WAVE-VLA: A WHOLE-BODY HUMANOID VLA WITH EXECUTABLE VISION–MOTION CODES LEARNED FROM HETEROGENEOUS HUMAN AND ROBOT DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Human video–motion offers behaviors beyond scarce humanoid demonstrations, but embodiment and supervision differ across sources. We present WAVE-VLA (*Whole-body Action Vocabulary for Execution*), which learns from task-unpaired G1 demonstrations and egocentric human video–motion. Both sources enter a common G1 action space and shared vision–motion codebook to extend one policy's repertoire while retaining robot skills. Current RGB, state, and language select body and hand codes that decode into synchronized references. A scene-conditioned Action Expert predicts a residual from the body reference, which zero residual preserves exactly. On HE and Nymeria development cohorts, refinement reduces mean forward-kinematics error relative to decoded references by 21.1% and 15.2%, respectively. The same frozen body–hand policy completes 39 of 60 real-G1 trials across four HE-covered and two Nymeria-covered tasks without task-specific parameter updates or retrieved trajectories. The HE-only baseline scores 27/40 on HE-covered and 0/20 on Nymeria-covered tasks, versus 25/40 and 14/20 for the joint policy in the reported robot trials.

INTRODUCTION

Whole-body humanoid VLAs turn egocentric observations and language into locomotion and manipulation, but their coverage depends on costly, narrow robot demonstrations Brohan et al. (2022); Kim et al. (2024); Bjorck et al. (2025). Egocentric human video–motion offers broader behavior Ding et al. (2025); Chen et al. (2026); Xie et al. (2026); Yue et al. (2026), yet independently collected sources can differ in tasks, embodiment, coordinates, and supervision. The challenge is to give this task-unpaired experience consistent target-robot semantics so one policy can acquire human behaviors while retaining robot skills.

Prior work uses shared corpora, related tasks, trajectory interfaces, language–motion alignment, or robot post-training Open X-Embodiment Collaboration (2024); Xu et al. (2023); Chi et al. (2024); Ding et al. (2025); Chen et al. (2026); Yue et al. (2026). Our independently collected sources have no explicit task correspondences and different supervision. Discrete action representations offer a bridge Lee et al. (2024); Ye et al. (2024); Pertsch et al. (2025); Kang et al. (2026); Song et al. (2025); Chen et al. (2026), but a common ID does not equate motions across bodies, and optional code conditioning need not govern a from-noise Action Expert. Transfer requires target-robot action semantics and a control interface committed to the selected plan.

Nymeria targets *behavioral coverage*: the policy should represent and select human-covered behaviors while maintaining HE-covered skills. Motion reconstruction tests support; current-only selection and real-robot execution test its use. HE-only robot trials assess added task coverage and HE retention.

WAVE-VLA maps native HE demonstrations and Nymeria human video–motion into one 93D G1 body/root space. Modality-routed training learns a visually grounded body codebook and synchronized, body-conditioned hand stream without fabricating absent labels. Current-only selectors decode aligned 16-frame body and hand plans; future-feature prediction supervises representation

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

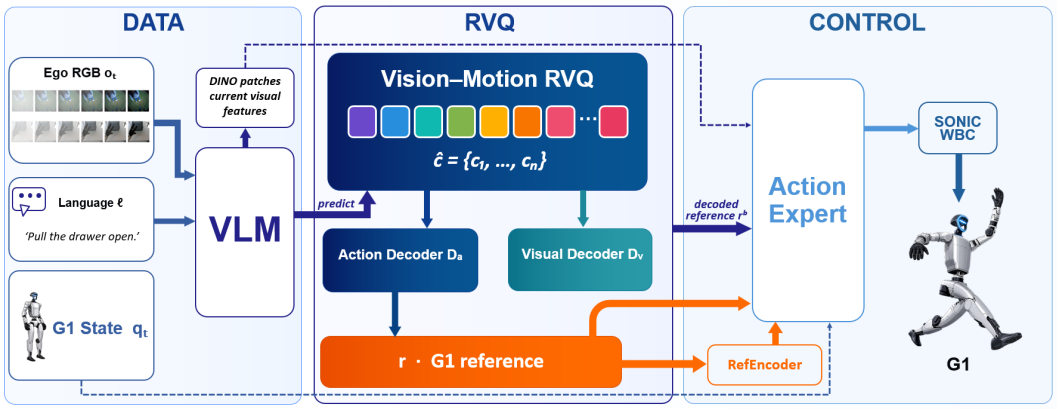


Figure 1: WAVE-VLA inference architecture. Current egocentric RGB, language, and G1 state predict discrete vision–motion codes. The frozen action decoder instantiates a G1 reference, while current DINO patches, state, and the decoded plan condition a reference-preserving Action Expert before tracking by the SONIC whole-body controller. The visual decoder predicts the anticipated visual feature for representation supervision and evaluation; it does not enter control. For clarity, the figure shows the body-reference path; the synchronized body-conditioned hand stream is detailed in the Methods.

learning but never enters control. Finally, *decode before control* makes the decoded trajectory the Action Expert's base action, leaving it to predict only a scene-conditioned residual.

We evaluate reconstruction, current-only selection, and residual control before physical execution. One codebook reconstructs held-out HE and Nymeria motion; refinement reduces decoded-reference FK error by 21.1%/15.2%. The frozen body–hand policy completes 39/60 real-G1 trials across four HE-covered and two Nymeria-covered tasks, while oracle hand codes reduce active-finger fingertip error by 26.0%. In robot tests, HE-only scores 27/40 on HE-covered tasks and 0/20 on Nymeria-covered tasks, versus 25/40 and 14/20 for the two-source policy.

Our contributions are:

- **Task-unpaired data unification.** We map native G1 egocentric video–action records and G1-retargeted egocentric human video–motion into one 93D G1 space. Human motion supplies additional target-robot behavior supervision without matched tasks; body supervision comes from both sources, while hand supervision comes from HE.
- **Visually grounded body–hand plan vocabulary.** We introduce a fixed-rate 112-bit body code that jointly represents target-G1 body/root motion and its egocentric visual consequence, together with a synchronized 112-bit hand stream conditioned on the predicted body plan. Leakage-safe selectors instantiate both as aligned 16-frame target-G1 trajectories from current observations, state, and language.
- **Decoded-plan residual control.** We introduce *decode before control*: predicted codes instantiate a coarse target-robot trajectory that becomes the Action Expert's base action, and the expert learns a current-scene-, state-, and language-conditioned residual. Zero correction exactly recovers the decoded plan, and matched evaluations demonstrate the benefit of refining this reference.

Related Work

VLA models. VLA policies map image–language context to robot actions through two dominant output interfaces. RT-1 and OpenVLA formulate control as prediction over quantized action symbols, whereas GR00T N1 and π_0 attach continuous generative action experts; Open X-Embodiment supplies a common data substrate across robot embodiments Brohan et al. (2022); Kim et al. (2024); Bjorck et al. (2025); Black et al. (2024); Open X-Embodiment Collaboration (2024). Humanoid systems add the requirement that semantic decisions become dynamically feasible whole-body motion.

Humanoid-VLA combines language–motion pre-alignment with egocentric adaptation, while LeVERB and WholeBodyVLA connect semantic policies to latent whole-body or loco-manipulation controllers; ZeroWBC further explores generalizable whole-body visuomotor control Ding et al. (2025); Xue et al. (2025); Jiang et al. (2025); Yang et al. (2026). Across these designs, a robot-compatible action or skill interface is specified before policy learning.

Human-to-humanoid learning. First-person human video is especially relevant to VLA because its observation geometry is closer to onboard robot sensing, while synchronized motion supplies full-body kinematics; neither alone resolves embodiment. Existing methods provide this bridge in three ways. Semantic alignment discovers shared visual skill prototypes or pre-aligns language and human motion Xu et al. (2023); Ding et al. (2025); interface alignment chooses transferable control coordinates or explicitly converts human behavior into robot trajectories Chi et al. (2024); Xie et al. (2026); and representation alignment learns a shared latent space or combines human pre-training with robot-specific adaptation Chen et al. (2026); Yue et al. (2026). These mechanisms address embodiment mismatch through different assumptions about shared semantics, coordinates, or downstream adaptation. Less studied is the regime in which task inventories are collected independently and modality availability differs by source, leaving no cross-source positive pairs from which to learn alignment.

Action tokenization and control. Latent-action methods differ principally in what grounds their bottleneck. Action-only tokenizers compress demonstrated control trajectories, with VQ-VLA further studying how such tokenizers scale with heterogeneous synthetic robot data Lee et al. (2024); Pertsch et al. (2025); Dong et al. (2026); Wang et al. (2025). Video-derived methods infer latent transitions without action labels Ye et al. (2024), while JALA aligns context-predictive states with inverse-dynamics latents and available actions to learn from heterogeneous human video Luo et al. (2026a). Multimodal methods additionally introduce reconstruction, foundation-model features, latent futures, or state-aware decoding Kang et al. (2026); Song et al. (2025); Zhao et al. (2025a); Luo et al. (2026b); Jiang et al. (2026). Among joint models, UniT uses cross-reconstruction to learn fused visual–action tokens with dual decoders Chen et al. (2026). This establishes joint discretization as an alignment mechanism, while leaving open how a decoded token should constrain a continuous controller.

A distinct issue is the physical role assigned to a predicted code. Existing tokenized or latent-plan VLAs primarily use code embeddings, token-supervised context, or predicted future representations as conditions for a continuous Action Expert, which still generates the executable chunk rather than inheriting a decoded trajectory Chen et al. (2026); Yue et al. (2026); Kang et al. (2026). Conditioning may bias generation, but it does not commit the controller to the motion represented by the code. Residual robot policies preserve a base controller by learning additive corrections Johannink et al. (2019), while Flow Matching permits non-Gaussian source distributions Lipman et al. (2023); Liu et al. (2023) and Action-to-Action Flow Matching initializes generation from action history Jia et al. (2026). WAVE-VLA instead decodes a current-only visual–motion code into a target-robot reference and learns only the residual required by the current scene. The codebook therefore selects the coarse skill, while the expert adapts its execution; zero residual recovers the selected plan exactly.

METHOD

WAVE-VLA separates discrete skill selection from scene-specific execution. Modality-routed RVQ learns a shared target-G1 motion vocabulary from HE and Nymeria, while a synchronized stream represents bilateral hand motion where labels are available. Given current egocentric RGB, body and hand state, and language, current-only selectors decode predicted codes into aligned 16-frame references. A reference-preserving Action Expert uses the body trajectory as its base action and predicts only the residual required by the current scene; the hand trajectory remains anchored to the selected body plan. Missing modalities are handled by availability masks, and future observations or actions appear only in supervision or oracle-evaluation branches.

Training proceeds sequentially: body codec $\rightarrow$ body selector $\rightarrow$ body-conditioned hand codec $\rightarrow$ hand selector $\rightarrow$ body residual expert. Each selector is trained with its corresponding codec frozen. Hand-codec training freezes the body codec and selector to provide predicted body context; hand-selector training also keeps these body modules frozen. Finally, residual-expert training freezes both

codecs and both selectors. The body and hand codecs are optimized in separate stages, without joint back-propagation through the body context.

A Common Target-Humanoid Interface

Task-unpaired records. The training mixture contains native-G1 egocentric demonstrations from Humanoid Everyday (HE) and paired human egocentric video–motion from Nymeria. Their task inventories are collected independently; WAVE-VLA requires neither matched human–robot tasks nor cross-source positive pairs. Task-unpaired describes the absence of pairing supervision, rather than requiring disjoint behavior inventories. A source-specific converter C_d maps every available motion stream to the same G1 kinematic tree, MuJoCo joint order, metric units, heading-local root convention, continuous-segment boundary, and 30 Hz clock. HE provides measured G1 state and action. Nymeria motion is retargeted by warm-started, joint-limited inverse kinematics and resampled jointly with its first-person video.

Body and root. Let $q_k \in \mathbf{R}^{29}$ be the absolute G1 joint configuration, $p_k \in \mathbf{R}^3$ the root position, and $\vartheta_k = (\psi_k, \phi_k, \theta_k)$ its yaw–roll–pitch angles. For each continuous segment we derive joint velocity $\dot{q}_k$, joint displacement Δq_k, heading-local root translation Δp_k, and wrapped root orientation increment $\Delta\vartheta_k$. A body/root frame is

$$a_k^b = [q_k, \dot{q}_k, \Delta q_k, \Delta p_k, \Delta\vartheta_k] \in \mathbf{R}^{93}. \tag{1}$$

Thus the learned action interface contains $29 + 29 + 29 + 3 + 3$ channels. Absolute q_k is used for kinematic evaluation; the remaining channels retain local dynamics and root progression. The synchronized current state is the pose at the chunk anchor. One train-split global mean and standard deviation normalize all body sources.

The hand stream contains seven joints per hand, $a_k^h \in \mathbf{R}^{14}$, in a fixed left–right joint layout. It shares the same 16-frame clock but remains a separate output stream because validated hand labels are available only in HE. Consequently, two-source learning applies to body/root motion, while hand learning uses HE supervision and predicted body context. Missing hand labels in Nymeria are masked rather than interpreted as physical zero targets.

Canonicalization gives all sources one *physical* decoder semantics; it does not assume that two sources use identical code IDs for semantically similar activities. Cross-source support is learned through shared robot-space reconstruction, whereas instruction grounding is learned only from sources that contain deployment-time visual and language context.

Modality-Routed Vision–Motion Tokenization

Body-plan routes. Figure 2 illustrates the shared body tokenizer. For a chunk $y_t^b = (a_{t+1}^b, \ldots, a_{t+16}^b)$, frozen DINOv2 features represent the current and endpoint egocentric frames as $z_t = \Phi(o_t)$ and $z_{t+16}^+ = \Phi(o_{t+16})$ Oquab et al. (2024). Visual, action, and fused encoders produce eight latent slots u_v, u_a, u_{va}. Binary encoder-input gates $p_v, p_a \in \{0, 1\}$ select fused $(1, 1)$, visual-only $(1, 0)$, or action-only $(0, 1)$ computation. These gates specify which modalities enter the encoder; supervision availability is determined separately by the record. HE and Nymeria support all three routes: visual-only encoding withholds the recorded action from the encoder, while action-only encoding withholds the image from the encoder. Both remain available as reconstruction targets. The degenerate $(0, 0)$ route is never quantized. Every valid route reconstructs canonical G1 motion and the endpoint feature using the available training targets. Language is not an input to the codec.

The body quantizer Q_b shares two successive 128-entry RVQ stages across all routes and applies them to eight latent slots, yielding $c^b \in \{0, \ldots, 127\}^{8\times 2}$, or 112 bits per 16-frame chunk. The state-conditioned decoder D_b reconstructs robot motion from the quantized slots and current pose. We write $\bar{r}^b$ for its normalized, 128-channel output and $r^b = \mathcal{N}_b^{-1}(\Pi_{93}(\bar{r}^b))$ for the physical trajectory, where Π_{93} removes padding and $\mathcal{N}_b$ is the global body normalization. The visual decoder D_v reconstructs the endpoint feature from the same slots and current image. Thus each code denotes both a pose-anchored motion operator and an egocentric transition; the visual branch predicts a feature rather than an RGB frame. True future features remain supervision or oracle evidence and are never supplied to the deployed controller. We call these codes executable because every code

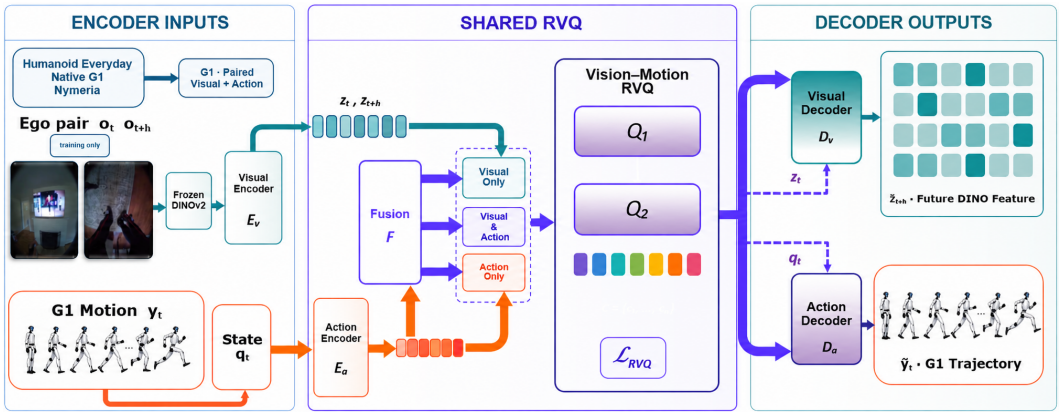


Figure 2: Shared two-source body vision–motion RVQ. Native-G1 HE and G1-retargeted Nymeria provide paired egocentric video and motion. Visual-only, fused, and action-only encoder routes share two RVQ stages and decode a 16-frame G1 trajectory and an endpoint DINO feature. The image and motion branches are supervised by the available training targets; the SONIC motion corpus is not used.

sequence deterministically decodes into a target-G1 reference for controller tracking; this interface property does not guarantee dynamic feasibility.

For each record, the body objective averages masked smooth-ℓ_1 action reconstruction, endpoint-feature cosine reconstruction, and standard straight-through RVQ codebook/commitment losses over its valid routes van den Oord et al. (2017). Routes are averaged before records are drawn with fixed source-sampling weights, preventing modality count from changing a record's aggregate weight. Sharing Q_b and D_b gives every selected entry one G1 decode path, but does not mathematically force source-invariant IDs.

Synchronized hand tokens. Let $h_t \in \mathbf{R}^{14}$ be the current hand state and $d^h_{t+k} = h_{t+k} - h_t$ the future residual. The hand codec groups 16 frames into eight two-frame slots and conditions each slot on the corresponding frames of the *predicted*, current-only physical body reference r^b, together with the global predicted body codes. Body tensors come from the frozen current-only body path and are detached before use, so no oracle future body trajectory enters the hand path. A separate hand RVQ applies two 128-entry stages to eight slots, yielding $c^h \in \{0, \dots, 127\}^{8\times 2}$: a second 112-bit interface synchronized with c^b.

The hand decoder predicts a standardized correction $\kappa^h \in \mathbf{R}^{16\times 14}$ and returns $r^h = h_t + m^h \odot (\sigma_h \odot \kappa^h)$, where $m^h \in \{0,1\}^{16\times 14}$ is the validity mask and σ_h is the train-only per-joint scale. Zero correction is exactly a current-hand hold; invalid elements receive this fallback and no loss. On valid HE labels, the hand objective combines masked Huber losses on pose and temporal differences with the hand-code commitment loss.

Current-Only Plan Selection

Codec encoders use future evidence and cannot select a plan at deployment. We train each selector after freezing its corresponding codec, restricting selector inputs to current RGB, current state, and language together with any upstream predictions derived from these inputs.

Current-only selector. Eight bridge embeddings appended to a frozen Qwen2.5-VL-3B prefix Bai et al. (2025) read (o_t, ℓ). A projection of normalized q_t is added to every bridge embedding, and two 128-way heads predict all body IDs without an autoregressive oracle prefix. If $c^{b\star}$ denotes the frozen fused-route teacher, training minimizes weighted cross-entropy over all slots and levels, with level weights $(1, \frac{1}{2})$. Only the bridge embeddings, state projector, and categorical heads are optimized. HE and Nymeria both shape the shared plan vocabulary and supervise current-only selection with their current RGB, state, and language. No source-specific selector or action vocabulary is introduced. Let $E_b(\hat{c}^b)$ denote lookup and summation in the frozen RVQ codebooks. Hard predicted

IDs decode to the normalized, padded coarse reference $\bar{r}^b = D_b(E_b(\widehat{c}^b), q_t)$ and the anticipated endpoint feature $\widehat{z}^+_{t+16} = D_v(z_t, E_b(\widehat{c}^b))$. The endpoint feature evaluates the visual consequence retained by the plan and is not fed to the controller.

Hand selector. The hand selector reuses the detached current VLM prefix and receives $(q_t, h_t, \widehat{c}^b, r^b)$. Eight time-aligned hand queries cross-attend to the current multimodal prefix and body-code memory; each query also reads the corresponding two frames of r^b. One categorical head per hand-code level predicts $\widehat{c}^h$. On valid hand slots, the selector minimizes level-weighted cross-entropy to frozen teacher IDs $c^{h\star}$. Future hand trajectories produce teacher IDs only. Training and inference use the same predicted body context, avoiding an oracle-body exposure mismatch.

Reference-Preserving Action Generation

Decoded plan as the base action. The body code first becomes the full 16-step reference $\bar{r}^b$; it is not reduced to eight optional prompt tokens. The frozen Action Expert backbone receives time-local embeddings of $\bar{r}^b$, embeddings of $\widehat{c}^b$, the current VLM prefix, and current state. In parallel, a lightweight visual adapter lets each reference step cross-attend to frozen DINO patches from the current egocentric image, using the time-aligned reference and current-state embedding as the query. The deployed adapter uses a typed current-feature copy in place of unavailable future patches, so refinement remains a deterministic function of the scene observed at execution time.

Every base trajectory is decoded by the shared HE–Nymeria RVQ, so the expert always consumes the same target-G1 reference interface rather than a source-specific action head. Both sources provide current visual–language context and aligned target actions for residual supervision.

The parent codecs, selectors, VLM, DINO encoder, and Action Expert backbone are frozen. Trainable residual adapters update action tokens inside the expert, and zero-initialized grouped heads predict corrections for joint, velocity, displacement, and root channels. Let $\bar{y} = \mathrm{Pad}_{128}(\mathcal{N}_b(y^b_t))$, let $m \in \{0,1\}^{16\times 128}$ mark its valid channels, and let γ be the RMS of $(\bar{y} - \bar{r}^b)$ computed once on the training split. One expert evaluation produces

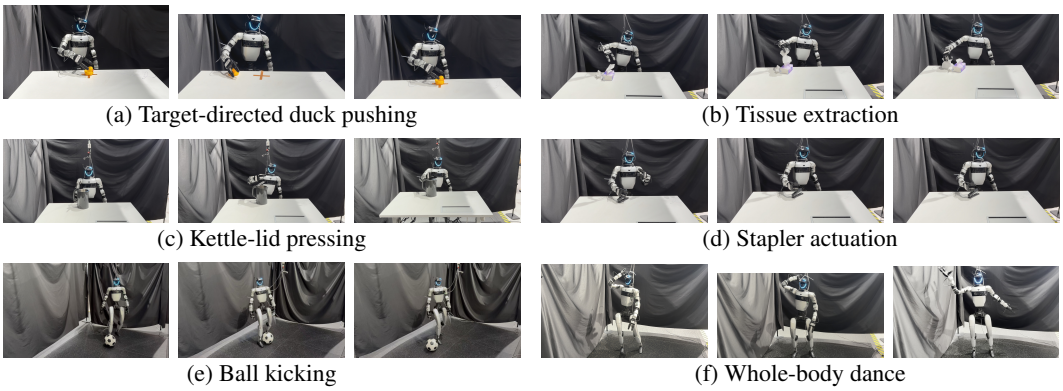
$$\begin{aligned} v_\theta &= R_\theta(\bar{r}^b, \widehat{c}^b, o_t, q_t, \ell), \\ \widehat{\bar{y}} &= \bar{r}^b + \gamma m \odot v_\theta. \end{aligned} \tag{2}$$

This is direct residual prediction with one network evaluation, not from-noise action generation or iterative flow sampling. The output identity $v_\theta = 0 \Rightarrow \widehat{\bar{y}} = \bar{r}^b$ holds elementwise, and the physical body trajectory is $\widehat{y}^b = \mathcal{N}_b^{-1}(\Pi_{93}(\widehat{\bar{y}}))$.

Physical refinement objective. For source d, A_d is the masked robust error over all 93 normalized channels. After applying $\mathcal{N}_b^{-1} \circ \Pi_{93}$, differentiable G1 FK maps the predicted and target absolute joint blocks to 14 registered body endpoints. We denote the group-balanced endpoint mean by M_d, the worst-decile endpoint error by T_d, and a one-sided safeguard $S_d = [M_d(\widehat{\bar{y}}, \bar{y}) - M_d(\bar{r}^b, \bar{y})]_+$. A mismatched same-source current image is paired with the unchanged reference, state, language, and code; A^-_d penalizes unsupported visual corrections. We minimize the source-averaged weighted sum of A_d, M_d, T_d, S_d, and A^-_d. Source-wise normalization and gradient balancing prevent the larger Nymeria residual scale from dominating HE. The matched terms teach a useful correction, while A^-_d suppresses unsupported image-dependent changes; no future frame enters the controller.

Body–hand composition and execution. Body and hand references share the same 16-frame clock and are selected from the same current context. The hand stream reads the frozen coarse body plan; the body residual expert does not rewrite hand codes or back-propagate into either codec. The hand stream reconstructs finger-joint motion relative to the current hand state, specifying articulation for grasping and interaction. The refined body trajectory determines wrist position and orientation; finger articulation is composed with that wrist pose through the robot's kinematic chain. Thus body refinement preserves the hand plan's synchronization and both zero-correction identities: zero body residual recovers r^b, while zero hand correction holds h_t. These are algebraic identities, not automatic failure detection or safety guarantees.

At each receding-horizon cycle, WAVE-VLA predicts both code streams, decodes (r^b, r^h), applies one body correction, and executes a fixed prefix $K_{\mathrm{exec}} \le 16$ before replanning. The learned model



(a) Target-directed duck pushing (b) Tissue extraction

(c) Kettle-lid pressing (d) Stapler actuation

(e) Ball kicking (f) Whole-body dance

Figure 3: Real-G1 executions of six skills represented in the HE–Nymeria training corpus. Each task is shown at three points in its rollout.

Task	Training-source coverage	HE-only	HE+Nymeria
Duck pushing	HE	7/10	6/10
Tissue extraction	HE	6/10	6/10
Kettle-lid pressing	HE	8/10	7/10
Stapler actuation	HE	6/10	6/10
Ball kicking	Nymeria	0/10	4/10
Whole-body dance	Nymeria	0/10	10/10
HE-covered subtotal	HE	27/40	25/40
Nymeria-covered subtotal	Nymeria	0/20	14/20
Overall	**Both sources**	**27/60**	**39/60**

Table 1: Real-G1 task completion: ten trials per task and policy under the same robot setup, SONIC control stack, and success criteria. Source coverage identifies which dataset contains each task.

outputs target-G1 references, not torques. Online 14D hand plans come from the hand selector and decoder, without prescribed task-specific gestures. The SONIC whole-body controller tracks the denormalized body/root and hand references Luo et al. (2025). This controller is used only for tracking; its motion corpus is not part of WAVE-VLA training. The deployment adapter maps joint order and resamples absolute joint positions from 30 to 50 Hz, then computes commanded joint velocities by finite differences of that position trajectory. Predicted velocity and displacement channels supply motion supervision, while the position-derived velocities maintain command consistency. This boundary keeps representation and policy claims distinct from low-level stability supplied by the controller.

EXPERIMENTS

We evaluate the proposed expansion of one policy's behavioral repertoire in stages: whether the shared codebook represents both motion sources, whether current observations select their plans, and whether decoded references can be refined and executed. Reconstruction on Nymeria tests motion support; real-robot tasks covered by each source test joint execution. Offline and robot HE-only comparisons assess added task coverage and HE retention.

EXPERIMENTAL PROTOCOL

Data and splits. The two-source training pool contains 4,064 HE episodes and 135,538 Nymeria episodes with 20,003,611 synchronized first-person frames. The HE episodes are divided into 3,224 training and 840 validation episodes and provide native G1 egocentric video, state, body/root actions, bilateral hand targets, and language Zhao et al. (2025b). Nymeria provides paired human first-person video–motion and language, with motion retargeted to G1 while preserving synchro-

Source	Oracle	Predicted reference	Refined	Improvement
HE	10.66	15.58	12.29	21.1%
Nymeria	52.26	72.86	61.80	15.2%

Table 2: Reference-preserving refinement on fixed architecture-development manifests. Values are mean FK error in millimeters. "Predicted reference" is the decoder-only trajectory from current-only predicted codes; improvement is measured against that reference on the same examples and targets.

nization Ma et al. (2024). Missing Nymeria hand labels are masked. We construct no cross-source correspondence by task, object, instruction, or scene.

Optimization and splits. The reported checkpoint is trained on HE and Nymeria. Tokenizer and selector training use eight H100s with global batch 16. Development splits contain 20 HE tasks and 20 Nymeria sessions. Tokenizer evaluation uses 20,000 chunks per source; selector evaluation uses 36,159 unique samples and 10,000 bootstrap resamples. The refiner is evaluated on a separate manifest of frozen decoded references. Offline HE-only body results and robot trials on all six tasks are reported below.

Metrics and controls. One MuJoCo FK pipeline evaluates inverse-normalized 93D trajectories over 16 frames and 14 bodies; we also report code usage and future-DINO similarity. Static-pose, constant-velocity, majority-code, current-feature, condition-only, and decoder-only controls isolate coverage, selection, and plan commitment. Future evidence is restricted to teacher/oracle controls: replacing future actions or injecting a goal image leaves predictor logits unchanged. The bracketed DINO-MSE result below uses a 95% hierarchical bootstrap, excluding training-seed variation.

A Shared Codebook for Robot and Human Motion

The shared 112-bit body codec reconstructs disjoint HE and Nymeria splits with mean FK errors of 11.42 and 65.82 mm, respectively, without source-specific codebooks; both RVQ levels use multiple entries (Appendix Table 5). Robot trials below test whether this shared representation supports deployable skills.

On HE dev, two-source training retains similar motion-prediction accuracy: physical full93 MAE changes from 0.04083 to 0.04151 and body FK14 from 0.01384 to 0.01407 (both about 1.7% higher; Table 6). Nymeria is included to extend the task repertoire, not to lower mean HE reconstruction error.

Body-Conditioned Hand Reconstruction

On an episode-disjoint 64-episode HE split, oracle hand-RVQ decoding improves active-finger FK over current-hand hold (Table 5); the RVQ levels use 45/98 entries, and MuJoCo validates all 14 hand-joint gradients. On identical HE dev and HE routes, three-seed means for HE-only/HE+Nymeria are 0.03961/0.04063 (+2.6%) in predicted-code hand FK14 error and 0.02707/0.02737 (+1.1%) in action reconstruction error. These mean differences suggest limited interference, not statistical equivalence.

Current-Only Plan Selection

Across 20 task-disjoint HE development tasks and 20 session-disjoint Nymeria development sessions, predicted codes recover 89%/98% of the oracle DINO-cosine gain. This fraction is the predicted-code cosine improvement over current-feature copying divided by the oracle-code improvement over the same baseline. Nymeria DINO MSE improves by 1.097 [0.849, 1.322], while HE matches current-feature copying. In an 88-example HE counterfactual with fixed image and state, mismatched instructions alter every code sequence (41.69% of code elements) and reduce token accuracy against the same frozen teacher labels by 2.699 percentage points. Correct language yields higher token accuracy on 9/11 tasks. Top-1 body-RVQ token accuracy against frozen fused-route teachers is 89.55%/87.32% on HE/Nymeria. Unlike the relative oracle DINO gains, these are

Policy	Norm. full93 MAE ↓ HE	Nymeria	Body FK14 (mm) ↓ HE	Nymeria
Code-conditioned direct	0.12662	0.39736	33.81	68.27
Decoded reference + residual	0.08688	0.36283	14.07	47.63

Table 3: Main full-policy comparison on identical V343 HE/Nymeria development sets (288 anchors per source), with predicted q0/q1 codes and common 93D/FK processing. Only the second policy uses the decoded 16-frame motion as the Action Expert's base. Table 2 uses a separate architecture-development manifest.

absolute per-token, not exact-sequence, scores; future evidence labels teachers only and is unavailable to the selector.

Reference-Preserving Continuous Refinement

To isolate the value of decoding the selected plan, we train a complete code-conditioned direct-action policy. It receives the same current observations, state, language, and predicted RVQ codes as WAVE-VLA, but no decoded 16-frame reference, and directly outputs 93D actions under matched training and inference conditions. Additional condition-only and Gaussian-residual diagnostics appear in Appendix .

Decoding the reference lowers normalized full93 MAE by 31.4%/8.7% and body FK14 by 58.4%/30.2% on HE/Nymeria relative to the complete code-conditioned policy (Table 3). These are offline motion comparisons, not a measured difference in real-robot success.

Table 7 isolates visual conditioning on a fixed development cohort with codes and decoded references held constant. Current-feature copying reduces mean FK relative to the base refiner by 43.0% on HE and 41.9% on Nymeria. Matched future features provide an oracle comparison; mismatched future features and shuffled bundles yield larger errors than current-feature copying.

On the separate final-refiner development cohort (Table 2), refinement lowers decoder-only error by 21.1%/15.2%. Matched decoder-only and refined pairs share codes, observations, and targets; only the residual expert changes. This component comparison is distinct from the V343 full-policy comparison above.

Real-Robot Execution

One frozen body–hand policy with SONIC completes 39/60 real-G1 trials without task-specific updates or retrieved/oracle references (Figure 3, Table 1). Each task uses ten consecutive trials with online body/14D hand plans. HE-only completes 27/40 HE-covered and 0/20 Nymeria-covered trials, versus 25/40 and 14/20 for joint training under the same setup. The 2/40 HE difference is not an equivalence test; kicking is only 4/10, and these two Nymeria tasks do not establish broad human-motion transfer (task criteria in Appendix).

Conclusion

WAVE-VLA unifies task-unpaired HE and Nymeria motion in a shared target-humanoid vocabulary and decodes codes before residual control. One frozen G1 policy executes both task groups: HE task success is 25/40 versus 27/40 for HE-only, while Nymeria task success is 14/20 versus 0/20.

AI Use Statement

Generative AI tools were used to assist with language editing and software development and debugging. The authors reviewed all AI-assisted text, verified the reported claims against the underlying experimental artifacts, and tested AI-assisted code before use. The authors take responsibility for the final content of this work, including text, claims, and software produced with the aid of generative AI.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025. URL https://arxiv.org/abs/2502.13923.

Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, et al. GR00T N1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. RT-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.

Boyu Chen, Yi Chen, Lu Qiu, Jerry Bai, Yuying Ge, and Yixiao Ge. UniT: Toward a unified physical language for human-to-humanoid policy learning and world modeling. *arXiv preprint arXiv:2604.19734*, 2026.

Cheng Chi, Zhenjia Xu, Chuer Pan, Eric Cousineau, Benjamin Burchfiel, Siyuan Feng, Russ Tedrake, and Shuran Song. Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots. In *Proceedings of Robotics: Science and Systems*, 2024. doi: 10.15607/RSS.2024.XX.045.

Pengxiang Ding, Jianfei Ma, Xinyang Tong, Binghong Zou, Xinxin Luo, Yiguo Fan, Ting Wang, Hongchao Lu, Panzhong Mo, Jinxin Liu, Yuefan Wang, Wenshuo Feng, Jiacheng Liu, Siteng Huang, and Donglin Wang. Humanoid-vla: Towards universal humanoid control with visual integration. *arXiv preprint arXiv:2502.14795*, 2025.

Zibin Dong, Yicheng Liu, Shiduo Zhang, Baijun Ye, Yifu Yuan, Fei Ni, Jingjing Gong, Xipeng Qiu, Hang Zhao, Yinchuan Li, and Jianye Hao. Actioncodec: What makes for good action tokenizers. *arXiv preprint arXiv:2602.15397*, 2026.

Jindou Jia, Gen Li, Xiangyu Chen, Tuo An, Yuxuan Hu, Jingliang Li, Xinying Guo, and Jianfei Yang. Action-to-action flow matching. *arXiv preprint arXiv:2602.07322*, 2026.

Haoran Jiang, Jin Chen, Qingwen Bu, Li Chen, Modi Shi, Yanjie Zhang, Delong Li, Chuanzhe Suo, Chuang Wang, Zhihui Peng, and Hongyang Li. WholeBodyVLA: Towards unified latent VLA for whole-body loco-manipulation control. *arXiv preprint arXiv:2512.11047*, 2025.

Tengyue Jiang, Chunpu Xu, Jiayue Kang, and Yao Mu. Sa-vla: State-aware tokenizer for improving vision-language-action models' performance. *arXiv preprint arXiv:2606.30113*, 2026.

Tobias Johannink, Shikhar Bahl, Ashvin Nair, Jianlan Luo, Avinash Kumar, Matthias Loskyll, Juan Aparicio Ojea, Eugen Solowjow, and Sergey Levine. Residual reinforcement learning for robot control. In *IEEE International Conference on Robotics and Automation*, pp. 6023–6029, 2019. doi: 10.1109/ICRA.2019.8794127.

Miracle Kang, Lights Shi, Lucy Liang, Roy Gan, Dongxiu Liu, Pushi Zhang, Sylas Chen, Shawn Qin, Yinan Zheng, Jinliang Zheng, Hao Wang, Xianyuan Zhan, and Hang Su. X-tokenizer: A multimodal action tokenizer for vision-language-action pretraining. *arXiv preprint arXiv:2606.14752*, 2026.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. OpenVLA: An open-source vision-language-action model. In *Conference on Robot Learning*, 2024.

Seungjae Lee, Yibin Wang, Haritheja Etukuru, H. Jin Kim, Nur Muhammad Mahi Shafiullah, and Lerrel Pinto. Behavior generation with latent actions. *arXiv preprint arXiv:2403.03181*, 2024. URL https://arxiv.org/abs/2403.03181.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023.

Hao Luo, Ye Wang, Wanpeng Zhang, Haoqi Yuan, Yicheng Feng, Haiweng Xu, Sipeng Zheng, and Zongqing Lu. Joint-aligned latent action: Towards scalable VLA pretraining in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 35047–35058, 2026a.

Hao Luo, Wanpeng Zhang, Yicheng Feng, Sipeng Zheng, Haiweng Xu, Chaoyi Xu, Ziheng Xi, Yuhui Fu, and Zongqing Lu. Being-H0.7: A latent world-action model from egocentric videos. *arXiv preprint arXiv:2605.00078*, 2026b.

Zhengyi Luo, Ye Yuan, Tingwu Wang, Chenran Li, Sirui Chen, Fernando Castañeda, Zi-Ang Cao, Jiefeng Li, David Minor, Qingwei Ben, et al. SONIC: Supersizing motion tracking for natural humanoid whole-body control. *arXiv preprint arXiv:2511.07820*, 2025.

Lingni Ma, Yuting Ye, Fangzhou Hong, Vladimir Guzov, Yifeng Jiang, Rowan Postyeni, Luis Pesqueira, Alexander Gamino, Vijay Baiyya, Hyo Jin Kim, et al. Nymeria: A massive collection of multimodal egocentric daily motion in the wild. *arXiv preprint arXiv:2406.09905*, 2024.

Open X-Embodiment Collaboration. Open x-embodiment: Robotic learning datasets and RT-X models. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6892–6903, 2024. doi: 10.1109/ICRA57147.2024.10611477.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Karl Pertsch, Kyle Stachowicz, Brian Ichter, Danny Driess, Suraj Nair, Quan Vuong, Oier Mees, Chelsea Finn, and Sergey Levine. FAST: Efficient action tokenization for vision-language-action models. *arXiv preprint arXiv:2501.09747*, 2025.

Wenxuan Song, Ziyang Zhou, Han Zhao, Jiayi Chen, Pengxiang Ding, Haodong Yan, Yuxin Huang, Feilong Tang, Donglin Wang, and Haoang Li. ReconVLA: Reconstructive vision-language-action model as effective robot perceiver. *arXiv preprint arXiv:2508.10333*, 2025.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Yating Wang, Haoyi Zhu, Mingyu Liu, Jiange Yang, Hao-Shu Fang, and Tong He. VQ-VLA: Improving vision-language-action models via scaling vector-quantized action tokenizers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11089–11099, 2025.

Tianyi Xie, Haotian Zhang, Jinhyung Park, Zi Wang, Bowen Wen, Jiefeng Li, Xueting Li, Qingwei Ben, Haoyang Weng, Yufei Ye, David Minor, Tingwu Wang, Chenfanfu Jiang, Sanja Fidler, Jan Kautz, Linxi Fan, Yuke Zhu, Zhengyi Luo, Umar Iqbal, and Ye Yuan. GRAIL: Generating humanoid loco-manipulation from 3d assets and video priors. *arXiv preprint arXiv:2606.05160*, 2026. URL https://arxiv.org/abs/2606.05160.

Mengda Xu, Zhenjia Xu, Cheng Chi, Manuela Veloso, and Shuran Song. XSkill: Cross embodiment skill discovery. In *Proceedings of the 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 3536–3555, 2023. URL https://proceedings.mlr.press/v229/xu23a.html.

Haoru Xue, Xiaoyu Huang, Dantong Niu, Qiayuan Liao, Thomas Kragerud, Jan Tommy Gravdahl, Xue Bin Peng, Guanya Shi, Trevor Darrell, Koushil Sreenath, and Shankar Sastry. LeVERB: Humanoid whole-body control with latent vision-language instruction. *arXiv preprint arXiv:2506.13751*, 2025.

Haoran Yang, Jiacheng Bao, Yucheng Xin, Haoming Song, Yuyang Tian, Bin Zhao, Dong Wang, and Xuelong Li. ZeroWBC: Learning natural visuomotor humanoid control directly from human egocentric video. *arXiv preprint arXiv:2603.09170*, 2026.

Seonghyeon Ye, Joel Jang, Byeongguk Jeon, Sejune Joo, Jianwei Yang, Baolin Peng, Ajay Mandlekar, Reuben Tan, Yu-Wei Chao, Bill Yuchen Lin, Lars Liden, Kimin Lee, Jianfeng Gao, Luke Zettlemoyer, Dieter Fox, and Minjoon Seo. Latent action pretraining from videos. *arXiv preprint arXiv:2410.11758*, 2024. URL https://arxiv.org/abs/2410.11758.

Junpeng Yue, Boyuan Li, Yuxuan Wang, Zepeng Wang, Yuhui Fu, Feiyang Xie, Yu Zhang, Jing Zhang, Jiangxing Wang, and Zongqing Lu. Being-M0.7: A latent world-action model for humanoid robots. BeingBeyond Technical Report, 2026. URL https://research.beingbeyond.com/being-m07/being-m07.pdf.

Qingqing Zhao, Yao Lu, Moo Jin Kim, Zipeng Fu, Zhuoyang Zhang, Yecheng Wu, Zhaoshuo Li, Qianli Ma, Song Han, Chelsea Finn, Ankur Handa, Tsung-Yi Lin, Gordon Wetzstein, Ming-Yu Liu, and Donglai Xiang. CoT-VLA: Visual chain-of-thought reasoning for vision-language-action models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1702–1713, 2025a.

Zhenyu Zhao, Hongyi Jing, Xiawei Liu, Jiageng Mao, Abha Jha, Hanwen Yang, Rong Xue, Sergey Zakharor, Vitor Guizilini, and Yue Wang. Humanoid everyday: A comprehensive robotic dataset for open-world humanoid manipulation. *arXiv preprint arXiv:2510.08807*, 2025b.

TASK-UNPAIRED DATA

Source	Observed	Role	Scale
HE (G1)	$V, q_t, h_t, A^{b,h}, \ell$	robot skills	3,224/840 ep.
Nymeria→G1	V, q_t, A^b, ℓ	human motion	135,538 ep.

Table 4: Task-unpaired data after G1 conversion. HE scale is train/validation. $V = (o_t^{\text{ego}}, o_{t+16}^{\text{ego}})$, $A^{b,h}$ are body/root and hand motion, and ℓ is language. All valid A^b train the shared body vocabulary; HE supplies hand supervision.

BODY AND HAND RVQ RECONSTRUCTION

Source	FK mean (mm)	RVQ-1 active	RVQ-2 active
HE	11.42	69	121
Nymeria	65.82	84	127

Hand metric	Static	Hand RVQ	Reduction
Fingertip FK, all (mm)	4.522	4.129	8.7%
Fingertip FK, active (mm)	12.609	9.328	26.0%
Joint MAE, active (rad)	0.1546	0.1145	25.9%

Table 5: Body and hand RVQ reconstruction. Top: body tokenizer on 20,000 chunks per source; active counts are reported separately for each 128-entry RVQ level. Bottom: oracle hand-code decoding on episode-disjoint HE; "active" denotes motion above 0.05 rad.

FULL-BODY SOURCE ABLATION

Model (route)	Dev	Norm. full93	Physical full93	q29	dq29	$\Delta q29$	root6	FK14
HE-only (HE)	HE	0.08573	0.04083	0.02687	0.10056	0.00336	0.000687	0.01384
HE+Nymeria (HE)	HE	0.08688	0.04151	0.02748	0.10208	0.00341	0.000678	0.01407
HE-only (forced HE)	Nymeria	0.35309	0.14730	0.07698	0.38127	0.01271	0.006795	0.04594
HE+Nymeria (Nymeria)	Nymeria	0.36283	0.15266	0.07975	0.39524	0.01318	0.006793	0.04763

Table 6: Offline HE-only versus two-source 93D body ablation (all errors ↓). FK14 evaluates 14 body points, not hand FK. HE rows share the HE route; Nymeria rows use the forced HE and Nymeria routes, respectively. These errors assess motion prediction, not causal gains in robot task success.

REAL-ROBOT TRIAL CRITERIA

Task completion criteria. Both policies use the same task-specific success criteria. Duck pushing succeeds when the duck reaches the designated target position; tissue extraction succeeds when a tissue is pulled out. Kettle-lid pressing and stapler actuation require pressing the respective target to its task-defined end position. Ball kicking succeeds when the robot's foot contacts the ball. For whole-body dance, success is judged visually by whether the robot's limb movements resemble the dance motions in Nymeria; this is a qualitative criterion, not a pose-error threshold. Trials not meeting their task criterion are scored as failures.

Visual condition / control	HE	Nymeria
Base refiner (V31)	5.096	13.807
Current-feature copy (deployable)	2.904	8.023
Matched future (oracle)	2.523	6.408
Wrong future (intervention)	4.660	9.678
Shuffled bundle (intervention)	4.882	12.831

Table 7: Visual-conditioning ablation: mean FK error in millimeters (↓) on a fixed V31 development cohort, with matched examples, body codes, and decoded references. Future-feature rows are diagnostic controls, not deployment inputs. This cohort differs from Table 2; absolute errors are compared only within each table.

VISUAL-CONDITIONING CONTROLS

ADDITIONAL REFINER DIAGNOSTICS

In an earlier condition-only diagnostic, replacing learned-null embeddings with predicted codes changes HE mean FK from 47.06 to 47.63 mm and increases error across all 20 Nymeria sessions (91.78 to 99.06 mm). An evaluated Gaussian-residual candidate also degrades 14.75/40.68 mm references while retaining its initial noise (cosine 0.9999). These probes motivated the decoded-reference residual interface; they are not the separately trained full code-conditioned policy compared in Table 3.

CROSS-ENTROPY CONTINUATION: SEED SENSITIVITY

We ran an exploratory q0-only cross-entropy (CE) continuation of the body selector's q0 head after fixed pretraining with three training seeds. The VLM features and RVQ decoder remain frozen. Predicted q0 is decoded with the second-stage RVQ residual set to the zero vector, without Action Expert refinement. All seeds use the V567 `selection_v2` fresh-episode cohort (16 episodes and 48 anchors per source). Table 8 reports body FK14 and its across-seed mean ± sample standard deviation. This q0-only diagnostic uses a different model path and evaluation manifest from the V343 full-policy comparison (Table 3) and the separate architecture-development refinement study (Table 2).

Source	Seed 56100	Seed 56501	Seed 56502	Mean ± SD
HE	12.314	12.302	12.239	12.285 ± 0.040
Nymeria	36.562	38.815	38.994	38.124 ± 1.355

Table 8: Exploratory q0-only CE-continuation body FK14 in millimeters (↓) on the V567 `selection_v2` fresh-episode cohort. The reported SD is across three training seeds, not a confidence interval; these results are not a full-policy comparison.