽에는 VLA 모델의네가지주요단계인멀티모달통합, 토큰화및표현, 학습패러다임, 적응제어및실시간실행을설명한플로우차트가있습니다.

3. **일반화및안전중요배포 (2025).** 최신시스템은강인성과인간정렬을우선시합니다. SafeVLA [274] 는위험인식의사결정을위해형식적검증을통합했고, Humanoid-VLA [53] 는계층적 VLA 를통한전신제어를입증했으며, Edge VLA [20] 는임베디드배포를위한계산효율성을최적화했습니다. CogACT [131] 는인과추론을위해신경-기호 (neuro-symbolic) 논리를결합했습니다. 또한 Chain-of-Affordance [129] 의어포던스 (행동유도성) 체이닝과 GR00T N1 [14] 의시뮬레이션-실제 (Sim-to-Real) 전이는신체간일반화문제를다루며, ShowUI [139] 는자연어그라운딩을통해 VLA 와인간개입 (Human-in-the-loop) 인터페이스를연결합니다.

그림 6 은 2022 년부터 2025 년까지개발된 45 개의 VLA 모델의진화과정을종합적으로보여줍니다. CLIPort [202], Gato [181], RT-1 [19], VIMA [112] 와같은초기 VLA 시스템은사전학습된비전-언어표현을작업조건부조작및제어정책과결합하여기초를마련했습니다. 이후 ACT [287], RT-2 [299], VoxPoser [100] 는시각적사고연쇄추론과어포던스그라운딩을통합했습니다. Diffusion Policy [40] 와 Octo [218] 는확률모델링과확장가능한데이터파이프라인을도입했습니다. 2024 년에는 Deer-VLA [269], ReVLA [48], Uni-NaVid [280] 가도메인특화와메모리효율적설계를추가했고, Occllama [239] 와 ShowUI [139] 는부분관측성과사용자상호작용을다루었습니다. 이러한흐름은 QUAR-VLA [54] 와 RoboMamba [143] 와같은로봇공학중심 VLA 로이어졌습니다. 최근에는 SafeVLA [274], Humanoid-VLA [53], Mo-ManipVLA [246] 가검증, 전신제어, 메모리시스템을통합하면서일반화와배포가능성을강조합니다. GR00T N1 [14] 과 SpatialVLA [175] 는시뮬레이션-실제 (Sim-to-Real) 전이와공간그라운딩을한층강화했습니다. 이연대표는 VLA 가모듈식학습에서범용적이고안전한체화된지능으로발전해온과정을보여줍니다.

2.2. 멀티모달통합*:* 고립된파이프라인에서통합에이전트로

VLA 모델의등장에서중심적인진전은멀티모달통합, 즉시각, 언어, 행동을통합된아키텍처내에서공동처리할수있다는점에있습니다. 전통적인로봇시스템은지각, 자연어이해, 제어를개별모듈로취급하며, 종종수동으로정의된인터페이스나데이터변환을통해연결되었습니다 [140, 21, 219]. 예를들어, 고전적인파이프라인기반프레임워크는상징적라벨을출력하는지각모델이필요했고, 이를플래너가도메인별수작업엔지니어링을통해특정행동에매핑했습니다 [178, 117]. 이러한접근법은적응력이부족했고, 모호하거나보이지않는환경에서는실패했으며, 미리정의된템플릿이상의지침을일반화할수없었습니다.

반면, 현대 VLA 는대규모사전학습인코더와트랜스포머기반아키텍처를사용하여다양한양식을종단간융합합니다 [244]. 이변화는모델이동일한계산공간내에서시각적관찰과언어적지시를해석할수있게하여유연하고맥락인식있는추론을가능하게합니다 [128]. 예를들어, 그림 5에나타난' 익은사과따기' 과제에서비전인코더는일반적으로비전트랜스포머 (ViT) 또는 ConvNeXt 로, 장면을파싱하여관련객체 (예: 과일, 잎사귀, 배경) 를위치화하고분류하며, 고정된색상가정이아닌학습된질감, 형태, 맥락적특징에기반해숙성도와관련된시각적단서를추론합니다 [243]. 한편, 언어모델은종종 T5, GPT, 또는 BERT 의변형으로, 명령어를고차원임베딩으로인코딩합니다. 이표현들은교차주의또는공동토큰화방식을통해융합되어, 액션정책을이해하는통합된잠재공간을만들어냅니다 [86].

이멀티모달시너지는 CLIPort 에서처음으로효과적으로입증되었습니다 [202] 이명령어는테이블탑장면의 RGB 이미지와자연어명령어 (예: ' 빨간칸위에파란블록을놓다') 를입

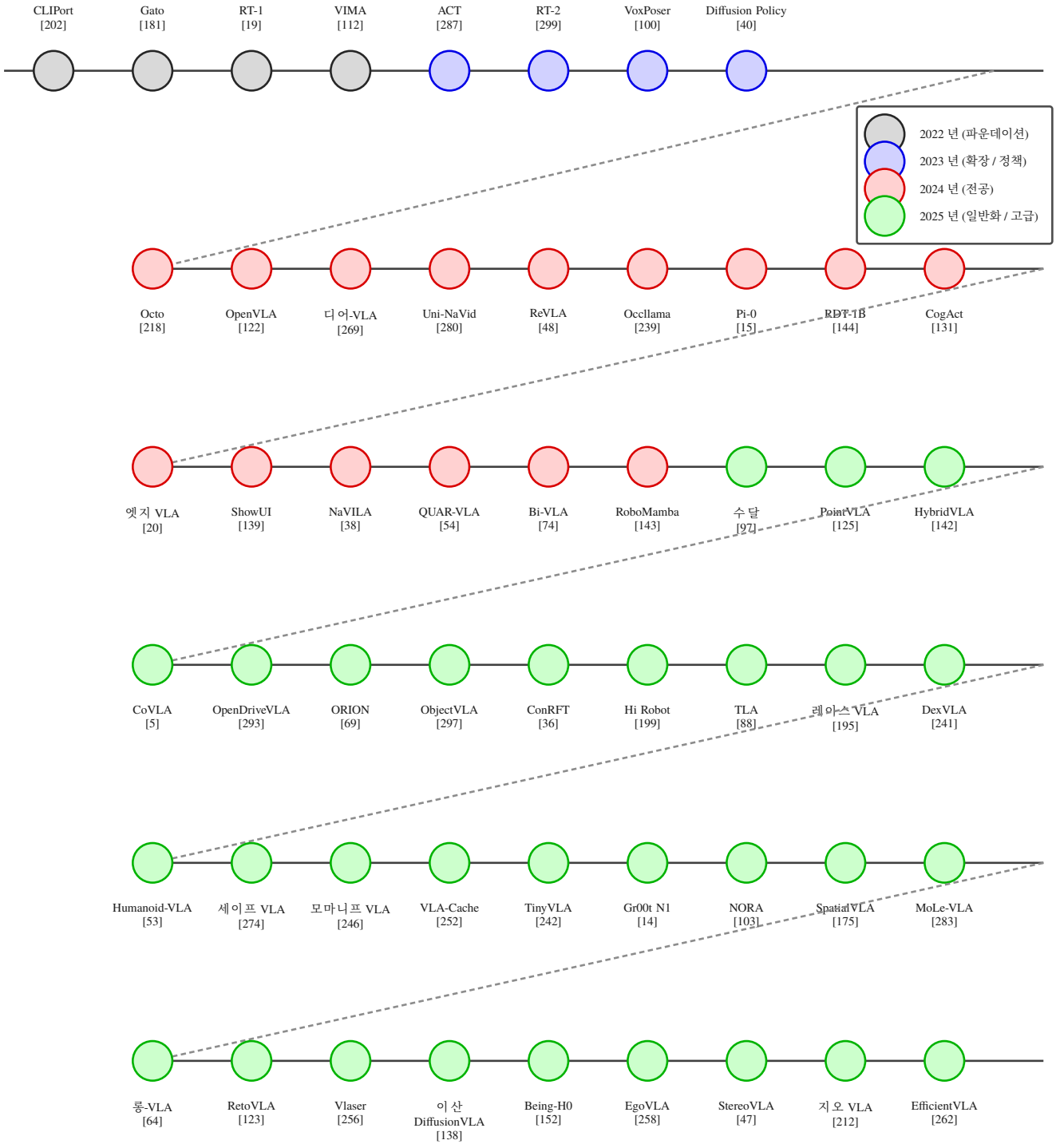

Figure 6: 2022 년부터 2025 년까지연도별로구성된 Vision-Language-Action(VLA) 모델의종합연대표로, 고대비색상코딩과주제별그룹화가적용되었습니다.

력으로받아의미론적그라운딩을위해 CLIP 으로인코딩하고, 합성곱전송디코더를통해픽셀단위의픽앤플레이스액션분포를출력합니다. 언어임베딩에시각-운동정책을직접조건화함으로써 CLIPort 는명시적언어구문분석을제거하고종단간언어조건조작을가능하게합니다. 마찬가지로 VIMA [112] 이접근법은트랜스포머인코더를사용하여객체중심시각토큰과명령어토큰을공동처리함으로써공간추론과제전반에걸쳐 few-shot 일반화를가능하게했습니다.

최근의발전은시간적·공간적기반을포함시켜이융합을더욱촉진합니다. VoxPoser [100] 사전학습된시각언어모델과고전모션플래너를구성하여 3D 객체선택의모호함을해결하기위해복셀수준의추론을사용하며, 특히과제별훈련데이터없이제로샷조작을달성합니다. 반면, RT-2 는 [299] 시각적언어토큰과액션표현을통합된트랜스포머내에서융합하며, 대규모인터넷시각언어말뭉치와 RT-1 데이터셋의 10 만개이상의실제로봇시연을공동학습하여보이지않는명령에대한

제로샷일반화를가능하게합니다. 또다른주목할만한기여는 Octo [218] 입니다. 이모델은 Open X-Embodiment 데이터셋을통해다양한로봇과환경에서수집된 400 만개이상의로봇궤적을학습한메모리증강트랜스포머를소개하며, 장기적의사결정을지원하고공동지각-언어-행동학습의확장성을입증합니다.

무엇보다도, VLA 는실제그라운딩에서직면한문제에대한강력한해결책을제공합니다. 예를들어, Occllama [239] 는주의기반메커니즘을통해가림객체참조를처리하며, ShowUI [139] 는비전문가사용자가음성이나타이핑된입력으로에이전트를지휘할수있는자연어인터페이스를시연합니다. 이러한능력들은통합이표면수준의융합에만국한되지않기때문에가능하며; 오히려양상전반에걸친의미적, 공간적, 시간적정렬을포착합니다.

2.3. 토큰화와표현: *VLA* 가세상을인코딩하는방법

VLA 모델을기존시각언어아키텍처와차별화하는핵심혁신은토큰기반표현프레임워크로, 지각보다전체적인추론을가능하게합니다 [161, 286], 언어적및물리적행동공간 [136]. 트랜스포머와같은자기회귀생성모델에서영감을받은현대 VLA 는모든양식 (시각, 언어, 상태, 행동) 를통합하는이산토큰을사용해세계를인코딩하여공유임베딩공간에통합합니다 [142]. 이를통해모델은" 무엇을해야하는가"(의미적추론) 뿐만아니라" 어떻게해야하는가"(제어정책실행) 를완전히학습가능하고조합적으로이해할수있게합니다 [248, 150, 221].

- **접두사토큰: 인코딩문맥및명령어:** 접두사토큰은 VLA 모델의맥락적중추역할을합니다 [252, 107]. 이토큰들은환경장면 (이미지또는비디오를통해) 과함께제공되는자연어지시를압축된임베딩으로인코딩하여모델내부표현을프라이밍합니다 [17].

  예를들어, 그림 7에나타난바와같이, " 빨간트레이위에초록색블록을쌓기" 와같은작업에서는복잡하게잡힌탁상이미지를 ViT 나 ConvNeXt 같은비전인코더를통해처리하고, 명령어는 LLM(예: T5 또는 LLaMA) 에의해임베드됩니다. 이토큰들은이후모델이목표와환경배치에대한초기이해를확립하는접두사토큰의연속으로변환됩니다. 이공유표현은교차모달그라운딩 (cross-modal grounding) 를가능하게하여, 시스템이공간참조 (예: " 왼쪽에", " 파란컵옆") 와객체의미론 (" 녹색블록") 을두양식모두에서해결할수있게합니다.

- **상태토큰: 로봇구성삽입:** 외부자극을인지하는것외에도, VLA 는자신의내부물리적상태를인지해야합니다 [242, 143]. 이는상태토큰을사용하여에이전트의구성관절위치, 힘-토크측정값, 그리퍼상태, 엔드이펙터자세, 심지어인근물체의위치에대한실시간정보를인코딩함으로써달성됩니다 [126]. 이토큰들은특히조작이나이동시상황인식과안전을보장하는데매우중요합니다 [211, 105].

  그림 8는 VLA 모델이상태토큰을활용하여조작과내비게이션모두에서동적이고맥락인식적인의사결정을가능하게하는방식을보여줍니다. 그림 8a 에서는로봇팔이부서지기쉬운물체근처에부분적으로뻗은모습이보여집니다. 이러한상황에서상태토큰은관절각도, 그리퍼자세, 종단효과기근접성과같은실시간고유수용성정보를인코딩함으로써중요한역할을합니다. 이토큰들은시각및언어기반접두사토큰과지속적으로융합되어트랜스포머가물리적제약을추론할수있게합니다. 따라서모델은충돌이임박했다고추론하고팔궤적을재조정하거나출력힘조절등모터명령을조정할수있습니다. 그림 8b 에묘사된이동식로봇플랫폼에서는상태토큰은주행거리, LiDAR

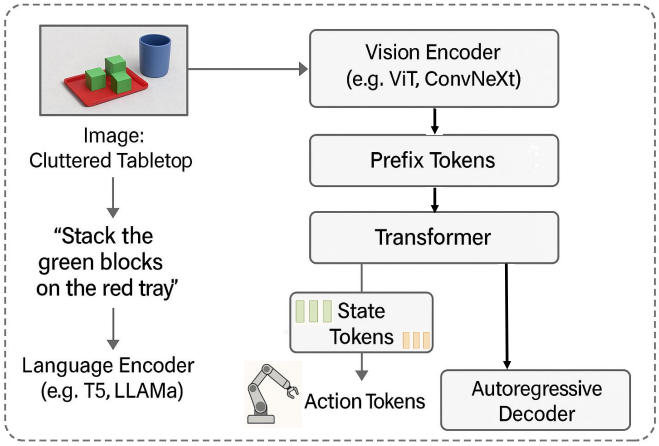

Figure 7: VLA 모델에서종단간토큰화및표현과정을보여주는다이어그램입니다. 시각적입력 (예: 복잡하게보이는탁상) 은비전인코더 (예: ViT) 에의해인코딩되며, 자연어명령어 (예: " 초록블록을쌓기") 는언어인코더 (예: T5) 로처리됩니다. 이시스템은접두사, 상태, 액션토큰을트랜스포머를통해융합하고, 모터동작을자기회귀적으로예측합니다.

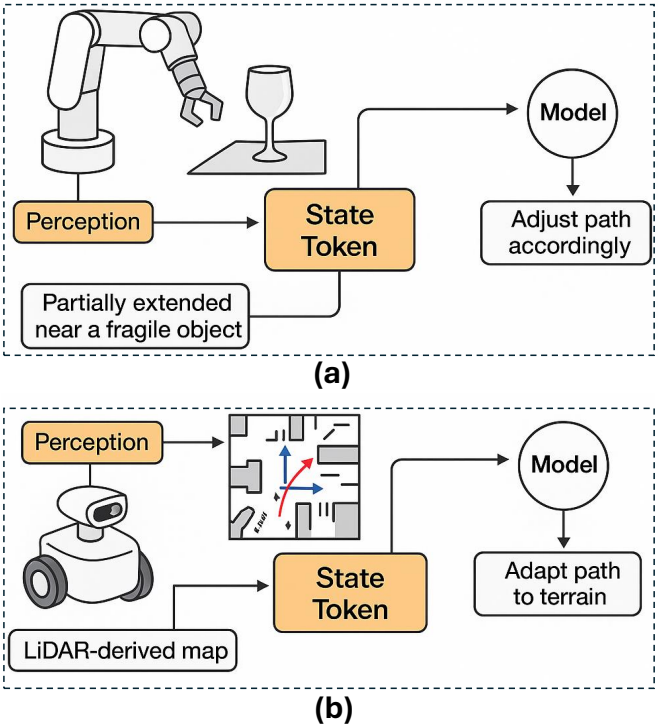

Figure 8: VLA 모델이실제상황에서접두사, 상태, 액션토큰을어떻게활용하는지보여줍니다. 로봇조작에서상태토큰은취약한물체근처에서팔확장을감지하여경로조정을가능하게합니다. 내비게이션에서는 LiDAR 와주행거리측정데이터를나타냅니다. 사과따기과제는접두사토큰이목표이해를안내하는반면, 액션토큰은목표잡기와실행을위한동작시퀀스를생성하는방식을보여줍니다.

스캔, 관성센서데이터와같은공간적특징을캡슐화합니다. 이장비들은지형에의식하는이동과장애물회피에필수적입니다. 트랜스포머모델은이상태표현을환경및교육맥락과통합하여변화하는환경에동적으로적응하는내비게이션동작을생성합니다. 복잡한환경에서물체를잡거나울퉁불퉁한지형을자율적으로이동하는경우, 상태토큰은상황인식을위한구조화된메커니즘을제공하여자동회귀디코더가내부로봇구성과외부감각데이터를모두반영하는정밀하고맥락기반액션시퀀스를생성할수있게합니다.

- **액션토큰: 자기회귀적제어생성:** VLA 토큰파이프라인의마지막계층은액션토큰을포함합니다 [121, 122], 이모델이자기회귀적으로생성하여운동제어의다음단계를나타냅니다 [242]. 각토큰은관절각도업데이트, 토크값, 바퀴속도, 고수준이동원시등저수준제어신호에대응합니다 [81]. 추론과정에서모델은접두사와상태토큰을조건으로한단계씩토큰을해독하여 VLA 모델을언어기반정책생성기로효과적으로전환합니다 [67, 209]. 이공식은실제작동시스템과의원활한통합을가능하게하며, 가변길이의동작시퀀스를지원합니다 [11, 99], 그리고강화또는모방학습프레임워크를통한모델미세조정을가능하게합니다 [285]. 특히 RT-2 같은모델이있습니다 [299] 그리고 PaLM-E [58] 이설계를예로들수있는데, 지각, 지시, 구현이통합된토큰스트림으로통합된것입니다.

  예를들어, 그림 9에묘사된사과따기과제에서모델은과수원이미지와텍스트명령어를포함하는접두사토큰을받을수있습니다. 상태토큰은로봇의현재팔자세와그리퍼가열려있는지닫혀있는지설명합니다. 그후액션토큰이단계별로예측되어로봇팔을사과쪽으로유도하고, 그리퍼방향을조정하며, 적절한힘으로잡기를실행합니다. 이접근법의장점은전통적으로텍스트생성에사용되던트랜스포머가문장을생성하는것과유사한방식으로물리적동작의연속을생성할수있게해준다는점입니다. 다만여기서문장은움직임입니다.

로봇공학에서 VLA 패러다임을구체화하기위해, 그림 9 는시각, 언어, 고유수용감각상태에특화된멀티모달정보가어떻게인코딩되고융합되어실행가능한액션시퀀스로변환되는지를보여주는구조화된파이프라인을제시합니다. 이종단간루프는로봇이"푸른잎근처의익은사과를따라"와같은복잡한작업을해석하고정밀하며상황민감적인조작을수행할수있게합니다. 시스템은먼저 **멀티모달입력획득**단계에서시작하며, 여기서시각적관찰 (예: RGB-D 프레임), 자연어명령, 실시간로봇상태정보 (예: 관절각도또는속도) 라는세가지데이터스트림이수집됩니다. 이들은사전학습된모듈을사용하여각각이산임베딩으로토큰화됩니다 [51, 282]. 그림 9에나타난바와같이, 이미지는 Vision Transformer(ViT) 백본을거쳐 비전토큰을생성하고, 명령어는 BERT 나 T5 와같은언어모델로파싱되어 언어토큰을생성하며, 상태입력은경량 MLP 인코더를통해압축된 상태토큰으로변환됩니다.

이토큰들은교차모달주의메커니즘을통해융합되며, 모델은객체의미, 공간배치, 물리적제약을공동으로추론합니다 [76]. 이렇게융합된표현은의사결정을위한맥락적기반을형성합니다 [96, 149]. 그림 9에서이는 **멀티모달퓨전 (multi-modal fusion)** 단계로표시됩니다. 융합임베딩은일반적으로일련의액션토큰을생성하는자기회귀디코더로전달됩니다.

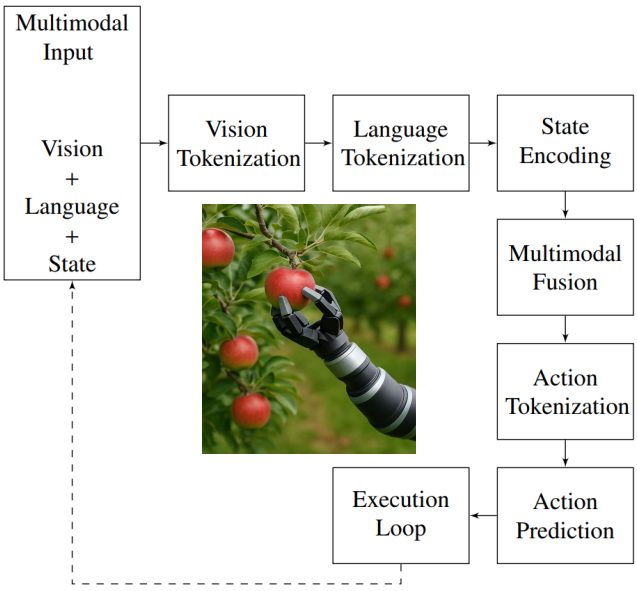

Figure 9: **VLA 가세상을어떻게인코딩하는과정을보여주는과정입니다.** VLA 는시각, 언어, 센서입력을토큰으로변환하고, 교차주의를통해융합하며, 트랜스포머를통한액션시퀀스를예측하고, 실시간피드백으로작업을실행하여로봇이장면을해석하고, 지시를따르며, 동적으로행동을적응할수있게합니다.

이토큰은관절변위, 그리퍼힘조절, 또는고수준모션프리미티브 (예: "파지자세로이동", "손목회전") 를나타낼수있습니다. 예측된액션토큰은이후저수준제어명령으로변환되고, 외부하드웨어의존실행루프를통해실행됩니다. 이루프는로봇컨트롤러와상호작용하여다음 VLA 추론단계에필요한갱신된상태관찰을되돌려줌으로써지각-행동사이클을닫습니다. 이러한폐루프메커니즘덕분에모델은교란, 물체이동, 가림에실시간으로동적으로적응할수있습니다 [276, 155, 251].

명확한구현세부사항을제공하기위해, 알고리즘 1은 VLA 토큰화과정을형식화합니다. RGB-D 프레임 I, 자연어명령 T, 관절각벡터 θ 가주어지면, 알고리즘은순차적으로실행가능한액션토큰집합을생성합니다. 이미지 I 는 ViT 를통해처리되어 400 개의시각토큰집합 V 로변환됩니다. 병렬적으로명령 T 는 BERT 모델에의해 12 개의의미언어토큰시퀀스 L 로인코딩됩니다. 동시에로봇상태 θ 는관절각, 엔드이펙터자세, 고유수용신호를포함하며, 다층퍼셉트론 (MLP) 에의해 64 차원상태임베딩 S 로인코딩됩니다. 이토큰들은교차주의모듈을통해융합되어공유 512 차원표현 F 를생성하며, 이표현은실제행동에필요한의미, 의도, 상황인식을포착합니다. 마지막으로 FAST 와같은정책디코더 [171] 는융합특징을 50 개의개별액션토큰으로매핑하며, 이토큰들은모터명령 $\tau_{1:N}$ 으로디코딩됩니다.

디코딩과정은트랜스포머기반아키텍처를사용하여구현되며, 이는코드스니펫에서확인할수있습니다. 액션예측코드. A **트랜스포머**디코더는 12 개의레이어, 모델차원 512, 그리고 8 개의어텐션헤드로구현되어있습니다. 융합된멀티모달토큰은컨텍스트로제공되며, 디코더는액션토큰을한단계씩자동회귀하여예측합니다. 각예측토큰은전체멀티모달컨텍스트와이전에생성된모든행동을조건으로한다음제어결정을나타냅니다. 결과적으로생성된액션토큰시퀀스는실행을위한

연속모터명령궤적으로디토큰화됩니다. 이구현은 LLM 에서텍스트생성이작동하는방식을반영하지만, 여기서" 문장" 은자연어생성기법을물리적행동합성을위해새롭게재해석한움직임궤적입니다.

함께, 그림 9, 알고리즘 1, 그리고의사코드는 VLA 가일관되고해석가능한토큰공간내에서지각, 지시, 구현을통합하는방식을보여줍니다. 이러한모듈화덕분에프레임워크는작업과로봇형태에걸쳐일반화할수있어, 사과따기, 가정작업, 모바일내비게이션과같은실제응용분야에신속히적용할수있습니다. 특히, 토큰화단계의명확성과분리가능성덕분에아키텍처가확장가능해져 VLA 시스템에서토큰학습, 계층적계획, 기호적기반에대한추가연구가가능해집니다.

Algorithm 1 VLA 토큰화파이프라인

Input: RGB-D frame I, text command T, joint angles θ
$V \leftarrow \text{ViT}(I)$ ▷ 400 vision tokens
$L \leftarrow \text{BERT}(T)$ ▷ 12 language tokens
$S \leftarrow \text{MLP}(\theta)$ ▷ 64-dim state encoding
$F \leftarrow \text{CrossAttention}(V, L, S)$ ▷ 512-dim fused token
$A \leftarrow \text{FAST}(F)$ ▷ 50 action tokens
Output: Motor commands $\tau_{1:N}$

액션예측코드

```
# Python-like pseudocode
def predict_actions(fused_tokens):
    transformer = Transformer(num_layers=12,
        d_model=512,
        nhead=8)
    action_tokens = transformer.decode(fused_tokens,
        memory=fused_tokens)
    return detokenize(action_tokens)
```

2.4. 학습패러다임*:* 데이터소스와교육전략

VLA 모델훈련은웹의의미지식과로봇데이터셋의작업기반정보를통합하는하이브리드학습패러다임을필요로합니다 [35]. 앞서보았듯이, VLA 의멀티모달아키텍처는언어이해, 시각인식, 운동제어를지원하는다양한형태의데이터에노출되어야합니다. 이는일반적으로두가지주요데이터소스를통해이루어집니다.

첫째, 그림 10에묘사된바와같이, 대규모인터넷기반말뭉치가모델의의미론적사전지식 (semantic prior) 의중추를형성합니다. 이데이터셋에는이미지-캡션쌍 (예: COCO, LAION-400M), 지시-따르기데이터셋 (예: HowTo100M, WebVid), 시각적질문답변말뭉치 (예: VQA, GQA) 가포함됩니다. 이러한데이터셋은시각및언어인코더의사전학습을가능하게하여모델이객체, 행동, 개념의일반적인표현을획득하는데도움을줍니다 [2]. 이단계는 CLIP 스타일의대조학습이나언어모델링손실과같은대조적이거나가스막힌모델링목표를사용하여공유임베딩공간내에서시각과언어양식을정렬하는데사용됩니다 [186, 260]. 중요한점은, 이단계가 VLA 에기본' 세계에대한이해' 를제공하여구성일반화, 물체그라운딩, 제로샷전송을용이하게한다는것입니다 [33, 16].

하지만의미론적이해만으로는물리적작업수행에충분하지않습니다 [44, 231, 137]. 따라서두번째단계는모델을체화된경험에기반시키는데초점을맞춥니다 [231]. 실제로봇이나고충실도시뮬레이터에서수집된로봇궤적데이터셋을사용해언어와지각이어떻게행동으로연결되는지모델을가르칩니다 [67]. 여기에는 RoboNet 과같은데이터셋이포함됩니다 [45], Bridge 데이터 [61], 그리고 RT-X [227] 이기능은자연어명령에따라비디오-액션쌍, 공동궤적, 환경상호작용을제공합니다 [159]. 시범데이터는운동감각교육, 원격조작, 스크립트정책등에서나올수있습니다 [115, 13]. 이단계는일반적으로감독학습 (예: 행동복제) 을사용합니다 [68], 강화학습 (RL), 또는모방학습을통해자기회귀정책디코더가융합된시각언어상태임베딩을기반으로액션토큰을예측하도록훈련시키는데사용됩니다 [83].

최근작품들은점점더다양한내용을채택하고있습니다. 다단계또는다중작업훈련전략. 예를들어, 모델은종종마스크언어모델링을이용해시각언어데이터셋에서사전학습된후, 토큰수준의자기회귀손실을이용해로봇시연데이터를미세조정하는경우가많습니다 [122, 295, 252]. 다른이들은커리큘럼학습을사용하는데, 간단한작업 (예: 객체밀기) 이더복잡한작업 (예: 다단계조작) 보다먼저진행됩니다 [288]. OpenVLA 와같이도메인적응을더욱활용하는방법도있습니다 [122] 또는합성및실제배포간의간극을메우기위한시뮬레이션-투리얼전송 [125]. 의미적사전과작업실행데이터를통합함으로써, 이러한학습패러다임은 VLA 모델이작업, 도메인, 구현체에걸쳐일반화할수있게하며, 이는견고한실제세계에서의동작이가능한확장가능한명령어추적에이전트의중추를형성합니다.

공동파인튜닝을통해이데이터셋들은정렬됩니다 [232, 63]. 모델은시각및언어적입력을적절한액션시퀀스로매핑하는것을학습합니다 [175]. 이훈련패러다임은모델이객체활용가능성 (예: 사과를잡을수있음) 과행동결과 (예: 들어올리는데힘과궤적이필요함) 을이해하는데도움을줄뿐만아니라, 새로운시나리오로의일반화를촉진합니다 [129]. 주방조작과제로훈련된모델은객체위치파악, 잡기, 언어지시사항준수등의일반원리를배웠다면야외과수원에서사과를따는방법을추론할수있을것입니다.

최근아키텍처, 예를들어구글딥마인드의 RT-2(Robotics Transformer 2) [299] 이원리를실제로입증했습니다. RT-2 는액션생성을텍스트생성의한형태로다루며, 각액션토큰이로봇의제어공간내개별명령에대응합니다. 이모델은웹규모의멀티모달데이터와수천개의로봇시연을모두기반으로학습되었기때문에, 새로운명령어를유연하게해석하고새로운객체와작업에대한제로샷일반화를수행할수있는데, 이는전통적인제어시스템이나초기멀티모달모델에서는거의불가능했던일입니다.

2.5. 적응형제어및실시간실행

VLA 의또다른강점은센서의실시간피드백을활용해실시간으로동작을조정하는적응형제어에있습니다 [195]. 이는과수원, 가정, 병원과같은역동적이고구조화되지않은환경에서특히중요하며, 예: 바람이사과를움직이거나, 조명변화, 인간의존재등으로인해과제매개변수가바뀔수있습니다. 실행중에상태토큰은실시간으로업데이트되어센서입력과공동피드백을반영합니다 [252]. 그후모델은계획된행동을그에맞게수정할수있습니다. 예를들어, 사과따기시나리오에서목표사과가약간이동하거나다른사과가시야에들어오면, 모델은장면을동적으로재해석하고잡는궤적을조정합니다. 이능력은

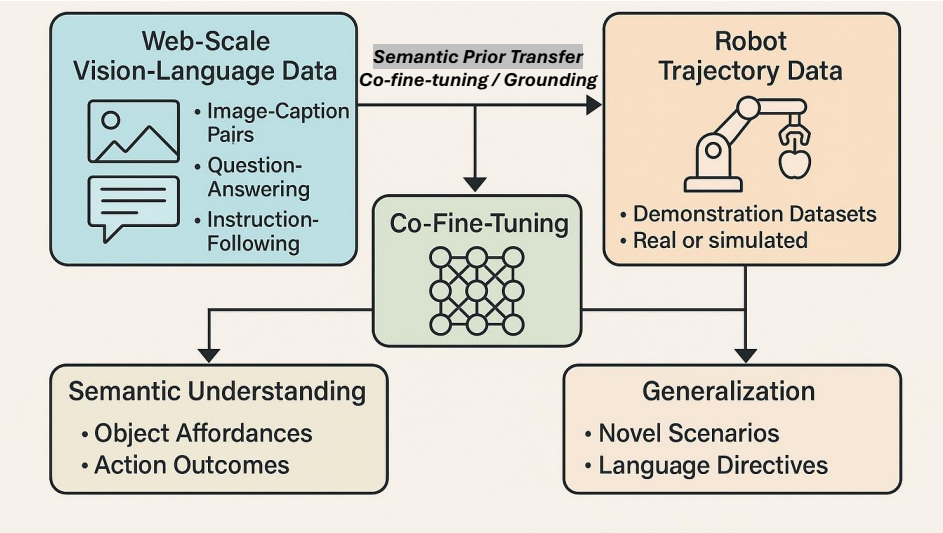

Figure 10: **학습패러다임: VLA 를위한데이터소스와훈련전략.**

인간과유사한적응력을모방하며, 파이프라인기반로봇대비 VLA 시스템의핵심장점입니다.

3. 비전-언어-액션모델의진전

VLA 모델의시작은 2022 년 11 월에출시된 ChatGPT 를비롯한트랜스포머기반 LLM 의놀라운성공에의해촉발되었습니다. ChatGPT 는전례없는의미추론능력을입증했습니다 [179]. 이돌파구는연구자들이언어모델을멀티모달영역으로확장하도록영감을주었으며, 로봇공학을위한지각과행동을통합했습니다. 2023 년까지 GPT-4 는텍스트와이미지를모두처리하는멀티모달기능을도입했고, 이는이후언어중심의멀티모달기초모델을물리적액션표현과제어인터페이스를통합하는노력을촉진했습니다 [1]. 동시에 CLIP(2022) 같은 VLM 도있습니다. [202] 그리고 Flamingo (2022) [3] 대조학습을통해견고한시각-텍스트정렬을확립하여제로샷객체인식을가능하게하고 CLIP 과같은 VLM 모델의기초를마련했습니다. 이모델은대규모라벨데이터셋을활용해이미지와텍스트설명을정렬했으며, 이는액션을통합하는데중요한전조였습니다.

중대한발전중하나는 RT-1 의 13 만개시연과같은대규모로봇데이터셋의생성으로, 시각, 언어, 행동구성요소의공동훈련에필수적인행동기반데이터를제공했습니다 [19]. 이데이터셋은다양한작업과환경을포착하여모델이일반화가능한행동을학습할수있게했습니다. 2023 년구글의 RT-2 로아키텍처적돌파구가이어졌습니다 [18] 이모델은시각, 언어, 액션토큰을통합하여로봇제어를자기회귀적수열예측작업으로다루는획기적인 VLA 모델입니다 (RT-2 블로그 (Source Link)). RT-2 는이산코사인변환 (DCT) 압축과바이트-쌍인코딩 (BPE) 을사용하여이산화동작을수행하며 63% 새로운물체에서의성능향상. 교차주의트랜스포머와같은멀티모달융합기술은 Vision Transformer(ViT) 처리된이미지 (예: 400 패치토큰) 와언어임베딩을통합하여, 로봇이" 그릇왼쪽에있는빨간컵을선택하세요" 와같은복잡한명령을실행할수있게합니다. 또한 UC 버클리의 Octo 모델 (2023) 은 OpenX-Embodiment 데이터셋에서 80 만건의로봇시연을기반으로 93M 매개변수와확산디코더를활용한오픈소스접근법을도입하여연구분야를더욱확장시켰습니다 [218].

3.1. VLA 모델의아키텍처혁신

2023 년부터 2024 년까지 VLA 모델은큰아키텍처발전과훈련방법론의정교화를거쳤습니다. 듀얼시스템아키텍처는핵심혁신으로부상했으며, 이는 NVIDIA 의 GR00T N1(2025) 에서대표적입니다 [14] 이프로그램은시스템 1(저수준제어를위한 10ms 지연시빠른 Diffusion Policy) 과시스템 2(고수준작업분해를위한 LLM 기반플래너) 를결합했습니다. 이러한분리는전략적계획과실시간실행간의효율적인조정을가능하게하여역동적인환경에서적응력을높였습니다. 스탠포드의 OpenVLA(2024) 같은다른모델도있습니다 [122], 97 만개의실제로봇시연을기반으로학습된 7B 매개변수오픈소스 VLA 를도입했으며, 이중비전인코더 (DINOv2 [164] 그리고 SigLIP [271]) 그리고 Llama 2 언어모델 [223], RT-2-X(55B) 와같은대형모델보다더좋은성능을보입니다 [122]. 교육패러다임은웹규모의시각언어데이터 (예: LAION-5B) 에대한공동파인튜닝을활용하도록진화했습니다 [194] 로봇궤적데이터 (예: RT-X) [227], 의미지식과물리적제약조건을일치시키는데 [194]. UniSim 과같은합성데이터생성도구는견고한학습에필수적인가린객체와같은사진처럼사실적인시나리오를만들어데이터부족문제를해결했습니다 (UniSim [264]). 파라미터효율성은미세조정을위한저랭크적응 (LoRA) 어댑터를통해향상되었습니다 [93] 이기술은완전한재학습없이도메인적응을가능하게하여 GPU 시간을 70 시간줄였습니다%. Physical Intelligence 의 pi 0 모델 (2024) 에서볼수있는확산기반정책의도입 [15] 행동다양성을개선했지만상당한계산자원이필요했습니다. 이러한발전은 VLA 기술을민주화하여협업을촉진하고혁신을가속화했습니다.

최근 VLA 모델은효율성, 모듈성, 견고성을균형있게갖춘세가지주요아키텍처패러다임으로수렴했습니다: 초기융합모델, 이중시스템아키텍처, 그리고자기보정프레임워크입니다. 이각각의혁신은실제로봇시스템에서그라운딩, 일반화, 동작신뢰성이라는구체적인도전과제를해결합니다.

1. 초기융합모델: VLA 접근법의한종류는입력단계에서비전과언어표현을융합한후정책모듈로전달하는데중점을

둡니다. Huang 등의 EF-VLA 모델 [96] 국제학습표현학회 (ICLR 2025) 에서발표된이문서는 CLIP 이확립한표현적정렬을유지함으로써이러한경향을잘보여줍니다 [202]. EF-VLA 는이미지-텍스트쌍을받아들여 CLIP 의동결인코더로인코딩하고, 동작예측전에트랜스포머백본초기에생성된임베딩을융합합니다. 이설계는 CLIP 사전학습과정에서학습된의미적일관성을유지하여과적합을줄이고일반화를향상시킵니다. 특히 EF-VLA 는구성조작과제에서 20% 의성능향상을기록했으며, 이전에본적없는목표설명에대해서도 85% 의성공률을달성했습니다. 시각언어중추를고정함으로써계산효율성을유지하고치명적인망각을방지하며, 도메인별훈련은경량정책또는액션모듈에한정되어모델의범용시각의미표현을희생하지않고작업적응을가능하게합니다.

2. 이중시스템아키텍처: 인간인지의이중과정이론에서영감을받은 NVIDIA 의 GR00T N1(2025) 모델이있습니다 [14] 두가지보완적인하위시스템을구현합니다: 빠른반응모듈 (시스템 1) 과느린추론계획기 (시스템 2). 시스템 1 은 10ms 지연시간으로동작하는확산기반제어정책을포함하며, 종단효과기안정화나적응형그래스핑과같은미세하고저수준제어에이상적입니다. 반면 System 2 는작업계획, 기술구성, 고수준순서작성에 LLM 을사용합니다. 플래너는장기목표 (예: " 테이블정리") 를원자하위작업으로분석하며, 저수준컨트롤러는실시간실행을보장합니다. 이분해는다시간척도의추론과향상된안전성을가능하게하며, 특히신속한반응과숙고가공존해야하는환경에서더욱그렇습니다. 다단계가정조작에대한벤치마크테스트에서 GR00T N1 은 RT-1, RT-2, OpenVLA 와같은모놀리식모델보다성공률에서 17% 더높은성능을보였으며, 충돌실패율을 28% 감소시켰습니다.

3. 자기교정프레임워크: 세번째아키텍처진화는기존의추론파이프라인에명시적인장애감지및복구메커니즘을추가하는자가수정 VLA 모델의등장입니다. SC-VLA(2024) 는이전의종단간또는계층적 VLA 설계와유사한표준빠른추론경로를유지하지만, 실행실패나불일치가감지될때선택적으로재평가하고복구조치를생성하기위해선택적으로활성화되는느린수정경로를추가로도입합니다. 이프레임워크에서는경량트랜스포머를사용해퓨즈드임베딩에서직접포즈나동작을예측하는것이기본동작입니다. 실패가감지되면, 예를들어실패한잡기나장애물충돌이발생하면, 모델은사고연쇄추론을수행하는 2 차과정을호출합니다 [281, 270]. 이경로는내부 LLM(또는외부전문가시스템) 에쿼리하여실패모드를진단하고수정전략을생성합니다 [59]. 예를들어, 로봇이가려진물체를반복적으로잘못식별하면, LLM 은능동적시점변경이나그리퍼재배치를제안할수있습니다. 폐루프실험에서 SC-VLA 는작업실패율을 35

4. VLA 모델의아키텍처디자인공간 VLA 모델은풍부한아키텍처설계와기능강조점을보여주며, 이는종단간파이프라인과모듈식파이프라인, 계층적또는평면정책구조, 저수준통제와고수준계획간의균형차원을따라체계적으로조직할수있습니다 (표참조) 1). CLIPort 와같은종단간 VLA [202], RT-1 [19], 그리고 OpenVLA [122] 원시감각입력을단일통합네트워크를통해직접모터명령으로처리합니다. 반면, VLATest 와같은컴포넌트중심모델은 [237] 그리고 Chain-of-Affordance [129] 지각, 언어기반, 행동모듈을분리하여개별하위모듈에서목표개선을가능하게합니다.

계층적아키텍처는전략적의사결정과반응적통제를분리하여복잡하고장기적인과제를해결하기위해등장했습니다. 예를들어, CogACT [131] 그리고 NaVILA [38] LLM 기반계획자가하위목표를저수준컨트롤러에게발행하는 2 단계계층구조를사용하여시스템 2 의추론과시스템 1 실행의강점을결합합니다. 마찬가지로, ORION 도 [69] 장기맥락집계를위한 QT-포머와생성궤적계획기를통합하여응집된프레임워크를만듭니다.

저수준정책강조는확산기반컨트롤러 (예: Pi-0 [15], DexGraspVLA [291]) 는부드럽고다양한모션분포를생성하는데뛰어나지만, 종종더많은계산비용이발생합니다. 반면, 고수준플래너 (예: FAST Pi-0 Fast [171], CoVLA [5] 빠른하위목표생성또는거친궤적예측에집중하며, 세밀한제어는전문모듈이나전통적인모션플래너에위임합니다. HybridVLA 와같은종단간듀얼시스템모델 [142] 그리고 Helix [217] 모듈해석가능성을유지하면서두구성요소를공동학습시켜이러한구분을흐리게합니다.

표 1 최근 VLA 이이러한균형을어떻게균형있게관리하는지도강조합니다. OpenDriveVLA 와같은모델 [293] 그리고 CombatVLA [34] 동적이고안전에중요한영역에서는계층적계획을우선시하는반면, Edge VLA 와같은경량의엣지타겟시스템은선호합니다 [20] 그리고 TinyVLA [242] 실시간저수준정책을강조하고고차원추론을희생하세요. 이분류프레임워크는 VLA 의설계공간을명확히할뿐만아니라, 임베디드배포에최적화된완전한종단투엔드계층모델과같은충분히탐구되지않은조합을정확히찾아내어로봇공학, 자율주행등다양한분야에서 VLA 시스템의기능과적용가능성을발전시킬것을약속함으로써향후개발을안내합니다.

표의분류 1 이는다양한 VLA 아키텍처를비교할수있는명확한틀을제공하고, 종단간통합과계층적분해같은설계선택이작업수행, 확장성, 적응성에어떤영향을미치는지강조하기때문에중요합니다. 모델을저수준정책실행과고수준계획등차원별로분류함으로써, 연구자들은기존접근법의강점과한계를보다명확히파악하고아키텍처혁신의기회를발견할수있습니다. 예를들어, 고속과일수확이나정밀살포와같은농업로봇작업은빠르고반응적인저수준제어기를강조하는아키텍처의이점을누리는반면, 과수원내비게이션, 다중행커버리지계획, 장기수작물모니터링과같은응용분야는더강력한고수준계획및추론능력이필요합니다. 따라서이분류법은특정사용사례에적합한 VLA 아키텍처를선택하는데도움을주며, 응답성과인지계획의균형을맞추는하이브리드시스템개발을이끌어궁극적으로구현 AI 의발전을가속화합니다.

또한, 최근 VLA 모델의발전을종합하기위해표 2 2022 년부터 2025 년까지개발된주목할만한시스템들을요약합니다. 초기융합, 이중시스템처리, 자기보정피드백루프와같은아키텍처혁신을바탕으로다양한설계철학과훈련전략을통합합니다. 각테이블항목은모델의아키텍처구성요소—비전인코더, 언어인코더, 액션디코더와함께모델기능을기반으로평가하는데사용되는주요학습데이터셋—을명시적으로열거합니다. 예를들어 *CLIPort* [202] 그리고 *RT-2* [299] 의미임베딩을액션정책과일치시켜초기기초를마련했으며, 최근의프레임워크는 *Pi-Zero*, *CogACT* [131], 그리고 *GR00T N1* [14] 확산기반또는고주파컨트롤러를통한확장가능한아키텍처를도입합니다. 여러모델은인터넷규모의시각언어말뭉치와로봇궤적데이터셋을활용한멀티모달사전학습을활용하여일반화와제로샷능력을향상시킵니다 [297, 291, 289, 257]. 이표비교는실제및시뮬레이션환경에서 VLA 설계의기능다양성, 도메인적용가능성, 그리고새로운트렌드를이해하려는연구자들의기준점이됩니다.

Table 1: 아키텍처패러다임과과학적우선순위에기반한구조적분류를보여주는 VLA 모델의분류학. 우리는모델을종단간실행, 계층적계획-통제분해, 구성요소중심모듈화를지원하는점과, 저수준운동정책과고수준작업계획자중하나로차별화합니다.

모델명	연도	종단간	계층형	구성요소중심	저수준정책	고수준플래너
CLIPort [202]	2022	✓	✗	✗	✓	✗
RT-1 [19]	2022	✓	✗	✗	✓	✗
Gato [181]	2022	✓	✗	✗	✓	✗
VIMA [112]	2022	✓	✗	✗	✓	✗
Diffusion Policy [40]	2023	✓	✗	✗	✓	✗
ACT [287]	2023	✓	✗	✗	✓	✗
VoxPoser [100]	2023	✓	✗	✗	✓	✗
Seer [80]	2023	✓	✗	✗	✓	✗
Octo [218]	2024	✓	✗	✗	✓	✗
OpenVLA [122]	2024	✓	✗	✗	✓	✗
CogACT [131]	2024	✗	✓	✗	✓	✓
VLATest [237]	2024	✗	✗	✓	✗	✗
NaVILA [38]	2024	✗	✓	✗	✓	✓
RoboNurse-VLA [132]	2024	✓	✗	✗	✓	✗
Mobility VLA [42]	2024	✗	✓	✗	✓	✓
RevLA [48]	2024	✗	✗	✓	✗	✗
Uni-NaVid [280]	2024	✗	✓	✗	✓	✓
RDT-1B [144]	2024	✓	✗	✗	✓	✗
RoboMamba [143]	2024	✓	✗	✗	✓	✗
Chain-of-Affordance [129]	2024	✗	✗	✓	✗	✗
Edge VLA [20]	2024	✗	✗	✓	✗	✗
ShowUI-2B [139]	2024	✓	✗	✗	✓	✗
Pi-0 [15]	2024	✓	✗	✗	✓	✗
FAST (Pi-0 Fast) [171]	2025	✗	✗	✓	✓	✗
OpenVLA-OFT [121]	2025	✓	✗	✗	✓	✗
CoVLA [5]	2025	✗	✓	✗	✓	✓
OpenDriveVLA [293]	2025	✗	✓	✗	✓	✓
ORION [69]	2025	✗	✓	✗	✓	✓
UAV-VLA [191]	2025	✗	✓	✗	✓	✓
CombatVLA [34]	2025	✓	✗	✗	✓	✗
HybridVLA [142]	2025	✗	✓	✗	✓	✓
NORA [103]	2025	✓	✗	✗	✓	✗
SpatialVLA [175]	2025	✗	✗	✓	✓	✗
MoLe-VLA [283]	2025	✗	✗	✓	✓	✗
JARVIS-VLA [130]	2025	✓	✗	✗	✓	✗
UP-VLA [279]	2025	✓	✗	✗	✓	✗
Shake-VLA [120]	2025	✗	✗	✓	✓	✗
DexGraspVLA [291]	2025	✗	✓	✗	✓	✓
DexVLA [241]	2025	✗	✓	✗	✓	✓
Humanoid-VLA [53]	2025	✓	✗	✗	✓	✗
ObjectVLA [297]	2025	✓	✗	✗	✓	✗
Long-VLA [64]	2025	✓	✗	✗	✓	✗
RetoVLA [123]	2025	✗	✗	✓	✓	✗
Vlaser [256]	2025	✗	✓	✗	✓	✓
Discrete Diffusion VLA [138]	2025	✓	✗	✗	✓	✗
Being-H0 [152]	2025	✓	✗	✗	✓	✗
EgoVLA [258]	2025	✓	✗	✗	✓	✗
StereoVLA [47]	2025	✓	✗	✗	✓	✗
GeoVLA [212]	2025	✗	✓	✗	✓	✓
EfficientVLA [262]	2025	✗	✗	✓	✓	✗

Table 2: **대표적인비전-언어-액션 (VLA) 모델의간결한요약.** 각행은주요인코더/디코더, 훈련데이터, 그리고주요고유기능을보고합니다.

모델 (참고문헌)	아키텍처 (비전 / 언어 / 액션)	학습데이터	핵심강점 / 차별성
CLIPort [202]	CLIP-ResNet50 + Transporter-ResNet / CLIP-GPT / LingUNet	Self-collected [SC]	의미적 CLIP 특징을트랜스포터공간추론과정렬하여정밀한 SE(2) 조작을지원합니다.
RT-1 [19]	EfficientNet / Universal Sentence Encoder / Transformer (이산화된동작)	RT-1-Kitchen [SC]	토큰화된행동을이용한다중작업주방조작을위한초기대규모트랜스포머정책.
RT-2 [299]	ViT-22B 또는 ViT-4B / PaLI-X 또는 PaLM-E / 심볼튜닝 (액션토큰)	VQA + RT-1-Kitchen	인터넷규모의 VQA 를로봇데이터와함께공동조정하여구현된작업에대한창발적일반화를가능하게합니다.
Gato [181]	ViT / SentencePiece / Transformer (통합토큰스트림)	Self-collected [SC]	공유토큰화와단일트랜스포머를통해로봇공학, 언어, 아타리를통합하는범용에이전트.
VIMA [112]	ViT + Mask R-CNN / T5 / 트랜스포머	VIMA-Data [SC]	여러구성작업유형 (6 가지프롬프트양식) 에서의프롬프트기반 VL 그라운딩.
ACT [287]	ResNet-18 / — / CVAE-트랜스포머	ALOHA [SC]	시간앙상블은세밀한제어로부드러운양손모방을가능하게합니다.
Octo [218]	CNN / T5-베이스 / 확산트랜스포머	Open X-Embodiment (OXE)	대규모다중로봇정책은 4M+ 궤적을따라여러로봇구현체에걸쳐훈련되었습니다.
VoxPoser [100]	ViLD + MDETR / GPT-4 / MPC (LLM 가이드 계획)	Zero-shot	작업특화교육없이제약인지모션계획을위한 LLM+VLM 을구성합니다.
Diffusion Policy [40]	ResNet-18 / — / U-Net 또는확산트랜스포머	Self-collected [SC]	확산모델링은견고한시각-운동조절을위한멀티모달액션분포를포착합니다.
OpenVLA [122]	DINOv2 + SigLIP / Prismatic-7B / 심볼튜닝	OXE + DROID	오픈소스 RT-2 유사 VLA; 효율적인 LoRA 적응과광범위한일반화를지원합니다.
π^0 (Pi-Zero) [15]	PaliGemma VLM / PaliGemma (멀티모달) / 300M 확산액션모델	Pi-Cross-Embodiment	경량일반로봇컨트롤러 (보고됨) ~ 총 3B 등급) 은강력한로봇간, 오픈월드일반화, 그리고양손능력을갖추고있습니다.
π^0-Fast [171]	PaliGemma VLM / PaliGemma / FAST 토큰화 가적용된자기회귀트랜스포머	Pi-Cross-Embodiment	압축된주파수-공간동작토큰을통한고주파실시간제어 (최대 15× 더빠른추론).
OpenVLA-OFT [121]	SigLIP + DINOv2 (멀티뷰) / Llama-2 7B / 병렬디코딩 + 액션청킹 (L1 회귀)	LIBERO; 양손 ALOHA	병렬디코딩과청크작업을통한레시피미세조정; 보고된 수치는 97.1% LIBERO 성공과 26× 고주파양손제어를위한더빠른추론.
RDT-1B [144]	멀티뷰 RGB 인코더 / 트랜스포머언어모듈 / 확산트랜스포머 (통합액션공간)	46 개의데이터셋 (>100 만에피소드) + ALOHA 파인튜닝	1.2B 확산기초모델은강한언어조건화와제로샷전송을통한민첩한양손조작을위한모델입니다.
Helix[1]	시스템 2: 멀티모달추론을위한오픈소스 VLM (7–9 Hz) / 통합의미론 / 시스템 1: 트랜스포머시각-운동정책 (200 Hz, 상체전체)	Figure robot E2E (pixels+language→actions)	휴머노이드중심이중속도 VLA 는실시간고자유도제어, 민첩성, 그리고제로샷일반화를통한다중로봇협력조작을가능하게합니다.
CogACT [131]	DINOv2 ViT-L/14 + SigLIP ViT-So400M/14 / Llama-2 Prismatic-7B / DiT-베이스 (300M 확산)	OXE 부분집합; Realman & Franka 과제	확산액션트랜스포머를가진구성요소기반 VLA; 보고된 +59.1% OpenVLA 에비해실제성공과보이지않는로봇/물체에대한강한적응.
Chain-of-Affordance (CoA) [129]	어포던스인식시각인코더 / 트랜스포머추론 프롬프트 / 자기회귀 + 확산정책 (어포던스조건부)	LIBERO; Real+Sim 조작	순차적어포던스 (행동유도성) 추론 (객체 → 파지 → 공간 → 동작은공간계획과장애물회피를향상시킵니다; OpenVLA 보다 LIBERO 성과가더강하다고보고했습니다.
Edge VLA (EVLA) [20]	SigLIP + DINOv2 / Qwen2 (0.5B) / 비자기회귀적공동제어예측	Bridge; OXE; 120 만개의텍스트-이미지쌍	엣지최적화 VLA(예: Jetson 급) 는 30-50 Hz 추론과 OpenVLA 와비교할만한저력성능.
ShowUI-2B [139]	UI 가이드시각적토큰선택 / 인터리브 V–L–A 스트리밍 / 트랜스포머 GUI 동작예측기	256K GUI 명령어팔로잉	디지털자동화를위한컴팩트 2B VLA; 강력한스크린샷그라운딩과 GUI/웹내비게이션, 효율적인토큰선택.
GR00T N1 [14]	NVIDIA Eagle-2 VLM / 통합고수준계획 / 확산트랜스포머 (DiT)	인간시연 + 로봇궤적 + 시뮬레이션 + 인터넷영상	계획과확산실행을결합한다단계제어와광범위한신체간일반화를위한범용휴머노이드이중시스템설계.
Seer [80]	그라운딩최적화시각백본 / 트랜스포머언어 / 자기회귀동작헤드	LIBERO	조작을위한강한시각적기반; LIBERO 에서는경쟁력이있으나, 일반적으로 OpenVLA-OFT 와같은최신파인튜닝변형모델보다는낮은성능을보입니다.
DiffusionVLA [240]	Transformer visual encoder / autoregressive reasoning / diffusion action head	LIBERO; 공장분류; 제로샷빈피킹	확산제어는견고성과해석가능성을향상시킵니다; 일부구성에서 CoA 보다공간일반화가약하다고보고되었습니다.
NaVILA [38]	CLIP + CNN / LLaMA-2 / 계층적제어: 위상계획 + RL 이동	실제다리로봇내비게이션데모	지형일반화를위한모듈식계층구조; 보고된 88% 자연어를통한실제내비게이션성공.
RoboNurse-VLA [132]	SAM2 + RGB-D / LLaMA-2 + 음성-텍스트변환/포즈회귀 + 그리퍼분류기	수술인수인계동영상 + 음성안내	실시간수술도구인수인계는새로운도구와역동적인수술실장면에강력하게대응합니다.

다음페이지에서계속

[1] https://www.figure.ai/news/helix

이전페이지에서계속

모델 (참고문헌)	아키텍처 (비전 / 언어 / 액션)	학습데이터	핵심강점 / 차별성
Mobility VLA [42]	롱컨텍스트 ViT + 목표이미지인코더 / T5 기반명령어인코더 / 그래프플래너 + 시각적목표현지화	MINT 데이터셋 (VL 명령어투어)	멀티모달투어에서얻은위상지도는넓고보이지않는공간에대한내비게이션일반화를가능하게합니다.
TinyVLA [242]	FastViT / 컴팩트 128D 언어 / 확산디코더 (50M)	Mini-ALOHA + SC tasks	대규모사전훈련이필요하지않습니다; 더빠른추론보고 (5×) 이함수는제한된컴퓨트예산내에서강력한정밀도로 이루어집니다.
QUAR-VLA [54]	CLIP + 고유수용성임베딩 / BERT + 그라운딩어댑터 / 트랜스포머전신디코더	QUART (이동 + 조작)	강력한시뮬레이션-실제 (Sim-to-Real) 전이과세밀한명령어정렬을가진쿼드루프중심 VLA 입니다.
ChatVLA [295]	위상정렬비전인코더 / Prismatic MoE LLM / 통합 V–L–A 플래너	통합채팅액션 (웹 + 로봇)	공동 VQA+ 기획; 잊어버리는것을줄이고조작을통해대화형작업실행을지원합니다.
PointVLA [125]	CLIP + 3D 포인트클라우드융합 / LLaMA-2 / 공간토큰융합을이용한트랜스포머	few-shot 공간작업 (실제 + 시뮬레이션)	사전학습된 2D 지식을유지하면서 3D 구조를주입하여 장기과제공간추론을향상시킵니다.
VLA-Cache [252]	SigLIP + 토큰메모리버퍼 / Prismatic-7B / 동적토큰재사용트랜스포머	ALOHA + sim/real 퓨전	효율성을위해정적시각적토큰을캐시합니다; 보고된 40–50% 성능저하가거의없는빠른추론.
HybridVLA [142]	CLIP + DINOv2 / LLaMA-2 / 하이브리드확산 + 자기회귀앙상블	RT-X + synthetic fusion	동적앙상블은다중암환경에서견고성을높이고더강력한 sim-to-real 일반화를지원합니다.
MoLe-VLA [283]	다단계 ViT + STAR 라우터 / CogKD 강화트랜스포머 / 희소트랜스포머 (동적라우팅)	RLBench + 실제조작	선택적층활성화는효율을제공합니다 (보고된 5.6× 속도증가) 및더높은성공률 (보고된 +8)%).
UAV-VLA [191]	항공이미지 / GPT 명령어파싱 / 트랜스포머경로계획용 ViT	위성 + UAV 영상지침	대규모미지도환경을위한확장가능한언어기반을갖춘제로샷항공임무계획.
DexGraspVLA [291]	객체중심공간 ViT / 트랜스포머그라스프추론 / 확산그라스프컨트롤러	민첩잡기벤치마크 (시뮬 + 실제)	보도 >90% 다양한물체에서제로샷성공; 조명/배경변화와보이지않는조건에견고합니다.
GraspVLA [46]	멀티뷰 DINOv2 + SigLIP / VLM 이박스 + 그라스프 / 플로우매칭액션전문가 (PAG) 를예측합니다	SynGrasp-1B; GRIT	합성사전학습잡기 VLA; 시뮬레이션-실제 (Sim-to-Real) 전이을개선하고롱테일객체클래스와선호도에대한제로/few-shot 일반화를지원합니다.
Interleave-VLA [62]	InternVL2.5 + OWLv2 / Qwen2.5 / 연속동작예측기 (OpenVLA+π^0-스타일, 확산컨트롤러)	Open Interleaved X-Embodiment (21만에피소드, 11 개데이터셋)	종단간인터리브이미지-텍스트지시가뒤따르는방식; 보고됨 2–3× 스케치와새로운멀티모달프롬프트를통한영역외이득과제로샷실행.
Long-VLA [64]	장기과제작업 / 위상인식입력마스킹 + 트랜스포머정책 (하위작업단계분할) 을위한종단간 VLA	장기과제다단계로봇조작시연 (작업시퀀스)	' 이동' 단계와' 상호작용' 단계를명시적으로분리하여장기작업수행을목표로하여, 하위작업호환성과장기작업범위에걸쳐견고함을향상시킵니다.
RetoVLA [123]	VLM 기반정책/재사용레지스터토큰을공간적맥락으로사용하여행동예측을지원합니다	7 도 DoF 팔에서의실제로봇조작 + 작업별시연	레지스터토큰을재활용하여경량공간추론업그레이드; 최소한의아키텍처오버헤드로복잡한조작에서상당한성공률향상을보고했습니다.
Vlaser [256]	VLM-to-VLA 파이프라인 / 시너지적체성추론 + 정책학습 (추론인지 VLA 미세조정)	Vlaser-6M 내체추론데이터셋 + VLA 미세조정데이터 (로봇시연)	Bridge 는추론과통제를구현했으며, 체화된기반/QA/기획전반에서강력한성과를보여주며, 도메인전환시정책학습으로의전환을개선했습니다.
Discrete Diffusion VLA [138]	단일트랜스포머 VLA / 이산화된액션청크 + 이산확산정제 (CE 교육; 재마스킹)	LIBERO + SimplerEnv(프랙탈/Bridge) 스타일의 VLA 벤치마크	분산스타일의정제를이산토큰인터페이스로통합합니다: 적응형디코딩순서, 리마스킹을통한오류수정, 그리고강한벤치마크성공률로자기회귀병목현상감소.
Being-H0 [152]	인간영상에서사전학습된민첩한 VLA / 명시적인손동작모델링 + 행동을위한 VL 그라운딩	대규모인간조작영상 (자기중심/3인칭) + 로봇제어로의전환	다양한인간비디오데이터를활용하여능숙한조작을확장합니다; 작은텔레옵스전용로봇데이터셋과비교해새로운작업/장면에대한일반화를향상시킵니다.
EgoVLA [258]	자기중심적인간조작 / 통합인간-로봇행동공간 + 로봇미세조정에관한 VLM 사전학습	대규모자기중심적인간영상 + 소규모로봇시연세트	확장가능한사전학습을위해풍부한자기중심적영상을활용한후, 통합행동공간을통해구현체를정렬하여제한된로봇데이터로실용적인로봇전송을가능하게합니다.
StereoVLA [47]	스테레오강화 VLA / 기하학-의미론적융합 스테레오쌍 (+ 보조깊이단서) 을통한행동디코딩	스테레오로봇조작시연 (언어조건 동작이포함된스테레오 RGB)	스테레오기하학을명시적으로활용하여공간정밀도 (깊이민감잡기/조작) 와시점/카메라변화에대한견고함을향상시킵니다.
EfficientVLA [262]	훈련없는 VLA 가속/압축/구조화된중복성제거 (VLA 파이프라인전반에걸친토큰/인터페이스최적화)	기존 VLA 에적용됨 (표준조작벤치마크평가; 예: SIMPLER 스타일설정)	실용적배포가능성: 대규모 VLA 정책의연산/메모리를최소화하면서정확도저하를최소화하여, 제한된하드웨어에서더근실한추론을가능하게합니다.

3.2. 비전-언어-액션모델에서의훈련효율성발전

VLA 모델은멀티모달입력을조화시키고, 계산요구량을줄이며, 실시간제어를가능하게하는훈련및최적화기법에서빠르게발전했습니다. 주요발전분야는다음과같습니다:

- **데이터효율적인학습.**
  - 코-미세튜닝방대한비전-언어말뭉치 (예: LAION-5B) 와로봇궤적집합 (예: Open X-Embodiment) 에서의미이해를운동능력과일치시킵니다. OpenVLA(7B 파라미터) 는 55B 매개변수 RT-2 변형보다 16.5% 더높은성공률을보이며, 공동파인튜닝이적은매개변수로강력한일반화를가능하게함을보여줍니다 [194, 227, 122] 모델크기확장만비교했을때와비교했을때입니다.
  - 합성데이터생성 Via UniSim 은오클루전과동적조명을포함한사진사실적인장면을생성하여드문예외상황을보강하며, 복잡한환경에서모델의강인성을 20% 이상향상시킵니다 [264, 218].
  - 자기감독사전훈련 CLIP 과같은대조적목적을채택하여동작미세조정전에시각적텍스트임베딩을학습하여작업별라벨에대한의존도를줄입니다. Qwen2-VL 은자체감독정렬을활용하여하위 pick-and-place 과제의수렴을 12% 가속화합니다 [177, 98].
- **매개변수효율적인적응.** 저랭크적응 (LoRA) 은경량어댑터매트릭스를동결된트랜스포머층에삽입하여성능을유지하면서도훈련가능한가중치를최대 70% 까지줄입니다 [93]. Pi-0 Fast 변형은정적백본위에 10M 어댑터매개변수만사용해정확도손실이거의없는 200Hz 연속제어를제공합니다 [171].
- **추론가속.**
  - 압축액션토큰 (빠르게) 그리고 병렬디코딩이중시스템프레임워크 (예: GR00T N1) 에서는최대 2.5 를달성합니다.× 정책추론속도증가, 단계당지연시간을 5 미만으로줄입니다 실시간조작및휴머노이드제어벤치마크를기준으로평가한단일정책 VLA 에서의표준자기회귀디코딩과 MS 의비교. 이가속은궤적의부드러움에약간의대가를치르며, 액션이산화오차가약간더크고고주파제어하에서미세한운동연속성이감소하는방식으로나타납니다 [14, 209].

이방법들은 VLA 를언어조건화되고시각유도작업을역동적이고현실세계에서처리할수있는실용적인에이전트로탈바꿈시켰습니다.

3.3. VLA 모델에서의매개변수효율적인방법과가속기법

데이터효율적인학습전략을넘어, VLA 모델에대한보완적인연구는모델크기, 메모리사용량, 추론지연을줄여제한된계산및전력자원을가진실제로봇플랫폼에배포할수있도록하는데중점을두고있습니다. 앞서논의한훈련중심효율성방법과달리, 이하위섹션의기법들은주로대상 적응시점의매개변수효율성그리고 정책추론중런타임가속.

1. **저랭크모듈을통한매개변수효율적인적응:** 완전한미세조정대신, 많은 VLA 는훈련효율성맥락에서이전에도입된 Low-Rank Adaptation(LoRA) 과같은매개변수효율적인적응메커니즘을채택합니다. 이절에서는그역할에중점을둡니다 배포중유효매개변수발자국감소. 예를들어, OpenVLA 는경량 LoRA 어댑터 (약 20 개) 를사용합니다. M 파라미터) 위에얼어붙은 7 위에있습니다 B-파라미터백본으로, 최소한의메모리오버헤드로작업적응을가능하게하고작업간전체모델가중치의중복을방지합니다 [93, 122]. 이설계는자원이제한된시스템에서여러작업특화된정책이공존하면서도사전학습과정에서학습된범용시각적·의미론적표현을유지할수있게합니다.
2. **엣지배포를위한양자화:** 모델양자화는추론처리량과메모리효율성을개선하기위해수치정밀도를낮춥니다. OpenVLA 실험결과, NVIDIA Jetson Orin 과같은임베디드플랫폼에서의 INT8 양자화는 pick-and-place 벤치마크에서완전정밀도대비약 97% 의작업성공률을보존했으며완전한정밀도로작업성공을거두었으며, 세밀하고민첩한조작에서는약간의저하만있었습니다 [194, 122]. 학습후양자화와채널별보정은고동적범위센서입력시정확도손실을더욱완화합니다 [164]. 이러한최적화는최대 30 개의지속제어주파수를가능하게합니다 이동식로봇의엄격한전력예산내에서 Hz 를사용합니다.
3. **모델가지치기와아키텍처슬림:** 구조적가지치기는어텐션헤드나피드포워드서브레이어와같은중복된아키텍처구성요소를제거하여메모리및컴퓨팅요구량을줄입니다. VLA 에서는독립적인시각또는언어모델에비해덜탐구되지만, 확산기반시력운동정책에관한초기연구들은합성곱시각인코더를최대 20% 까지줄여도잡기안정성저하가거의없음을보여줍니다 [40]. 트랜스포머기반 VLA 에적용되는유사한가지치기전략 (예: RDT-1B) 은메모리사용량을약 25% 줄이면서도과제성공률감소를 2% 미만으로유지할수있습니다, Sub-4 허용 GB 배포 [144, 131].
4. **압축액션토큰화:** 장기과제제어시퀀스에서발생하는추론병목현상을해결하기위해압축액션표현이제안되었습니다. FAST 는연속동작경로를컴팩트한주파수영역토큰으로재구성하여디코딩길이를크게단축합니다. Pi-0 Fast 변형은최대 15 개까지달성할수있습니다 × 1000 을압축하여더빠른추론 MS 액션창은 16 개의개별토큰으로나뉘어최대 200 의제어속도를지원합니다 데스크톱 GPU 에서의 Hz [171]. 이방법은최소한의궤적세분성을대가로대가로큰속도향상을제공하여, 고주파의반응적인조작작업에적합합니다.
5. **병렬디코딩과액션청킹:** 표준자기회귀 VLA 는동작을순차적으로디코딩하여누적지연시간을발생시킵니다. GR00T N1 과같은이중시스템아키텍처에서사용되는병렬디코딩전략은시공간액션토큰그룹을동시에생성하여약 2.5 개의액션토큰을달성합니다.× 100 도에서작동하는 7-DoF 로봇팔의종단간추론지연감소 Hz [14, 209]. 액션청킹은다단계루틴 (예: 픽앤플레이스) 단일고수준토큰으로변환하여추론단계를최대 40% 까지줄입니다 주방워크플로우와같은장기과제조작작업에서 [112].
6. **하드웨어인식컴파일및런타임최적화:** 마지막으로, 하드웨어인지최적화는컴파일러수준의그래프재작성, 커널융합, 가속기특화원시요소를활용하여처리량을극대화합니다. TensorRT-LLM 과같은프레임워크는텐서코

어, 퓨즈드어텐셔닝커널, 파이프라인메모리전송을활용하여트랜스포머추론과확산샘플링을모두가속화합니다. OpenVLA-OFT 에서는이러한최적화가추론지연시간을약 30% 줄이고추론당에너지소비를 25% 줄입니다 RTX 급 GPU 와표준 PyTorch 실행비교 [121]. 이러한시스템수준의최적화는엄격한전력제약이있는이동식로봇, 항공플랫폼, 휴머노이드시스템에서실시간 VLA 를배포하는데매우중요합니다.

토론: 매개변수효율적인적응및추론가속기법은 VLA 배포를함께민주화합니다:

- LoRA 와양자화는소규모연구소와연구개발프로그램이소비자급하드웨어에서수십억파라미터의 VLA 를미세조정하고작동할수있게하여로봇의최첨단의미론적이해를열어줍니다 [93, 122].
- 가지치기와 FAST 토큰화는모델및액션표현을압축하여 4GB 이하, 5ms 미만의제어루프를가능하게하며, 민첩한작업에서정밀도를희생하지않습니다 [144, 171].
- 병렬디코딩과액션청킹은자기회귀정책의순차적병목현상을극복하여민첩한조작과다리이동에필요한 100–200 Hz 의사결정률을지원합니다 [14, 209].
- 하이브리드 RL-SL 훈련은복잡한환경에서탐색을안정화하며, 하드웨어인식컴파일은엣지가속기에서실시간성능을보장합니다 [159, 121].

이러한발전들은산업용조작기, 보조드론, 소비자로봇전반에 VLA 모델을내장하는것이실용적이어서, 연구프로토타입에서실제자율성으로의간극을메우는데기여하고있습니다.

3.4. 비전-언어-액션모델의응용

VLA 모델은지각, 자연어이해, 운동제어를통합한통합아키텍처내에서체화된지능의기초구성요소로빠르게부상하고있습니다. 시각적및언어적양식을공유된의미공간에인코딩하고맥락에기반한행동을생성함으로써, VLA 모델은에이전트와그환경간의원활한상호작용을가능하게합니다 [131, 293]. 시각적및언어적양상을공유된의미공간에인코딩하고맥락에기반한행동을생성함으로써, VLA 모델은에이전트와그환경간의원활한상호작용을가능하게합니다. 이러한멀티모달역량은 VLA 를다양한실제응용분야에서변혁적에이전트로자리매김하게했습니다. 휴머노이드로봇공학에서는 Helix 와 RoboNurse-VLA 와같은시스템이시각, 언어, 그리고능숙한조작을결합하여가정업무와외과수술을지원하며, 실시간추론과안전인식제어를보여줍니다 [132, 242]. 자율주행차에서는 OpenDriveVLA 와 ORION 과같은모델이동적시각스트림과자연어지시를처리하여복잡한도시환경에서투명하고적응적인운전결정을내립니다 [69, 293]. 산업배포는고정밀조립, 검사및협업제조를위해 VLA 아키텍처를활용합니다 [131]. 농업분야에서는 VLA 기반로봇시스템이시각유도과일수확, 식물모니터링, 이상탐지를가능하게하여노동의존도를줄이고지속가능성을높일수있습니다. 더나아가, 최근인터랙티브증강현실시스템의발전은 VLA 모델을활용해실시간언어조건공간내비게이션을제공하며, 음성이나시각적단서를바탕으로실내외환경에서사용자를안내합니다 [191, 74]. 이들영역전반에걸쳐 VLA 는견고하고적응력있으며의미적으로정렬된작업실행을위한통합프레임워크를제공하며, 이는체화된범용에이전트로의중대한전환을의미합니다.

표 3 최근 VLA 모델의방법론, 응용분야, 주요혁신을요약하여보여줍니다.

다음하위섹션들은그림에나타난대로적용분야를시간순서대로심층적으로탐구합니다 11.

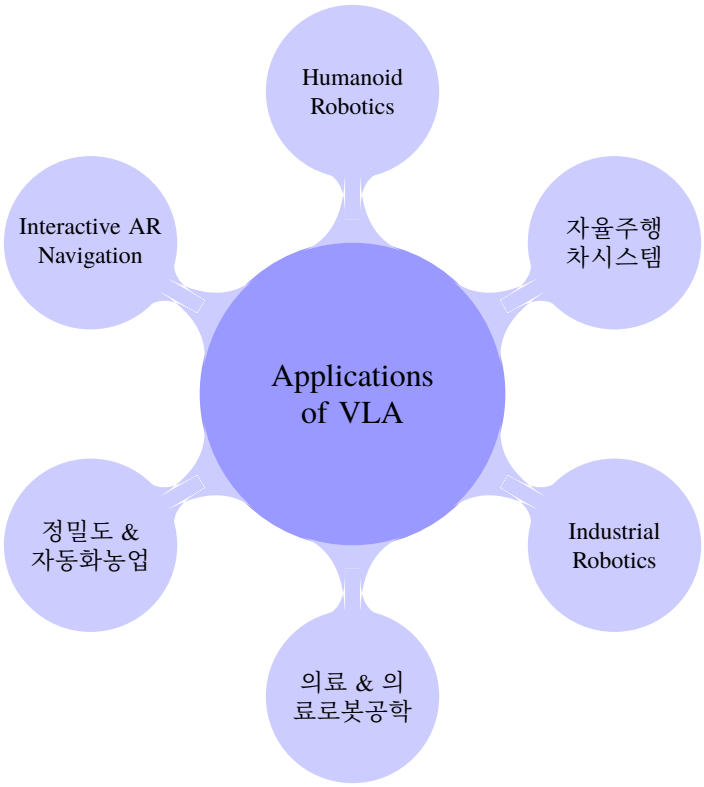

Figure 11: 비전-언어-액션모델의응용영역을마인드맵으로구성하며, 휴머노이드로봇공학이맨위에위치하고나머지영역들은이섹션의논의순서에맞춰시계방향으로배열되어있습니다.

3.4.1. 휴머노이드로봇

인간의신체형태와기능을모방하도록설계된휴머노이드로봇은 VLA 모델배포에있어가장까다롭지만영향력있는분야중하나를대표합니다. 이플랫폼들은복잡한환경을원활하게인지하고, 말하거나쓰여진자연어를이해하며, 인간수준의민첩성으로복잡한신체작업을수행해야합니다 [184, 23]. VLA 모델의핵심강점은지각, 인지, 제어를단일종단간훈련가능한프레임워크로통합할수있다는점에있습니다. 이를통해휴머노이드로봇이시각적입력 (예: 복잡한장면의 RGB-D 이미지) 을해석하고, 언어적지시 (예: " 숟가락을서랍에넣기") 를이해하며, 정밀한운동경로를생성할수있습니다 [151, 296].

최근의발전은휴머노이드로봇분야에서 VLA 의도입을크게가속화했습니다. 예를들어, Helix(Helix)[2]Figure AI 가개발한휴머노이드로봇은완전통합 VLA 모델을활용해전신조작을고주파로수행하며, 팔, 손, 몸통, 심지어미세한손가락움직

[2]https://www.figure.ai/news/helix

Table 3: 대표적인비전-언어-액션 (VLA) 방법론, 적용분야및주요혁신비교.

참고문헌 (연도)	VLA 방법론	응용분야	강점 / 핵심혁신
CogACT [131] (2024)	확산트랜스포머를기반으로한특수액션모듈을갖춘모듈형 VLA 입니다.	산업로봇공학; 언어유도조작.	빠른적응과강력한일반화를갖춘견고한행동모델링을통해다양한산업환경에서과제성공률을향상시킵니다.
VLATest [237] (2024)	VLA 조작평가를위한자동화된대규모테스트프레임워크.	로봇조작벤치마킹 (견고성/신뢰성).	체계적인다중장면및다중작업평가를통해강인성격차를드러내고목표 VLA 개선을지원합니다.
NaVILA [38] (2024)	2 단계 VLA: 고수준시각언어는중간수준의항법명령을생성합니다; RL 로코모션은실행됩니다.	어지러운현실세계장면에서자연어를통한다리내비게이션.	모듈식고·저수준분할방식으로, 강한일반화와도전적인지형에서의높은실제성공률을갖추고있습니다.
RoboNurse-VLA [132] (2024)	실시간음성투액션파이프라인을지원하는비전모듈 (SAM2) + 언어모듈 (Llama2).	수술보조 (기구잡기및인수).	정확한실시간지원, 보이지않는도구에대한견고함, 그리고동적인수술실조건에서의안정적인성능.
Mobility VLA [42] (2024)	목표위치및위상그래프탐색을위한긴컨텍스트 VLM 을가진계층적 VLA 입니다.	시연투어가포함된멀티모달교수법내비게이션 (MINT).	대규모공간에서의확장가능한내비게이션을위한시연을활용합니다; 새로운언어 + 이미지쿼리에견고합니다.
CoVLA [5] (2025)	CLIP 기반비전, Llama-2 언어, 행동을위한궤적예측.	자율주행 (데이터셋 + VLA 훈련).	해석가능한장면이해와견고한경로계획을가능하게하는대규모풍부하게주석이달린데이터셋.
OpenDriveVLA [293] (2025)	2D/3D 시각토큰과언어임베딩의계층적정렬; 자기회귀에이전트-환경-자아모델링.	종단간자율주행.	복잡한교통현장에서계획및질의응답성능을향상시키는통합의미공간및상호작용모델링.
ORION [69] (2025)	QT-Former 는과거문맥, LLM 추론, 생성계획기획기 (generative planner) 를사용해과거문맥, LLM 추론, 궤적예측을위한도구입니다.	종합적인자율주행시스템.	추론및행동공간을 VQA 및계획의통합최적화와일치시켜더강력한폐쇄루프성능을제공합니다.
QUAR-VLA [54] (2025)	QUART 기반시각과언어의융합을통해액션생성을지원합니다.	네발달린내비게이션, 조작, 전신작업.	레그드플랫폼에대한 VL-A 결합과명령어정렬, 강력한시뮬레이션-실물일반화.
TinyVLA [242] (2025)	Diffusion Policy 디코더가포함된컴팩트멀티모달백본.	빠르고데이터효율적인조작제어.	강한일반화로추론비용감소; 많은사전훈련없이효율성을향상시킵니다.
UAV-VLA [191] (2025)	모듈식파이프라인: 목표추출을위한 GPT, 객체탐색을위한 VLM, 액션생성을위한 GPT.	언어 + 위성이미지를기반으로한 UAV 임무계획.	최소한의사전훈련으로효율적인비행및작전계획; 직관적인인간-UAV 상호작용및벤치마킹을지원합니다.
Bi-VLA [74] (2025)	시각, 언어, 양손행동모듈을연결하는멀티모달트랜스포머.	양손가정조작.	명시적인액션-모듈통합을통한높은적응력과견고한실제양손실행.
ChatVLA [295] (2025)	VL-A 통합을위한 Mixture-of-Experts 와의단계별정렬교육.	통합된멀티모달이해와로봇제어.	잊어버림/간결을줄여 VQA 와조작모두를효율적인전문화로향상시킵니다.
RoboMamba [143] (2025)	Mamba/SSM 기반 VLA 와공동학습된비전인코더와 SE(3) 동작모델링을포함합니다.	효율적인로봇추론과조작.	선형복잡도추론과최소한의미세조정은빠른포즈인식추론과효율적인제어를가능하게합니다.
OTTER [97] (2025)	고정된사전학습된 VLM 을이용한텍스트인식특징추출.	제로샷일반화를이용한조작.	VLM 의의미적정렬을 VLM 의미세조정없이도과제관련특징선택을통해강력한제로샷전송을보장합니다.
PointVLA [125] (2025)	모듈식스킵블록을통해동결된 VLA 에 3D 포인트클라우드특징을주입합니다.	공간적추론; 소수의샷과장거리지평선조작.	2D 지식을유지하면서 3D 기하학을추가하여완전한재훈련없이도공간적기반을개선합니다.
VLA-Cache [252] (2025)	정적시각토큰을선택적으로재사용하는적응형토큰캐싱.	실시간효율적인조작추론.	레이어별토큰재사용은정확도손실을최소화하면서큰속도향상을제공하며, 로봇내배포에실용적입니다.
CombatVLA [34] (2025)	빠른추론을위한단축된 AoT 를적용한비디오액션 AoT 훈련.	3D 게임에서의실시간전투의사결정.	대규모추론가속과향상된전술적추론, 실시간상호작용환경에서의높은성공률.
HybridVLA [142] (2025)	협업확산및자기회귀액션정책이포함된통합 LLM.	시뮬레이션과실제작업모두에서단일또는이중암조작.	적응행동앙상블은복잡한조작에대한견고성과일반화를향상시킵니다.
NORA [103] (2025)	Qwen-2.5-VL-3B 백본과 FAST+ 토큰라이저를사용하는 3B 매개변수 VLA 입니다.	범용몸체로봇공학 (시뮬레이터 + 실제).	낮은오버헤드와강한시각적추론, 빠른액션디코딩; 더큰 VLA 과경쟁할수있습니다.
SpatialVLA [175] (2025)	공간인식 VLA 를위한 Ego3D 위치인코딩및적응형동작격자.	크로스로봇, 멀티태스킹, 제로샷조작.	명시적 3 차원공간통합과적응적이산화는전이와일반화를개선합니다; 오픈소스입니다.
MoLe-VLA [283] (2025)	라우터와증류를통한동적레이어건너뛰기를통한레이어혼합.	RLBench + 실제로봇에서의효율적인조작.	선택적층활성화는인지를유지하면서성공률을높이면서강력한속도향상을제공합니다.
JARVIS-VLA [130] (2025)	VL 가이던스와키보드/마우스제어용액션헤드가있는대형 VLM 을후처리한모델입니다.	오픈월드비주얼게임 (예: 마인크래프트), 1k+ 과제등이있습니다.	자기지도후교육은강력한일반화와세계지식을제공합니다; 오픈소스작업스위트와모델.
UP-VLA [279] (2025)	공동멀티모달이해와미래예측목표를가진통합 VLA.	정밀한공간추론을통한신체적조작.	공동의미론 + 동역학학습은세밀한공간제어가필요한장기과제의수행능력을향상시킵니다.
Shake-VLA [120] (2025)	비전, 음성변환, RAG, 이상감지, 양손팔을갖춘모듈식스택.	양손칵테일이잡음과잡음을섞는것.	신뢰할수있는재료취급, 레시피적응, 실제작업완료를포함한견고한종단간배포.
MoRE [285] (2025)	LoRA 모듈과 RL 기반 Q-함수훈련을활용한희소한 MoE.	네발로다기능이동, 항법, 조작을모두수행했습니다.	혼합품질데이터에대한확장가능한미세조정과시뮬과실에서강력한다중작업및 OOD 일반화를제공합니다.
DexGraspVLA [291] (2025)	계층적계획기 (사전학습된 VL) + 확산저수준컨트롤러.	다양한조건에서의일반적인손잡이능력.	도메인불변표현은객체, 조명, 배경전반에걸쳐강력한제로샷일반화를지원합니다.
DexVLA [241] (2025)	구현된커리큘럼학습을갖춘플러그인확산행동전문가.	일반로봇조작: 단일팔, 양손, 손재주.	신체간행동모델링과빠른적응은과제별튜닝없이도강력한장기성적성능을가능하게합니다.

임까지실시간으로제어합니다. 아키텍처는이중시스템설계를따르며, 멀티모달트랜스포머는언어명령과비전스트림같은입력을처리하고, 실시간모터정책은 200Hz 의밀집된동작벡터를출력합니다. 이로인해 Helix 는이전에못한객체와작업에걸쳐일반화할수있으며, 작업별재학습없이변화하는환경에유연하게적응할수있습니다.

휴머노이드시스템에서 VLA 의주요장점은공유표현을통해다양한작업에확장할수있다는점입니다 [8]. 작업별프로그래밍이나모듈식파이프라인에의존하는전통적인로봇시스템과달리, VLA 기반휴머노이드는통합토큰기반프레임워크하에서작동합니다. 시각입력은 DINOv2 나 SigLIP 와같은사전학습된시각언어모델로인코딩되며, 명령어는 LLaMA 나 GPT 스타일인코더와같은 LLM 을사용해처리됩니다. 이표현들은장면과작업의전체맥락을포착하는접두사토큰으로융합됩니다. 액션토큰은언어디코딩과유사한자동회귀방식으로생성되지만, 로봇의관절과최종이펙터에대한운동명령을나타냅니다.

이능력덕분에휴머노이드로봇은가정, 병원, 소매환경등인간중심공간에서효과적으로작동할수있습니다. 가정환경에서는 VLA 기반로봇이음성명령을해석하는것만으로도표면을청소하거나간단한식사를준비하거나물건을정리할수있습니다 [151, 296]. 의료분야에서는 RoboNurse-VLA 와같은시스템이있습니다 [132] 실시간음성및시각적신호를이용해외과의사에게정밀한기기인수인계를수행할수있는능력을입증했습니다. 소매업에서는 VLA 가장착된휴머노이드플랫폼이고객문의, 진열대재입고, 매장배치탐색을명시적인사전프로그래밍없이지원할수있습니다 [8].

현대 Humanoid-VLA 를구별하는점은임베디드저전력하드웨어에서동작할수있어실제배포가가능하다는점입니다. 예를들어, TinyVLA 와같은시스템들이있습니다 [242] 그리고 MoManipVLA [246] Jetson 급 GPU 에서실행되는효율적인추론파이프라인을보여주어성능저하없이모바일배포를가능하게합니다. 이모델은확산기반정책, LoRA 기반미세조정, 동적토큰캐싱과같은기법을활용하여계산비용을최소화하면서도높은정밀도와일반화를유지합니다.

물류및제조분야에서는 VLA 지원휴머노이드가이미상업적영향력을발휘하고있습니다. 그림 01 과같은로봇은창고에서인간작업자와함께피킹, 분류, 선반정리와같은반복적이고육체적으로집약적인작업을수행합니다. 새로운객체범주와역동적으로변화하는장면을다루는능력은지속적인학습과견고한멀티모달기반덕분에강화됩니다 [252, 131].

VLA 모델이다양한액션생성, 공간추론, 실시간적응능력을계속발전함에따라, 휴머노이드로봇은가정, 산업환경, 공공장소에서매우유능한조수로부상하고있습니다. 이들의강점은토큰기반아키텍처를통해지각, 언어이해, 운동제어를통합하는능력에있으며, 이를통해비구조화된인간환경에서매끄럽고맥락인식있는행동을가능하게합니다.

예를들어, 그림에묘사된것처럼 12 차세대 VLA 모델을갖춘최첨단휴머노이드로봇'Helix' 를생각해보십시오. " 냉장고에서물병을꺼내주세요" 라는구두지시를받으면 Helix 는통합지각시스템을활성화하며, 기초시각언어모델 (예: SigLIP 또는 DINOv2) 이시각장면을분할하여냉장고, 손잡이, 병을식별합니다. 언어입력은 LLaMA-4 와같은 LLM 에의해처리되며, 명령어를토큰화하고시각적맥락과융합합니다. 이융합된표현은계층적컨트롤러로전달됩니다: 상위정책담당자는작업시퀀스 (손잡이위치, 문당기기, 병식별, 잡기 등) 를계획하고, 중간단계계획자는잡기유형과관절궤적과같은모터원시요소를정의합니다. 저수준 VLA 컨트롤러는종종 Diffusion Policy 네트워크를기반으로하며, 이동작들을초미만의지연시간으로실행합니다. 변이 (예: 기울어진병이나미끄러운그립) 를만나면, Helix 의에이전트틱 AI 모듈은실시간으로미세정책을정제하여피드백에따라그립을조정합니다. 이예시는 VLA 가적용된휴머노이드의변혁적잠재력을보여줍니다. 주방부터클리닉에이르기까지, 이시스템들은복잡한지시를해석하고신체작업을능숙하게수행할뿐만아니라환경의예측불가능성에도적응합니다. 에이전트추론과안전정렬메커니즘을내장함으로써, VLA 로구동되는현대휴머노이드로봇은좁은작업수행자에서범용적이고신뢰할수있는협력자로전환하고있습니다. TinyVLA 와 MoManipVLA 와같은에너지효율모델이성숙함에따라, 모바일의저전력플랫폼배포가점점더실용적이되어사회적으로연계된새로운 AI 시대를열고있습니다.

3.4.2. 자율주행차시스템

자율주행차 (AV), 자율주행차, 트럭, 공중드론을포함한자율주행차는 VLA 모델의최전선적응용분야를대표하며, 안전에중요한의사결정이인지, 의미이해, 실시간액션생성이긴밀하게결합되어야합니다. 인식, 계획, 제어를명시적으로분리하는전통적인모듈형 AV 파이프라인과달리, VLA 프레임워크는시각적관찰, 고수준의의미론적단서, 내부상태표현을통합된모델내에서공동으로처리하여보다긴밀한아키텍처결합을탐구합니다. 이러한엔드투엔드공식화는시뮬레이션과명령조건내비게이션및추론의제어벤치마크에서유망한결과를보여주었지만, 대규모상업용시스템 (예: 테슬라오토파일럿) 은Source Link) 는모듈형또는하이브리드파이프라인에계속의존하며, 최근산업계에서는안전중요운전에서 VLA 스타일의액션생성전면도입보다는비전언어추론구성요소통합에집중하고있습니다.

VLA 모델은 AV 가픽셀수준객체인식을넘어서는복잡한환경을이해할수있도록지원합니다. 예를들어, 도시환경을이동하는자율주행차는교통표지판을감지하고, 보행자행동을이해하며, " 주유소다음두번째우회전" 같은내비게이션명령을해석해야합니다. 이작업들은시각및언어신호를융합하여공간적관계를이해하고, 의도를예측하며, 맥락인식주행행동을생성하는것을포함합니다. VLA 는이정보를토큰기반표현을통해인코딩하며, 시각적인코더 (예: ViT, CLIP), 언어모델 (예: LLaMA-4), 궤적디코더가일관된의미공간에서작동하여고수준목표를추론하고이를저수준움직임으로변환할수있게합니다.

이방향에서주목할만한기여는다음과같습니다 **CoVLA** [5] 는 80 시간이상의실제주행영상과동기화된센서스트림 (예: LiDAR, 주행거리), 상세한자연어주석, 고해상도주행경로를결합한포괄적인데이터셋을제공합니다. 이데이터셋은 VLA 모델을학습시켜시각및언어적특징을신체행동과일치시킬수있게합니다. CoVLA 는시각적그라운딩에 CLIP, 명령어임베딩을위해 LLaMA-2, 그리고움직임예측을위해궤적디코더를사용합니다. 이구성은 AV 가음성신호 (예: " 구급차에양보") 와환경조건 (예: 교통합류) 을해석하여투명하고안전한운전결정을내릴수있게합니다.

OpenDriveVLA [293] 는 2D/3D 다중뷰시각토큰과자연어입력의계층적정렬을통합하여 VLA 모델링기술을발전시킵니다. 이아키텍처는자기중심적공간인식과외부장면이해를모두활용하여동적인에이전트-환경-자아상호작용모델을구축합니다. 자기회귀디코딩을통해 OpenDriveVLA 는조

Figure 12: 이그림은 VLA 프레임워크를사용해집안일을수행하는차세대휴머노이드로봇'Helix' 를보여줍니다. 구두명령을받으면 Helix 는시각언어모델 (예: SigLIP) 과언어모델 (예: LLaMA-4) 을통합하여환경을함께인지하고해석합니다. 계층적 VLA 컨트롤러는냉장고열기, 병잡기등하위작업을계획하고실행하며, 에이전트 AI 모듈은실시간으로동작을적응합니다. 이연구는동적과제적응과안전하고의미론적으로기반한조작을갖춘 VLA 기반범용로봇공학을시연합니다.

작계획 (예: 조향각도, 가속도) 과인간이해석할수있는궤적시각화를생성합니다. 이종단간프레임워크는 nuScenes 와 Waymo Open Motion 데이터셋의계획및궤적예측과제뿐만아니라운전장면의비전-언어질문-응답벤치마크를포함한공공자율주행벤치마크에서선도적인성능을발휘하며, 도시내비게이션과행동예측에서강력한강인성을입증합니다.

또다른기념비적인모델, **ORION** [69] 은 QT-Former 를통합하여장기시각맥락을유지하고, LLM 을통해교통내러티브를추론하며, 생성궤적계획기를도입하여폐쇄루프자율주행의경계를넓혔습니다. ORION 은시각언어모델의이산추론공간을 AV 모션의연속적제어공간과일치시키는데뛰어납니다. 이통합최적화는정확한시각적질문답변 (VQA) 과궤적계획으로이어지며, 모호한인간지시나가려진장애물 (예: " 빨간트럭뒤출구로나가라") 이필요한상황에서매우중요합니다.

예를들어, 그림 13에묘사된것처럼, 차세대 VLA 아키텍처를사용하여밀집된도시환경에서운행하는자율운송차량인"AutoNav" 를고려해보십시오. AutoNav 가클라우드기반명령어로" 빵집옆빨간차양근처에소포를내려놓고, 공사구역을피하며기지로돌아오라" 는명령을받으면, 온보드 VLM(예: CLIP 또는 SigLIP) 이여러카메라의시각스트림을분석하여빵집간판, 빨간차양, 교통콘등동적랜드마크를식별합니다. 동시에, LLaMA-4 에그라운딩된 LLM 모듈은명령어를해독하여 LiDAR, GPS, 관성주행측정등실시간감각맥락과융합합니다. 계층적제어스택은자기중심적관점과세계중심지도를통합하여적응경로를계획하는자기회귀 VLA 디코더를통해이러한멀티모달신호를처리합니다. 차량이배송위치에접근하면, 예상치못한보행자활동이강화학습기반정책개선루틴을사용해에이전트형서브모듈을촉발합니다. 동시에 AutoNav 는보행자에게경고를내고안전마진을유지하기위해속도를재조정합니다. 이러한의미적이해, 지각기반구축, 적응적제어의상호작용은 VLA 기반시스템이안전이중요한상황에서해석가능하고인간친화적인행동을달성하는데얼마나강력한지를보여줍니다.

이시나리오는긴밀히통합된 VLA 아키텍처가종단간의미론적추론, 빠른모듈간적응, 해석가능한의사결정을가능하게하여전통적인인식-계획-제어파이프라인을능가할수있음을보여줍니다. 모듈식파이프라인에서는인식출력, 계획자업데이트, 제어조정이느슨하게결합된구성요소에의해제한된의미적피드백을받지만, VLA 기반시스템은언어의도, 시각적맥락, 체화된상태를함께추론하여동적으로경로를재계획하고, 안전관련의도를인간에게전달하며, 실시간으로제어정책을조정할수있습니다. 그결과, 시스템은더큰자율성, 인간이해석할수있는출력을통한투명성향상, 그리고안전이중요한환경에서의사결정의민첩성을향상시킵니다.

항공로봇공학에서 VLA 는드론이나 UAV 의전달및기타임무능력을향상시킵니다. 예를들어 **UAV-VLA** [191] 위성이미지, 자연어임무설명, 온보드감지를결합하여고수준명령 (예: " 파란방수포와함께옥상패드에전달") 을실행합니다. 이시스템들은모듈식 VLA 아키텍처를사용하여비전언어플래너가전지구적맥락을해석하고비행관제사가정확한웨이포인트를실행하여물류, 재난대응, 군사정찰분야의응용을지원합니다.

자율시스템이점점더비구조화환경에서작동함에따라, VLA 는전통적인파이프라인에대한확장가능하고해석가능하며데이터효율적인대안을제공합니다. 대규모멀티모달데이터셋에서학습하고의사결정을토큰예측으로모델링함으로써, VLA 는인간수준의의미론을로봇움직임과일치시켜더안전하고스마트한자율주행및내비게이션기술의길을열어줍니다.

3.4.3. 산업로봇

산업로봇공학은 VLA 모델의통합으로패러다임전환을겪고있으며, 고수준추론, 유연한작업수행, 인간조작자와의자연스러운소통이가능한새로운세대의지능형로봇을가능하게하고있습니다 [32, 7]. 전통적인산업용로봇은일반적으로엄

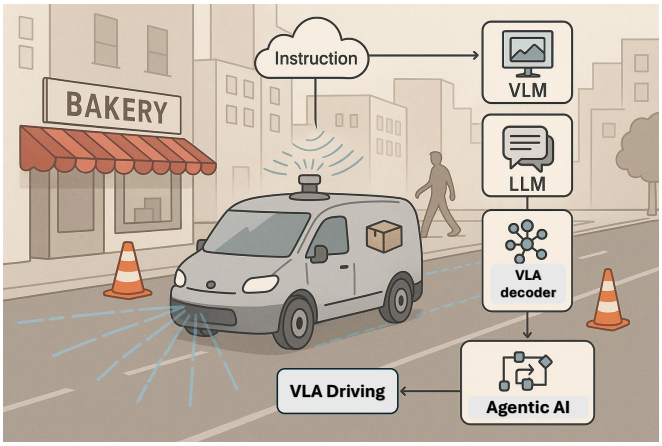

Figure 13: 이그림은 VLA 시스템으로구동되는자율운송차량을묘사하며, 시각적그라운딩용 VLM, 명령어파싱용 LLM, 경로계획용 VLA 디코더를통합합니다. 에이전트 AI 는동적환경에서적응형궤적정제를가능하게하며, 멀티모달통합이실제내비게이션작업에서안전하고해석가능하며자율적인의사결정을어떻게이끄는지를보여줍니다.

격한프로그래밍을사용하여고도로구조화된환경에서작동하며, 새로운조립라인이나제품변형에적응할때광범위한재구성과수동개입이필요합니다 [6, 182]. 이러한시스템은현대의동적제조환경에필요한의미론적기반과적응력이부족합니다.

반면 VLA 모델은더인간이해석하기쉽고일반화가능한프레임워크를제공합니다. 시각적입력 (예: 구성요소배치또는컨베이어벨트상태), 자연어명령어 (예: " 빨간모듈의나사를조여라"), 로봇상태를공동삽입함으로써 VLA 는맥락을추론하고적절한제어명령을실시간으로실행할수있습니다 [135, 72, 156]. 비전트랜스포머 (예: ViT, DINOv2), LLM(예: LLaMA-4), 그리고자기회귀또는확산기반액션디코더가이시스템의중추를이루며, 로봇이멀티모달명령어를해석하고환경에기반한작업을수행할수있게합니다.

이분야에서가장중요한기여중하나는 **CogACT** [131] 산업용로봇조작을위해명시적으로설계된모듈형 VLA 프레임워크입니다. 초기 VLA 가고정된언어-시각임베딩과직접액션양자화에의존하는것과달리, CogACT 는액션시퀀스를보다견고하고적응적으로모델링하는확산기반행동트랜스포머를도입합니다. 이시스템은시각언어인코더 (예: Prismatic-7B)를사용하여고수준장면및명령어임베딩을추출하고, 이를확산트랜스포머 (DiT-Base) 로전달하여미세한모터동작을생성합니다. 이러한모듈식분리는보이지않는도구, 부품, 레이아웃에대한우수한일반화를가능하게하면서도실제제약조건하에서도해석가능성과견고성을유지합니다.

더나아가, CogACT 는 6-DoF 팔이나양손시스템등다양한로봇구현체에대한빠른적응을효율적인미세조정을통해입증하여, 이기종적인공장환경전반에배치할수있도록합니다 [131]. 실증평가에따르면 CogACT 는 OpenVLA 와같은이전모델보다 28% 이상우수한성능을보입니다실제과제성공률에서 [122] 특히다단계조립, 나사고정, 부품분류와같은복잡하고고정밀작업에서그렇습니다 [131, 262].

제조업이인더스트리 4.0 패러다임으로전환함에따라, VLA 는프로그래밍오버헤드를줄이고, 음성명령로봇프로그래밍을지원하며, 혼합이니셔티브작업에서실시간인간-로봇협업을촉진할것을약속합니다. 실행정밀도, 안전보장, 지연최적화분야는여전히활발히연구되고있지만, VLA 모델의활용은자율적이고지능적이며적응력있는로봇으로나아가공장을변화시키는중요한진전입니다.

3.4.4. 헬스케어및의료로봇

의료및의료로봇공학은정밀함, 안전성, 적응력이 VLA 모델이점점더잘제공하는중요한특성인고위험분야를대표합니다 [132, 193]. 전통적인의료로봇시스템은원격조작이나미리프로그래밍된행동에크게의존합니다 [166, 204] 이는역동적인외과또는돌봄환경에서자율성과반응성을제한합니다. 반면, VLA 모델은실시간시각지각, 언어이해, 세밀한운동제어를통합하는유연한프레임워크를제공하여의료로봇이고수준지시를이해하고복잡한절차나보조작업을자율적으로수행할수있게합니다 [131, 54, 225].

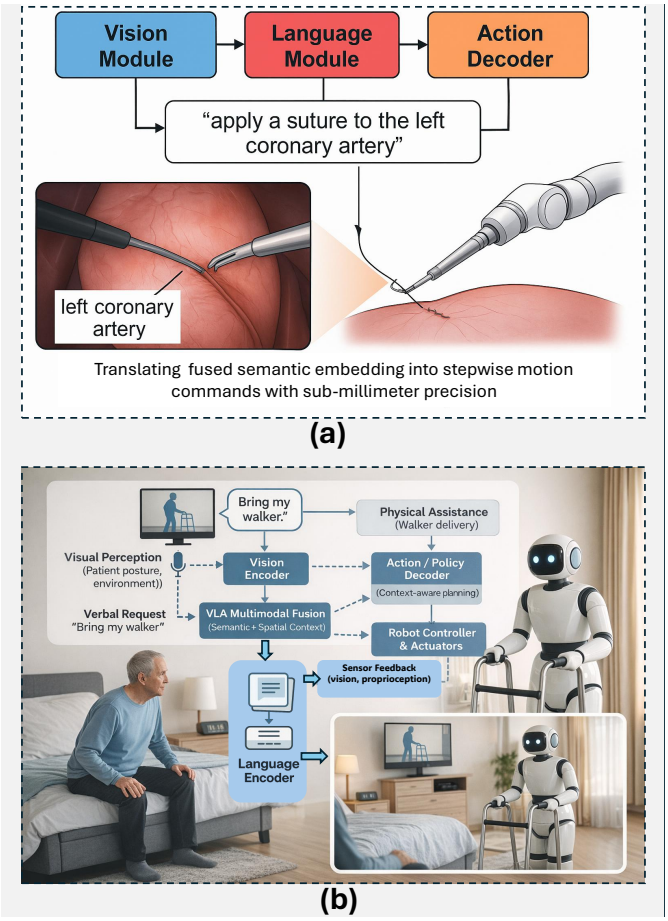

Figure 14: a) 이그림은 VLA 수술시스템이" 왼쪽관상동맥에봉합사를삽입" 하는작업을수행하고있음을보여줍니다. 비전모듈은해부학적표적을식별하고, 언어모델은지시를해석하며, 액션디코더는정밀한운동명령을생성하여적응형도구제어, 실시간피드백, 안전한자율작동을가능하게합니다; b) VLA 기반보조로봇은환자행동을인지하고, 구두요청 (예: " 내워커를가져와줘") 을처리하며, 상황인식동작계획을자율적으로실행하여노인돌봄, 재활, 병원물류분야에서사전정의된스크립트나수동감독에의존하지않고실시간지원을가능하게합니다.

수술로봇공학에서 VLA 는최소침습수술의능력을획기적으로향상시킬수있습니다 [50, 230]. 이시스템들은복강경비디오피드를융합할수있습니다 [127], 해부학지도 [147, 50] 음성명령은비전인코더 (예: ViT, SAM-2) 와언어모델 (예:

LLaMA, T5) 을사용하여통합된토큰화된표현으로변환합니다 [236]. 예를들어, 그림에나타난바와같이 14A, ” 왼쪽관상동맥에봉합사를붙이기” 와같은작업에서는시각모듈이해부학적대상을식별하고, 언어모듈은지시를맥락화합니다. 액션디코더는융합된의미임베딩을서브밀리미터정밀도로단계별모션명령으로변환합니다. 시각적인식, 언어기반의도, 행동수준제어의이러한폐쇄루프융합은로봇이도구를적응적으로재배치하고, 동적포스피드백을적용하며, 중요한해부학적구조를회피함으로써외과의의세세한관리가필요없이줄어들고인간의실수위험을최소화할수있게합니다.

수술실외에도, VLA 모델은노인돌봄, 재활, 병원물류분야에서새로운세대의환자보조로봇을구동하고있습니다. 이시스템들은환자행동을자율적으로인지하고, 말이나몸짓입력을이해하며, 약물수거, 이동보조기구안내, 응급상황시돌봄제공자알림과같은반응적인작업을수행할수있습니다. 예를들어, 그림 14 에묘사된것처럼,b, VLA 지원로봇은환자가침대에서일어나려는것을시각적으로감지하고, ” 내워커를가져오라” 와같은구두요청을해석하며, 사전정해진대본이나지속적인감독없이도상황에맞는동작계획을생성할수있습니다.

RoboNurse-VLA 와같은최근 VLA 프레임워크 [132] 이접근법의현실적실현가능성을강조합니다. RoboNurse 는의미론적장면분할을위해 SAM-2 를, 명령이해를위해 LLaMA-2 를사용하며, 이를실시간음성투액션파이프라인에통합하여수술실내수술기구인수인계를로봇이지원할수있게합니다 [132]. 이시스템은다양한도구, 다양한조명조건, 소음환경에견고함을입증하며, 이는임상환경에서흔히발생하는도전과제입니다.

또한, VLA 아키텍처는규제의료분야에서매우중요한설명가능성과감사가능성측면에서장점을제공합니다 [224, 146]. 장면지상및궤적예측은사후에시각화하고검토할수있습니다 [278] 임상신뢰를촉진하고 FDA 식검증파이프라인을가능하게할수있습니다. LoRA 기반미세조정은최소한의데이터와계산인프라로특정병원환경이나절차적워크플로우에적응할수있게해줍니다 [9, 229, 147].

중요하게도, VLA 모델의멀티모달기반은도메인간전환성을가능하게합니다: 수술도구조작에대해훈련된동일한모델이적당한재교육만으로도환자이동성작업에적응할수있습니다 [56]. 이러한모듈화는작업별자동화시스템에비해개발시간과비용을크게줄여줍니다 [94]. 의료로봇공학이원격조종보조에서반자율및협업시스템으로전환함에따라, VLA 모델은이변화의핵심에자리잡고있습니다.

앞서다른응용분야에서논의되었듯이, VLA 가고수준의미론적이해와저수준제어를결합하는능력은확장가능하고인간친화적이며적응형로봇의료를위한통합솔루션을제공하는데필수적입니다 [249, 295, 279]. 의료시스템이수요와인력부족증가에직면함에따라, VLA 기반로봇공학은의료정밀도, 운영효율성, 환자중심진료를향상시키는데중요한역할을할것입니다.

3.4.5. 정밀및자동화농업

그림 15에나타난바와같이, VLA 모델은다양한농업환경에서노동집약적작업에지능적이고적응력있는솔루션을제공하는정밀및자동화농업분야에서혁신적도구로부상하고있습니다 [70, 191]. 전통적인농업자동화시스템은각작업이나환경변화에따라수동으로재프로그래밍해야하는경직되고센서기반파이프라인에의존합니다 [220, 108], VLA 는멀티모달지각, 자연어이해, 실시간액션생성을통합된프레임워크내에통합합니다 [168, 84]. 이통합된멀티모달통합은자율지상로봇과드론이복잡한밭장면을해석하고, 음성또는텍스트기반농업지시를따르며, 선택적과일수확이나적응적관개같은상황인식행동을생성할수있게합니다. VLA 는폐색, 지형불규칙성, 조명변동성, 다양한작물유형에동적으로적응할수있는능력과합성되고사실적인데이터셋을이용한학습을결합하여작물유형, 지리, 계절전반에걸쳐일반화할수있습니다. 액션토큰화를활용함으로써 [245], 트랜스포머기반정책생성 [12, 85], 그리고 LoRA 미세조정과같은기법들 [93] 이시스템들은지속가능하고정밀한농업을위한농업로봇의확장성과지능을재정의하고있습니다.

현대과수원과기타작물밭에서는 VLA 가 RGB-D 카메라, 다중분광센서, 드론등에서오는시각적입력을처리하여식물성장을모니터링하고, 질병을감지하며, 영양소결핍을식별할수있습니다. 시각트랜스포머 (예: ConvNeXt, DINOv2) 는시각적장면에서공간적·의미론적정보를인코딩하는반면, LLM(예: T5, LLaMA) 은” 동쪽구역에흰가루병검사” 나” 관개도랑근처에서익은사과수확하기” 와같은자연어명령을구문분석합니다. 토큰융합을통해이러한양식들은공유표현공간에정렬되어, 로봇이정밀하고맥락인식있는동작을수행할수있게합니다.

예를들어, 그림 15 의과일따기과제에서볼수있듯이, VLA 가장착된지상로봇은이미지기반숙성신호를사용해익은농산물을식별하고, ”A 등급과일만따라” 와같은사용자가지정한기준을해석하며, 최종이펙터를제어하는액션토큰을통해동작시퀀스를실행할수있습니다. 이방법은작물피해를최소화하고, 수확률을최적화하며, 가림이나지형이동과같은예상치못한변수에실시간으로적응할수있게합니다. 관개관리에서는 VLA 모델에의해유도된드론이현장지도와구두지시를해석하여수분부담지역을선택적으로적용하여물사용량을줄일수있습니다.

즉각적인작업실행을넘어, VLA 모델은폐쇄루프피드백메커니즘을통한동적재구성과평생학습을지원할것으로기대됩니다. 특히, 실행결과, 감각관찰, 배치중수집된작업성공신호는기록되어정기적으로 VLA 정책의오프라인또는점진적업데이트에반영될수있습니다. 작물환경의사진사실시뮬레이션 (예: 3D 과수원렌더링) 에서생성된합성학습데이터와결합하면, 이러한피드백기반적응은광범위한수동주석없이도새로운작물품종, 해충상태, 계절변화에대한견고성을점진적으로향상시킬수있게할수있습니다. LoRA 어댑터와확산기반정책정제와같은매개변수효율적인기법이계산오버헤드를최소화하면서이러한지속적인적응을촉진하는데핵심적인역할을할것으로기대됩니다.

전반적으로 VLA 모델을농업작업흐름에통합하는것은숙련된수작업의존도감소, 목표기반개입을통한수확량향상, 최적화된투입사용을통한환경지속가능성증대등여러장기적이점을제공할것으로기대됩니다. 전세계식품시스템이점점더기후변동성과자원제약에직면함에따라, VLA 기반농업기술은실제세계의복잡성을더잘수용하는확장가능하고지능적이며맥락인식있는농업관행에기여할것으로기대됩니다.

3.4.6. 비전-언어-액션모델을이용한인터랙티브 *AR* 내비게이션

인터랙티브증강현실 (AR) 내비게이션은 VLA 모델이실시간으로지능적이고맥락인식적인안내를제공하여인간과환경상호작용을크게향상시킬수있는최전선을나타냅니

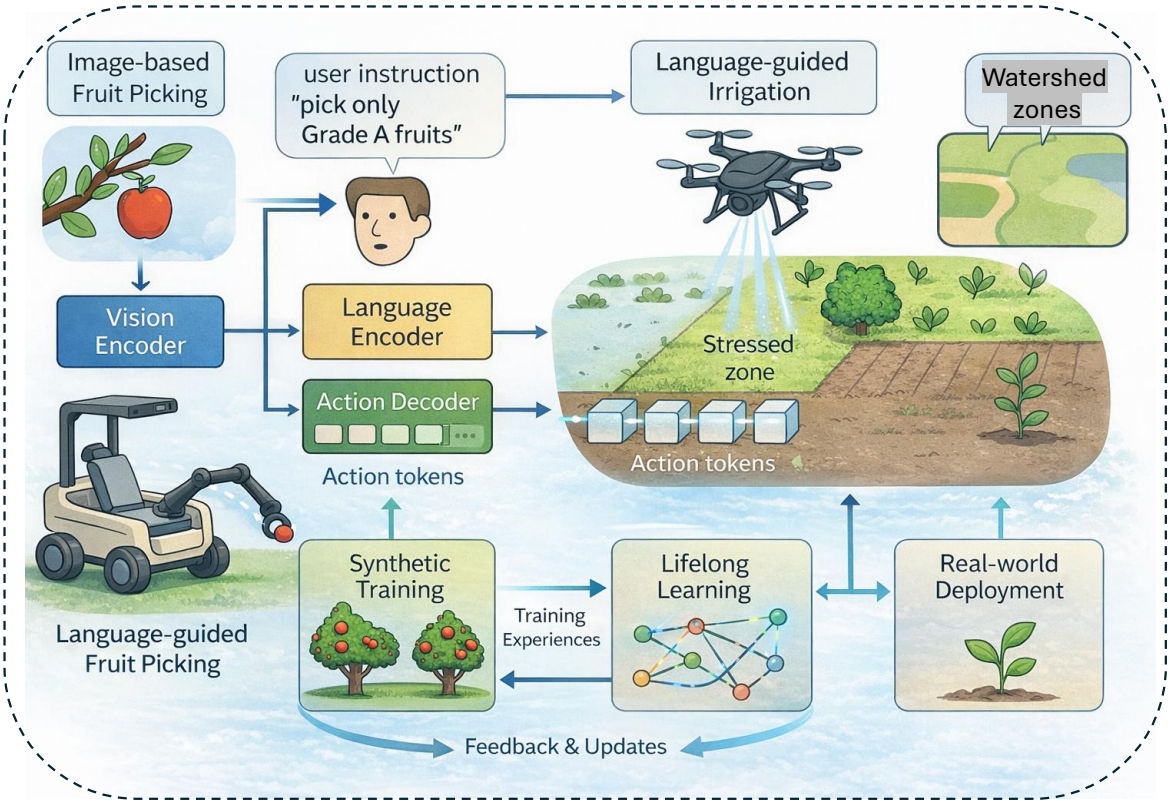

Figure 15: 정밀및자동화농업에서의 VLA 모델의개념적설명. 지상로봇은시각인코더와언어명령 (예: "A 등급과일만따라") 을결합해무손상수확을위한액션 토큰을생성하며, 항공로봇은언어유도관개를위해 VLA 추론을사용합니다. 합성교육, LoRA 기반적응, 평생피드백을통해작물, 환경, 지리적전반에걸친일반화가가능하며, 지속가능하고데이터기반농업자동화를지원합니다.

다 [31, 104, 255]. 이패러다임에서 VLA 는스마트글라스나스마트폰과같은 AR 지원장치에서발생하는시각적데이터의연속적인스트림을자연어쿼리와함께처리하여, 사용자의물리적세계시점에직접겹쳐놓은동적내비게이션단서를생성합니다. 전통적인 GPS 기반시스템과달리, 엄격한지도와제한된사용자입력에의존하는것과다릅니다 [29, 207], VLA 기반 AR 요원들은교차로, 실내복도, 표지판등복잡한시각적장면을해석하고, " 휠체어경사로가있는가장가까운약국으로데려다주세요" 또는" 회의실로가는가장조용한경로를안내해주세요" 와같은자유로운형태의지시에응답합니다.

기술적으로이모델은 RGB 카메라프레임에서장면표현을추출하는비전인코더 (예: ViT, DINOv2), 사용자프롬프트나음성명령을처리하는언어인코더 (예: T5 또는 LLaMA), 그리고방향오버레이, 웨이포인트, 음성지시와같은토큰화된내비게이션신호를예측하는액션디코더를통합합니다. 트랜스포머기반아키텍처는이러한양식를융합하여공간적배치와의미적의도를모두추론하여, AR 에이전트가사용자의시야내에서경로, 랜드마크, 위험요소를적응적으로하이라이트할수있게합니다 [211, 165]. 예를들어, 그림 16 에나타난바와같이혼잡한공항에서는 VLA 상담원이에스컬레이터, 게이트, 수하물찾는곳을시각적으로식별하고, " 계단없이 22 번게이트에어떻게도달하나요?" 와같은질문을이해하고, 실시간인원상황과장애물에맞춰경로를조정할수있습니다.

VLA 는사용자가먼저고수준명령어 (예: ' 약국으로이동') 를내린후, ' 붐비는구역피하기' 나' 경치좋은길로이동하기' 와같은추가제약으로이를세정할수있는대화형명령어루프를지원할것으로기대됩니다. 맥락인식피드백과반복적인명확화를통해이러한상호작용패러다임은시각장애인이나인지장애인의접근성과사용성을향상시킬수있습니다. 물류및실내내비게이션에서는 IoT 센서와디지털트윈과통합되어창고작업자, 유지보수팀또는배달로봇이복잡한환경에서안내할수있습니다. 또한, VLA 모델이사용자의선호도와지역공간배치를시간에따라학습하는지속적인미세조정을통해개인화된내비게이션이가능합니다.

AR 하드웨어가점점더저렴해지고일상생활에통합됨에따라, VLA 기반내비게이션시스템은공공, 산업, 보조환경에서원활한공간이해, 멀티모달상호작용, 자율안내를가능하게하여인간이물리적공간을인식하고탐험하며상호작용하는방식을재정의할것입니다.

4. 비전-언어-액션모델의도전과한계

VLA 모델은연구프로토타입에서견고한실제시스템으로의전환을방해하는상호연관된다양한도전과제에직면해있습니다. 첫째, 실시간자원인지추론을달성하는것은여전히어렵습니다: DeeR-VLA 와같은모델은동적조기종료아키텍처를활용해조작벤치마크에서 5–6× 계산을줄이면서도정확도를유지하지만, 복잡한상황에서는그효과가줄어듭니다 [269]. 마찬가지로 Uni-NaVid 는 5Hz 내비게이션을위해자기중심적인비디오토큰을압축하지만, 매우모호한지시와긴지평선아래에서는여전히어려움을겪습니다 [280]. 더불어, 이러한효율성중심설계는계산속도와표현커버리지사이의절충관계

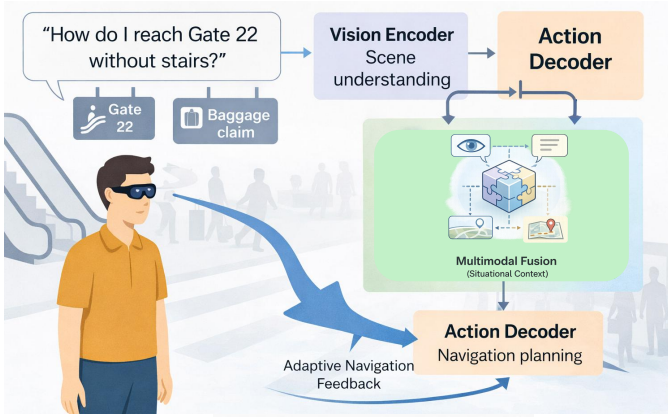

Figure 16: VLA 모델이실시간시각인식, 언어이해, 행동계획을융합하여인터랙티브 AR 내비게이션을가능하게하는방식을보여줍니다. 공항과같은동적환경에서는 VLA 가"22 번게이트로계단피하라" 와같은사용자쿼리를해석하고, 시각적장면 (예: 에스컬레이터감지) 을분석하며, 이에따라내비게이션경로를조정하여개인화되고접근가능하며상황인식있는이동안내를지원합니다.

를자주드러냅니다. 공격적인압축이나조기종료제약하에서작동할때, 고급하이브리드비전-언어기반기법조차도제한된객체일반화를보입니다; 예를들어, ObjectVLA 는새로운객체의 64% 에대해서만일반화되며, 이는실시간최적화가오픈월드강인성의격차를얼마나크게확대할수있는지를보여줍니다 [297].

둘째, 최소한의감독으로 VLA 모델을적응시키고, 드물고잡음이많은데이터에서안정적인정책업데이트를보장하는것은쉽지않습니다. ConRFT 는행동복제와 Q-학습을인간개입미세조정 (Human-in-the-loop fine-tuning) 과결합하여, 접촉이많은 8 개과제에서 96.3% 의성공률에빠르게수렴했지만, 전문가의개입과보상형성에크게의존합니다 [36]. Hi Robot 과같은계층적프레임워크는고수준추론과저수준실행을분리하여명령어충실도를높이지만, 이러한모듈을조정하고모호한피드백을안정시키는것은여전히어렵습니다 [199]. 마찬가지로, 촉각스트림과언어명령어를융합한촉각-언어-액션모델은보이지않는 peg-in-hole 작업에서 85% 이상의성공률을달성했지만, 데이터셋의폭과실시간다단계디코딩은여전히더넓은일반화에한계가있습니다 [88].

더나아가, 동적인환경에서안전성, 일반화및종단간신뢰성을보장하기위해서는새로운모델링및평가기준이필요합니다. OccLLaMA 와같은점유-언어-행동모델은 3D 장면이해와행동계획을통합하지만, 양식전반에걸쳐더풍부한장면역학과의미적일관성을위해확장되어야합니다 [239]. RaceVLA 는양자화된반복제어루프를통해고속드론내비게이션을발전시킵니다; 하지만더큰 VLA 와전용추론모델에비해시각적·물리적일반화가제한적이어서보이지않거나빠르게변화하는환경에서의안전우려가제기됩니다 [195]. ReVLA 의모델병합전략은영역외시각적강인성을잃었던것을회복하여방향대상감지 (OOD) 잡기성공률을최대 77% 까지향상시킵니다하지만추가적인계산과복잡성을도입합니다 [48]. 마지막으로, SafeVLA 는제약마르코프의사결정과정을통해안전하지않은행동을 80% 이상줄이기위한제약조건을공식화하지만, 다양한실제업무에대한포괄적이고제한없는안전규칙을정의하는것은여전히미해결문제입니다 [274]. 이러한교차하는한계를해결하는것은 VLA 모델이실제로봇공학의전체복잡성하에서신뢰할수있고자율적으로작동할수있도록매우중요합니다.

위에서언급한중요한한계를바탕으로, 각도전과제를목표완화전략과매핑하고시스템수준의영향을평가하는것이필수적입니다. 표 4 는핵심한계를식별하고, 최근발전에서도출한잠재적기술적해결책이며, 실제 VLA 배포에대한기대되는이점을명확히설명합니다. 예를들어, 실시간추론제약문제를해결하는방법은병렬디코딩과양자화된트랜스포머파이프라인을하드웨어가속 (예: TensorRT) 으로활용하여드론과매니퓰레이터에서제어루프속도를유지합니다 [129, 122, 75, 142]. 하이브리드확산-자기회귀정책을통한멀티모달액션표현은복잡한작업에대해다양하고맥락민감한운동명령을생성하는모델을풍부하게만듭니다 [171, 156]. 오픈월드의안전을보장하기위해동적위험평가모듈과적응형계획계층을통합하여예측불가능한환경에서강력한비상정지행동을보장할수있습니다 [183, 233, 113]. 마찬가지로, 데이터셋편향과그라운딩는선별된편향제거말뭉치와고급대조미세조정을통해최소화할수있으며, 새로운객체와장면에일반화할때공정성과의미적정확성을강화할수있습니다 [185, 17, 175]. 이러한전략들과시뮬레이션-실제 (Sim-to-Real) 전이, 촉각통합, 에너지효율아키텍처를아우르는다양한접근법이결합되어 VLA 연구를신뢰할수있고확장가능한자율성으로전환하는포괄적인로드맵을형성합니다.

이섹션의나머지부분은다섯개의집중된하위섹션으로구성되어있으며, 각섹션은문헌에서확인된 VLA 문제의뚜렷한군집을살펴봅니다. 먼저, 실시간추론제약과이를해결하기위한신흥방법을분석합니다. 다음으로, 오픈월드환경에서의안전성보장과함께멀티모달액션표현을탐구합니다. 그후데이터셋편향, 기반전략, 보이지않는작업에대한일반화에대해논의하고, 이어서시스템통합복잡도와계산요구량에대해탐구합니다. 마지막으로, 우리는견고성과실제응용분야에서 VLA 배포의윤리적함의를고려합니다.

4.1. 실시간추론제약

최근진전에도불구하고, 지연이중요한환경에서 VLA 모델을배치하는것은특히로봇조작, 자율주행, 항공제어와같은응용분야에서실시간추론요구에의해제약을받고있습니다. VLA 는일반적으로이전예측을바탕으로순차적으로액션토큰을생성하는자기회귀디코딩전략에의존합니다. 많은작업에효과적이지만, 이자기회귀복호화패러다임은추론속도를크게제한하여보통 3-5 에불과합니다 표준 GPU 기반연구플랫폼 (예: 단일고급소비자용또는데이터센터 GPU) 에서종단간 VLA 추론을위해배포할때 Hz [67]. 이속도는반응적이고안정적인로봇작동에일반적으로요구되는제어주파수보다상당히낮으며, 고수준계획은수십헤르츠, 저수준피드백제어는더높은업데이트속도까지다양하며, 이는작업및하드웨어플랫폼에따라다릅니다. 예를들어, 로봇팔이섬세한물체를조작할때는정확성을유지하고손상을방지하기위해자주위치정보를업데이트하는것이필수적입니다. OpenVLA 와같은모델 [122] 그리고 Pi-0 [15] 이순차적토큰생성방식은본질적인어려움에직면해동적환경에서의효율성이제한됩니다.

병렬디코딩과같은신흥솔루션으로, NVIDIA 의 GR00T N1 모델이대표적입니다 [14], 여러토큰을동시에예측하여추론을가속화하는것을목표로합니다. GR00T N1 은전통적인디코딩방식보다약 2.52× 속도향상을달성합니다; 하지만이러한병렬성은종종궤적의부드러움에상충을초래하여로봇의움직임이최적이되지않습니다. 이러한움직임은정밀함, 적응

Table 4: 도전과제, 잠재적해결책, 그리고 VLA 모델의예상영향.

과제 / 한계	잠재적해결책	기대효과
실시간추론제약	병렬디코딩, 양자트랜스포머, 하드웨어가속 (예: TensorRT) [129, 122]; 자기회귀오버헤드감소 [75, 142].	지연이중요한영역에서실시간제어및배포를가능하게합니다 [251, 191] (예: UAV, 조작기).
멀티모달액션표현	확산과자기회귀정책을결합한하이브리드토큰화 [171]; 다양한시연과멀티모달산출물에대해훈련합니다 [156].	복잡하고동적인조작에서다중유효한해법모드를통해성능을향상시킵니다 [74].
오픈월드에서의안전보장	동적위험평가모듈 [183, 233]; 저지연비상정지및적응형계획계층 [113].	가정, 공장, 의료등예측불가능한환경에서신뢰성과안전성을향상시킵니다; 사용자수용도를향상시킵니다.
데이터셋편향과그라운딩	다양하고편향이없는데이터셋을선별하세요 [185]; 더강한그라운딩 (예: 하드네거티브를이용한 CLIP 미세조정) [282, 17].	공정성과의미정확성을향상시킵니다 [109], 그리고새로운실제입력에대한일반화를강화합니다 [227, 288, 175].
제한된 3D 지각과추론	깊이/LiDAR 통합; 3D 인식아키텍처를개발하고; 포인트클라우드와비전—언어특징을융합한다.	복잡한환경에서조작과내비게이션을위한더강한공간적추론을가능하게합니다 [129].
신체간일반화	다양한형태를거치며훈련합니다; 형태중립적행동추상화를배우고; 도메인간적응적용 [266].	로봇플랫폼및구성간정책이전을용이하게합니다 [279, 122].
주석복잡성및비용	수작업라벨링감소를위한약한감독, 능동적학습, 합성데이터생성 [148].	개발비용을줄이고새로운작업/도메인으로의확장을가속화합니다 [233, 286].
시뮬레이션-실제 (Sim-to-Real) 전이갭	도메인적응, 시뮬레이션-실-미세조정, 그리고실제세계보정 [210, 134].	시뮬레이션이후로배포할때신뢰성과일관성을향상시킵니다 [4, 66].
물리적지식의통합	물리사전, 시뮬레이션환경, 그리고학습파이프라인에서의동역학모델링 [53].	실제물리적제약조건하에서예측및계획을개선합니다 [106].
멀티모달통합 (촉각, 오디오)	촉각/오디오와시각및언어를융합합니다 [114]; 멀티모달트랜스포머퓨전을확장하세요.	폐쇄/모호성하에서도견고성을높이고작업범위를확장합니다 [76, 136, 91].
장기과제다단계과제	계층적정책, 메모리증강네트워크, 그리고궤적계획모듈 [136].	순차적계획, 기억, 구성적실행을향상시킵니다 [131, 286, 227, 175].
시스템통합복잡도	통합트랜스포머백본 [289]; 시간적정렬과시뮬레이션-실제 (Sim-to-Real) 전송전략 [163, 277].	더엄격한계획수립가능—제어조정및물리적로봇으로의보다견고한이송 [187, 208].
에너지및컴퓨팅수요	가지치기, LoRA, 양자화인식학습, 저전력가속기.	효율적인임베디드/모바일배포가능하게합니다 [252, 283, 125, 246].
보이지않는작업으로의일반화	구성일반화, few-shot 메타러닝, 그리고작업에구애받지않는사전학습 [173, 146].	과적합을줄이고제로/few-shot 적응을강화합니다 [97, 242, 286].
환경변동성에대한견고성	도메인무작위화, 센서융합, 광범위한훈련데이터셋, 온라인재보정 [156].	변화하는조명, 잡동사니, 장면역학상황에서의안정성향상 [285, 242].
윤리적·사회적함의	기기내처리/익명화를통한프라이버시 [149, 198, 252, 34]; 공정성감사; 규제및신뢰프레임워크.	사회, 의료, 노동분야에서공정하고신뢰할수있는채택을촉진합니다 [160, 180, 217, 172].

성, 유연성이가장중요한수술용로봇과같은민감한응용분야에서는바람직하지않습니다. 따라서출력품질을저해하지않고빠른추론을달성하는것은여전히열린도전과제로남아있습니다.

게다가하드웨어의한계는실시간추론제약을더욱악화시킵니다. 예를들어, 일반적으로 512 차원 400 개이상의비전토큰을포함하는고차원시각임베딩처리는약 1.2 GB/s 메모리대역폭을필요로합니다. 이수요는현재임베디드시스템이나 NVIDIA Jetson 플랫폼과같은엣지 AI 하드웨어의용량을상당히초과하여실질적인배포를제한합니다 [82, 275]. 메모리제약을완화하기위해부동소수점연산의정밀도를낮추는효율적인양자화기법을사용하더라도, 특히양손로봇조작이나의료로봇공학과같이서브밀리미터정밀도가요구되는작업에서는정확도가자주저하됩니다.

4.2. 멀티모달액션표현및안전보장

멀티모달액션표현: 현재 VLA 모델의중요한한계중하나는특히연속적이고세밀한제어가필요한시나리오에서멀티모달동작을정확히표현하는것입니다 [62, 46]. 256 개의뚜렷한빈으로행동을나누는전통적인이산토큰화방식은본질적으로정밀도가부족하여섬세한로봇잡기나복잡한수술절차와같은세밀한작업에서상당한오류를초래합니다 [171]. 예를들어, 조립작업에서정밀한로봇조작과정에서이산표현은정렬이맞지않거나부정확한동작을초래하여성능과신뢰성을저해할수있습니다. 반면, 연속다층퍼셉트론 (MLP) 기반접근법은모드붕괴위험에직면해있습니다 [162, 232] 모델이여러실행가능한경로가있음에도불구하고단일동작궤적에조기수렴하는경우입니다. 이로인해매우역동적인환경에서적응적의사결정을위한유연성이줄어듭니다. Pi-Zero 와 RDT-1B 모델로대표되는확산기반정책의신흥 [144] 는다양한액션가능성을포착할수있는더풍부한멀티모달액션표현을제공합니다. 하지만기존트랜스포머기반디코더의약세배에달하는상당한계산오버헤드때문에실시간배포에는실용적이지않습니다. 따라서 VLA 모델은현재밀집된공간에서의로봇내비게이션이나정교한양손조작과같은복잡한동적작업에어려움을겪고있습니다 [74, 247], 여기서여러전략적행동이동등하게유효하고상황에따라달라질수있습니다.

오픈월드의안전보장: VLA 가직면한또다른중요한과제는실제상황에서특징적인역동적이고예측불가능한환경에서견고한안전성을보장하는것입니다 [39, 274]. 현재많은구현은미리정의된하드코딩된힘과토크임계값에크게의존하여, 예기치못한장애물이나갑작스러운환경변화와같은새로운조건에대비한적응력을크게제한합니다 [156]. 충돌예측에사용되는모델은일반적으로복잡하고역동적인공간에서약 82% 의정확도에그치는경우가많아안전마진이거의없는창고물류나가정용로봇같은응용분야에서심각한위험을초래합니

다 [288, 122]. 더불어, 긴급정지와같은필수안전메커니즘은포괄적인안전검증으로인해 200 에서 500 밀리초사이의상당한지연시간을포함합니다 [170, 122]. 이지연은겉보기에는미미해보일수있지만, 고속운행이나자동화운전, 긴급로봇대응과같은중요한개입상황에서는위험할수있습니다.

4.3. 데이터셋편향, 그라운딩, 그리고보이지않는과제로의일반화

VLA 모델의효과를제한하는주요장애물은데이터셋편향과그라운딩결함의만연한존재입니다. 현재학습데이터셋은주로웹크롤저장소에서수집되며, 종종내재된편향을보입니다 [214, 119]. 연구에따르면약 17% 표준데이터셋내연관성은' 의사' 와같은용어를남성인물과불균형적으로연관짓는등고정관념적해석에치우쳐있습니다 [222, 124]. 이러한편향은훈련을통해전파되어, 다양한환경에서사용될때의미적으로어긋나거나맥락상부적절한반응을내는 VLA 를초래합니다. 예를들어, OpenVLA 와같은모델은새로운환경에서객체참조의약 23% 를누락하는것으로보고되어, 명령어의정확한해석이중요한실제응용분야에서는실용성을크게제한합니다 [122]. 이그라운딩문제는조합일반화에서도확장되는데, VLA 는훈련말뭉치에서"yellow horse" 같은구절을해석하는등드물거나비전통적인조합을만날때종종실패하는경우가많습니다. 이러한한계는신중하게선별되고균형잡힌포괄적도메인특화데이터셋의긴급한필요성을강조하며, 다양한맥락에서편향을완화하고의미적정렬을강화하기위한고급그라운딩알고리즘과결합되어야합니다.

데이터셋편향이제기하는도전과더불어, VLA 의실질적배포에중요한장벽인보이지않는작업으로의일반화문제가더넓게존재합니다. 기존모델은익숙한환경이나훈련시나리오와유사한작업에서숙련도를보여도, 성능이크게저하되어최대 40% 까지저하되는경우가많으며, 완전히새로운과제나낯선변형을마주칠때입니다. 예를들어, 가정용작업에특화된 VLA 는산업또는농업환경에도입될때주로객체유형, 환경역학, 운영제약등의차이로인해어려움을겪거나실패할수있습니다. 이한계는주로좁은범위의훈련분포에과적합되고다양한과제표현에대한노출이부족하기때문입니다. 따라서현재 VLA 는제로샷또는 few-shot 학습시나리오에서일반화가제한적이어서적응성과확장성을저해합니다.

4.4. 시스템통합복잡성과계산요구사항

고수준인지계획 (시스템 2) 과실시간물리제어 (시스템 1) 를결합한이중시스템아키텍처내에 VLA 모델을통합하는것은로봇응용에서상당한복잡성을야기합니다. 주요도전과제는이두시스템간의시간적불일치에서발생합니다. 일반적으로 System 2 는복잡한작업분해와전략적기획을위해 GPT 나 LLaMA-4 와같은 LLM 을활용합니다. 이러한모델은상당한계산요구량때문에종종다음과같은수준의추론지연을겪습니다 ~800 표준 GPU 기반추론플랫폼 (예: 단일고급소비자용또는데이터센터 GPU) 에서실행될경우 MS 이상을기록하며, 이는 LLM 배포에흔히사용됩니다. 반대로, 시스템도 저수준모터실행을담당하는 1 구성요소는일반적으로실시간 CPU, 마이크로컨트롤러또는전용로봇컨트롤러에구현된엄격하게제한된제어루프내에서실행되며, 플랫폼과작업에따라수밀리초단위의업데이트간격이흔히발생합니다. 이러한극명한운영주기차이는동기화에어려움을초래하여지연과잠재적으로비효율적인실행경로를초래합니다. 예를들어, NVIDIA 의 GR00T N1 모델은이두시스템의효과적인통합을보여주지만, 비동기상호작용으로인해움직임중가끔씩끊김이발생해이본질적인도전과제를부각시킵니다.

더욱이, Vision Transformers(ViT) 와같은고차원비전인코더와저차원액션디코더간의특징공간불균형은통합복잡성을악화시킵니다. 이러한이질적인임베딩을조화시키려할때, 지각이해와실행가능한명령간의일관성이크게저하될수있습니다. OpenVLA [122] 그리고 RoboMamba [143] 트랜스포머기반시각처리와후속액션디코딩을활용하는이통합문제는시뮬레이션환경에서물리적하드웨어배포로이식할때성능저하로이어지는문제를보여줍니다. 이러한차이는주로시뮬레이션된동역학과실제센서노이즈또는보정문제간의불일치로인해성능저하로이어질수있습니다 [122, 49, 92].

에너지및컴퓨팅수요는특히자율드론, 모바일로봇, 웨어러블로봇시스템등엣지컴퓨팅환경에서 VLA 도입에또다른중요한장벽을형성합니다. 고급 VLA 에서흔히볼수있는방대한매개변수수 (예: 70 억개이상의매개변수를가진모델) 는기본형태에서 28GB 를초과하는 VRAM 의연산자원을요구합니다. 이러한요구사항은대부분의최신엣지지향프로세서와 GPU 의성능보다훨씬높아, 특수고자원환경외부에서정교한 VLA 의실용적적용을제한합니다.

4.5. VLA 배포의견고성과윤리적도전과제

VLA 모델의실제배치에가장큰장애물은환경변동성에대한견고성이제한적이라는점에있으며, 이는중요한윤리적및안전문제를제기합니다. 환경적견고성은동적으로변화하고부분적으로관찰가능한조건하에서시스템이신뢰할수있는인식, 추론, 액션생성을지속할수있는능력을의미합니다. 실제로실제환경은조명변화, 악천후, 센서잡음, 물체가림과같은요인으로상당한불확실성을초래합니다.

경험적증거는여러 VLA 구성요소에서이러한한계를강조합니다. 예를들어, OpenDriveVLA 와같은시스템에서사용되는비전모듈 [293] 정확도가약 20–30% 저하되는현상저대비또는그림자가지배적인장면에서현재시각인코더의민감한조명조건을반영합니다. 마찬가지로, CoVLA 와같은 VLA 에서의언어이해도마찬가지입니다 [5] 음향적으로잡음이크거나의미가모호한환경에서악화되며, 명령어오해가잘못된동작실행으로이어질수있습니다. 조작중심시나리오에서는 RoboMamba 와같은 VLA 지원로봇시스템이있습니다 [143] 어수선한환경에서어려움을겪으며, 부분적으로가려진물체의자세나방향을자주잘못판단하여작업성공률을떨어뜨립니다.

이러한견고성제한은안전에중요한배포에서직접적인윤리적영향을미치는데, 실제변동성하에서성능저하가의도치않은동작, 신뢰성저하, 사용자신뢰상실로이어질수있기때문입니다. 따라서견고성문제는단순한기술적도전이아니라인간중심환경에서 VLA 시스템을책임감있고윤리적으로배포하기위한전제조건이기도합니다.

5. 토론

그림에나와같이 17VLA 모델은알고리즘, 계산, 윤리적차원에걸쳐다면적인도전과제에직면해있습니다. 첫째, 자원이제한된하드웨어에서실시간추론을달성하는것은자기회귀디코더의순차적특성과멀티모달입력의고차원성때문에여전히어렵습니다. 둘째, 비전, 언어, 행동을일관된정책으로융합

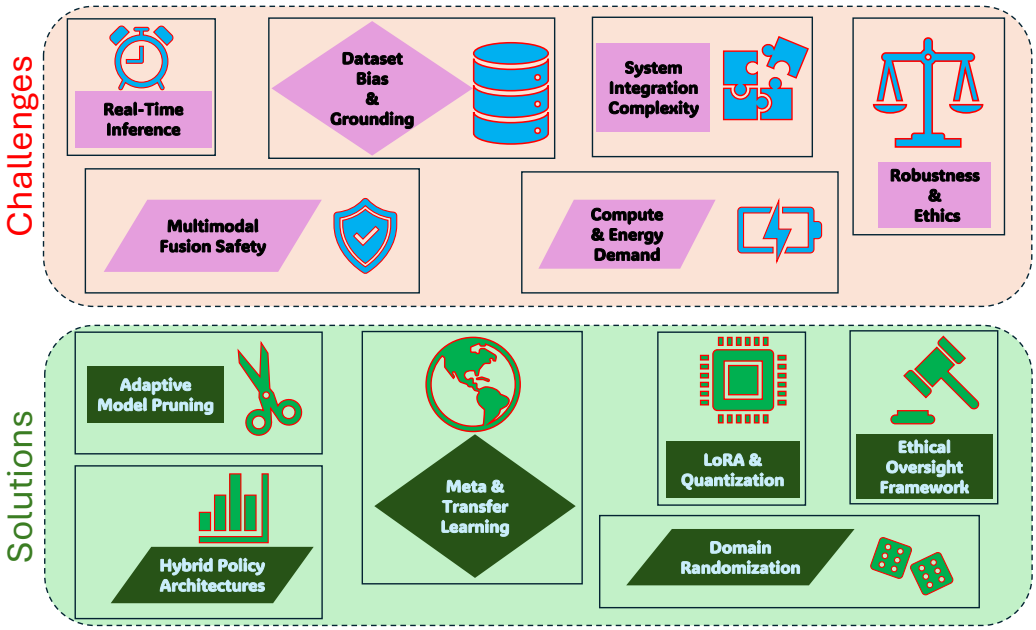

Figure 17: 그림은여섯가지핵심 VLA 도전과제를지도화합니다; 즉, 실시간추론, 멀티모달융합안전성, 데이터셋편향, 통합복잡도, 계산요구, 강인성/윤리성 등이포함됩니다; 적응형가지치기, 하이브리드정책아키텍처, 메타/전이학습, LoRA/양자화, 도메인무작위화, 윤리적감독등 6 가지목표해결책에대항합니다. 이체계적인정렬은더넓은실제로봇영역에서견고하고효율적이며안전한 VLA 배포경로를명확히합니다.

하면예상치못한환경변화에직면했을때안전취약점을초래합니다. 셋째, 데이터셋편향과그라운딩오류가일반화를저해하여종종분포외작업에서모델이실패하는원인이됩니다. 넷째, 다양한구성요소인인식, 추론, 제어를통합하면최적화및유지가어려운복잡한아키텍처가만들어집니다. 다섯째, 대규모 VLA 시스템의에너지및컴퓨팅요구는임베디드또는모바일플랫폼에서의배포를방해합니다.

마지막으로, 환경변동성에대한견고성이제한되면안전하지않거나신뢰할수없는행동이발생할수있으며, 이는안전보장, 책임성, 프라이버시, 편향완화와관련된윤리적·규제적우려를야기합니다. 이러한한계들은실제로봇공학, 자율시스템, 인터랙티브응용분야에서 VLA 모델의실질적채택을제한합니다. 이러한도전에대한잠재적해결책은아래에서논의됩니다.

5.1. 잠재적해결책

- **실시간추론제약.** 향후연구는지연시간, 처리량, 작업특화의정확성을조화시키는 VLA 아키텍처를개발해야합니다. 유망한방향중하나는 FPGA 기반비전프로세서와희소행렬연산에최적화된텐서코어와같은특수하드웨어가속기의통합으로, 서브밀리초규모의합성곱및변환계층을실행하는것입니다 [122, 129]. 로우랭크적응 (LoRA) 과같은모델압축기법 [93] 지식증류는매개변수를최대 90 까지줄일수있습니다% 메모리사용량과추론시간을모두줄이면서도 95 이상을유지합니다% 벤치마크작업에서의원래성과에대한평가를받았습니다. 혼합정밀도산술 (예: FP16/INT8) 과블록별보정을결합한점진적양자화전략은정확도손실을최소화하면서계산량을 2–4× 더줄일수있습니다 [121]. 입력복잡도에따라네트워크깊이또는폭을동적으로조절하는적응형추론아키텍처, DeeR-VLA 의조기종료분기와유사합니다 [269] 시각적장면이나언어명령이단순할때트랜스포머계층을

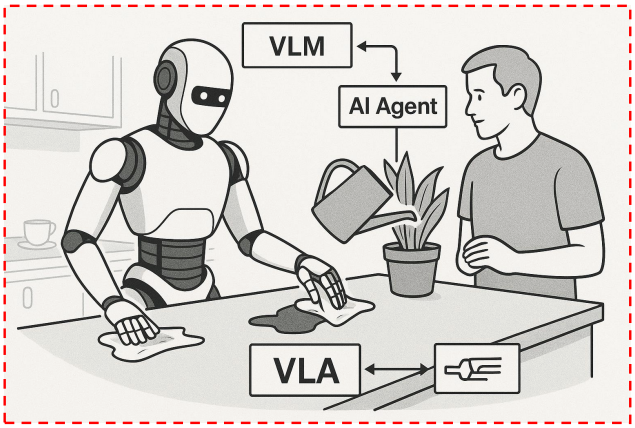

Figure 18: 이개념적일러스트는시각언어모델 (VLM), VLA 프레임워크, 에이전트 AI 시스템으로구동되는미래의휴머노이드어시스턴트"Eva" 를소개합니다. VLM 은의미론적장면이해와객체어포던스예측을가능하게하며, VLA 는언어기반명령어를계층적운동계획으로번역합니다. 에이전트 AI 모듈은개방형환경에서적응형학습, 자기정제, 상호작용적의사결정을보장합니다. 이구성요소들은지각, 언어이해, 계획, 안전한자율행동이현실세계의사회적인식과제에서융합되는로봇공학내인공일반지능 (AGI) 의기초청사진을제공합니다.

선택적으로우회하여평균계산량을줄일수있습니다. 마지막으로, 서브워드패치임베딩과동적어휘할당을활용한효율적인토큰화방식은시각적·언어적입력을압축하여의미론적풍부함을희생하지않으면서토큰수를최소화할수있습니다 [171]. 이러한혁신들은일반엣지 GPU 에서 50ms 미만의종단간추론을가능하게하여, 자율드론비행, 실시간원격조작, 협업제조등지연에민감한응용분야로발전할수있습니다.

- **멀티모달액션표현및안전보장.** 멀티모달액션표현과견

고한안전성을해결하려면엄격한안전제약하에서지각, 추론, 제어를통합하는종단간프레임워크가필요합니다. 저수준모션원시요소를위한확산기반샘플링을결합한하이브리드정책아키텍처 [40] 자기회귀적고수준계획자와함께 [242] 다양한액션궤적을압축된확률적표현으로하여동적환경에서의적응력을향상시킵니다. 안전은시각, 깊이, 고유수용성데이터를포함한다중센서융합스트림을실시간으로수집하여충돌확률과관절스트레스임계값을예측하고, 미리정해진안전한계를위반할때비상정지회로를작동시키는실시간위험평가모듈을통해강화할수있습니다 [183, 233]. 제약최적화가추가된강화학습알고리즘 (예: SafeVLA 의라그랑지안방법) [274]) 안전제약을엄격히준수하면서작업성공률을극대화하는정책을학습할수있습니다. 규칙기반 RL(GRPO) 과직접선호최적화 (DPO) 와같은온라인모델적응기법은새로운환경조건에서행동선택을더욱정교하게조정하여시나리오전반에걸쳐일관된안전성능을보장합니다 [113]. 무엇보다도, 실행전에계획출력을상징적으로분석하는형식적검증계층을내장하면신경망기반제어장치에서도안전불변량준수를보장할수있습니다. 이러한방법론을통합하면복잡하고멀티모달동작을실행할뿐만아니라비구조화된실제환경에서안전성이입증된 VLA 시스템을만들어낼것입니다.

- **데이터셋편향, 그라운딩, 그리고보이지않는과제에대한일반화.** 견고한일반화는확장된데이터다양성과고급학습패러다임을모두요구합니다. LAION-5B 와같은웹규모이미지-텍스트말뭉치를결합한대규모편향없는멀티모달데이터셋을큐레이션합니다 [194] Open X-Embodiment 와같은로봇중심궤적아카이브를포함해 [227] 이는공정한의미론적그라운딩을위한토대를마련합니다. 하드네거티브샘플링과비전-언어골격의대조적미세조정 (예: CLIP 변형) 은허위상관관계를완화하고의미충실도를향상시킬수있습니다 [17, 282]. 메타러닝프레임워크는시각언어로봇내비게이션모델에서입증된바와같이과제군간에공유된사전학습을통해새로운과제에빠르게적응할수있도록합니다 [175]. 재생버퍼와정규화전략을가진연속학습알고리즘은기존지식을보존하면서새로운개념을통합하여 VLA 모델에서의치명적인망각문제를해결합니다 [48]. 3D 지각영역에서의학습전이 (예: 3D-VLA 에서의포인트클라우드추론) [288]) 는모델에더강한공간귀납편향을부여하여분산외상황에대한견고성을향상시킬수있습니다. 마지막으로, 시뮬레이션-투리얼 (sim2real) 미세조정과도메인무작위화, 동적조명, 텍스처, 물리변화같은실제보정을통해합성환경에서배운정책이물리적로봇에효과적으로전달되도록보장합니다 [4, 66]. 이러한통합전략들은 VLA 가실제배치에서보이지않는물체, 장면, 과제에대해자신있게일반화할수있도록힘을실어줄것입니다.

- **시스템통합복잡도와계산요구.** 좁은컴퓨팅예산하에서멀티모달파이프라인의복잡한조정을관리하기위해, 연구자들은모델모듈화와하드웨어-소프트웨어공동설계를수용해야합니다. 저랭크적응 (LoRA) 어댑터는사전학습된트랜스포머계층에주입할수있어, 코어가중치를변경하지않고도작업별미세조정이가능합니다 [93]. 대규모' 교사' VLA 에서경량' 학생' 네트워크로지식을정제하면서, 학생이교사의중간표상과행동분포를맞추도록유도하는상호정보기반목표에따라 5-10 규모의간결한모델을만듭니다 × 매개변수는적으면서도 90-95 를유지한다% 과제수행에관한내용 [121]. 혼합정밀도양자화와양자화인식훈련을결합하면가중치를 4–8 비트로압축하여메모리대역폭과에너지소비를 60 비트이상줄일수있습니다% [122]. VLA 워크로드에맞춘하드웨어가속기는희소텐서연산, 동적토큰라우팅, 융합된비전-언어커널을지원하며, 20–30W 전력범위내에서지속적인 100+ TOPS 처리량을제공함으로써임베디드로봇플랫폼의요구를충족할수있습니다 [171, 242]. TensorRT-LLM 과같은툴체인 [129] TVM 은특정엣지장치에맞춰종단간 VLA 그래프를최적화하여계층을융합하고정적부분그래프를사전계산할수있습니다. TinyVLA 와같은신흥아키텍처는 1B 이하매개변수 VLA 가실시간추론을통해조작벤치마크에서최첨단성능을달성할수있음을보여주며, 자원이제한된환경에서광범위한배포의길을개척합니다.

- **VLA 배포의환경변동성에대한견고성.** 실제환경에서견고한 VLA 성능을보장하려면환경불확실성과장기적인시스템드리프트를다루기위한목표지향적인기술적개입이필요합니다. UniSim 의폐쇄루프센서시뮬레이터와같은도메인무작위화및합성데이터증강파이프라인은조명, 폐쇄, 센서노이즈에서사실적인변화를생성하여분포변화에대한회복력을향상시킵니다 [264]. 또한, 실시간피드백을기반으로인지의임계값을동적으로조정하고이득을제어하는적응형재보정모듈은센서노후화나작동조건변화로인한성능저하를완화할수있습니다. 이러한접근법들은다양하고변화하는배치시나리오에서 VLA 시스템의안정성과신뢰성을향상시키는것을목표로합니다.

- **VLA 배포시윤리적, 개인정보보호, 사회적고려사항.** 기술적견고성을넘어, VLA 시스템도입은거버넌스지향적해결책이필요한중요한윤리적·사회적도전을제기합니다. 편향감사도구가학습데이터에서왜곡된인구통계학적또는의미적분포를식별하고, 그다음에는적대적편향해소및반사실데이터증강과같은교정전략이필요합니다 [185, 282]. 기기내처리, 민감한데이터스트림에대한동동형암호화, 교육중차등프라이버시보호메커니즘은의료및스마트홈등분야에서사용자데이터를보호하는데매우중요합니다 [160, 180]. 더나아가, 투명한영향평가, 이해관계자참여, 인력역량강화이니셔티브는사회경제적영향을관리하는데도움이되며, 규제체계와업계표준은책임성과책임있는 VLA 채택을보장하는데필수적입니다.

5.2. 미래로드맵

VLA 기반시스템의미래는점점더강력한멀티모달기반, 에이전트추론, 그리고구현된지속적학습이교차하는지점에서진화할것으로예상됩니다. 앞으로 10 년간우리는 VLA 가유능하지만취약한작업전문가에서신뢰할수있고범용적인로봇지능으로전환하는여러수렴추세가나타날것으로예상합니다. 그러나이궤적은앞서강조된지속적인제약및한계에의해형성될것입니다: (i) 폐쇄루프제어에서의실시간추론병목, (ii) 불완전한멀티모달동작표현과약한안전성보장, (iii) 분포이동시데이터셋편향및그라운딩실패, (iv) 인지-메모리-추

론-제어전반에걸친통합복잡성, (v) 엣지배포를방해하는높은컴퓨팅및에너지요구, 그리고 (vi) 견고성과투명성, 그리고오픈월드환경에서의윤리적문제들. 이문제들을전체적으로해결하기위해, Figure 19 시스템수준의연구로드맵을요약하며, 그림 18 VLM, VLA 아키텍처, 에이전트 AI 모듈이로봇공학에서구현된 AGI 로어떻게공진화할수있는지에대한직관적인개념적예시를제공합니다.

멀티모달기초모델은체화된지각의’ 피질’ 으로보입니다:. 오늘날의 VLA 스택은종종비전언어백본과작업별정책헤드에의존하여일반지식의재사용이제한되고도메인간재학습비용이증가합니다. 그럴듯한다음단계는웹규모의이미지, 비디오, 텍스트, 상호작용/어포던스흔적을기반으로훈련된통합멀티모달기초모델을구축하여정적의미론뿐아니라역학, 접촉사전, 상식적물리적지식까지인코딩하는공유’ 피질’ 로기능하는것입니다 [295, 289, 272]. 이러한피질은언어를객체중심표현과지속적인장면구조에그라운딩함으로써피상적상관관계로인한고장모드를줄일수있습니다 [206]. 그림 18에서강조된바와같이, 이기초모델피질은로봇이환경을실행가능한개체 (객체, 영역, 어포던스) 로분할하고, 하위계획자와컨트롤러에게안정적인의미적앵커를제공할수있게할것입니다. 그러나과신과환각적그라운딩를방지하기위해, 이러한모델은보정된불확실성과증거연계추론을포함하여인식기반계획이가림, 혼잡, 모호한지시아래에서도검증가능하도록해야합니다 [145].

에이전트형, 자기감독적, 평생학습및지속적인적응:. 현재 VLA 의결정적인한계는정적인특성입니다: 한번훈련된정책은비고정환경에서작동하더라도변경되지않고배포됩니다. 미래의 VLA 는모델이탐색목표를제안하고, 결과를가설하며, 시뮬레이션및실제롤아웃을통해자기수정하는에이전트학습루프를채택하여수개월또는수년에걸쳐지속적인기술향상을가능하게해야합니다 [43, 90]. 이방향은분포이동, 데이터셋편향, 장기과제취약성을완화하여모델이시간이지남에따라적응할수있도록할것으로기대됩니다; 하지만지속적인정책업데이트는새로운위험을초래합니다. 특히, 반복되는온라인또는점진적학습은이전에습득한역량을덮어쓰는치명적망각, 의도치않은행동회귀를유도하고, 학습된정책을손상시킬수있는소음적이고적대적이거나의도치않은환경피드백에취약해질수있습니다 [292, 263]. 따라서평생학습은재생및안전인식업데이트, 모듈형어댑터, 검증기반정책개정과함께이루어져야합니다 [121]. 그림 19의로드맵에서, 이에이전트형평생학습패러다임은자연스럽게 신뢰가능하고안전한지능과 통합시스템 & 거버넌스의교차점에위치하며, 지속적인학습을임시로미세조정하는단계가아니라통제되고감사가능한생애주기과정으로다루는것입니다.

확장성과해석가능성을위한계층적이고신경상징적인계획:. 저수준모터프리미티브에서장기과제목표로확장하려면명시적인계층구조가필요합니다 [261, 265]. 차세대 VLA 시스템은언어기반플래너 (LLM 스타일모듈, 유연성과제약조건에맞게미세조정됨) 를사용할가능성이높으며, 목표를구조화된하위작업으로분해한뒤, 중간수준의기술정책과준수동작을보장하는저수준컨트롤러로연결합니다 [94, 253]. 이러한신경-상징적혼합은디버깅, 모니터링, 인증이더쉬운인터페이스를부여함으로써통합복잡성을연결하는데도움을줍니다 [201]. 중요한점은계층구조가선택적검증을가능하게한다는것입니다: 상위계획은제약위반 (안전하지않은단계, 금지영역) 을확인할수있습니다 [192, 71] 저수준궤적은제어배리어기능, MPC, 런타임안전모니터로차폐할수있습니다 [188, 196]. 이구성요소들은그림 19의 신뢰가능하고안전한지능축과일치합니다. 여기서안전성은사후평가가아닌계획시간제약과실행시간가드를통해모두강화됩니다 [101, 274].

세계모델과물리적/인과적추론을통한실시간적응:. 비구조화된환경에서견고한배포를위해서는 VLA 가객체, 접촉, 동적요소에대한내부예측모델을유지해야합니다. 단기상태전이와실패가능성을예측하는세계모델은반사실적평가 (’ 여기서밀면무엇이충돌하는가?’) 와현실이기대에서벗어날 때 (예: 미끄러짐, 예상치못한마찰) 신속한교정조치를지원할수있습니다 [203]. 이기능은안전한조작, 항법, 인간-로봇상호작용의핵심으로, 작은오류가빠르게누적되는부분입니다 [156]. 하지만세계모델은탑재시충분히효율적이어야하며, 다중센서증거와일치해야합니다. 따라서하드웨어인식과메모리효율적인예측모델링 (예: 시간토큰압축) 이주요미래방향입니다 [267, 216], 이벤트기반상태업데이트 [250, 226], 그리고미분가능한물리시뮬레이터와학습된동역학을결합하여제어관련업데이트속도를가능하게하는하이브리드물리학습모델입니다 [102, 52]. 그림 19에서이러한필요성은 효율적인배포축 (실시간제약) 과 신뢰가능하고안전한지능축 (그라운딩된의사결정을위한물리·인과추론) 에함께제시됩니다.

효율성과확장성: 일반성과엣지배포의다리:. VLA 채택의중심장벽은대형멀티모달백본의계산용량과폐쇄루프제어의지연/에너지제약간의불일치입니다 [268, 238]. 향후 VLA 는일반화를유지하면서추론비용을줄이면서매개변수효율적인설계 (구조적희소성, 저랭크적응, 모듈러전문가) 를우선시해야합니다 [65, 197]. 훈련시간효율성을넘어, 언제/조기종료정책은계산을적응적으로할당하여안전중요단계가높은충실도를유지하는반면, 일상단계는더저렴한경로를사용할수있습니다 [41, 274]. 동등하게중요한것은액션공간효율성입니다: 컴팩트한액션토큰화와청크된제어표현은자기회귀구간을단축하여시간적부드러움을희생하지않으면서도더높은제어속도를가능하게합니다 [171]. 이러한모델수준의선택은 GPU/NPU/엣지가속기전반에걸친하드웨어인식컴파일과함께이루어져야하며, 양자화인식스케줄링과메모리최적화주의커널등이포함됩니다 [65, 167]. 컴퓨팅인식캐싱과에피소드메모리는중복순방향패스를줄이고장기과제작업에서반응성을향상시킵니다 [27, 133]. 이러한방향은그림 19의 효율적인배포축을구체화하며앞서강조한컴퓨팅/에너지한계를직접해결할수있습니다.

신체간전이및형태중립적기술표현:. 각로봇형태에대해별도의 VLA 를훈련시키는방식은확장성이낮을가능성이큽니다. 미래의핵심주제는형태중립적정책학습으로, 기술이추상적행동공간 (예: 접촉목표, affordance-point 조작, 작업공간제약) 에서표현되어바퀴달린플랫폼, 네발보행, 휴머노이드간에전이됩니다 [275, 118]. 메타러닝과 few-shot 보정은몇주간의훈련이아닌몇분의데이터로신형로봇을빠르게부트스트랩할수있게해줍니다 [35]. 이방향은구현체와환경간에불변성을강제함으로써데이터셋편향을완화하지만, 표준화된표현, 공통인터페이스, 재현가능한평가프로토콜을요구합니다 [118, 275]. 그림 19에서이형태중립적기술학습패러다임

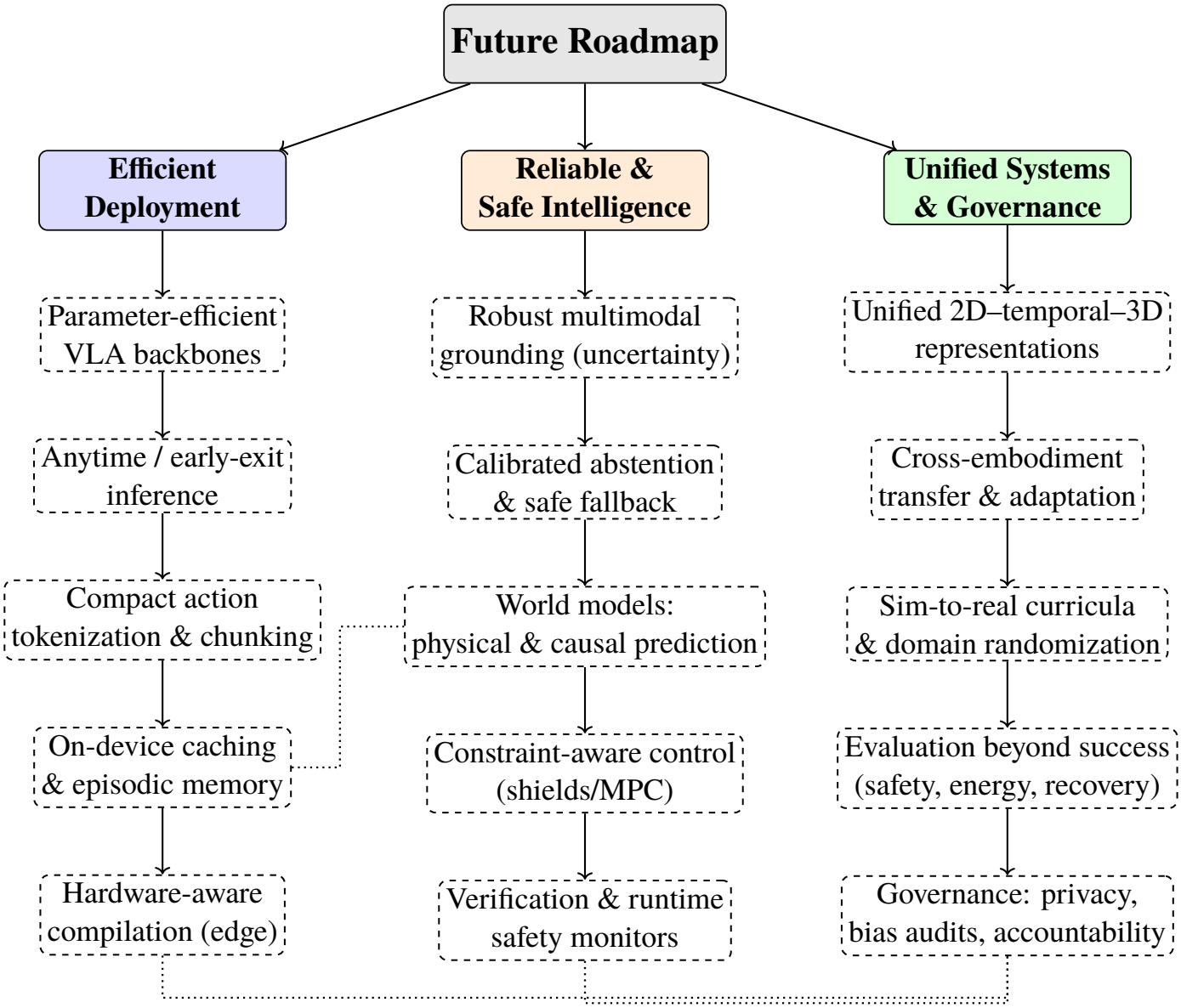

Figure 19: **비전-언어-액션 (VLA) 모델의향후연구로드맵 (그림 19).** 점선상자는이작업에서확인된 VLA 모델의핵심한계를해결하기위한우선순위방향을나타냅니다: (i) 효율적인배포 (지연, 계산, 에너지), (ii) 신뢰가능하고안전한지능 (그라운딩, 불확실성, 안전보장), 그리고 (iii) 통합시스템 *&* 거버넌스 (2D-시간-3D 통합, 이전, 평가및책임있는배포).

은 통합시스템 *&* 거버넌스에속하며, 아키텍처통합과원칙적전이및벤치마킹을연결합니다.

과제성공을넘어선평가*:* 안전, 회복, 자원인지지표*:*. VLA 의발전은배포현실을반영하는측정이필요합니다. 과제성공만으로는실패의심각성, 시간적불일치, 위험한근접실패, 에너지비효율을명확히나타내지않습니다. 향후벤치마크는인간의제약하에서안전위반, 불확실성보정, 회복행동, 시간적일관성, 에너지소비, 그리고다운스트림효용를정량화해야합니다 [274, 78]. 더불어, 평가는공정한비교와편향기반성과를진단하기위해컴퓨팅예산, 데이터셋구성, 배포조건을보고해야합니다 [215, 273, 118]. 이러한측정은단순한과학적위생이아니라감사와거버넌스의기초이며, 시스템이대규모로책임감있게배포될수있는지를결정합니다 [274, 275]. 이동기는그림평가분기의근간입니다 19 그리고그림에나오는개념적배치서사를보완합니다 18.

안전, 윤리, 인간중심의정렬을일류설계목표로삼는것*:*. VLA 가자율성을얻으면서내장된안전성과가치정렬이매우중요해집니다. 향후시스템에는고위험행동을실행하기전에잠재적피해를평가하는실시간위험추정기를통합하고, 모호성있는자연어확인을요청하며, 책임성을위한투명한로그를유지해야합니다 [274, 118]. 특히보조로봇과안전에중요한자율성에있어프라이버시인식감지, 편향감사, 그리고인간개입 (Human-in-the-loop) 감독이생명주기에통합되어야합니다 [284, 235]. 규제부합의평가프로토콜과표준화노력이 VLA 의진전을신뢰할수있는실제시스템에적용하는데필수적입니다 [169, 294, 234]. 이거버넌스관점은그림에서명확히담겨있습니다 19 그리고이는그림에묘사된사회적인식이강한인간형배경에암묵적으로반영된다 18.

교차주제*:* 지속적인학습*,* 고장복구*,* 상호작용*,* 그리고제어충실도*:*. 그림에나오는모든축 19 여러교차주제들이향후 10 년간 VLA 연구를형성할것으로예상됩니다. 첫째, 지속적이고평생학습은안전하고감사가능하며잊히지않아야하며, 배포된시스템을불안정하게하지않으면서장기적인적응을가능하게해야합니다 [290]. 둘째, 실패감지및복구는내성적모니터링, 불확실성인식, 실행결과와다르게발생할때구조화된복구행동을포함하는일류역량으로다뤄져야합니다 [116, 259]. 셋째, 액션생성의정밀도와신뢰성향상이여전히중요합니다. VLA 기반플래너는유연하고언어조건화된의사결정을가능하게하지만, 궤적정확도와제어안정성은현재모델예측제어, 샘플링기반모션계획, 피드백선형제어등기존의분석접근법에비해뒤처져있습니다. 그결과, VLA 기반고수준계획과고전적또는학습된저수준제어기를결합한하이브리드아키텍처는의미론적유연성과제어수준의정밀도를모두달성하는데중심적인역할을할것으로기대됩니다. 마지막으로, 인간의정렬과상호작용은의도명확화, 공유자율성, 설명가능한행동그라운딩을위한메커니즘을필요로하며, 이는다양한실제환경에서신뢰와사용성을지원합니다 [111, 110].

요약하자면, 그림 19 실험실시연과견고한실제배치간의격차를해소하려면효율성, 안전성, 데이터및일반화, 시스템통합, 평가및거버넌스에서조정된발전이필요함을강조합니다. 보완적으로그림 18 이러한발전이모바일로봇, 조작기, 보조시스템, 휴머노이드등다중플랫폼에서멀티모달지각, 계층적계획, 지속적인적응, 인간친화적안전이통합된정보스택내에통합된범용체화된에이전트로수렴할수있음을보여줍니다. 이방향들을함께해결하면 VLA 가유망한연구프로토타입에서단일로봇형태에국한된솔루션이아닌신뢰할수있고폭넓게적용가능한내장시스템으로변화할것으로기대됩니다.

6. 결론

이종합리뷰에서우리는지난 3 년간발표된비전-언어-액션 (VLA) 모델의최근발전, 방법론및적용을체계적으로평가했습니다. 우리의분석은 VLA 의기초개념에서시작하여, 물리적또는시뮬레이션환경에서시각지각, 자연어이해, 액션생성을통합하는멀티모달시스템으로서이들의역할을정의했습니다. 우리는이들의진화와연대를추적하며, 고립된지각-행동모듈에서완전히통합된지시를따르는로봇에이전트로의전환을나타내는주요이정표들을상세히설명했습니다. 우리는멀티모달통합이느슨하게결합된파이프라인에서양식간원활한조정을가능하게하는트랜스포머기반아키텍처로성숙해졌다는점을강조했습니다.

다음으로, VLA 가행동원시및공간의미를포함한시각적및언어적정보를어떻게인코딩하는지에초점을맞춰토큰화와표현기법을살펴보았습니다. 학습패러다임을탐구하며, VLA 성능을형성한학습대상학습, 모방학습부터강화학습, 멀티모달사전학습에이르기까지다양한데이터셋과훈련전략을상세히설명했습니다. ’ 적응형제어및실시간실행’ 섹션에서는현대 VLA 가동적환경에최적화되고, 지연시간에민감한작업을지원하는정책을분석했습니다. 그후주요아키텍처혁신을분류하여 50 개이상의최근 VLA 모델을조사했습니다. 이논의에는모델설계, 메모리시스템, 상호작용충실도의발전이포함되었습니다. 또한 LoRA, 양자화, 모델가지치기와같은매개변수효율적인방법과병렬디코딩, 하드웨어인식추론과같은가속기법을포함한훈련효율성향상전략을연구했습니다. 실제응용분석은휴머노이드로봇, 자율주행차, 산업자동화, 의료, 농업, 증강현실 (AR) 내비게이션등 6 개분야에서 VLA 모델의가능성과현재의한계를모두강조했습니다. 이러한환경에서 VLA 는고수준의미론적추론, 지시-따르기, 과제일반화에서특히구조화된또는부분적으로통제된환경에서강력한능력을보여주었습니다. 그러나실시간추론지연, 환경변동성에서의견고성제한, 그리고기존의분석계획및제어파이프라인에비해장기또는안전에중요한제어의정밀도가떨어지는점이효과적으로제약을받았습니다. 더불어, 안정적인성능을달성하기위해애플리케이션별적응과광범위한데이터선별이자주필요했는데, 이는확장성과배포의어려움을부각시켰습니다. 이결과들은 VLA 가의미론적의사결정과유연한작업명세에적합하지만, VLA 추론과고전적또는학습된저수준제어기를통합한하이브리드아키텍처가실용적이고현실적인작동에여전히필수적임을시사합니다.

도전과한계를해결하기위해우리는실시간추론, 멀티모달동작표현및안전성, 편향과일반화, 시스템통합및계산제약, 윤리적배포라는다섯가지핵심영역에집중했습니다. 우리는모델압축, 교차모달그라운딩, 도메인적응, 에이전트학습프레임워크등최신문헌에서잠재적해결책을제안했습니다. 마지막으로, 우리의논의와향후로드맵에서는 VLM, VLA 아키텍처, 에이전트 AI 시스템의융합이로봇공학을인공일반지능 (AGI) 으로이끌고있음을명확히했습니다. 이검토는 VLA 발전에대한통합된이해를제공하고, 해결되지않은도전과제를식별하며, 미래에지능적이고체체적이며인간친화적에이전트를개발하기위한체계적인경로를제시합니다.

자금지원선언

이연구는미국국립과학재단 (NSF) 과미국농무부 (USDA), 국립식품농업연구소 (NIFA) 의’ 인공지능 (AI) 농업연구소’ 프로그램을통해’ 로봇소프트매니퓰레이터를이용한로봇꽃솎음’ 프로젝트의 AWD003473 및 AWD004595 및 USDA-NIFA 등록번호 1029004 지원을받았습니다. 추가지원은’ExPanding UCF AI 연구를새로운농업공학응용으로확장 (PARTNER)’ 프로젝트하에 USDA/NIFA 보조금번호 2024-67022-41788, 등록번호 1031712 를통해제공되었습니다.

선언문

저자들은이해상충이없음을선언합니다.

AI 글쓰기지원에관한성명서

ChatGPT 와 Perplexity 는문법정확성을높이고문장구조를다듬기위해활용되었으며; 모든 AI 생성수정본은관련성을위해철저히검토되고편집되었습니다.

References

[1] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al., 2023. Gpt-4 technical report. arXiv preprint arXiv:2303.08774 .

[2] Agarwal, L., Verma, B., 2024. From methods to datasets: A survey on image-caption generators. Multimedia Tools and Applications 83, 28077–28123.

[3] Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al., 2022. Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems 35, 23716–23736.

[4] Anderson, P., Shrivastava, A., Truong, J., Majumdar, A., Parikh, D., Batra, D., Lee, S., 2021. Sim-to-real transfer for vision-and-language navigation, in: Conference on Robot Learning, PMLR. pp. 671–681.

[5] Arai, H., Miwa, K., Sasaki, K., Watanabe, K., Yamaguchi, Y., Aoki, S., Yamamoto, I., 2025. Covla: Comprehensive vision-language-action dataset for autonomous driving, in: 2025 IEEE/ CVF Winter Conference on Applications of Computer Vision (WACV), IEEE. pp. 1933–1943.

[6] Asif, S., Bueno, M., Ferreira, P., Anandan, P., Zhang, Z., Yao, Y., Ragunathan, G., Tinkler, L., Sotoodeh-Bahraini, M., Lohse, N., et al., 2025. Rapid and automated configuration of robot manufacturing cells. Robotics and Computer-Integrated Manufacturing 92, 102862.

[7] Assres, G., Bhandari, G., Shalaginov, A., Gronli, T.M., Ghinea, G., 2025. State-of-the-art and challenges of engineering ml-enabled software systems in the deep learning era. ACM Computing Surveys .

[8] Asuzu, K., Singh, H., Idrissi, M., 2025. Human–robot interaction through joint robot planning with large language models. Intelligent Service Robotics , 1–17.

[9] Ayaz, M., Khan, M., Saqib, M., Khelifi, A., Sajjad, M., Elsaddik, A., 2024. Medvlm: Medical vision-language model for consumer devices. IEEE Consumer Electronics Magazine .

[10] Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J., 2025. Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 .

[11] Bartoccioni, F., Ramzi, E., Besnier, V., Venkataramanan, S., Vu, T.H., Xu, Y., Chambon, L., Gidaris, S., Odabas, S., Hurych, D., et al., 2025. Vavim and vavam: Autonomous driving through video generative modeling. arXiv preprint arXiv:2502.15672 .

[12] Bathula, N.V., Paleti, I., Pagidi, S., Akkumahanthi, S.S., Guduru, N.T., 2024. Policy learning-based image captioning with vision transformer, in: 2024 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS), IEEE. pp. 1–6.

[13] Belkhale, S., Ding, T., Xiao, T., Sermanet, P., Vuong, Q., Tompson, J., Chebotar, Y., Dwibedi, D., Sadigh, D., 2024. Rt-h: Action hierarchies using language. arXiv preprint arXiv:2403.01823 .

[14] Bjorck, J., Castañeda, F., Cherniadev, N., Da, X., Ding, R., Fan, L., Fang, Y., Fox, D., Hu, F., Huang, S., et al., 2025. Gr00t n1: An open foundation model for generalist humanoid robots. arXiv preprint arXiv:2503.14734 .

[15] Black, K., Brown, N., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., Groom, L., Hausman, K., Ichter, B., et al., 2024. Pi-0: A vision-language-action flow model for general robot control. arXiv preprint arXiv:2410.24164 .

[16] Bolya, D., Huang, P.Y., Sun, P., Cho, J.H., Madotto, A., Wei, C., Ma, T., Zhi, J., Rajasegaran, J., Rasheed, H., et al., 2025. Perception encoder: The best visual embeddings are not at the output of the network. arXiv preprint arXiv:2504.13181 .

[17] Bordes, F., Pang, R.Y., Ajay, A., Li, A.C., Bardes, A., Petryk, S., Mañas, O., Lin, Z., Mahmoud, A., Jayaraman, B., et al., 2024. An introduction to vision-language modeling. arXiv preprint arXiv:2405.17247 .

[18] Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Chen, X., Choromanski, K., Ding, T., Driess, D., Dubey, A., Finn, C., et al., 2023. Rt-2: Vision-language-action models transfer web knowledge to robotic control. arXiv preprint arXiv:2307.15818 .

[19] Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., Finn, C., Gopalakrishnan, K., Hausman, K., Herzog, A., Hsu, J., et al., 2022. Rt-1: Robotics transformer for real-world control at scale. arXiv preprint arXiv:2212.06817 .

[20] Budzianowski, P., Maa, W., Freed, M., Mo, J., Xie, A., Tipnis, V., Bolte, B., 2024. Edgevla: Efficient vision-language-action models. environments 20, 3.

[21] Cangelosi, A., Metta, G., Sagerer, G., Nolfi, S., Nehaniv, C., Fischer, K., Tani, J., Belpaeme, T., Sandini, G., Nori, F., et al., 2010. Integration of action and language knowledge: A roadmap for developmental robotics. IEEE Transactions on Autonomous Mental Development 2, 167–195.

[22] Cao, J., Gan, Z., Cheng, Y., Yu, L., Chen, Y.C., Liu, J., 2020. Behind the scene: Revealing the secrets of pre-trained vision-and-language models, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16, Springer. pp. 565–580.

[23] Cao, L., 2024. Ai robots and humanoid ai: Review, perspectives and directions. arXiv preprint arXiv:2405.15775 .

[24] Cao, Y., Ju, Y., Xu, D., 2024a. 3dgs-det: Empower 3d gaussian splatting with boundary guidance and box-focused sampling for 3d object detection. arXiv preprint arXiv:2410.01647 .

[25] Cao, Y., Zeng, Y., Xu, H., Xu, D., 2023. Coda: Collaborative novel box discovery and cross-modal alignment for open-vocabulary 3d object detection, in: NeurIPS.

[26] Cao, Y., Zeng, Y., Xu, H., Xu, D., 2024b. Collaborative novel object discovery and box-guided cross-modal alignment for open-vocabulary 3d object detection. arXiv preprint arXiv:2406.00830 .

[27] Chang, C., Shi, Y., Cao, D., Yang, W., Hwang, J., Wang, H., Pang, J., Wang, W., Liu, Y., Peng, W.C., et al., 2025. A survey of reasoning and agentic systems in time series with large language models. arXiv preprint arXiv:2509.11575 .

[28] Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., et al., 2024. A survey on evaluation of large language models. ACM transactions on intelligent systems and technology 15, 1–45.

[29] Chatzopoulos, D., Bermejo, C., Huang, Z., Hui, P., 2017. Mobile augmented reality survey: From where we are to where we go. Ieee Access 5, 6917–6950.

[30] Chen, B., Xu, Z., Kirmani, S., Ichter, B., Sadigh, D., Guibas, L., Xia, F., 2024a. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14455–14465.

[31] Chen, H., Hou, L., Wu, S., Zhang, G., Zou, Y., Moon, S., Bhuiyan, M., 2024b. Augmented reality, deep learning and vision-language query system for construction worker safety. Automation in Construction 157, 105158.

[32] Chen, H., Li, S., Fan, J., Duan, A., Yang, C., Navarro-Alarcon, D., Zheng, P., 2025a. Human-in-the-loop robot learning for smart manufacturing: A human-centric perspective. IEEE Transactions on Automation Science and Engineering .

[33] Chen, H., Liu, B., Wang, S., Wang, X., Han, W., Zhu, Y., Wang, X., Bi, Y., 2025b. Language modulates vision: Evidence from neural networks and human brain-lesion models. arXiv preprint arXiv:2501.13628 .

[34] Chen, P., Bu, P., Wang, Y., Wang, X., Wang, Z., Guo, J., Zhao, Y., Zhu, Q., Song, J., Yang, S., et al., 2025c. Combatvla: An efficient vision-language-action model for combat tasks in 3d action role-playing games. arXiv preprint arXiv:2503.09527 .

[35] Chen, X., Xu, W., Kan, S., Zhang, L., Jin, Y., Cen, Y., Li, Y., 2025d. Vision-semantics-label: A new two-step paradigm for action recognition with large language model. IEEE Transactions on Circuits and Systems for Video Technology .

[36] Chen, Y., Tian, S., Liu, S., Zhou, Y., Li, H., Zhao, D., 2025e. Conrft: A reinforced fine-tuning method for vla models via consistency policy. arXiv preprint arXiv:2502.05450 .

[37] Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al., 2024c. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 24185–24198.

[38] Cheng, A.C., Ji, Y., Yang, Z., Gongye, Z., Zou, X., Kautz, J., Bıyık, E., Yin, H., Liu, S., Wang, X., 2024a. Navila: Legged robot vision-language-action model for navigation. arXiv preprint arXiv:2412.04453 .

[39] Cheng, H., Xiao, E., Yu, C., Yao, Z., Cao, J., Zhang, Q., Wang, J., Sun, M., Xu, K., Gu, J., et al., 2024b. Manipulation facing threats: Evaluating physical vulnerabilities in end-to-end vision language action models. arXiv preprint arXiv:2409.13174 .

[40] Chi, C., Xu, Z., Feng, S., Cousineau, E., Du, Y., Burchfiel, B., Tedrake, R., Song, S., 2023. Diffusion policy: Visuomotor policy learning via action diffusion. The International Journal of Robotics Research , 02783649241273668.

[41] Chi, H., Gao, H.a., Liu, Z., Liu, J., Liu, C., Li, J., Yang, K., Yu, Y., Wang, Z., Li, W., et al., 2025. Impromptu vla: Open weights and open data for driving vision-language-action models. arXiv preprint arXiv:2505.23757 .

[42] Chiang, H.T.L., Xu, Z., Fu, Z., Jacob, M.G., Zhang, T., Lee, T.W.E., Yu, W., Schenck, C., Rendleman, D., Shah, D., et al., 2024. Mobility vla: Multimodal instruction navigation with long-context vlms and topological graphs. arXiv preprint arXiv:2407.07775 .

[43] Chowa, S.S., Alvi, R., Rahman, S.S., Rahman, M.A., Raiaan, M.A.K., Islam, M.R., Hussain, M., Azam, S., 2026. From language to action: a review of large language models as autonomous agents and tool users. Artificial Intelligence Review .

[44] Dang, R., Yuan, Y., Zhang, W., Xin, Y., Zhang, B., Li, L., Wang, L., Zeng, Q., Li, X., Bing, L., 2025. Ecbench: Can multi-modal foundation models understand the egocentric world? a holistic embodied cognition benchmark. arXiv preprint arXiv:2501.05031 .

[45] Dasari, S., Ebert, F., Tian, S., Nair, S., Bucher, B., Schmeckpeper, K., Singh, S., Levine, S., Finn, C., 2019. Robonet: Large-scale multi-robot learning. arXiv preprint arXiv:1910.11215 .

[46] Deng, S., Yan, M., Wei, S., Ma, H., Yang, Y., Chen, J., Zhang, Z., Yang, T., Zhang, X., Cui, H., Zhang, Z., Wang, H., 2025a. Graspvla: a grasping foundation model pre-trained on billion-scale synthetic action data. URL: https://arxiv.org/abs/2505.03233, arXiv:2505.03233.

[47] Deng, S., Yan, M., Zheng, Y., Su, J., Zhang, W., Zhao, X., Cui, H., Zhang, Z., Wang, H., 2025b. Stereovla: Enhancing vision-language-action models with stereo vision. arXiv preprint arXiv:2512.21970 .

[48] Dey, S., Zaech, J.N., Nikolov, N., Van Gool, L., Paudel, D.P., 2024. Revla: Reverting visual domain limitation of robotic foundation models. arXiv preprint arXiv:2409.15250 .

[49] Din, M.U., Akram, W., Saoud, L.S., Rosell, J., Hussain, I., 2025. Vision language action models in robotic manipulation: A systematic review. arXiv preprint arXiv:2507.10672 .

[50] Ding, D., Yao, T., Luo, R., Sun, X., 2025a. Visual question answering in robotic surgery: A comprehensive review. IEEE Access .

[51] Ding, J., Zhang, Y., Shang, Y., Zhang, Y., Zong, Z., Feng, J., Yuan, Y., Su, H., Li, N., Sukiennik, N., et al., 2024a. Understanding world or predicting future? a comprehensive survey of world models. arXiv preprint arXiv:2411.14499 .

[52] Ding, M., Chen, Z., Du, T., Luo, P., Tenenbaum, J., Gan, C., 2021. Dynamic visual reasoning by learning differentiable physics models from video and language. Advances in Neural Information Processing Systems 34, 887–899.

[53] Ding, P., Ma, J., Tong, X., Zou, B., Luo, X., Fan, Y., Wang, T., Lu, H., Mo, P., Liu, J., et al., 2025b. Humanoid-vla: Towards universal humanoid control with visual integration. arXiv preprint arXiv:2502.14795 .

[54] Ding, P., Zhao, H., Zhang, W., Song, W., Zhang, M., Huang, S., Yang, N., Wang, D., 2024b. Quar-vla: Vision-language-action model for quadruped robots, in: European Conference on Computer Vision, Springer. pp. 352–367.

[55] Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T., 2015. Long-term recurrent convolutional networks for visual recognition and description, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2625–2634.

[56] Dong, H., Liu, M., Zhou, K., Chatzi, E., Kannala, J., Stachniss, C., Fink, O., 2025. Advances in multimodal adaptation and generalization: From traditional approaches to foundation models. arXiv preprint arXiv:2501.18592 .

[57] Doveh, S., Arbelle, A., Harary, S., Schwartz, E., Herzig, R., Giryes, R., Feris, R., Panda, R., Ullman, S., Karlinsky, L., 2023. Teaching structured vision & language concepts to vision & language models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2657–2668.

[58] Driess, D., Xia, F., Sajjadi, M.S., Lynch, C., Chowdhery, A., Wahid, A., Tompson, J., Vuong, Q., Yu, T., Huang, W., et al., 2023. Palm-e: An embodied multimodal language model. Openreview .

[59] Duan, J., Pumacay, W., Kumar, N., Wang, Y.R., Tian, S., Yuan, W., Krishna, R., Fox, D., Mandlekar, A., Guo, Y., 2024. Aha: A vision-language-model for detecting and reasoning over failures in robotic manipulation. arXiv preprint arXiv:2410.00371 .

[60] Duarte, N.F., Raković, M., Tasevski, J., Coco, M.I., Billard, A., Santos-Victor, J., 2018. Action anticipation: Reading the intentions of humans and robots. IEEE Robotics and Automation Letters 3, 4132–4139.

[61] Ebert, F., Yang, Y., Schmeckpeper, K., Bucher, B., Georgakis, G., Daniilidis, K., Finn, C., Levine, S., 2021. Bridge data: Boosting generalization of robotic skills with cross-domain datasets. arXiv preprint arXiv:2109.13396 .

[62] Fan, C., Jia, X., Sun, Y., Wang, Y., Wei, J., Gong, Z., Zhao, X., Tomizuka, M., Yang, X., Yan, J., Ding, M., 2025a. Interleave-vla: Enhancing robot manipulation with interleaved image-text instructions. URL: https://arxiv.org/abs/2505.02152, arXiv:2505.02152.

[63] Fan, L., Chen, K., Xu, Z., Yuan, M., Huang, P., Huang, W., 2024. Language reasoning in vision-language-action model for robotic grasping, in: 2024 China Automation Congress (CAC), IEEE. pp. 6656–6661.

[64] Fan, Y., Ding, P., Bai, S., Tong, X., Zhu, Y., Lu, H., Dai, F., Zhao, W., Liu, Y., Huang, S., et al., 2025b. Long-vla: Unleashing long-horizon capability of vision language action model for robot manipulation. arXiv preprint arXiv:2508.19958 .

[65] Fang, H., Liu, Y., Du, Y., Du, L., Yang, H., 2025a. Sqap-vla: A synergistic quantization-aware pruning framework for high-performance vision-language-action models. arXiv preprint arXiv:2509.09090 .

[66] Fang, Y., Yang, Y., Zhu, X., Zheng, K., Bertasius, G., Szafir, D., Ding, M., 2025b. Rebot: Scaling robot learning with real-to-sim-to-real robotic video synthesis. arXiv preprint arXiv:2503.14526 .

[67] Firoozi, R., Tucker, J., Tian, S., Majumdar, A., Sun, J., Liu, W., Zhu, Y., Song, S., Kapoor, A., Hausman, K.,

et al., 2023. Foundation models in robotics: Applications, challenges, and the future. The International Journal of Robotics Research , 02783649241281508.

[68] Foster, D.J., Block, A., Misra, D., 2024. Is behavior cloning all you need? understanding horizon in imitation learning. arXiv preprint arXiv:2407.15007 .

[69] Fu, H., Zhang, D., Zhao, Z., Cui, J., Liang, D., Zhang, C., Zhang, D., Xie, H., Wang, B., Bai, X., 2025. Orion: A holistic end-to-end autonomous driving framework by vision-language instructed action generation. arXiv preprint arXiv:2503.19755 .

[70] Gao, B., Liu, Y., Li, Y., Li, H., Li, M., He, W., 2025a. A vision-language model for predicting potential distribution land of soybean double cropping. Frontiers in Environmental Science 12, 1515752.

[71] Gao, C., Liu, Z., Chi, Z., Huang, J., Fei, X., Hou, Y., Zhang, Y., Lin, Y., Fang, Z., Jiang, Z., et al., 2025b. Vlaos: Structuring and dissecting planning representations and paradigms in vision-language-action models. arXiv preprint arXiv:2506.17561 .

[72] Gao, J., Belkhale, S., Dasari, S., Balakrishna, A., Shah, D., Sadigh, D., 2025c. A taxonomy for evaluating generalist robot policies. arXiv preprint arXiv:2503.01238 .

[73] Gao, S.H., Cheng, M.M., Zhao, K., Zhang, X.Y., Yang, M.H., Torr, P., 2019. Res2net: A new multi-scale backbone architecture. IEEE TPAMI .

[74] Gbagbe, K.F., Cabrera, M.A., Alabbas, A., Alyunes, O., Lykov, A., Tsetserukou, D., 2024. Bi-vla: Vision-language-action model-based system for bimanual robotic dexterous manipulations, in: 2024 IEEE International Conference on Systems, Man, and Cybernetics (SMC), IEEE. pp. 2864–2869.

[75] Geens, R., 2024. Bringing generative ai to edge devices through interoperable compute cores, in: Flanders AI Research Day, Location: Ghent.

[76] Ghosh, A., Acharya, A., Saha, S., Jain, V., Chadha, A., 2024. Exploring the frontier of vision-language models: A survey of current methodologies and future directions. arXiv preprint arXiv:2404.07214 .

[77] Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., et al., 2018. Recent advances in convolutional neural networks. Pattern recognition 77, 354–377.

[78] Gu, Q., Ju, Y., Sun, S., Gilitschenski, I., Nishimura, H., Itkina, M., Shkurti, F., 2025a. Safe: Multitask failure detection for vision-language-action models. arXiv preprint arXiv:2506.09937 .

[79] Gu, Q., Su, J., Yuan, L., 2021. Visual affordance detection using an efficient attention convolutional neural network. Neurocomputing 440, 36–44. doi:https://doi.org/10.1016/j.neucom.2021.01.018.

[80] Gu, X., Wen, C., Ye, W., Song, J., Gao, Y., 2023. Seer: Language instructed video prediction with latent diffusion models. arXiv preprint arXiv:2303.14897 .

[81] Gu, Z., Li, J., Shen, W., Yu, W., Xie, Z., McCrory, S., Cheng, X., Shamsah, A., Griffin, R., Liu, C.K., et al., 2025b. Humanoid locomotion and manipulation: Current progress and challenges in control, planning, and learning. arXiv preprint arXiv:2501.02116 .

[82] Guan, W., Hu, Q., Li, A., Cheng, J., 2025. Efficient vision-language-action models for embodied manipulation: A systematic survey. arXiv preprint arXiv:2510.17111 .

[83] Guo, Y., Zhang, J., Chen, X., Ji, X., Wang, Y.J., Hu, Y., Chen, J., 2025. Improving vision-language-action model with online reinforcement learning. arXiv preprint arXiv:2501.16664 .

[84] Guruprasad, P., Sikka, H., Song, J., Wang, Y., Liang, P.P., 2024. Benchmarking vision, language, & action models on robotic learning tasks. arXiv preprint arXiv:2411.05821 .

[85] Haldar, S., Peng, Z., Pinto, L., 2024. Baku: An efficient transformer for multi-task policy learning. arXiv preprint arXiv:2406.07539 .

[86] Han, S., Wang, M., Zhang, J., Li, D., Duan, J., 2024. A review of large language models: Fundamental architectures, key technological evolutions, interdisciplinary technologies integration, optimization and compression techniques, applications, and challenges. Electronics 13, 5040.

[87] Hanson, A., Riseman, E., 2014. The visions image-understanding system, in: Advances in Computer Vision. Psychology Press, pp. 1–114.

[88] Hao, P., Zhang, C., Li, D., Cao, X., Hao, X., Cui, S., Wang, S., 2025. Tla: Tactile-language-action model for contact-rich manipulation. arXiv preprint arXiv:2503.08548 .

[89] He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, pp. 770–778.

[90] Holldack, F., Banh, L., Strobel, G., 2026. Agentic information systems. Electronic Markets 36, 5.

[91] Hong, Y., 2025. Building 3D Foundation Models for the Embodied Minds. Ph.D. thesis. University of California, Los Angeles.

[92] Hou, Z., Zhang, T., Xiong, Y., Duan, H., Pu, H., Tong, R., Zhao, C., Zhu, X., Qiao, Y., Dai, J., et al., 2025. Dita: Scaling diffusion transformer for generalist vision-language-action policy. arXiv preprint arXiv:2503.19757 .

[93] Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al., 2022. Lora: Low-rank adaptation of large language models. ICLR 1, 3.

[94] Hu, Y., Tang, J., Gong, X., Zhou, Z., Zhang, S., Elvitigala, D.S., Mueller, F., Hu, W., Quigley, A.J., 2025. Vision-based multimodal interfaces: A survey and taxonomy for enhanced context-aware system design. arXiv preprint arXiv:2501.13443 .

[95] Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q., 2017. Densely connected convolutional networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4700–4708.

[96] Huang, H., Liu, F., Fu, L., Wu, T., Mukadam, M., Malik, J., Goldberg, K., Abbeel, P., 2024. Early fusion helps vision language action models generalize better, in: 1st Workshop on X-Embodiment Robot Learning.

[97] Huang, H., Liu, F., Fu, L., Wu, T., Mukadam, M., Malik, J., Goldberg, K., Abbeel, P., 2025a. Otter: A vision-language-action model with text-aware visual feature extraction. arXiv preprint arXiv:2503.03734 .

[98] Huang, S., Dong, L., Wang, W., Hao, Y., Singhal, S., Ma, S., Lv, T., Cui, L., Mohammed, O.K., Patra, B., et al., 2023a. Language is not all you need: Aligning perception with language models. Advances in Neural Information Processing Systems 36, 72096–72109.

[99] Huang, W., Gu, Q., Ye, N., 2025b. Decision spikeformer: Spike-driven transformer for decision making. arXiv preprint arXiv:2504.03800 .

[100] Huang, W., Wang, C., Zhang, R., Li, Y., Wu, J., Fei-Fei, L., 2023b. Voxposer: Composable 3d value maps for robotic manipulation with language models. arXiv preprint arXiv:2307.05973 .

[101] Huang, Y., Hua, H., Zhou, Y., Jing, P., Nagireddy, M., Padhi, I., Dolcetti, G., Xu, Z., Chaudhury, S., Rawat, A., et al., 2025c. Building a foundational guardrail for general agentic systems via synthetic data. arXiv preprint arXiv:2510.09781 .

[102] Huang, Z., Chen, F., Pu, Y., Lin, C., Su, H., Gan, C., 2023c. Diffvl: Scaling up soft body manipulation using vision-language driven differentiable physics. Advances in Neural Information Processing Systems 36, 29875–29900.

[103] Hung, C.Y., Sun, Q., Hong, P., Zadeh, A., Li, C., Tan, U., Majumder, N., Poria, S., et al., 2025. Nora: A small open-sourced generalist vision language action model for embodied tasks. arXiv preprint arXiv:2504.19854 .

[104] Ikeda, B., Gramopadhye, M., Nekervis, L., Szafir, D., 2025. Marcer: Multimodal augmented reality for composing and executing robot tasks, in: 2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI), IEEE. pp. 529–539.

[105] Imran, A., Gopalakrishnan, K., 2025. Foundation models in robotics, in: AI for Robotics: Toward Embodied and General Intelligence in the Physical World. Springer, pp. 139–210.

[106] Intelligence, P., Black, K., Brown, N., Darpinian, J., Dhabalia, K., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., et al., 2025. pi0.5: a vision-language-action model with open-world generalization. arXiv preprint arXiv:2504.16054 .

[107] Jeong, H., Lee, H., Kim, C., Shin, S., 2024. A survey of robot intelligence with large language models. Applied Sciences 14, 8868.

[108] Jha, K., Doshi, A., Patel, P., Shah, M., 2019. A comprehensive review on automation in agriculture using artificial intelligence. Artificial Intelligence in Agriculture 2, 1–12.

[109] Jiang, J., Xiao, W., Lin, Z., Zhang, H., Ren, T., Gao, Y., Lin, Z., Cai, Z., Yang, L., Liu, Z., 2024. Solami: Social vision-language-action modeling for immersive interaction with 3d autonomous characters. arXiv preprint arXiv:2412.00174 .

[110] Jiang, J., Xiao, W., Lin, Z., Zhang, H., Ren, T., Gao, Y., Lin, Z., Cai, Z., Yang, L., Liu, Z., 2025a. Solami: Social vision-language-action modeling for immersive interaction with 3d autonomous characters, in: Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 26887–26898.

[111] Jiang, S., Huang, Z., Qian, K., Luo, Z., Zhu, T., Zhong, Y., Tang, Y., Kong, M., Wang, Y., Jiao, S., et al., 2025b. A survey on vision-language-action models for autonomous driving. arXiv preprint arXiv:2506.24044 .

[112] Jiang, Y., Gupta, A., Zhang, Z., Wang, G., Dou, Y., Chen, Y., Fei-Fei, L., Anandkumar, A., Zhu, Y., Fan, L., 2022. Vima: General robot manipulation with multimodal prompts. arXiv preprint arXiv:2210.03094 2, 6.

[113] Jiang, Y., Zhang, R., Wong, J., Wang, C., Ze, Y., Yin, H., Gokmen, C., Song, S., Wu, J., Fei-Fei, L., 2025c. Behavior robot suite: Streamlining real-world whole-body manipulation for everyday household activities. arXiv preprint arXiv:2503.05652 .

[114] Jones, J., Mees, O., Sferrazza, C., Stachowicz, K., Abbeel, P., Levine, S., 2025. Beyond sight: Finetuning generalist robot policies with heterogeneous sensors via language grounding. arXiv preprint arXiv:2501.04693 .

[115] Karamcheti, S., Zhai, A.J., Losey, D.P., Sadigh, D., 2021. Learning visually guided latent actions for assistive teleoperation, in: Learning for dynamics and control, PMLR. pp. 1230–1241.

[116] Karli, U.B., Kurumisawa, T., Fitzgerald, T., 2025. Ask before you act: Token-level uncertainty for intervention in vision-language-action models, in: Second Workshop on Out-of-Distribution Generalization in Robotics at RSS 2025.

[117] Katiyar, N., 2023. A Model-Driven Framework for Domain-Specific Adaptation of Time Series Forecasting Pipeline. McGill University (Canada).

[118] Kawaharazuka, K., Oh, J., Yamada, J., Posner, I., Zhu, Y., 2025. Vision-language-action models for robotics: A review towards real-world applications. IEEE Access .

[119] Kelly, C., Hu, L., Yang, B., Tian, Y., Yang, D., Yang, C., Huang, Z., Li, Z., Hu, J., Zou, Y., 2024. Visiongpt: Vision-language understanding agent using generalized multimodal framework. arXiv preprint arXiv:2403.09027 .

[120] Khan, M.H., Asfaw, S., Iarchuk, D., Cabrera, M.A., Moreno, L., Tokmurziyev, I., Tsetserukou, D., 2025. Shake-vla: Vision-language-action model-based system for bimanual robotic manipulations and liquid mixing. arXiv preprint arXiv:2501.06919 .

[121] Kim, M.J., Finn, C., Liang, P., 2025. Fine-tuning vision-language-action models: Optimizing speed and success. arXiv preprint arXiv:2502.19645 .

[122] Kim, M.J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E., Lam, G., Sanketi, P., et al., 2024. Openvla: An open-source vision-language-action model. arXiv preprint arXiv:2406.09246 .

[123] Koo, J., Cho, T., Kang, H., Pyo, E., Oh, T.G., Kim, T., Choi, A.J., 2025. Retovla: Reusing register tokens for spatial reasoning in vision-language-action models. arXiv preprint arXiv:2509.21243 .

[124] Lee, N., Bang, Y., Lovenia, H., Cahyawijaya, S., Dai, W., Fung, P., 2023. Survey of social bias in vision-language models. arXiv preprint arXiv:2309.14381 .

[125] Li, C., Wen, J., Peng, Y., Peng, Y., Feng, F., Zhu, Y., 2025a. Pointvla: Injecting the 3d world into vision-language-action models. arXiv preprint arXiv:2503.07511 .

[126] Li, D., Jin, Y., Sun, Y., Yu, H., Shi, J., Hao, X., Hao, P., Liu, H., Sun, F., Zhang, J., et al., 2024a. What foundation models can bring for robot learning in manipulation: A survey. arXiv preprint arXiv:2404.18201 .

[127] Li, J., Skinner, G., Yang, G., Quaranto, B.R., Schwaitzberg, S.D., Kim, P.C., Xiong, J., 2024b. Llava-surg: towards multimodal surgical assistant via structured surgical video learning. arXiv preprint arXiv:2408.07981 .

[128] Li, J., Wei, P., Han, W., Fan, L., 2023. Intentqa: Context-aware video intent reasoning, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 11963–11974.

[129] Li, J., Zhu, Y., Tang, Z., Wen, J., Zhu, M., Liu, X., Li, C., Cheng, R., Peng, Y., Feng, F., 2024c. Improving vision-language-action models via chain-of-affordance. arXiv preprint arXiv:2412.20451 .

[130] Li, M., Wang, Z., He, K., Ma, X., Liang, Y., 2025b. Jarvis-vla: Post-training large-scale vision language models to play visual games with keyboards and mouse. arXiv preprint arXiv:2503.16365 .

[131] Li, Q., Liang, Y., Wang, Z., Luo, L., Chen, X., Liao, M., Wei, F., Deng, Y., Xu, S., Zhang, Y., et al., 2024d. Cogact: A foundational vision-language-action model for synergizing cognition and action in robotic manipulation. arXiv preprint arXiv:2411.19650 .

[132] Li, S., Wang, J., Dai, R., Ma, W., Ng, W.Y., Hu, Y., Li, Z., 2024e. Robonurse-vla: Robotic scrub nurse system based on vision-language-action model. arXiv preprint arXiv:2409.19590 .

[133] Li, S., Wu, H., Shao, J., Ma, Y., Gan, Y., Luo, Y., Wang, Y., Nie, D., Wang, L., Wu, W., et al., 2025c. Seeing the forest and the trees: Query-aware tokenizer for long-video multimodal language models. arXiv preprint arXiv:2511.11910 .

[134] Li, Y., Deng, Y., Zhang, J., Jang, J., Memmel, M., Yu, R., Garrett, C.R., Ramos, F., Fox, D., Li, A., et al., 2025d. Hamster: Hierarchical action models for open-world robot manipulation. arXiv preprint arXiv:2502.05485 .

[135] Li, Y., Gong, Z., Li, H., Huang, X., Kang, H., Bai, G., Ma, X., 2025e. Robotic visual instruction. arXiv preprint arXiv:2505.00693 .

[136] Li, Y., Lai, Z., Bao, W., Tan, Z., Dao, A., Sui, K., Shen, J., Liu, D., Liu, H., Kong, Y., 2025f. Visual large language models for generalized and specialized applications. arXiv preprint arXiv:2501.02765 .

[137] Li, Z., Wu, X., Du, H., Nghiem, H., Shi, G., 2025g. Benchmark evaluations, applications, and challenges of large vision language models: A survey. arXiv preprint arXiv:2501.02189 1.

[138] Liang, Z., Li, Y., Yang, T., Wu, C., Mao, S., Nian, T., Pei, L., Zhou, S., Yang, X., Pang, J., et al., 2025. Discrete diffusion vla: Bringing discrete diffusion to action decoding in vision-language-action policies. arXiv preprint arXiv:2508.20072 .

[139] Lin, K.Q., Li, L., Gao, D., Yang, Z., Wu, S., Bai, Z., Lei, W., Wang, L., Shou, M.Z., 2024. Showui: One vision-language-action model for gui visual agent. arXiv preprint arXiv:2411.17465 .

[140] Lin, Y., Zhou, H., Chen, M., Min, H., 2019. Automatic sorting system for industrial robot with 3d visual perception and natural language interaction. Measurement and Control 52, 100–115.

[141] Liu, H., Yao, R., Liu, W., Huang, Z., Shen, S., Ma, J., 2025a. Codrivevlm: Vlm-enhanced urban cooperative dispatching and motion planning for future autonomous mobility on demand systems. URL: https://arxiv.org/abs/2501.06132, arXiv:2501.06132.

[142] Liu, J., Chen, H., An, P., Liu, Z., Zhang, R., Gu, C., Li, X., Guo, Z., Chen, S., Liu, M., et al., 2025b. Hybridvla: Collaborative diffusion and autoregression in a unified vision-language-action model. arXiv preprint arXiv:2503.10631 .

[143] Liu, J., Liu, M., Wang, Z., An, P., Li, X., Zhou, K., Yang, S., Zhang, R., Guo, Y., Zhang, S., 2024a. Robomamba: Efficient vision-language-action model for robotic reasoning and manipulation. Advances in Neural Information Processing Systems 37, 40085–40110.

[144] Liu, S., Wu, L., Li, B., Tan, H., Chen, H., Wang, Z., Xu, K., Su, H., Zhu, J., 2024b. Rdt-1b: a diffusion foundation model for bimanual manipulation. arXiv preprint arXiv:2410.07864 .

[145] Liu, S., Yang, S., Fang, D., Jia, S., Tang, Y., Su, L., Peng, R., Yan, Y., Zou, X., Hu, X., 2026. Vision-language introspection: Mitigating overconfident hallucinations in mllms via interpretable bi-causal steering. arXiv preprint arXiv:2601.05159 .

[146] Liu, Y., Cao, X., Chen, T., Jiang, Y., You, J., Wu, M., Wang, X., Feng, M., Jin, Y., Chen, J., 2025c. From screens to scenes: A survey of embodied ai in healthcare. arXiv preprint arXiv:2501.07468 .

[147] Liu, Y., Cao, X., Chen, T., Jiang, Y., You, J., Wu, M., Wang, X., Feng, M., Jin, Y., Chen, J., 2025d. A survey of embodied ai in healthcare: Techniques, applications, and opportunities. arXiv preprint arXiv:2501.07468 .

[148] Liu, Z., Liang, H., Huang, X., Xiong, W., Yu, Q., Sun, L., Chen, C., He, C., Cui, B., Zhang, W., 2024c. Synthvlm: High-efficiency and high-quality synthetic data for vision language models. arXiv preprint arXiv:2407.20756 .

[149] Lu, H., Li, H., Shahani, P.S., Herbers, S., Scheutz, M., 2025. Probing a vision-language-action model for symbolic states and integration into a cognitive architecture. arXiv preprint arXiv:2502.04558 .

[150] Lu, J., Clark, C., Lee, S., Zhang, Z., Khosla, S., Marten, R., Hoiem, D., Kembhavi, A., 2024. Unified-io 2: Scaling autoregressive multimodal models with vision language audio and action, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 26439–26455.

[151] Lu, Y., Liao, Z., 2023. Towards happy housework: Scenario-based experience design for a household cleaning robotic system. EAI Endorsed Transactions on Scalable Information Systems 10.

[152] Luo, H., Feng, Y., Zhang, W., Zheng, S., Wang, Y., Yuan, H., Liu, J., Xu, C., Jin, Q., Lu, Z., 2025. Being-h0: vision-language-action pretraining from large-scale human videos. arXiv preprint arXiv:2507.15597 .

[153] Luo, J., Xu, C., Wu, J., Levine, S., 2024. Precise and dexterous robotic manipulation via human-in-the-loop reinforcement learning. arXiv preprint arXiv:2410.21845 .

[154] Lyre, H., 2024. "understanding ai": Semantic grounding in large language models. URL: https://arxiv.org/abs/2402.10992, arXiv:2402.10992.

[155] Lyu, J., Li, Z., Shi, X., Xu, C., Wang, Y., Wang, H., 2025. Dywa: Dynamics-adaptive world action model for generalizable non-prehensile manipulation. arXiv preprint arXiv:2503.16806 .

[156] Ma, Y., Song, Z., Zhuang, Y., Hao, J., King, I., 2024. A survey on vision-language-action models for embodied ai. arXiv preprint arXiv:2405.14093 .

[157] Misra, I., Girdhar, R., Joulin, A., 2021. An end-to-end transformer model for 3d object detection, in: ICCV.

[158] Mohammed, M.Q., Chung, K.L., Chyi, C.S., 2020. Review of deep reinforcement learning-based object grasping: Techniques, open challenges, and recommendations. Ieee Access 8, 178450–178481.

[159] Moroncelli, A., Soni, V., Shahid, A.A., Maccarini, M., Forgione, M., Piga, D., Spahiu, B., Roveda, L., 2024. Integrating reinforcement learning with foundation models for autonomous robotics: Methods and perspectives. arXiv preprint arXiv:2410.16411 .

[160] Mumuni, A., Mumuni, F., 2025. Large language models for artificial general intelligence (agi): A survey of foundational principles and approaches. arXiv preprint arXiv:2501.03151 .

[161] Ni, F., Hao, J., Wu, S., Kou, L., Yuan, Y., Dong, Z., Liu, J., Li, M., Zhuang, Y., Zheng, Y., 2024. Peria: Perceive, reason, imagine, act via holistic language and vision planning for manipulation. Advances in Neural Information Processing Systems 37, 17541–17571.

[162] Nie, Y., Li, L., Gan, Z., Wang, S., Zhu, C., Zeng, M., Liu, Z., Bansal, M., Wang, L., 2021. Mlp architectures for vision-and-language modeling: An empirical study. arXiv preprint arXiv:2112.04453 .

[163] Noorani, E., Serlin, Z., Price, B., Velasquez, A., 2025. From abstraction to reality: Darpa's vision for robust sim-to-real autonomy. arXiv preprint arXiv:2503.11007 .

[164] Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al., 2023. Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 .

[165] Pang, J., Zheng, P., Fan, J., Liu, T., 2025. Towards cognition-augmented human-centric assembly: A visual computation perspective. Robotics and Computer-Integrated Manufacturing 91, 102852.

[166] Pantalone, D., Faini, G.S., Cialdai, F., Sereni, E., Bacci, S., Bani, D., Bernini, M., Pratesi, C., Stefàno, P., Orzalesi, L., et al., 2021. Robot-assisted surgery in space: pros and cons. a review from the surgeon's point of view. npj Microgravity 7, 56.

[167] Park, S., Kim, H., Kim, S., Jeon, W., Yang, J., Jeon, B., Oh, Y., Choi, J., 2025. Saliency-aware quantized imitation learning for efficient robotic control, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 13140–13150.

[168] Park, S.M., Kim, Y.G., 2023. Visual language integration: A survey and open challenges. Computer Science Review 48, 100548.

[169] Pasas-Farmer, S., Jain, R., 2025. From discovery to delivery: Governance of ai in the pharmaceutical industry. Green Analytical Chemistry 13, 100268.

[170] Patel, D., Eghbalzadeh, H., Kamra, N., Iuzzolino, M.L., Jain, U., Desai, R., 2023. Pretrained language models as visual planners for human assistance, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 15302–15314.

[171] Pertsch, K., Stachowicz, K., Ichter, B., Driess, D., Nair, S., Vuong, Q., Mees, O., Finn, C., Levine, S., 2025. Fast: Efficient action tokenization for vision-language-action models. arXiv preprint arXiv:2501.09747 .

[172] Plaat, A., van Duijn, M., van Stein, N., Preuss, M., van der Putten, P., Batenburg, K.J., 2025. Agentic large language models, a survey. arXiv preprint arXiv:2503.23037 .

[173] Polubarov, A., Lyubaykin, N., Derevyagin, A., Zisman, I., Tarasov, D., Nikulin, A., Kurenkov, V., 2025. Vintix: Action model via in-context reinforcement learning. arXiv preprint arXiv:2501.19400 .

[174] Qi, C.R., Litany, O., He, K., Guibas, L.J., 2019. Deep hough voting for 3d object detection in point clouds, in: ICCV.

[175] Qu, D., Song, H., Chen, Q., Yao, Y., Ye, X., Ding, Y., Wang, Z., Gu, J., Zhao, B., Wang, D., et al., 2025. Spatialvla: Exploring spatial representations for visual-language-action model. arXiv preprint arXiv:2501.15830 .

[176] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al., 2021. Learning transferable visual models from natural language supervision, in: ICML, PMLR.

[177] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al., 2018. Improving language understanding by generative pre-training .

[178] Rawal, P.K., 2025. An Intelligent Versatile Pipeline for 6D Localization of Industrial Components in a Production Environment. Ph.D. thesis. Fraunhofer Verlag.

[179] Ray, P.P., 2023. Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet of Things and Cyber-Physical Systems 3, 121–154.

[180] Raza, S., Qureshi, R., Zahid, A., Fioresi, J., Sadak, F., Saeed, M., Sapkota, R., Jain, A., Zafar, A., Hassan, M.U., et al., 2025. Who is responsible? the data, models, users or regulations? responsible generative ai for a sustainable future. arXiv preprint arXiv:2502.08650 .

[181] Reed, S., Zolna, K., Parisotto, E., Colmenarejo, S.G., Novikov, A., Barth-Maron, G., Gimenez, M., Sulsky, Y., Kay, J., Springenberg, J.T., et al., 2022. A generalist agent. arXiv preprint arXiv:2205.06175 .

[182] Rodriguez-Guerra, D., Sorrosal, G., Cabanes, I., Calleja, C., 2021. Human-robot interaction review: Challenges and solutions for modern industrial environments. Ieee Access 9, 108557–108578.

[183] Rodriguez-Juan, J., Ortiz-Perez, D., Garcia-Rodriguez, J., Tomás, D., Nalepa, G.J., 2025. Integrating advanced vision-language models for context recognition in risks assessment. Neurocomputing 618, 129131.

[184] Roychoudhury, A., Khorshidi, S., Agrawal, S., Bennewitz, M., 2023. Perception for humanoid robots. Current Robotics Reports 4, 127–140.

[185] Sahili, Z.A., Patras, I., Purver, M., 2025. Scaling for fairness? analyzing model size, data composition, and multilinguality in vision-language bias. arXiv preprint arXiv:2501.13223 .

[186] Sameni, S., Kafle, K., Tan, H., Jenni, S., 2024. Building vision-language models on solid foundations with masked distillation, in: Proceedings of the IEEE/CVF

Conference on Computer Vision and Pattern Recognition, pp. 14216–14226.

[187] Samson, M., Muraccioli, B., Kanehiro, F., 2025. Scalable, training-free visual language robotics: a modular multi-model framework for consumer-grade gpus, in: 2025 IEEE/SICE International Symposium on System Integration (SII), IEEE. pp. 193–198.

[188] Sanyal, S., Roy, K., 2025. Asma: An adaptive safety margin algorithm for vision-language drone navigation via scene-aware control barrier functions. IEEE Robotics and Automation Letters .

[189] Sapkota, R., Karkee, M., 2025. Object detection with multimodal large vision-language models: An in-depth review. Available at SSRN 5233953 .

[190] Sapkota, R., Roumeliotis, K.I., Cheppally, R.H., Calero, M.F., Karkee, M., 2025. A review of 3d object detection with vision-language models. arXiv preprint arXiv:2504.18738 .

[191] Sautenkov, O., Yaqoot, Y., Lykov, A., Mustafa, M.A., Tadevosyan, G., Akhmetkazy, A., Cabrera, M.A., Martynov, M., Karaf, S., Tsetserukou, D., 2025. Uav-vla: Vision-language-action system for large scale aerial mission generation. arXiv preprint arXiv:2501.05014 .

[192] Schakkal, A., Zandonati, B., Yang, Z., Azizan, N., 2025. Hierarchical vision-language planning for multi-step humanoid manipulation. arXiv preprint arXiv:2506.22827 .

[193] Schmidgall, S., Cho, J., Zakka, C., Hiesinger, W., 2024. Gp-vls: A general-purpose vision language model for surgery. arXiv preprint arXiv:2407.19305 .

[194] Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al., 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in neural information processing systems 35, 25278–25294.

[195] Serpiva, V., Lykov, A., Myshlyaev, A., Khan, M.H., Abdulkarim, A.A., Sautenkov, O., Tsetserukou, D., 2025. Racevla: Vla-based racing drone navigation with human-like behaviour. arXiv preprint arXiv:2503.02572 .

[196] Shadab Siddiqui, M., Rabbi, M., Islam, M.J., Ahmed, R.U., 2025. Comparison of different controller architectures for autonomous driving and recommendations for robust and safe implementations. Journal of Advanced Transportation 2025, 9995539.

[197] Shao, R., Li, W., Zhang, L., Zhang, R., Liu, Z., Chen, R., Nie, L., 2025. Large vlm-based vision-language-action models for robotic manipulation: A survey. arXiv preprint arXiv:2508.13073 .

[198] Sharshar, A., Khan, L.U., Ullah, W., Guizani, M., 2025. Vision-language models for edge networks: A comprehensive survey. arXiv preprint arXiv:2502.07855 .

[199] Shi, L.X., Ichter, B., Equi, M., Ke, L., Pertsch, K., Vuong, Q., Tanner, J., Walling, A., Wang, H., Fusai, N., et al., 2025. Hi robot: Open-ended instruction following with hierarchical vision-language-action models. arXiv preprint arXiv:2502.19417 .

[200] Shin, H.C., Roth, H.R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., Summers, R.M., 2016. Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. IEEE transactions on medical imaging 35, 1285–1298.

[201] Shivadekar, S., 2025. Artificial Intelligence for Cognitive Systems: Deep Learning, Neuro-symbolic Integration, and Human-Centric Intelligence. Deep Science Publishing.

[202] Shridhar, M., Manuelli, L., Fox, D., 2022. Cliport: What and where pathways for robotic manipulation, in: Conference on robot learning, PMLR. pp. 894–906.

[203] Shukor, M., Aubakirova, D., Capuano, F., Kooijmans, P., Palma, S., Zouitine, A., Aractingi, M., Pascal, C., Russi, M., Marafioti, A., et al., 2025. Smolvla: A vision-language-action model for affordable and efficient robotics. arXiv preprint arXiv:2506.01844 .

[204] Si, W., Wang, N., Yang, C., 2021. A review on manipulation skill acquisition through teleoperation-based learning from demonstration. Cognitive Computation and Systems 3, 1–16.

[205] Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 .

[206] Singh, G., 2025. Neural Object-Centric Scene Representation and Generation. Ph.D. thesis. Rutgers The State University of New Jersey, School of Graduate Studies.

[207] Singh, S., Singh, J., Shah, B., Sehra, S.S., Ali, F., 2022. Augmented reality and gps-based resource efficient navigation system for outdoor environments: Integrating device camera, sensors, and storage. Sustainability 14, 12720.

[208] Song, M., Deng, X., Zhou, Z., Wei, J., Guan, W., Nie, L., 2025a. A survey on diffusion policy for robotic manipulation: Taxonomy, analysis, and future directions. Authorea Preprints .

[209] Song, W., Chen, J., Ding, P., Zhao, H., Zhao, W., Zhong, Z., Ge, Z., Ma, J., Li, H., 2025b. Accelerating vision-language-action model integrated with action chunking via parallel decoding. arXiv preprint arXiv:2503.02310 .

[210] Sun, H., Wang, H., Ma, C., Zhang, S., Ye, J., Chen, X., Lan, X., 2025a. Prism: Projection-based reward integration for scene-aware real-to-sim-to-real transfer with few demonstrations. arXiv preprint arXiv:2504.20520 .

[211] Sun, J., Mao, P., Kong, L., Wang, J., 2025b. A review of embodied grasping. Sensors (Basel, Switzerland) 25, 852.

[212] Sun, L., Xie, B., Liu, Y., Shi, H., Wang, T., Cao, J., 2025c. Geovla: Empowering 3d representations in vision-language-action models. arXiv preprint arXiv:2508.09071 .

[213] Sutskever, I., Martens, J., Hinton, G.E., 2011. Generating text with recurrent neural networks, in: Proceedings of the 28th international conference on machine learning (ICML-11), pp. 1017–1024.

[214] Szot, A., Mazoure, B., Agrawal, H., Hjelm, R.D., Kira, Z., Toshev, A., 2024. Grounding multimodal large language models in actions. Advances in Neural Information Processing Systems 37, 20198–20224.

[215] Taherin, A., Lin, J., Akbari, A., Akbari, A., Zhao, P., Chen, W., Kaeli, D., Wang, Y., 2025. Cross-platform scaling of vision-language-action models from edge to cloud gpus. arXiv preprint arXiv:2509.11480 .

[216] Tan, X., Yang, Y., Ye, P., Zheng, J., Bai, B., Wang, X., Hao, J., Chen, T., 2025. Think twice, act once: Token-aware compression and action reuse for efficient inference in vision-language-action models. arXiv preprint arXiv:2505.21200 .

[217] Team, G.R., Abeyruwan, S., Ainslie, J., Alayrac, J.B., Arenas, M.G., Armstrong, T., Balakrishna, A., Baruch, R., Bauza, M., Blokzijl, M., et al., 2025. Gemini robotics: Bringing ai into the physical world. arXiv preprint arXiv:2503.20020 .

[218] Team, O.M., Ghosh, D., Walke, H., Pertsch, K., Black, K., Mees, O., Dasari, S., Hejna, J., Kreiman, T., Xu, C., et al., 2024. Octo: An open-source generalist robot policy. arXiv preprint arXiv:2405.12213 .

[219] Tellex, S., Gopalan, N., Kress-Gazit, H., Matuszek, C., 2020. Robots that use language. Annual Review of Control, Robotics, and Autonomous Systems 3, 25–55.

[220] Tian, H., Wang, T., Liu, Y., Qiao, X., Li, Y., 2020. Computer vision technology in agricultural automation—a review. Information processing in agriculture 7, 1–19.

[221] Tian, K., Jiang, Y., Yuan, Z., Peng, B., Wang, L., 2024. Visual autoregressive modeling: Scalable image generation via next-scale prediction. Advances in neural information processing systems 37, 84839–84865.

[222] Torres, N., Ulloa, C., Araya, I., Ayala, M., Jara, S., 2024. A comprehensive analysis of gender, racial, and prompt-induced biases in large language models. International Journal of Data Science and Analytics , 1–38.

[223] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al., 2023. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 .

[224] Trivedi, C., Bhattacharya, P., Prasad, V.K., Patel, V., Singh, A., Tanwar, S., Sharma, R., Aluvala, S., Pau, G., Sharma, G., 2024. Explainable ai for industry 5.0: vision, architecture, and potential directions. IEEE Open Journal of Industry Applications .

[225] Verbaan, L., 2024. Perception and control with large language models in robotic manipulation. TU Delft Library .

[226] Vinod, K., Ramesh, P.J., Chakravarthi, B., et al., 2025. Sebvs: Synthetic event-based visual servoing for robot navigation and manipulation. arXiv preprint arXiv:2508.17643 .

[227] Vuong, Q., Levine, S., Walke, H.R., Pertsch, K., Singh, A., Doshi, R., Xu, C., Luo, J., Tan, L., Shah, D., et al., 2023. Open x-embodiment: Robotic learning datasets and rt-x models, in: Towards Generalist Robots: Learning Paradigms for Scalable Skill Acquisition@ CoRL2023.

[228] Waite, J.R., Hasan, M.Z., Liu, Q., Jiang, Z., Hegde, C., Sarkar, S., 2025. Rls3: Rl-based synthetic sample selection to enhance spatial reasoning in vision-language models for indoor autonomous perception, in: Proceedings of the ACM/IEEE 16th International Conference on Cyber-Physical Systems (with CPS-IoT Week 2025), Association for Computing Machinery, New York, NY, USA. doi:10.1145/3716550.3722033.

[229] Wang, G., Bai, L., Nah, W.J., Wang, J., Zhang, Z., Chen, Z., Wu, J., Islam, M., Liu, H., Ren, H., 2024a. Surgical-lvlm: Learning to adapt large vision-language model for grounded visual question answering in robotic surgery. arXiv preprint arXiv:2405.10948 .

[230] Wang, H., Xing, Z., Wu, W., Yang, Y., Tang, Q., Zhang, M., Xu, Y., Zhu, L., 2024b. Non-invasive to invasive: Enhancing ffa synthesis from cfp with a benchmark dataset and a novel network, in: Proceedings of the 1st International Workshop on Multimedia Computing for Health and Medicine, pp. 7–15.

[231] Wang, J., Guo, D., Liu, H., 2025a. Where to learn: Embodied perception learning planned by vision-language models. IEEE Transactions on Cognitive and Developmental Systems .

[232] Wang, S., 2025. Roboflamingo-plus: Fusion of depth and rgb perception with vision-language models for enhanced robotic manipulation. arXiv preprint arXiv:2503.19510 .

[233] Wang, T., Han, C., Liang, J.C., Yang, W., Liu, D., Zhang, L.X., Wang, Q., Luo, J., Tang, R., 2024c. Exploring the adversarial vulnerabilities of vision-language-action models in robotics. arXiv preprint arXiv:2411.13587 .

[234] Wang, Y., Liu, Q., Jiang, Z., Wang, T., Jiao, J., Chu, H., Gao, B., Chen, H., 2025b. Rad: Retrieval-augmented decision-making of meta-actions with vision-language models in autonomous driving, in: Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 3838–3848.

[235] Wang, Y., Niu, X., Ba, J., Su, Z., Du, L., 2025c. Navigating embodied intelligence: Enabling technologies, security and privacy, and emerging trends. IEEE Internet of Things Journal .

[236] Wang, Y., Wu, S., Zhang, Y., Yan, S., Liu, Z., Luo, J., Fei, H., 2025d. Multimodal chain-of-thought reasoning: A comprehensive survey. arXiv preprint arXiv:2503.12605 .

[237] Wang, Z., Zhou, Z., Song, J., Huang, Y., Shu, Z., Ma, L., 2024d. Towards testing and evaluating vision-language-action models for robotic manipulation: An empirical study. arXiv preprint arXiv:2409.12894 .

[238] Wei, C., Guo, C., Zhang, J., Shan, H., Xu, Y., Zhang, Z., Liu, Y., Wang, Q., Zhou, C., Li, H., et al., 2025. Focus: A streaming concentration architecture for efficient vision-language models. arXiv preprint arXiv:2512.14661 .

[239] Wei, J., Yuan, S., Li, P., Hu, Q., Gan, Z., Ding, W., 2024. Occllama: An occupancy-language-action generative world model for autonomous driving. arXiv preprint arXiv:2409.03272 .

[240] Wen, J., Zhu, M., Zhu, Y., Tang, Z., Li, J., Zhou, Z., Li, C., Liu, X., Peng, Y., Shen, C., et al., 2024. Diffusion-vla: Scaling robot foundation models via unified diffusion and autoregression. arXiv preprint arXiv:2412.03293 .

[241] Wen, J., Zhu, Y., Li, J., Tang, Z., Shen, C., Feng, F., 2025a. Dexvla: Vision-language model with plug-in diffusion expert for general robot control. arXiv preprint arXiv:2502.05855 .

[242] Wen, J., Zhu, Y., Li, J., Zhu, M., Tang, Z., Wu, K., Xu, Z., Liu, N., Cheng, R., Shen, C., et al., 2025b. Tinyvla: Towards fast, data-efficient vision-language-action models for robotic manipulation. IEEE Robotics and Automation Letters .

[243] Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I.S., Xie, S., 2023. Convnext v2: Co-designing and scaling convnets with masked autoencoders, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 16133–16142.

[244] Wu, J., Zhong, M., Xing, S., Lai, Z., Liu, Z., Chen, Z., Wang, W., Zhu, X., Lu, L., Lu, T., et al., 2024a. Visionllm v2: An end-to-end generalist multimodal large language model for hundreds of vision-language tasks. Advances in Neural Information Processing Systems 37, 69925–69975.

[245] Wu, W., Feng, X., Gao, Z., Kan, Y., 2024b. Smart: scalable multi-agent real-time motion generation via next-token prediction. Advances in Neural Information Processing Systems 37, 114048–114071.

[246] Wu, Z., Zhou, Y., Xu, X., Wang, Z., Yan, H., 2025. Momanipvla: Transferring vision-language-action models for general mobile manipulation. arXiv preprint arXiv:2503.13446 .

[247] Xiang, T.Y., Jin, A.Q., Zhou, X.H., Gui, M.J., Xie, X.L., Liu, S.Q., Wang, S.Y., Duang, S.B., Wang, S.C., Lei, Z., et al., 2025. Vla model-expert collaboration for bi-directional manipulation learning. arXiv preprint arXiv:2503.04163 .

[248] Xiong, J., Liu, G., Huang, L., Wu, C., Wu, T., Mu, Y., Yao, Y., Shen, H., Wan, Z., Huang, J., et al., 2024. Autoregressive models in vision: A survey. arXiv preprint arXiv:2411.05902 .

[249] Xu, D., Chen, Y., Wang, J., Huang, Y., Wang, H., Jin, Z., Wang, H., Yue, W., He, J., Li, H., et al., 2024a. Mlevlm: Improve multi-level progressive capabilities based on multimodal large language model for medical visual question answering, in: Findings of the Association for Computational Linguistics ACL 2024, pp. 4977–4997.

[250] Xu, F., Zhai, G., Kong, X., Fu, T., Gordon, D.F., An, X., Busam, B., 2025a. Stare-vla: Progressive stage-aware reinforcement for fine-tuning vision-language-action models. arXiv preprint arXiv:2512.05107 .

[251] Xu, J., Sun, Q., Han, Q.L., Tang, Y., 2025b. When embodied ai meets industry 5.0: human-centered smart manufacturing. IEEE/CAA Journal of Automatica Sinica 12, 485–501.

[252] Xu, S., Wang, Y., Xia, C., Zhu, D., Huang, T., Xu, C., 2025c. Vla-cache: Towards efficient vision-language-action model via adaptive token caching in robotic manipulation. arXiv preprint arXiv:2502.02175 .

[253] Xu, Y., Liu, G., Kompella, R.R., Hu, S., Huang, T., Ilhan, F., Tekin, S.F., Yahn, Z., Liu, L., 2025d. Language-vision planner and executor for text-to-visual reasoning. arXiv preprint arXiv:2506.07778 .

[254] Xu, Z., Wu, K., Wen, J., Li, J., Liu, N., Che, Z., Tang, J., 2024b. A survey on robotics with foundation models: toward embodied ai. arXiv preprint arXiv:2402.02385 .

[255] Xue, H., Ren, J., Chen, W., Zhang, G., Fang, Y., Gu, G., Xu, H., Lu, C., 2025. Reactive diffusion policy: Slow-fast visual-tactile policy learning for contact-rich manipulation. arXiv preprint arXiv:2503.02881 .

[256] Yang, G., Zhang, T., Hao, H., Wang, W., Liu, Y., Wang, D., Chen, G., Cai, Z., Chen, J., Su, W., et al., 2025a. Vlaser: Vision-language-action model with synergistic embodied reasoning. arXiv preprint arXiv:2510.11027 .

[257] Yang, R., Chen, G., Wen, C., Gao, Y., 2025b. Fp3: A 3d foundation policy for robotic manipulation. arXiv preprint arXiv:2503.08950 .

[258] Yang, R., Yu, Q., Wu, Y., Yan, R., Li, B., Cheng, A.C., Zou, X., Fang, Y., Cheng, X., Qiu, R.Z., et al., 2025c. Egovla: Learning vision-language-action models from egocentric human videos. arXiv preprint arXiv:2507.12440 .

[259] Yang, S., Li, Y.L., Wang, S., 2025d. Upl-net: Uncertainty-aware prompt learning network for semi-supervised action recognition. Neurocomputing 619, 129126.

[260] Yang, Y., Huang, W., Wei, Y., Peng, H., Jiang, X., Jiang, H., Wei, F., Wang, Y., Hu, H., Qiu, L., et al., 2023a. Attentive mask clip, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 2771–2781.

[261] Yang, Y., Sun, J., Kou, S., Wang, Y., Deng, Z., 2025e. Lohovla: A unified vision-language-action model for long-horizon embodied tasks. arXiv preprint arXiv:2506.00411 .

[262] Yang, Y., Wang, Y., Wen, Z., Zhongwei, L., Zou, C., Zhang, Z., Wen, C., Zhang, L., 2025f. Efficientvla: Training-free acceleration and compression for vision-language-action models. arXiv preprint arXiv:2506.10100 .

[263] Yang, Y., Zhou, J., Ding, X., Huai, T., Liu, S., Chen, Q., Xie, Y., He, L., 2025g. Recent advances of foundation language models-based continual learning: A survey. ACM Computing Surveys 57, 1–38.

[264] Yang, Z., Chen, Y., Wang, J., Manivasagam, S., Ma, W.C., Yang, A.J., Urtasun, R., 2023b. Unisim: A neural closed-loop sensor simulator, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1389–1399.

[265] Yang, Z., Garrett, C., Fox, D., Lozano-Pérez, T., Kaelbling, L.P., 2025h. Guiding long-horizon task and motion planning with vision language models, in: 2025 IEEE International Conference on Robotics and Automation (ICRA), IEEE. pp. 16847–16853.

[266] Ye, S., Jang, J., Jeon, B., Joo, S., Yang, J., Peng, B., Mandlekar, A., Tan, R., Chao, Y.W., Lin, B.Y., et al., 2024. Latent action pretraining from videos. arXiv preprint arXiv:2410.11758 .

[267] Ye, Y., Ma, J., Cen, J., Lu, Z., 2025. Token expand-merge: Training-free token compression for vision-language-action models. arXiv preprint arXiv:2512.09927 .

[268] Yu, Z., Wang, B., Zeng, P., Zhang, H., Zhang, J., Gao, L., Song, J., Sebe, N., Shen, H.T., 2025. A survey on efficient vision-language-action models. arXiv preprint arXiv:2510.24795 .

[269] Yue, Y., Wang, Y., Kang, B., Han, Y., Wang, S., Song, S., Feng, J., Huang, G., 2024. Deer-vla: Dynamic inference of multimodal large language models for efficient robot execution. Advances in Neural Information Processing Systems 37, 56619–56643.

[270] Zawalski, M., Chen, W., Pertsch, K., Mees, O., Finn, C., Levine, S., 2024. Robotic control via embodied chain-of-thought reasoning. arXiv preprint arXiv:2407.08693 .

[271] Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L., 2023. Sigmoid loss for language image pre-training, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 11975–11986.

[272] Zhan, Z., Chen, Y., Zhou, J., Lv, Q., Liu, H., Wang, K., Lin, L., Wang, G., 2026. Stable language guidance for vision-language-action models. arXiv preprint arXiv:2601.04052 .

[273] Zhang, B., Li, J., Shen, J., Cai, Y., Zhang, Y., Chen, Y., Dai, J., Ji, J., Yang, Y., 2025a. Vla-arena: An open-source framework for benchmarking vision-language-action models. arXiv preprint arXiv:2512.22539 .

[274] Zhang, B., Zhang, Y., Ji, J., Lei, Y., Dai, J., Chen, Y., Yang, Y., 2025b. Safevla: Towards safety alignment of vision-language-action model via safe reinforcement learning. arXiv preprint arXiv:2503.03480 .

[275] Zhang, D., Sun, J., Hu, C., Wu, X., Yuan, Z., Zhou, R., Shen, F., Zhou, Q., 2025c. Pure vision language action (vla) models: A comprehensive survey. arXiv preprint arXiv:2509.19012 .

[276] Zhang, H., Ding, P., Lyu, S., Peng, Y., Wang, D., 2025d. Gevrm: Goal-expressive video generation model for robust visual manipulation. arXiv preprint arXiv:2502.09268 .

[277] Zhang, H., Yu, H., Zhao, L., Choi, A., Bai, Q., Yang, B., Xu, W., 2025e. Slim: Sim-to-real legged instructive manipulation via long-horizon visuomotor learning. arXiv preprint arXiv:2501.09905 .

[278] Zhang, H., Zantout, N., Kachana, P., Wu, Z., Zhang, J., Wang, W., 2024a. Vla-3d: A dataset for 3d semantic scene understanding and navigation. arXiv preprint arXiv:2411.03540 .

[279] Zhang, J., Guo, Y., Hu, Y., Chen, X., Zhu, X., Chen, J., 2025f. Up-vla: A unified understanding and prediction model for embodied agent. arXiv preprint arXiv:2501.18867 .

[280] Zhang, J., Wang, K., Wang, S., Li, M., Liu, H., Wei, S., Wang, Z., Zhang, Z., Wang, H., 2024b. Uninavid: A video-based vision-language-action model for unifying embodied navigation tasks. arXiv preprint arXiv:2412.06224 .

[281] Zhang, K., Yin, Z.H., Ye, W., Gao, Y., 2024c. Learning manipulation skills through robot chain-of-thought with sparse failure guidance. arXiv preprint arXiv:2405.13573 .

[282] Zhang, K., Yun, P., Cen, J., Cai, J., Zhu, D., Yuan, H., Zhao, C., Feng, T., Wang, M.Y., Chen, Q., et al., 2025g. Generative artificial intelligence in robotic manipulation: A survey. arXiv preprint arXiv:2503.03464 .

[283] Zhang, R., Dong, M., Zhang, Y., Heng, L., Chi, X., Dai, G., Du, L., Wang, D., Du, Y., Zhang, S., 2025h. Mole-vla: Dynamic layer-skipping vision language action model via mixture-of-layers for efficient robot manipulation. arXiv preprint arXiv:2503.20384 .

[284] Zhang, Z., Bao, C., Pan, X., Chang, C.M., Igarashi, T., Zhang, G., 2025i. Through the lens of privacy: Exploring privacy protection in vision-language model interactions on smart glasses, in: Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, pp. 1–8.

[285] Zhao, H., Song, W., Wang, D., Tong, X., Ding, P., Cheng, X., Ge, Z., 2025a. More: Unlocking scalability in reinforcement learning for quadruped vision-language-action models. arXiv preprint arXiv:2503.08007 .

[286] Zhao, Q., Lu, Y., Kim, M.J., Fu, Z., Zhang, Z., Wu, Y., Li, Z., Ma, Q., Han, S., Finn, C., et al., 2025b. Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. arXiv preprint arXiv:2503.22020 .

[287] Zhao, T.Z., Kumar, V., Levine, S., Finn, C., 2023. Learning fine-grained bimanual manipulation with low-cost hardware. arXiv preprint arXiv:2304.13705 .

[288] Zhen, H., Qiu, X., Chen, P., Yang, J., Yan, X., Du, Y., Hong, Y., Gan, C., 2024. 3d-vla: A 3d vision-language-action generative world model. arXiv preprint arXiv:2403.09631 .

[289] Zheng, J., Li, J., Liu, D., Zheng, Y., Wang, Z., Ou, Z., Liu, Y., Liu, J., Zhang, Y.Q., Zhan, X., 2025. Universal actions for enhanced embodied foundation models. arXiv preprint arXiv:2501.10105 .

[290] Zheng, J., Shi, C., Cai, X., Li, Q., Zhang, D., Li, C., Yu, D., Ma, Q., 2026. Lifelong learning of large language model based agents: A roadmap. IEEE Transactions on Pattern Analysis and Machine Intelligence .

[291] Zhong, Y., Huang, X., Li, R., Zhang, C., Liang, Y., Yang, Y., Chen, Y., 2025. Dexgraspvla: A vision-language-action framework towards general dexterous grasping. arXiv preprint arXiv:2502.20900 .

[292] Zhou, D.W., Zhang, Y., Wang, Y., Ning, J., Ye, H.J., Zhan, D.C., Liu, Z., 2025a. Learning without forgetting for vision-language models. IEEE Transactions on Pattern Analysis and Machine Intelligence .

[293] Zhou, X., Han, X., Yang, F., Ma, Y., Knoll, A.C., 2025b. Opendrivevla: Towards end-to-end autonomous driving with large vision language action model. arXiv preprint arXiv:2503.23463 .

[294] Zhou, Z., Cai, T., Zhao, S.Z., Zhang, Y., Huang, Z., Zhou, B., Ma, J., 2025c. Autovla: A vision-language-action model for end-to-end autonomous driving with adaptive reasoning and reinforcement fine-tuning. arXiv preprint arXiv:2506.13757 .

[295] Zhou, Z., Zhu, Y., Zhu, M., Wen, J., Liu, N., Xu, Z., Meng, W., Cheng, R., Peng, Y., Shen, C., et al., 2025d. Chatvla: Unified multimodal understanding and robot control with vision-language-action model. arXiv preprint arXiv:2502.14420 .

[296] Zhu, D.H., Chang, Y.P., 2020. Robot with humanoid hands cooks food better? effect of robotic chef anthropomorphism on food quality prediction. International Journal of Contemporary Hospitality Management 32, 1367–1383.

[297] Zhu, M., Zhu, Y., Li, J., Zhou, Z., Wen, J., Liu, X., Shen, C., Peng, Y., Feng, F., 2025. Objectvla: End-to-end open-world object manipulation without demonstration. arXiv preprint arXiv:2502.19250 .

[298] Zhu, Y., Zhou, Y., Wang, C., Cao, Y., Han, J., Hou, L., Xu, H., 2024. Unit: Unifying image and text recognition in one vision encoder, in: NeurIPS.

[299] Zitkovich, B., Yu, T., Xu, S., Xu, P., Xiao, T., Xia, F., Wu, J., Wohlhart, P., Welker, S., Wahid, A., et al., 2023. Rt-2: Vision-language-action models transfer web knowledge to robotic control, in: Conference on Robot Learning, PMLR. pp. 2165–2183.