RESURSLARI CHEKLANGAN ROBOTLARDA TABIIY TIL BUYRUQLARINI TUSHUNISH: YENGIL MODELLAR VA RESURSLARI CHEKLANGAN QURILMALARDA ISHLASH USULLARINING QIYOSIY TAHLILI

Murodjonov Lutfulla Murodjon o‘g‘li
Muhammad al-Xorazmiy nomidagi Toshkent axborot texnologiyalari universiteti, Robototexnika va intellektual tizimlar kafedrasi
murodjonovlutfulla@gmail.com
Ochilov Mannon Musinovich
Muhammad al-Xorazmiy nomidagi Toshkent axborot texnologiyalari universiteti, Robototexnika va intellektual tizimlar kafedrasi
ochilov.mannon@tuit.uz
Mirzahalilov Sanjar Serkabay o‘g‘li
Muhammad al-Xorazmiy nomidagi Toshkent axborot texnologiyalari universiteti, Robototexnika va intellektual tizimlar kafedrasi
mirzaxalilov86@tuit.uz

Annotatsiya. *Maqolada robotlarga berilgan og‘zaki va yozma buyruqlarni kam quvvatli qurilmalarda tushunishga qaratilgan 2024-2026-yillardagi tadqiqotlar qiyosiy tahlil qilinadi. Kvantlash, pruning, bilim distillyatsiyasi va ixcham arxitekturalar, shuningdek, ularni mikrokontroller, Raspberry Pi va Jetson platformalarida ishlatish natijalari yagona mezonlar - xotira, kechikish, sifatning saqlanishi, qayta o‘qitish talabi, apparatga bog‘liqlik va kam resursli tillarga moslik - bo‘yicha baholanadi. Tahlil past bitli kvantlash xotirani 5-11 marta kamaytirib, sifatning 90-95% ini saqlashini, biroq mikrokontrollerlarda hozircha faqat kalit so‘zlarni aniqlash amaliy ekanini ko‘rsatadi. Umumiy kechikishni baholashning yagona usuli yo‘qligi va o‘zbek tilida robot buyruqlari korpusining mavjud emasligi asosiy ochiq muammolar sifatida asoslanadi.*

Kalit so‘zlar: *robot buyruqlari, tabiiy tilni tushunish, resurslari cheklangan qurilmalar, kvantlash, distillyatsiya, mikrokontroller, kam resursli tillar.*

Аннотация. *В статье проводится сравнительный анализ исследований 2024-2026 гг., посвящённых пониманию устных и письменных команд роботов на маломощных устройствах. Квантование, прунинг, дистилляция знаний и компактные архитектуры, а также результаты их запуска на микроконтроллерах, Raspberry Pi и Jetson оцениваются по единым критериям: память, задержка, сохранение качества, потребность в переобучении, зависимость от аппаратуры и применимость к малоресурсным*

языкам. Анализ показывает, что низкоразрядное квантование сокращает память в 5-11 раз при сохранении 90-95 % качества, однако на микроконтроллерах пока практически применимо лишь распознавание ключевых слов. Отсутствие единой методики оценки общей задержки и корпуса команд роботов на узбекском языке обосновываются как основные открытые проблемы.

Ключевые слова: *команды роботов, понимание естественного языка, периферийные устройства, квантование, дистилляция, микроконтроллер, малоресурсные языки.*

Abstract. *The paper presents a comparative analysis of 2024-2026 studies on understanding spoken and written robot commands on low-power devices. Quantization, pruning, knowledge distillation and compact architectures, together with results of running them on microcontrollers, Raspberry Pi and Jetson platforms, are assessed against uniform criteria: memory, latency, quality retention, retraining needs, hardware dependence and suitability for low-resource languages. The analysis shows that low-bit quantization reduces memory by 5-11× while retaining 90-95% of quality, yet on microcontrollers only keyword spotting is currently practical. The absence of a unified method for measuring end-to-end latency and the lack of an Uzbek robot command corpus are substantiated as the main open problems.*

Keywords: *robot commands, natural language understanding, edge devices, quantization, distillation, microcontroller, low-resource languages.*

Kirish

So‘nggi yillarda tabiiy til yordamida robotlarni boshqarish texnologiyalari jadal rivojlanib, inson va robot o‘rtasidagi o‘zaro aloqani yanada qulay va intuitiv tashkil etish imkoniyatlarini kengaytirmoqda. An’anaviy robot boshqaruv tizimlarida buyruqlar, odatda, oldindan belgilangan algoritmlar, maxsus boshqaruv interfeyslari yoki qat’iy buyruqlar to‘plami orqali beriladi. Tabiiy tilni tushunish texnologiyalari esa robotga insonning kundalik tildagi ko‘rsatmalarini tahlil qilish, ularning mazmunini aniqlash va tegishli harakatlarni bajarish imkonini beradi. Bugungi ko‘rish, til va harakatni birlashtiruvchi (Vision-Language-Action - VLA) sun’iy intellekt modeli “qizil kubni qutiga sol” kabi erkin buyruqni to‘g‘ridan-to‘g‘ri robot harakatiga aylantira oladi, lekin buning narxi yuqori: 7,2 mlrd parametrli OpenVLA kuchli A6000 GPU’da ham bitta harakatga 292 ms sarflaydi[1]. Batareyada ishlaydigan mobil robot yoki dron uchun bunday hisoblash quvvati odatda mavjud emas. Bulut esa to‘liq yechim emas: tarmoq kechikishi, aloqa uzilishi va ovozli ma’lumotlarning maxfiyligi muammo tug‘diradi, oflayn ishlaydigan

kvantlangan 7B model esa mobil robotda bulutdagi modeldan ishonchlilik bo‘yicha sezilarli orqada qolgan [2].

Ushbu maqolaning maqsadi robototexnik tizimlarda tabiiy til buyruqlarini resurs tejamkor qayta ishlash va tushunishga qaratilgan zamonaviy ilmiy yondashuvlarni tizimli ravishda tahlil qilish, ularning afzalliklari va cheklovlarini aniqlash hamda kelgusidagi tadqiqotlar uchun istiqbolli yo‘nalishlarni belgilashdan iborat. Shu maqsadda ixcham modellarni yaratish va siqish usullari, ularni turli hisoblash platformalarida amalga oshirish imkoniyatlari hamda moslashuvchan hisoblash mexanizmlari ko‘rib chiqiladi [3]. Tahlil davomida xotira sarfi, bajarilish vaqti, vazifa sifatining saqlanishi, energiya samaradorligi, apparat platformasiga bog‘liqlik va kam resursli tillarga tatbiq etish imkoniyatlari asosiy baholash mezonlari sifatida qaraladi.

Maqolaning ilmiy-amaliy ahamiyati mavjud tadqiqot natijalarini yagona tahliliy doirada birlashtirish, turli apparat platformalarida qo‘llaniladigan usullarning imkoniyatlarini taqqoslash va hal etilmagan muammolarni aniqlash bilan belgilanadi. Olingan tahlil natijalari kelgusida o‘zbek tilidagi tabiiy til buyruqlarini tushunishga mo‘ljallangan, cheklangan hisoblash resurslarida ishlay oladigan, tezkor va ishonchli robot boshqaruv tizimlarini ishlab chiqish uchun nazariy asos bo‘lib xizmat qilishi mumkin.

Tadqiqot usuli.

Tadqiqotning birinchi bosqichida robototexnik tizimlarda tabiiy tilni tushunish, ixcham neyron tarmoqlar, modelni siqish va qurilmada bevosita hisoblashga oid ilmiy manbalarni aniqlash ko‘zda tutiladi.

Manbalarni izlash uchun arXiv, IEEE Xplore, Scopus, ACM Digital Library, ScienceDirect va Google Scholar kabi ilmiy platformalar bazalaridan foydalanildi. Tadqiqotning asosiy vaqt oralig‘i 2024-2026-yillarni qamrab oladi. Shu bilan birga, zarur hollarda asosiy algoritmlar va metodlarning nazariy asoslarini yorituvchi avvalgi ilmiy ishlar ham ko‘rib chiqildi.

Manbalardan natural language understanding, robotic command understanding, resource-efficient neural networks, model quantization, knowledge distillation, pruning, adaptive inference va edge deployment kabi kalit so‘zlar asosidagi so‘rovlar bilan izlandi. Tabiiy til kirishini qayta ishlaydigan, robot yoki o‘rnatilgan qurilmaga mo‘ljallangan va kamida bitta samaradorlik ko‘rsatkichini miqdoriy bergan ishlar tanlandi. Taqqoslash mezonlari: **M1** - xotira; **M2** - kechikish; **M3** - sifatning saqlanishi; **M4** - qayta o‘qitish talabi; **M5** - apparatga bog‘liqlik; **M6** - mikrokontrollerga moslik va kam resursli tillarga tatbiq etilishi. Raqamlar mualliflar e’lon qilgan ko‘rinishda olindi va turli benchmarklarga tegishli bo‘lgani uchun faqat yo‘nalish sifatida taqqoslanadi.

Natijalar va tahlil.

Yengil modellar. Kvantlash robot modellarining xotira sarfini kamaytirishning samarali usullaridan biridir. BitVLA modelida og‘irliklarni 1,58 bitgacha qisqartirish natijasida xotira hajmi 15,4 GB dan 1,4 GB gacha kamaygan va LIBERO sinovida 94,8% muvaffaqiyat ko‘rsatkichi saqlangan. Biroq, uzoq davom etadigan vazifalarda model samaradorligi pasaygan [4]. ActQuant usuli esa harakat natijasiga ko‘proq ta’sir qiluvchi og‘irliklarni hisobga olib, 3 bitdan past aniqlikda ham model sifatining taxminan 95 foizini saqlash imkonini bergan [5]. Real robotda o‘tkazilgan tajribada til modelining muhim qatlamlarini 8 bitda saqlash natijani 80,0% dan 92,5% gacha oshirgan [6]. Nutqni tanishda ham kvantlash darajasi muhim: Whisper-small modeli 8 bitli formatda sifatini saqlagan, 4 bitli formatda esa shovqinli nutq uchun xatolik darajasi 11,88 dan 14,09 gacha oshgan [7].

1-jadval. Yengil modellar bo‘yicha asosiy natijalar

Ish	Usul	Natija	Cheklov
BitVLA [4]	Model og‘irliklarini 1,58 bitli ifodalash	Model hajmi 15,4 GB dan 1,4 GB gacha kamaygan; LIBERO sinovida 94,8% natija qayd etilgan	Uzoq vazifalarda samaradorlik pasayishi mumkin; maxsus hisoblash yadrolari talab qilinishi mumkin
ActQuant [5]	≤3 bit PTQ	Model sifati taxminan 95% darajada saqlangan	2,5 bitli holatda ko‘rsatkich 90,1% gacha tushgan
FoldQuantVLA [6]	W4A4 + aralash aniqlik	Real robot tajribasida muvaffaqiyat 80,0% dan 92,5% gacha oshgan; NVIDIA Orin’da 1,2 - 1,3 baravar tezlashish qayd etilgan	Tezlanish mo‘tadil
Whisper-small [7]	INT8 / INT4	INT8: hajm - 57%, WER o‘zgarmagan	INT4: shovqinli nutqda WER oshadi
EfficientVLA [8]	Qatlam va token pruning	1,93× tez; - 0,6% muvaffaqiyat	Asosan GPU hisoblashiga mo‘ljallangan
PRISM [9]	Distillyatsiya	3B model: GPT-4o sifatining >93%	Apparat o‘lchovlari yo‘q
TinyVLA [1]	Ko‘rish, til va harakatni birlashtiruvchi ixcham model	1,3 milliard parametr; 14 ms ishlash vaqti; 94,0% muvaffaqiyat ko‘rsatkichi	O‘lchovlar GPU’da

MCUBERT [13]	MCU uchun BERT	256 KB dan kam xotirada 512 tadan ortiq token bilan ishlagan	Faqat juda kichik modellar

Modelni ixchamlashtirishda ortiqcha qatlam va ma’lumot qismlarini olib tashlash hamda bilimlarni katta modeldan kichik modelga o‘tkazish usullari qo‘llaniladi. EfficientVLA keraksiz qatlamlar va tasviriy belgilarni qisqartirish orqali ishlash tezligini 1,93 baravar oshirgan, muvaffaqiyat ko‘rsatkichi esa atigi 0,6 foizga kamaygan [8]. PRISM usuli sintetik ma’lumotlar yordamida 3 milliard parametrli modelning rejalashtirish sifatini GPT-4o ko‘rsatkichining 93 foizidan yuqori darajaga yetkazgan [9]. Kam resursli tillar uchun taklif qilingan moslashuvchan bilim uzatish usuli parametrlar sonini 80,3% ga kamaytirgan, ammo kichik ma’lumotlar to‘plamida oddiy usul yaxshiroq natija bergan [10].

Ixcham modellar orasida TinyVLA (1,3 milliard parametr, 14 ms) [1], SmolVLA (450 million parametr) [11], Whisper’ga qaraganda besh baravar kam hisoblash resursini talab qiladigan Moonshine [12] hamda 256 KB dan kam xotirada ishlaydigan MCUBERT [13] alohida ahamiyatga ega. Ushbu natijalar modelni tanlashda faqat aniqlikni emas, balki xotira sarfi, hisoblash tezligi va qurilma imkoniyatlarini ham hisobga olish zarurligini ko‘rsatadi (1-jadval).

Hisoblash imkoniyatlari cheklangan qurilmalarda qo‘llash.

Robotlarda tabiiy til buyruqlarini qayta ishlashda qurilmaning texnik imkoniyatlari muhim rol o‘ynaydi. Masalan, mikrokontroller, Raspberry Pi 5 va NVIDIA Jetson Orin bir xil murakkablikdagi vazifalarni bajara olmaydi. Ularning xotirasi va hisoblash quvvati turlicha bo‘lgani uchun modelni tanlashda bajariladigan vazifaning murakkabligini ham hisobga olish kerak (2-jadval).

2-jadval. Apparat sinflari va ularda bajariladigan vazifalar

Apparat sinfi	Vazifalar	Adabiyotdagi natija	Cheklov
ESP32 / Cortex-M	Kalit so‘z, qat’iy buyruqlar	93,70%; 34,9 KB RAM [14]	Ochiq lug‘at yo‘q
Raspberry Pi 5 / ARM	INT8 ASR, 1-3B SLM	RTF: 0,121 dan 0,077 gacha [7]	Xotira o‘tkazuvchanligi
Jetson Orin / Thor	VLA, rejalashtirish	Chastota: 0,70 Hz dan 6,93 Hz gacha [16]	Narx va quvvat
Edge + bulut	Murakkab buyruqlar	Yuklamaning 84,53% edge’da [17]	Tarmoqqa bog‘liq

Mikrokontrollerlarda xotira hajmi cheklanganligi sababli murakkab, keng lug‘atga ega tabiiy til buyruqlarini qayta ishlash qiyin. Shunga qaramay, kalit so‘zlarni aniqlash kabi sodda vazifalarni bajarish mumkin. Pham va hammualliflari

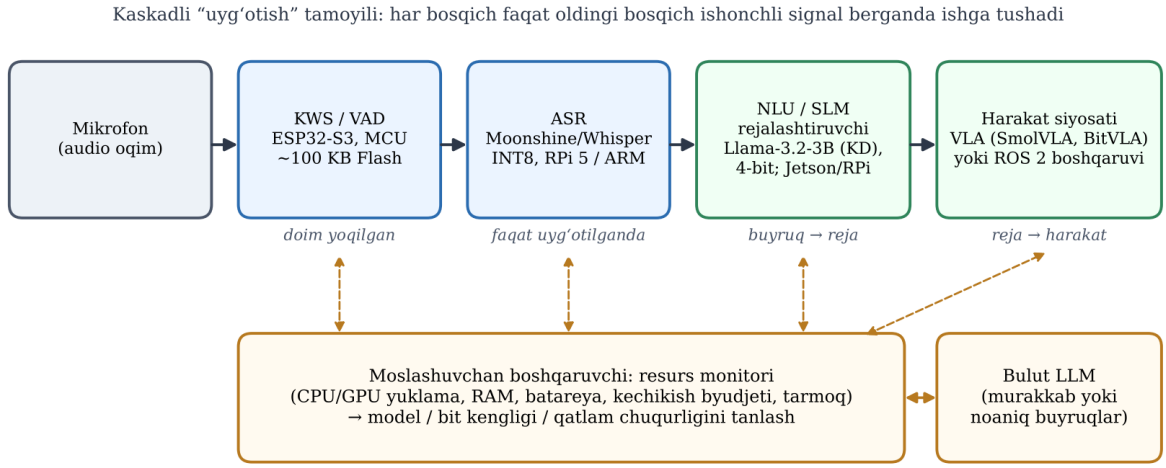
taklif qilgan model bunga misol bo‘la oladi: u 98,7 KB Flash va 34,9 KB RAM xotiradan foydalanib, 93,70% aniqlikka erishgan [14].

Raspberry Pi 5 kabi qurilmalarda imkoniyatlar kengroq. Ularda INT8 formatida kvantlangan nutqni tanish modellari va 1-3 milliard parametrli kichik til modellarini ishga tushirish mumkin. Biroq modelning katta bo‘lishi uning tez ishlashini anglatmaydi. Xususan, xotira o‘tkazuvchanligi va hisoblash resurslari inferensiya tezligiga ta’sir qiladi [15].

NVIDIA Jetson Orin esa murakkabroq modellar bilan ishlash uchun ko‘proq imkoniyat beradi. Masalan, [16]-manbada asinxron inferensiyadan foydalanish π0.5 modelining boshqaruv chastotasini 0,70 Hz dan 6,93 Hz gacha oshirgani keltirilgan. Bu misol modelni optimallashtirish bilan bir qatorda, hisoblash jarayonini qanday tashkil etish ham robotning ishlash tezligiga ta’sir qilishini ko‘rsatadi.

Demak, robot uchun hisoblash platformasini tanlashda faqat modelning aniqligiga qarash yetarli emas. Uning xotira talabi, javob berish vaqti va real vaqt rejimida ishlash imkoniyati ham muhim. Ayniqsa, ovozli buyruqlarni qabul qilib, ularni robot harakatiga aylantiradigan tizimlarda ushbu ko‘rsatkichlar o‘rtasidagi muvozanatni ta’minlash zarur.

Bu natijalar asosida kaskadli yondashuv eng maqbul ko‘rinadi: ESP32 doimo yoqilgan “darvoza” sifatida uyg‘otuvchi so‘zni kutadi, Raspberry Pi nutqni matnga o‘giradi va buyruqni tahlil qiladi, murakkab holatlardagina Jetson yoki bulutga murojaat qilinadi (1-rasm).



1-rasm. Nutq buyrug‘ini tushunishning kaskadli arxitekturasi

Moslashuvchan inferensiya. Robot resurslari o‘zgaruvchan bo‘lgani uchun hisoblash hajmini ish vaqtida moslashtirish muhim. DeeR-VLA erta chiqish orqali hisoblashni 5,2-6,5 marta kamaytirgan [18], MoLe-VLA qatlamlarni dinamik tanlab, hisoblashni 5,6 martagacha qisqartirgan va muvaffaqiyatni 8% ga oshirgan [19]. Any-Precision LLM bitta modeldan turli bit kengligida foydalanish imkonini

beradi [20], CE-CoLLM esa faqat ishonch past bo‘lganda bulutga murojaat qiladi [17]. Biroq bu usullarning hech biri batareya yoki harorat kabi real robot signallariga bog‘lab sinalmagan.

Yagona mezonlar bo‘yicha taqqoslash.

3-jadval. Yondashuvlarning M1-M6 mezonlari bo‘yicha taqqoslanishi

Mezon	Kvantlash	Pruning	Distillyatsiya	Ixcham arxitektura	Moslashuvchan inferensiya
M1. Xotira	Yuqori (5-11×)	O‘rta	Yuqori	Yuqori	Past
M2. Kechikish	O‘rta	O‘rta (≈2×)	Kichik modelda yuqori	Yuqori	5,2-6,5× kam hisoblash
M3. Sifat	Odatda yaxshi saqlanadi. 90-95%	≈99%	O‘qitishga bog‘liq. >93%	Teng yoki yuqori	Yo‘qotishsiz
M4. Qayta o‘qitish	PTQ: yo‘q	Ko‘pincha yo‘q	Ha	Ha	Ha
M5. Apparatga bog‘liqlik	Yuqori	O‘rta	Nisbatan past	Past	O‘rta
M6. MCU / kam resursli til	Faqat INT8 / kalibrlash kerak	Cheklangan / kam o‘rganilgan	Mahalliy ma’lumotlar zarur.	Kichik modellar uchun mos. KWS, BERT-tiny / korpus kerak	Kam o‘rganilgan, Qo‘shimcha sinovlar zarur

Taqqoslash natijalari uchta muhim jihatni ko‘rsatadi. Birinchidan, eng yuqori ko‘rsatkichlar asosan grafik protsessorlardan foydalanilgan platformalarda qayd etilgan. Mikrokontrollerlarda esa atigi ikkita tadqiqot amaliy natijalarni ko‘rsatgan [13, 14]. Ikkinchidan, modelni siqish natijasida yuzaga keladigan sifat yo‘qotilishi murakkab sharoitlarda yaqqolroq namoyon bo‘ladi. Bu holat uzoq davom etadigan robot vazifalarida [4] va shovqinli nutqni tanishda [7] kuzatilgan. Uchinchidan, mavjud tadqiqotlarning aksariyati ingliz tilida olib borilgan. O‘zbek tilidagi ochiq nutq ma’lumotlari esa cheklangan bo‘lib, umumiy mavzudagi taxminan 105 soatlik audio yozuvlarni o‘z ichiga oladi [21].

Ochiq muammolar.

4-jadval. Asosiy ochiq muammolar va yo‘nalishlar

Muammo	Dalil	Yo‘nalish

Umumiy buyruqdan harakatgacha kechikishi o‘lchanmaydi	KWS (0,359 s) [14] va VLA reaksiyasi (144,3 ms) [16] alohida o‘lchangan	Butun jarayon uchun yagona benchmark yaratish
O‘zbek tilida robot buyruqlari korpusi yo‘q	Mavjud korpus umumiy mavzuli [21]	Ovozli robot buyruqlari to‘plamini yaratish
Siqilgan modellar ishonchliligi	Bir xil W4A4: 80,0%, aralash: 92,5% [6]	Xavfsizlikni hisobga oluvchi aralash kvantlash
Moslashuv real signallarga bog‘lanmagan	Usullar GPU va simulyatsiyada sinalgan [18, 19]	Batareya va haroratga bog‘langan boshqaruv

Mavjud tadqiqotlarda modelni siqish natijalari ko‘rsatilgan bo‘lsa-da, ularning resurslari keskin cheklangan mikrokontrollerlarda real vaqt rejimida ishlash imkoniyatlari yetarlicha asoslanmagan. Buyruqni tanib olishdan robotning fizik harakatni bajarishigacha bo‘lgan umumiy kechikish va tizim ishonchliligi yagona mezonlar asosida baholanmagan. O‘zbek tilidagi robot buyruqlari uchun maxsus korpusning yo‘qligi esa ishlab chiqilayotgan algoritmlarni mahalliy til sharoitida sinash va taqqoslashni cheklaydi. Bundan tashqari, modelning hisoblash yuklamasini qurilmaning joriy xotirasi, energiya sarfi va haroratiga mos ravishda boshqarish masalasi ham yetarlicha o‘rganilmagan.

Xulosa.

Tahlil natijalari kvantlash, bilimlarni uzatish va ixcham arxitekturalar robotlarda tabiiy til buyruqlarini tushunish tizimlarini resurslari cheklangan qurilmalarda ishlatish imkonini kengaytirayotganini ko‘rsatadi. Ayrim tadqiqotlarda xotira sarfi 5-11 baravar kamayib, model sifatining 90-95 foizi saqlangan. Biroq mikrokontrollerlarda asosan kalit so‘zlarni aniqlash usullari amaliy qo‘llanmoqda, buyruqdan harakatgacha bo‘lgan umumiy kechikish esa yetarlicha baholanmagan. Bundan tashqari, o‘zbek tilidagi robot buyruqlari uchun maxsus korpus mavjud emas. Shu sababli keyingi tadqiqotlarda yagona baholash mezonlarini ishlab chiqish, o‘zbek tilidagi buyruqlar korpusini yaratish va qurilma resurslariga moslashuvchi kaskadli tizimlarni ishlab chiqishga ustuvor ahamiyat berish lozim.

FOYDALANILGAN ADABIYOTLAR

1. Wen J., Zhu Y., Li J. [et al.]. TinyVLA: Toward Fast, Data-Efficient Vision-Language-Action Models for Robotic Manipulation. IEEE Robotics and Automation Letters. - 2025. - Vol. 10, No. 4. - P. 3988-3995.

2. Sikorski P., Schrader L., Yu K. [et al.]. Deployment of Large Language Models to Control Mobile Robots at the Edge. 2025, 3rd International Conference on Mechatronics, Control and Robotics (ICMCR). - Piscataway : IEEE, 2025. - P. 19-24. - DOI: 10.1109/ICMCR64890.2025.10963303.

3. Yu Z., Wang B., Zeng P. [et al.]. A Survey on Efficient Vision-Language-Action Models. IEEE Transactions on Pattern Analysis and Machine Intelligence. - 2026.

4. Wang H., Xiong C., Wang R., Chen X. BitVLA: 1-bit Vision-Language-Action Models for Robotics Manipulation. Conference on Robot Learning (CoRL 2026).

5. Akbari A., Akbari A., Eskandar M. [et al.]. ActQuant: Sub-4-bit Action-Guided Quantization for Vision-Language-Action Models : preprint. - 2026.

6. Ho H.T., Nguyen K.D., Nguyen Q.D. [et al.]. FoldQuantVLA: Native Low-Bit Quantization of Vision-Language-Action Models via Consistent Folding : preprint. - 2026. - arXiv : 2609.24433.

7. Söhler A., Irigoyen J., Kirkedal A.S. Quantizing Whisper-small: How Design Choices Affect ASR Performance. SPEAKABLE Workshop at LREC 2026. - 2026. arXiv:2511.08093.

8. Yang Y., Wang Y., Wen Z. [et al.]. EfficientVLA: Training-Free Acceleration and Compression for Vision-Language-Action Models. Advances in Neural Information Processing Systems. - 2025. - Vol. 38.

9. Ravichandran Z., Hounie I., Cladera F. [et al.]. Distilling On-device Language Models for Robot Planning with Minimal Human Intervention // Proceedings of the 9th Conference on Robot Learning (CoRL 2025). - PMLR, 2025. - P. 4859-4884.

10. Rastogi P., Kumar P., Guru Prasad M.S., Prithviraj. Optimizing NLP Models for Low-Resource Languages Using Knowledge Distillation. 2026 International Conference on Intelligent Computing and Automation for Sustainable Solutions (ICASS). - Piscataway : IEEE, 2026.

11. Shukor M., Aubakirova D., Capuano F. [et al.]. SmolVLA: A Vision-Language-Action Model for Affordable and Efficient Robotics. preprint. - 2025. - arXiv:2506.01844.

12. Jeffries N., King E., Kudlur M. [et al.]. Moonshine: Speech Recognition for Live Transcription and Voice Commands : preprint. - 2024. - arXiv:2410.15608.

13. Yang Z., Chen R., Wu T. [et al.]. MCUBERT: Memory-Efficient BERT Inference on Commodity Microcontrollers. Proceedings of the 43rd IEEE/ACM International Conference on Computer-Aided Design (ICCAD ’24). - New York : ACM, 2024. - Art. 52. - P. 1-9.

14. Pham H.V., Vu T.P., Nguyen H.T., Le M. Lightweight Keyword Spotting with Inter-Domain Interaction and Attention for Real-Time Voice-Controlled Robotics. EAI Endorsed Transactions on AI and Robotics. - 2025. - Vol. 4.

15. Islam M.R., Deng B., Dhar N. [et al.]. Characterizing and Understanding Energy Footprint and Efficiency of Small Language Model on Edges. 2025 IEEE International Conference on Mobile Ad-Hoc and Smart Systems (MASS). - Piscataway : IEEE, 2025.

16. Yang Z., Wang Q., Wang Y. [et al.]. EagleVLA: Towards Onboard Real-Time Robot Control via Foresight-Aligned Asynchronous Inference. Conference on Robot Learning (CoRL 2026). - 2026. arXiv:2607.12659.

17. Jin H., Wu Y. CE-CoLLM: Efficient and Adaptive Large Language Models Through Cloud-Edge Collaboration. 2025 IEEE International Conference on Web Services (ICWS). - Piscataway : IEEE, 2025. - P. 316-323.

18. Yue Y., Wang Y., Kang B. [et al.]. DeeR-VLA: Dynamic Inference of Multimodal Large Language Models for Efficient Robot Execution. Advances in Neural Information Processing Systems. - Red Hook : Curran Associates, 2024. - Vol. 37. - P. 56619-56643.

19. Zhang R., Dong M., Zhang Y. [et al.]. MoLe-VLA: Dynamic Layer-skipping Vision Language Action Model via Mixture-of-Layers for Efficient Robot Manipulation . Proceedings of the AAAI Conference on Artificial Intelligence. - 2026. - Vol. 40, No. 22. - P. 18764-18772.

20. Park Y., Hyun J., Cho S. [et al.]. Any-Precision LLM: Low-Cost Deployment of Multiple, Different-Sized LLMs. Proceedings of the 41st International Conference on Machine Learning (ICML 2024). - PMLR, 2024. - Vol. 235. - P. 39682-39701.

21. Musaev M., Mussakhojayeva S., Khujayorov I. [et al.]. USC: An Open-Source Uzbek Speech Corpus and Initial Speech Recognition Experiments. Speech and Computer : SPECOM 2021. - Cham : Springer, 2021. - P. 437-447. - (Lecture Notes in Computer Science ; Vol. 12997).