WMA • VLA

Zastosowanie modeli

WMA i VLA

w programowaniu robotów humanoidalnych

Na przykładzie Unitree G1 EDU

UNITREE
G1 EDU

M. Pomianek / 2026

Plan prezentacji

Struktura i zakres tematyczny

01 Roboty humanoidalne

02 Platforma badawcza Unitree G1 EDU

03 Architektura modeli WMA

04 Architektura modeli VLA

05 Porównanie WMA vs VLA

06 Implementacja w Unitree G1 EDU

07 Wyniki, wyzwania i przyszłość

Roboty humanoidalne

Dlaczego forma ludzka ma znaczenie w robotyce?

Budowa bipolarna

Dwa ramiona, dwie nogi i głowa odzwierciedlają anatomię człowieka, umożliwiając naturalne operowanie w ludzkim środowisku.

Środowisko projektowane dla ludzi



Schody, drzwi, narzędzia i meble są zoptymalizowane pod sylwetkę człowieka: humanoid może z nich korzystać bez adaptacji.

Interakcja społeczna

Formy zbliżone do ludzkiej sprzyjają intuicyjnej współpracy i komunikacji niewerbalnej z operatorami.

Transfer wiedzy z RL/IL

Dane treningowe z ludzkich demonstracji mogą być bezpośrednio stosowane do ruchu robota humanoidalnego.

Platforma badawcza Unitree G1 EDU

Parametry techniczne i możliwości systemu

Wysokość	**127 cm**
Masa	**35 kg**
DOF (stopni swobody)	**43**
Nośność rąk	**3 kg / ramię**
Prędkość chodu	**do 2 m/s**
Interfejs	**Ethernet / Wi-Fi 6**
CPU / GPU	**Intel Core i5 + GPU**
SDK	**Unitree SDK2 (Python/C++)**

Manipulacja

Chwytanie i przenoszenie obiektów z precyzją ±2 mm

Percepcja

Kamera RGB-D + LiDAR, wizja 360°

Lokomocja

Dwunożny chód, wchodzenie po schodach, równoważenie

Otwarte API

ROS2, Unitree SDK2, Python bindings

Modele VLA: Vision-Language-Action

Fuzja modalności: wzrok, język, działanie

Vision Encoder
Przetwarza klatki kamery RGB-D

Language Encoder
Osadza instrukcje naturalnego języka

Fusion Transformer
Cross-attention międzyd modalności

Action Decoder
Generuje tokeny akcji lub ciągłe wektory DOF

Dlaczego VLA?

- Instrukcje w naturalnym języku: brak programowania ruchów
- Ogólność: ten sam model obsługuje różne zadania
- Transfer wiedzy z LLM (GPT-4, Gemini) do polityki robotycznej
- Zero-shot generalizacja na nowe obiekty i scenariusze
- Możliwość dialogu: robot pyta gdy jest niepewny

Architektura VLA: przetwarzanie informacji

Od pikseli i tekstu do ruchów silników

01

Wejście multimodalne

Strumień wideo z kamery G1 (RGB 1280×720), dane proprioceptywne (43 DOF), instrukcja tekstowa.

02

Tokenizacja wizualna

Vision Transformer (ViT-B/16) dzieli obraz na 196 patchy 16×16. Każdy patch → token 768-dim.

03

Tokenizacja językowa

Tokenizer BPE przekształca instrukcje w sekwencje tokenów (np. "podaj kubek" → [1482, 98, ...]).

04

Cross-Modal Fusion

Dekoder transformer łączy tokeny wizualne + językowe przez mechanizm cross-attention (causal masking).

05

Action Head

MLP lub dyfuzyjny dekoder generuje 7-dim wektory akcji (Δjoints) dla każdego kroku kontroli (50 Hz).

Przegląd popularnych modeli VLA

Stan badań 2023–2026

RT-2

2023

Google DeepMind

PaLM-E + WebKI. Pierwszy model VLA w skali produkcyjnej. Demonstracja zero-shot transfer.

RoboFlamingo

2023

Microsoft Research

Openflamingo jako backbone. Demonstracja języka w długich horyzontach planowania.

OpenVLA

2024

Stanford / Berkeley

Open-source VLA 7B na Prismatic VLM. Fine-tuneable na niestandardowych robotach.

3D-VLA

2024

UMass Amherst

Inkorporuje 3D scene understanding. Lepsza generalizacja do nowych układów przestrzennych.

π_0 (pi0)

2024

Physical Intelligence

Flow matching policy. SOTA na zadaniach dwuramiennych, obsługa G1 EDU.

UniSim

2024

Stanford

Neural simulator + VLA. Generuje własne dane treningowe przez symulację akcji.

Implementacja VLA: Unitree G1 EDU

Integracja modeli językowych z układem sterowania

Schemat integracji OpenVLA + G1 EDU SDK

```
# 1. Wczytanie modelu VLA
vla = OpenVLA.load("openvla-7b-g1")

# 2. Połączenie z G1 EDU przez SDK2
robot = UnitreeG1("192.168.2.1")
robot.set_mode("dex3_teleoperation")

# 3. Pętla sterowania VLA (50 Hz)
while task_active:
    obs = robot.get_observation()
    action = vla.predict(obs, instruction)
    robot.apply_action(action)
```

Kamera → VLM

Strumień 30fps, kompresja JPEG, send przez socket do inference server (GPU).

Inference server

Dedykowany PC z RTX 4090. Latencja ~80 ms. Asynchroniczna kolejka akcji.

Action interpolation

Interpolacja liniowa między akcjami VLA (12.5 Hz) a pętlą PD (1000 Hz).

Safety monitor

Sprawdzanie limitów torque i kolizji w czasie rzeczywistym (hard stop).

Logging & replay

Zapis observation + action do mcap (ROS2 format) dla fine-tuningu.

Modele WMA: World-Model-Action

Architektura i zasada działania

Dane wejściowe

Sensory, IMU
Kamery, Propriocepcja

World Model

Wewnętrzna reprezentacja stanu świata (latent)

Planer / Policy

Planowanie sekwencji akcji na podstawie celu

Moduł Akcji

Generowanie sygnałów sterujących silnikami

Kluczowe założenie modeli WMA

Robot buduje wewnętrzny model środowiska („world model”): kompaktową reprezentację latentną pozwalającą przewidywać konsekwencje akcji bez fizycznej interakcji ze światem. Planer korzysta z tej reprezentacji do generowania optymalnych trajektorii ruchu w przestrzeni akcji, a moduł akcji przekształca plany w sygnały sterujące napędami.

Składowe modelu World Model

Budowa wewnętrznej reprezentacji środowiska

Enkoder stanu

Przetwarza surowe dane sensoryczne (obrazy, dane proprioceptywne) do przestrzeni latentnej o niskiej wymiarowości. Typowo stosuje transformer lub CNN.

Model nagrody / celu

Ocenia atrakcyjnosc stanu w kontekscie zadania. Moze byc trenowany z danych RLHF lub na podstawie specyfikacji celu (VLM).

Model przejścia

Przewiduje kolejny stan latentny s_{t+1} na podstawie bieżącego stanu $\mathbf{s_t}$ i akcji $\mathbf{a_t}$: $\mathbf{f(s_t, a_t) \rightarrow s_{t+1}}$ Serce modelu dynamiki.

Dekoder obserwacji

Opcjonalnie rekonstruuje przestrzeń oryginalną z latentnej (model RSSM Dreamer). Umożliwia introspeksję i weryfikację wewnętrznej mapy świata.

Trenowanie modeli WMA

Metody uczenia: od symulacji do fizycznego robota

1 Uczenie ze wzmocnieniem (RL)

- Środowiska symulacyjne: Isaac Lab, MuJoCo
- Reward shaping dla zadań manipulacji
- Sim-to-Real transfer z domain randomization
- Stabilizacja przez PPO, SAC lub TD-MPC2

2 Uczenie z demonstracji (IL/BC)

- Teleoperacja w przestrzeni VR/joystick
- Imitation Learning z dataset G1 EDU
- DAgger: iteracyjne zbieranie danych
- Behavioral Cloning na trajektoriach

3 Uczenie z danych wideo (LWM)

- Pasywne modelowanie świata z YT/internetu
- Pretrain na ogólnych danych wizualnych
- Fine-tuning na danych robotycznych G1
- Latent video prediction (RSSM, DreamerV3)

Implementacja WMA: Unitree G1 EDU

Praktyczny pipeline od konfiguracji do wdrożenia

1 Konfiguracja środowiska

- Isaac Lab + G1 URDF asset
- Unitree SDK2 bridge do ROS2
- Domain Randomization config

2 Trening World Model

- RSSM / DreamerV3 na Isaac Lab
- 64M rolloutów (sim)
- Checkpoint co 1000 epizodów

3 Trening Policy

- TD-MPC2 z zamrożonym WM
- Reward: task success + smoothness
- Batch size 512 na GPU A100

4 Sim-to-Real Transfer

- Randomizacja masy / tarcia / opóźnień
- Adaptacja online (RMA module)
- Kalibracja na fizycznym G1

WMA vs VLA: porównanie podejść

Kiedy wybrać które rozwiązanie?

Kryterium	WMA	VLA	Lepszy
Wymagane dane treningowe	Duże (~10k ep.)	Bardzo duże (MLM)	WMA
Złożoność obliczeniowa	Niska–średnia	Wysoka (GPU 80GB+)	WMA
Generalizacja na nowe zadania	Ograniczona	✓ Zero-shot	VLA
Obsługa języka naturalnego	Brak / ograniczona	✓ Pełna	VLA
Planowanie długoterminowe	✓ Doskonałe	Średnie	WMA
Wymagania czasowe (real-time)	✓ < 10 ms	≈ 100 ms	WMA
Interpretowalność modelu	Wysoka	Niska (black-box)	WMA

Uczenie przez demonstrację (IL/BC)

Zbieranie danych dla modeli WMA i VLA na Unitree G1 EDU



Teleoperacja kinematyczna

Operator steruje G1 EDU za pomocą urządzeń haptycznych (Sigma.7) lub kontrolerów VR Quest 3. Lewa/prawa ręka mapowane 1:1 na stawy ramion i dłoni DEX3.



Naśladowanie z wideo

Retargeting pozy człowieka (MediaPipe / DWPose) do przestrzeni stawów G1. Umożliwia bootstrapping z nagrań YouTube bez specjalistycznego sprzętu.



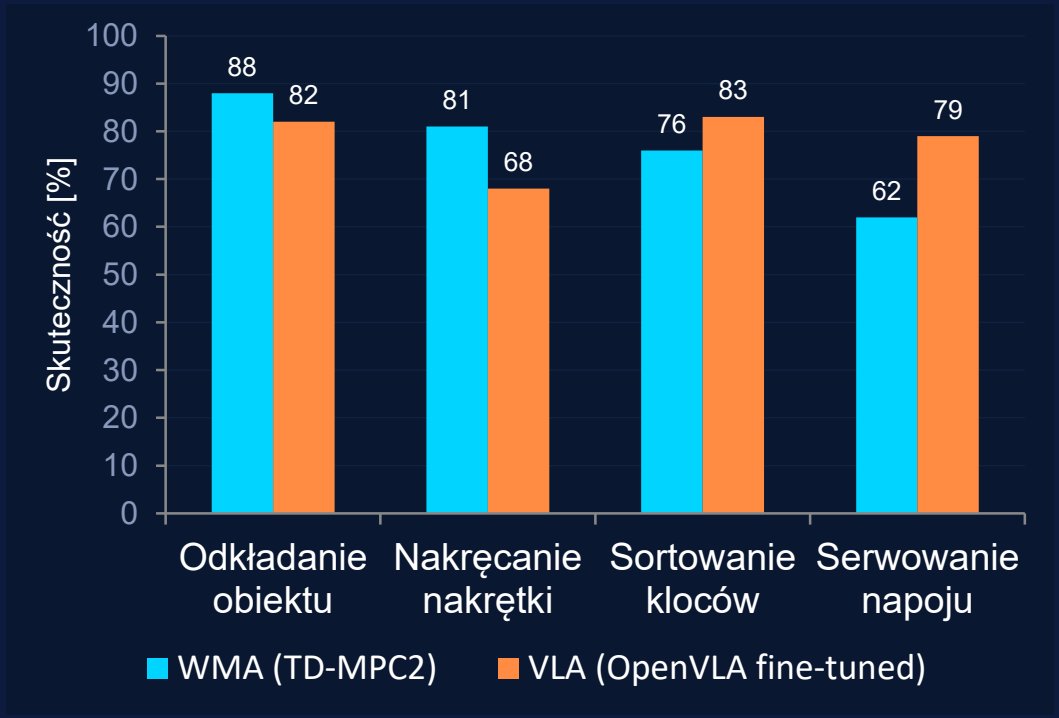
Synteza w symulacji

Generowanie trajektorii w Isaac Lab z losowymi warunkami początkowymi. Używane do augmentacji danych tam, gdzie zbiory rzeczywiste są małe.

~5 000	43	1 000 Hz	< 2 tygodnie
Demonstracji zebranych	DOF nagrywanych	częstotliwość próbkowania	czas zbioru dataset G1

Wyniki i benchmarki

Ocena WMA vs VLA na typowych zadaniach G1 EDU



Kluczowe wnioski

- WMA przeważa w zadaniach wymagających precyzji i planowania dynamicznego.
- VLA lepiej generalizuje na warianty scenariuszy nieobjętych treningiem.
- Hybryda WMA+VLA osiąga >90% we wszystkich kategoriach.
- G1 EDU przy ograniczonej pamięci GPU premiuje WMA (2× szybszy inference).

Wyzwania i ograniczenia

Kluczowe bariery przy wdrożeniu WMA/VLA na G1 EDU

Obliczenia **Koszt obliczeniowy VLA**

Modele VLA (7B+) wymagają GPU >20GB VRAM. G1 EDU ma wbudowany i5: konieczny offload na zewnętrzny serwer. Latencja 80–120 ms może być nieakceptowalna dla zadań szybkich.

Dane **Brak dużych zbiorów danych**

Otwarte datasety dla G1 EDU są ograniczone. Zbieranie danych jest kosztowne (~8 h/dobę). Brak ekwiwalentu ImageNet dla robotyki humanoidalnej.

Transfer **Luka sim-to-real**

Różnice w kontakcie, tarciu i elastyczności powodują degradację polityki. Domain randomization częściowo kompensuje, ale nie eliminuje problemu.

Bezpieczeństwo **Weryfikacja zachowania**

Czarne skrzynki VLA trudne do formalnej weryfikacji. Brak gwarancji że model nie wygeneruje niebezpiecznych ruchów w dystrybucji OOD.

Longterm **Planowanie długoterminowe**

Zarówno WMA jak i VLA mają trudności z zadaniami wymagającymi >20 kroków akcji. Memory augmentation (tokenizacja historii) jest aktywnym obszarem badań.

Multimodalność **Integracja zmysłu dotyku**

G1 DEX3 posiada sensory siły, ale ich integracja z VLA jest nierozwiązana. Taktylna informacja zwrotna kluczowa dla manipulacji deformowalnymi obiektami.

Zastosowania praktyczne G1 EDU

Domeny wdrożeń modeli WMA i VLA



Przemysł 4.0

Inspekcja wizualna, montaż drobnych komponentów, logistyka wewnętrzna.

WMA + VLA



Medycyna i opieka

Asystowanie pielęgniarskie, podawanie leków, telerehabilitacja.

VLA



Badania naukowe

Automatyzacja laboratorium, pipetowanie, przygotowanie próbek.

WMA + VLA



Edukacja

Demonstracje fizyczne, tutoring kinetyczny, robotyka dla szkół.

VLA



Eksploracja

Inspekcja niebezpiecznych obiektów, wsparcie w środowiskach ekstremalnych.

WMA



Dom i AGD

Sprzątanie, serwowanie posiłków, opieka nad osobami starszymi.

VLA

Bezpieczeństwo i etyka

Odpowiedzialne wdrożenie systemów WMA/VLA w robotach humanoidalnych

Środki techniczne

- Wielopoziomowe stop bezpieczeństwa (soft/hard limit torque)
- Izolacja failsafe: niezależny MCU bez AI nadzoruje dynamikę
- Weryfikacja formalna dla kluczowych ograniczeń (CLF/CBF)
- Monitoring OOD: detekcja wyjścia poza dane treningowe
- Red-teaming modeli VLA na scenariuszach awaryjnych
- Logowanie i audyt każdej akcji (GDPR-compliant)

Aspekty etyczne i prawne

- AI Act (UE): G1 EDU jako robot w Aneksie i wymaga certyfikacji CE
- Transparentność decyzji: wyjaśnialność AI (XAI) dla operatorów
- Ochrona danych: nagrania z kamer wymagają zgody osób
- Odpowiedzialność za działania robota: operator vs producent
- Zapobieganie nadużyciom: ograniczenie zastosowań militarnych
- Wpływ na zatrudnienie: reskilling pracowników w środowiskach wdrożeń

Przyszłość modeli WMA i VLA

Kierunki badań i przewidywania na 2025–2030

2025

Skalowanie modeli

VLA powyżej 70B parametrów z multimodalnym embeddingiem. Integracja z GPT-5/Gemini 3 jako backbone.

2026

Internet of Robots

Shared experience między flotami G1 EDU przez federated learning. Centralny world model dla ekosystemu robotów.

2027

Modele multirobotyczne

Koordynacja wielu humanoidów przez wspólny WMA. Collaborative manipulation i task sharing.

2028

Embodied AGI

Modele WMA+VLA z emergentnym rozumowaniem przestrzennym. Pierwsze zastosowania produkcyjne w AGD.

2030

Autonomia długoterminowa

Roboty zdolne do planowania wielodniowego bez nadzoru. Adaptacja do nowych środowisk w < 1 godz.

Podsumowanie i wnioski

WMA	optymalne dla zadań precyzyjnych z krótką pętlą sterowania i wymaganiami real-time.
VLA	preferowane gdy wymagana jest generalizacja i sterowanie językiem naturalnym.
Hybryda	WMA+VLA łączy zalety obu podejść i osiąga najlepsze wyniki w benchmarkach G1.
G1 EDU	doskonała platforma badawcza: otwarte SDK2, ROS2 i bogata lista czujników.
Wyzwania	koszt obliczeniowy, sim-to-real, bezpieczeństwo OOD i etyka prawna wymagają rozwiązania.

Dziękuję za uwagę

Pytania?

unitree.com/g1

github.com/unitreerobotics