

Nabila Wulandari

Kecerdasan Buatan Multimodal dan Robotika: Keterbatasan Model Visi-Bahasa (VLM) dalam Persepsi Visual serta Integras...

 Quick Submit Quick Submit Universitas 17 Agustus 1945 Semarang

Document Details

Submission ID

trn:oid:::1:3645305876

Submission Date

Sep 10, 2026, 9:57 PM GMT+7

Download Date

Sep 10, 2026, 10:01 PM GMT+7

File Name

Nabila_Wulandari_-_jurnal.docx

File Size

122.3 KB

11 Pages

2,880 Words

20,450 Characters

17% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Top Sources

- 17%  Internet sources
- 0%  Publications
- 0%  Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Top Sources

17%  Internet sources
0%  Publications
0%  Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Internet	jurnal.muaraedukasi.id	7%
2	Internet	arxiv.org	2%
3	Internet	www.mdpi.com	1%
4	Internet	mjbass.com	<1%
5	Internet	docplayer.info	<1%
6	Internet	ges.research.ncsu.edu	<1%
7	Internet	repositori.ukwms.ac.id	<1%
8	Internet	dikti.go.id	<1%
9	Internet	m.tribunnews.com	<1%
10	Internet	ouci.dntb.gov.ua	<1%
11	Internet	repository.unas.ac.id	<1%

12	Internet	journal.unpas.ac.id	<1%
13	Internet	ojs.unida.info	<1%
14	Internet	text-id.123dok.com	<1%
15	Internet	www.neliti.com	<1%
16	Internet	www.unipa.it	<1%
17	Internet	thesesups.ups-tlse.fr	<1%
18	Internet	www.medrxiv.org	<1%
19	Internet	batampos.co.id	<1%
20	Internet	core.ac.uk	<1%
21	Internet	researchportal.rma.ac.be	<1%



Kecerdasan Buatan Multimodal dan Robotika: Keterbatasan Model Visi-Bahasa (VLM) dalam Persepsi Visual serta Integrasi Sensorik dan Pembelajaran Mendalam untuk Sistem Robot Cerdas

Nabila Wulandari^{1*}, sekar ayu pramesti²

^{1*2} Universitas Tadulako, Indonesia

Alamat: Jalan Soekarno Hatta KM. 9, Tondo, Kecamatan Mantikulore, Kota Palu, Sulawesi Tengah

Email ; nabilawulandari@untad.ac.id , sekarayupramesti@untad.ac.id

Korespondensi penulis: nabilawulandari@untad.ac.id

Abstract The integration of multimodal artificial intelligence in robotics has opened new opportunities for intelligent robotic systems capable of understanding their environment and interacting naturally with humans. However, fundamental limitations in Vision-Language Models (VLMs) regarding visual perception pose significant obstacles to realizing truly autonomous and adaptive robotic systems. This study examines the limitations of VLMs in spatial reasoning and visual perception, while exploring sensorimotor integration strategies and deep learning approaches to overcome these limitations. Through a systematic literature review of recent research, the findings reveal that VLMs face significant challenges in understanding fine-grained spatial relationships, particularly in the context of robotic manipulation and human hand interaction. The HandVQA benchmark reveals that VLM models tend to hallucinate hand parts, make geometric errors, and exhibit poor generalization on tasks involving complex hand poses. The phenomenon of "modality collapse" is also identified, where strong visual priors override sparse linguistic signals, causing robotic agents to ignore language commands and rely solely on visual cues. To address these limitations, multimodal sensorimotor integration approaches combining visual, tactile, audio, and proprioceptive data as well as deep learning architectures that integrate multimodal representations show significant potential. This study concludes that although VLMs have fundamental limitations in visual perception, intelligent sensorimotor integration and deep learning offer a promising pathway toward more robust and adaptive intelligent robotic systems.

Keywords: multimodal artificial intelligence, vision-language models, robotics, visual perception, sensorimotor integration, deep learning, intelligent robotic systems

Abstrak Integrasi kecerdasan buatan multimodal dalam robotika telah membuka peluang baru bagi sistem robot cerdas yang mampu memahami lingkungan dan berinteraksi secara alami dengan manusia. Namun, keterbatasan mendasar pada Model Visi-Bahasa (Vision-Language Models/VLM) dalam persepsi visual menjadi hambatan signifikan dalam mewujudkan sistem robot yang benar-benar otonom dan adaptif. Penelitian ini mengkaji keterbatasan VLM dalam penalaran spasial dan persepsi visual, serta mengeksplorasi strategi integrasi sensorik dan pembelajaran mendalam untuk mengatasi keterbatasan tersebut. Melalui tinjauan literatur sistematis terhadap penelitian terkini, ditemukan bahwa VLM menghadapi tantangan signifikan dalam memahami hubungan spasial yang halus, terutama dalam konteks manipulasi robotik dan interaksi tangan-manusia. Benchmark HandVQA mengungkapkan bahwa model VLM cenderung mengalami halusinasi bagian tangan, kesalahan interpretasi geometris, dan generalisasi yang buruk pada tugas-tugas yang melibatkan pose tangan yang kompleks. Fenomena "modality collapse" juga teridentifikasi, di mana prior visual yang kuat mengalahkan sinyal linguistik yang sparse, menyebabkan agen robot mengabaikan perintah bahasa dan hanya mengandalkan isyarat visual. Untuk mengatasi keterbatasan ini, pendekatan integrasi sensorik multimodal menggabungkan data visual, taktil, audio, dan proprioceptive serta arsitektur pembelajaran mendalam yang menggabungkan representasi multimodal menunjukkan potensi besar. Penelitian ini menyimpulkan bahwa meskipun VLM memiliki keterbatasan fundamental dalam persepsi visual, integrasi sensorik yang cerdas dan pembelajaran mendalam menawarkan jalur yang menjanjikan menuju sistem robot cerdas yang lebih robust dan adaptif.

Kata Kunci: kecerdasan buatan multimodal, model visi-bahasa, robotika, persepsi visual, integrasi sensorik, pembelajaran mendalam, sistem robot cerdas

Received: April 19, 2026; Revised: Mei 10, 2026; Accepted: Mei 25, 2026; Online Available: juni 15, 2026; Published: Juli 10, 2026

1. LATAR BELAKANG

1.1 Latar Belakang

Perkembangan kecerdasan buatan (AI) multimodal telah membawa perubahan signifikan dalam bidang robotika, memungkinkan robot untuk memahami dan berinteraksi dengan lingkungan melalui berbagai modalitas sensorik. Model Visi-Bahasa (Vision-Language Models/VLM) dan Model Visi-Bahasa-Aksi (Vision-Language-Action/VLA) telah menunjukkan kemampuan yang mengesankan dalam tugas-tugas kontrol robotik yang digeneralisasi. Model-model ini menggabungkan pemahaman visual, pemrosesan bahasa alami, dan generasi aksi, memungkinkan robot untuk mengikuti perintah manusia dan beroperasi di lingkungan yang kompleks.

Namun, di balik kemajuan ini, terdapat keterbatasan mendasar yang menghambat penerapan luas sistem robot cerdas. Penelitian terkini mengungkapkan bahwa VLM dan VLA sering kali mengembangkan "visi shortcut" atau "modality collapse," di mana model mengabaikan perintah bahasa dan hanya mengandalkan isyarat visual. Fenomena ini menyebabkan kegagalan dalam skenario di mana tugas bersifat ambigu atau lingkungan berubah.

1.2 Rumusan Masalah

Berdasarkan latar belakang di atas, penelitian ini merumuskan pertanyaan penelitian sebagai berikut:

- Apa saja keterbatasan fundamental VLM dalam persepsi visual dan penalaran spasial untuk aplikasi robotik?
- Bagaimana fenomena "modality collapse" mempengaruhi kemampuan robot dalam mengikuti perintah bahasa?
- Bagaimana integrasi sensorik multimodal dan pembelajaran mendalam dapat mengatasi keterbatasan VLM dalam sistem robot cerdas?

1.3 Tujuan Penelitian

Penelitian ini bertujuan untuk:

- Mengidentifikasi dan menganalisis keterbatasan VLM dalam persepsi visual dan penalaran spasial
- Mengeksplorasi fenomena "modality collapse" dan dampaknya terhadap kontrol robotik
- Mengkaji strategi integrasi sensorik dan pembelajaran mendalam untuk mengembangkan sistem robot cerdas yang lebih robust.

2. KAJIAN TEORITIS

2.1 Model Visi-Bahasa (VLM) dan Arsitektur Multimodal

Model Visi-Bahasa (VLM) adalah sistem AI yang menggabungkan pemahaman visual dan bahasa untuk melakukan berbagai tugas, termasuk menjawab pertanyaan visual, menghasilkan deskripsi gambar, dan mendukung pengambilan keputusan berbasis visual. Arsitektur VLM modern biasanya terdiri dari encoder visual (seperti CNN atau Vision Transformer) dan model bahasa besar (LLM) yang dihubungkan melalui mekanisme cross-attention atau proyeksi linier.

Dalam konteks robotika, VLM berfungsi sebagai fondasi untuk Model Visi-Bahasa-Aksi (VLA), yang memperluas kemampuan VLM dengan menambahkan modul generasi aksi. VLA memungkinkan robot untuk menerjemahkan pemahaman visual-linguistik menjadi tindakan fisik, menjadikannya komponen kunci dalam sistem robot cerdas.

2.2 Pembelajaran Sensorimotor dan Integrasi Multimodal

Pembelajaran sensorimotor mengacu pada proses di mana agen belajar menghubungkan persepsi sensorik dengan tindakan motorik. Dalam konteks robotika, ini melibatkan integrasi data dari berbagai sensor visual, taktil, proprioceptif, dan audio untuk membentuk representasi lingkungan yang kaya dan memungkinkan pengambilan keputusan yang tepat.

Integrasi multimodal adalah proses menggabungkan informasi dari berbagai modalitas sensorik untuk menghasilkan pemahaman yang lebih lengkap dan robust tentang lingkungan. Pendekatan ini dapat dilakukan pada berbagai tingkat: early fusion (menggabungkan data mentah), late fusion (menggabungkan keputusan), atau intermediate fusion (menggabungkan representasi fitur).

3. METODE PENELITIAN

3.1 Jenis Penelitian

Penelitian ini menggunakan metode studi literatur sistematis (*systematic literature review*). Metode ini dipilih karena memungkinkan peneliti untuk mengumpulkan, mengevaluasi, dan mensintesis secara komprehensif berbagai temuan ilmiah tentang keterbatasan VLM dalam persepsi visual serta integrasi sensorik dan pembelajaran mendalam untuk sistem robot cerdas.

Pendekatan ini cocok karena:

- Topik bersifat lintas disiplin (AI, robotika, ilmu kognitif)
- Informasi tersebar di berbagai jurnal dan prosiding konferensi
- Membutuhkan sintesis dari berbagai perspektif teoretis dan empiris

3.2 Sumber Data

Penelitian ini menggunakan tiga jenis sumber data:

3.2.1 Literatur Akademik

- Artikel jurnal peer-review dari IEEE Xplore, ScienceDirect, SpringerLink, dan arXiv
- Prosiding konferensi internasional (CVPR, ICLR, ICML, ACL)
- Buku dan bab buku tentang AI multimodal dan robotika

3.2.2 Literatur Teknis

- Laporan teknis dari laboratorium penelitian dan industri
- Dokumentasi benchmark dan dataset (HandVQA, MV-RoboBench)
- Paten dan aplikasi paten terkait

3.2.3 Literatur Pendukung

- Artikel tentang pembelajaran sensorimotor dan integrasi multimodal
- Publikasi tentang arsitektur deep learning untuk robotika
- Standar dan pedoman pengembangan sistem robot cerdas

Tabel 1 : Kriteria Inklusi dan Eksklusi

Kriteria Inklusi	Kriteria Eksklusi
Membahas VLM atau VLA dalam konteks robotik	Hanya membahas NLP atau computer vision tanpa aplikasi robotik
Fokus pada persepsi visual, integrasi sensorik, atau pembelajaran mendalam	Tidak membahas sistem robot cerdas
Diterbitkan dalam 5 tahun terakhir (2021-2026)	Diterbitkan sebelum 2021
Artikel peer-review atau prosiding konferensi	Publikasi non-akademik tanpa peer-review

4. HASIL DAN PEMBAHASAN

4.1 Keterbatasan VLM dalam Persepsi Visual dan Penalaran Spasial

4.1.1 Kelemahan dalam Memahami Pose Tangan dan Hubungan Spasial Halus

Penelitian terbaru mengungkapkan keterbatasan signifikan VLM dalam memahami hubungan spasial yang halus, terutama dalam konteks pose tangan yang kompleks. Benchmark HandVQA, yang dikembangkan oleh peneliti di UNIST dan diterima di CVPR 2026, menggunakan lebih dari 1,6 juta pertanyaan untuk menguji pemahaman VLM tentang anatomi tangan.

Hasil evaluasi mengungkapkan bahwa tanpa pelatihan khusus, beberapa VLM terkemuka (termasuk LLaVA, DeepSeek, dan Qwen-VL) berkinerja mendekati chance level pada beberapa pertanyaan spasial. Model-model ini mengalami kesulitan khusus dalam menilai jarak antar sendi, membuat kesalahan geometris, dan dalam beberapa kasus, merujuk pada bagian jari yang tidak ada dalam gambar (halusinasi).

Tabel 4.1 Kelemahan dalam Memahami Pose Tangan dan Hubungan Spasial Halus

Temuan HandVQA	Deskripsi
Halusinasi bagian jari	Model merujuk pada bagian tangan yang tidak ada
Kesalahan geometris	Kesalahan dalam menilai sudut dan jarak antar sendi
Generalisasi buruk	Performa menurun pada tugas yang tidak terlihat selama pelatihan
Akurasi jarak sendi	Hanya 16,20% sebelum fine-tuning, meningkat menjadi 90,79% setelah fine-tuning

4.2 Fenomena "Modality Collapse" dan "Visual Shortcut" dalam VLA

4.2.1 Definisi dan Mekanisme

Fenomena "*modality collapse*" atau "*visual shortcut*" adalah masalah fundamental dalam VLA di mana model mengembangkan ketergantungan berlebihan pada isyarat visual, mengabaikan perintah bahasa yang seharusnya menjadi panduan utama. Penelitian oleh Huazhong University of Science and Technology, Harbin Institute of Technology, dan HKUST mengidentifikasi bahwa dataset pelatihan robotik sering menciptakan bias di mana bahasa dapat diprediksi sepenuhnya dari observasi visual.

Dari perspektif Bayesian, strategi VLA dapat didekomposisi menjadi:

- a. $p(a|v)$: Prior visual murni (aksi apa yang mungkin dalam skenario ini?)
- b. $p(a|v, \ell)$: Posterior yang mempertimbangkan Bahasa

Ketika prior visual sangat tajam, model dapat memprediksi aksi hanya dari visual tanpa memperhatikan bahasa. Akibatnya, posterior menjadi degenerasi menjadi prior, dan model mengabaikan perintah Bahasa.

4.2.2 Bukti Empiris

Eksperimen pada model Qwen3VL-4B-GR00T menunjukkan bahwa model yang dilatih hanya dengan observasi visual (tanpa bahasa) mencapai tingkat keberhasilan 44,6% pada 24 tugas RoboCasa, sangat dekat dengan baseline dengan bahasa (47,8%). Ini menunjukkan bahwa model dapat berhasil tanpa ketergantungan pada perintah Bahasa.

Penelitian oleh Zhan et al. (2026) mengidentifikasi fenomena serupa, di mana "prior visual yang kuat mengalahkan sinyal linguistik yang sparse, menyebabkan agen overfit pada frasa instruksi tertentu sambil mengabaikan maksud semantik yang mendasari".

4.2.3 Solusi: LangForce dan Residual Semantic Steering

Untuk mengatasi masalah ini, peneliti mengembangkan LangForce, metode yang memperkuat ketergantungan model pada bahasa melalui pengenalan loss log-likelihood-ratio. Metode ini secara eksplisit mendorong model untuk menggunakan bahasa dalam memilih tindakan dengan membandingkan apa yang akan disimpulkan robot dari gambar saja dengan apa yang harus dilakukan ketika instruksi juga dipertimbangkan.

Residual Semantic Steering (RSS) adalah pendekatan lain yang menggunakan kerangka probabilistik untuk memisahkan physical affordance dari semantic execution. RSS memperkenalkan dua inovasi teoritis: (1) Monte Carlo Syntactic Integration, yang mengaproksimasi posterior semantik sejati melalui ekspansi distribusi yang digerakkan LLM, dan (2) Residual Affordance Steering, mekanisme decoding dual-stream yang secara eksplisit mengisolasi pengaruh kausal bahasa dengan mengurangi prior affordance visual.

4.3 Integrasi Sensorik Multimodal untuk Sistem Robot Cerdas

4.3.1 Integrasi Visual-Taktil untuk Manipulasi Robotik

Penelitian tentang event-driven visual-tactile sensing mengembangkan arsitektur Spiking Neural Network (SNN) yang menggabungkan representasi visual dan taktil untuk tugas klasifikasi dan manipulasi. Sistem ini menggunakan sensor taktil NeuTouch dengan 39 taxel dan kamera berbasis event untuk menangkap data visual dan taktil secara bersamaan.

Hasil eksperimen menunjukkan bahwa model yang dilatih dengan kedua modalitas (visual-tactile) mencapai akurasi klasifikasi yang lebih tinggi dibandingkan model yang hanya menggunakan satu modalitas saja. Integrasi ini memungkinkan robot untuk membedakan objek dengan lebih akurat, terutama dalam skenario di mana isyarat visual saja tidak cukup (misalnya, objek dengan penampilan serupa tetapi berat berbeda).

4.3.2 Fusion dengan Confidence-Weighting untuk Kolaborasi Manusia-Robot

Dalam konteks kolaborasi manusia-robot (HRC), penelitian mengembangkan arsitektur fusion yang menggunakan pembobotan berbasis kepercayaan untuk menggabungkan input dari berbagai modalitas. Pendekatan ini mengevaluasi reliabilitas sinyal audio dan visual sebelum membuat keputusan grounding, memungkinkan sistem untuk mengabaikan modalitas yang terdegradasi (misalnya, karena noise mesin) dan fokus pada modalitas yang lebih dapat diandalkan.

Modul confidence-weighted fusion menghitung bobot per-modalitas dari sinyal kualitas eksplisit melalui kombinasi linier, softmax yang dikontrol suhu, dan clipping terbatas, sehingga setiap keputusan dapat dilacak kembali ke sinyal yang menghasilkannya. Pendekatan ini menawarkan transparansi dan interpretabilitas yang lebih baik dibandingkan metode fusion berbasis pembelajaran murni.

4.4 Pembelajaran Mendalam untuk Persepsi dan Kontrol Robotik

4.4.1 Arsitektur Berbasis Transformer dan Graph Neural Networks

Pembelajaran mendalam modern untuk robotika increasingly menggunakan arsitektur berbasis Transformer dan Graph Neural Networks (GNN) untuk menangani kompleksitas data multimodal. Mekanisme self-attention memungkinkan model untuk menangkap interaksi cross-modal yang kompleks dan dependensi jangka panjang, sementara GNN memungkinkan reasoning tentang hubungan spasial-temporal dalam operasi assembly atau disassembly.

Pendekatan ini telah menunjukkan hasil yang menjanjikan dalam berbagai tugas robotik, termasuk navigasi, manipulasi, dan interaksi manusia-robot. Namun, tantangan tetap ada dalam hal generalisasi ke lingkungan yang tidak terstruktur dan kemampuan untuk belajar dari sedikit contoh.

4.4.2 Pembelajaran Sensorimotor Berbasis Thousand Brains Theory

Pendekatan alternatif untuk pembelajaran robotik adalah Thousand Brains Theory, yang mengusulkan bahwa korteks manusia terdiri dari ribuan "learning modules" yang masing-masing membangun model objek melalui eksplorasi sensorimotor.

Sistem Monty, yang diimplementasikan berdasarkan teori ini, menunjukkan kemampuan untuk melakukan pengenalan objek 3D yang robust dan deteksi pose melalui interaksi aktif dengan lingkungan. Sistem ini menggunakan sensor modules untuk memproses input mentah, learning modules untuk membangun model objek dan menginferensi identitas serta pose, dan motor system untuk menghasilkan aksi berdasarkan state dan goal yang diterima.

Keunggulan pendekatan ini adalah kemampuannya untuk belajar secara efisien dari sedikit contoh dan beradaptasi dengan cepat terhadap objek baru, berbeda dengan pendekatan deep learning konvensional yang memerlukan data pelatihan dalam jumlah besar.

4.5 Tantangan dan Arah Pengembangan Masa Depan

4.5.1 Tantangan Integrasi

Meskipun kemajuan signifikan telah dicapai, integrasi sensorik multimodal untuk sistem robot cerdas masih menghadapi berbagai tantangan:

- Kompleksitas Kalibrasi:** Mengintegrasikan sensor dengan karakteristik berbeda (visual, taktil, audio) memerlukan kalibrasi yang cermat dan sinkronisasi temporal yang presisi.
- Keterbatasan Komputasi:** Pemrosesan data multimodal secara real-time memerlukan sumber daya komputasi yang signifikan, terutama untuk aplikasi di edge.
- Generalisasi Domain:** Model yang dilatih pada satu domain sering kali gagal ketika diterapkan pada domain lain, memerlukan pendekatan transfer learning yang canggih.

- d. Keamanan dan Keandalan: Sistem robot cerdas yang beroperasi di lingkungan manusia harus memenuhi standar keamanan yang ketat, memerlukan mekanisme fail-safe dan verifikasi formal.

4.5.2 Arah Pengembangan Masa Depan

Berdasarkan analisis literatur, beberapa arah pengembangan masa depan dapat diidentifikasi:

- a. Integrasi VLM dengan Sensor Taktil dan Proprioceptif: Menggabungkan kemampuan VLM dengan data sensorik tambahan untuk mengatasi keterbatasan visual dan meningkatkan robustness.
- b. Pembelajaran Berbasis Interaksi Aktif: Mengadopsi pendekatan sensorimotor aktif di mana robot secara aktif mengeksplorasi lingkungan untuk membangun model yang lebih akurat.
- c. Arsitektur Hybrid Neuro-Symbolic: Menggabungkan pembelajaran mendalam dengan reasoning simbolik untuk meningkatkan interpretabilitas dan kemampuan reasoning.
- d. Transfer Learning dan Meta-Learning: Mengembangkan metode yang memungkinkan robot untuk belajar dari sedikit contoh dan beradaptasi dengan cepat terhadap tugas dan lingkungan baru.

5. KESIMPULAN DAN SARAN

5.1 Simpulan

Berdasarkan hasil tinjauan sistematis, dapat disimpulkan bahwa:

- a. VLM memiliki keterbatasan fundamental dalam persepsi visual dan penalaran spasial, terutama dalam memahami hubungan spasial yang halus seperti pose tangan dan jarak antar sendi. Benchmark HandVQA mengungkapkan bahwa model VLM cenderung mengalami halusinasi dan kesalahan geometris pada tugas-tugas yang memerlukan pemahaman anatomi yang detail.
- b. Fenomena "modality collapse" atau "visual shortcut" merupakan masalah signifikan dalam VLA, di mana prior visual yang kuat mengalahkan sinyal linguistik, menyebabkan robot mengabaikan perintah bahasa dan hanya mengandalkan isyarat visual. Hal ini menyebabkan kegagalan dalam skenario di mana tugas bersifat ambigu atau lingkungan berubah.
- c. Integrasi sensorik multimodal menawarkan solusi yang menjanjikan untuk mengatasi keterbatasan VLM. Dengan menggabungkan data visual, taktil, audio, dan proprioceptif, sistem robot dapat membentuk representasi

lingkungan yang lebih kaya dan robust, memungkinkan pengambilan keputusan yang lebih akurat.

- d. Pembelajaran mendalam dengan arsitektur hybrid menggabungkan Transformer, Graph Neural Networks, dan pendekatan neuro-symbolic menunjukkan potensi besar untuk mengatasi kompleksitas data multimodal dan meningkatkan kemampuan reasoning robotik.

5.2 Implikasi

Bagi Peneliti:

- a. Mengembangkan benchmark yang lebih komprehensif untuk mengevaluasi keterbatasan VLM dalam konteks robotik
- b. Mengeksplorasi arsitektur integrasi sensorik yang lebih efisien dan robust
- c. Mengembangkan metode pembelajaran yang memungkinkan robot untuk belajar dari sedikit contoh dan beradaptasi dengan cepat

Bagi Praktisi:

- a. Mengadopsi pendekatan integrasi sensorik multimodal untuk meningkatkan robustness sistem robot
- b. Mempertimbangkan arsitektur fusion yang transparan dan interpretabel untuk aplikasi yang memerlukan akuntabilitas
- c. Mengembangkan sistem yang dapat mendeteksi dan mengatasi kegagalan modalitas (misalnya, sensor yang terdegradasi)

DAFTAR REFERENSI

- 2 Lian, S., Yu, B., Lin, X., Yang, L., Shen, Z., Wu, C., Miao, Y., Huang, C., & Chen, K. (2026). LangForce: Bayesian Decomposition of Vision Language Action Models via Latent Action Queries. ICML 2026.
- Sayem, M. K. C., Chowdhury, M. T., & Baek, S. (2026). HandVQA: A Large-Scale Diagnostic Benchmark for Fine-Grained Hand Understanding in Vision-Language Models. CVPR 2026.
- 3 MDPI. (2026). A Digital Twin Framework for Multimodal Operator-Centered Human-Cobot Collaboration in Assembly Tasks. *Machines*, 14(7), 780.
- Frontiers. (2026). Multimodal perception and information fusion for embodied agents. *Frontiers in Neurorobotics*.
- 3 MDPI. (2026). Mobile Robot Localization and SLAM: A Critical Review of Sensors, Multi-Sensor Fusion, and Neural Representations. *Robotics*, 15(8), 142.

Zhan, Z. (2026). Residual Semantic Steering for Vision-Language-Action Models. ACL 2026.

2 Tsinghua University. (2026). MV-RoboBench: Benchmarking Spatial Reasoning of Vision-Language Models in Robotic Scenes. ICLR 2026.

17 National University of Singapore. (2025). Event-Driven Visual-Tactile Sensing and Learning for Robots. US Patent Application y2025/0214255.

Wiley. (2026). Intelligent Maintenance Review for Robots: Multimodal Information, Deep Diagnosis and Embodied Artificial Intelligence. Advanced Robotics Research.

10 MIT Press. (2026). Thousand-Brains Systems: Sensorimotor Intelligence for Rapid, Robust Learning and Inference. Neural Computation, 38(6), 845.

6 Zhou, Y.-H., & Sun, G. (2025). Detecting Algorithmic Bias in ICU Clinical Decision: A Doubly-Robust Framework for Auditing Treatment Recommendation Disparities. medRxiv, 2025.11.16.25340359. <https://doi.org/10.1101/2025.11.16.25340359>

18 2 Zhang, E. (2026). Uncovering Latent Bias in LLM-Based Emergency Department Triage Through Proxy Variables. arXiv preprint, arXiv:2601.15306. <https://arxiv.org/abs/2601.15306>

Ginkgo Bioworks. (2025). Virtual Cell Pharmacology Initiative (VCPI): Building a Community-Driven Data Standard for AI Drug Discovery. Ginkgo Datapoints. <https://thevirtualcell.com>

Kementerian Pendidikan Tinggi, Sains, dan Teknologi. (2025). Kemdiktisaintek dan Komdigi Bahas Kolaborasi Riset dan AI. kemdiktisaintek.go.id. <https://kemdiktisaintek.go.id>

13 8 Kementerian Pendidikan Tinggi, Sains, dan Teknologi. (2026). 18.215 Riset Kampus Terima Pendanaan Rp1,7 Triliun, Kemdiktisaintek Dorong Solusi Nyata bagi Masyarakat. kemdiktisaintek.go.id. <https://kemdiktisaintek.go.id>

RRI.co.id. (2026). Kemendiktisaintek Dorong Konsorsium Riset Genomik Nasional. RRI.co.id. <https://berita.rri.co.id>

4 Springer. (2025). Identifying E-Commerce Fraud Through User Behavior Data: Observations and Insights. Data Science and Engineering, 10(1), 24-39