EgoLite-G: Restoring the Physical Gauge in Egocentric Token World Models

Anonymous authors
Paper under double-blind review

Abstract

Token-based egocentric world models canonicalize every modality in its own first-frame camera frame. We show that this design silently removes the physical gauge—the gravity direction and the height of the wearer above the support surface—on which world-frame body pose depends. We introduce EgoLite-G, a gauge-state egocentric world model that keeps the gauge as an explicit, persistent state: gravity from the head-worn IMU and a take-level support height estimated from the head trajectory alone. The state enters a masked-token world model through zero-initialized gauge embeddings and is trained with privileged, noise-perturbed heights so that the same network is deployed with legal estimates. On the official EE4D-Motion validation set (99 unseen participants, 44,046 windows, scored once after all predictions were frozen), EgoLite-G lowers world MPJPE to 130.4 mm: 17.2 mm below a budget-matched fine-tuned ReViV, 19.3 mm below the strongest pre-registered baseline, and 5.6 mm below EgoAllo given the very same floor estimate, while matching EgoAllo supplied with a label-derived floor. It improves 91% of the unseen participants. Controlled studies show that the gain is caused by the gauge itself: shuffling gravity at inference costs 15.9 mm, the gain grows with head-pitch deviation, and body error decreases monotonically as the gauge becomes more accurate.

1 Introduction

Egocentric world models that jointly represent video, camera motion and the wearer's body as discrete tokens (Lyu et al., 2026; Mizrahi et al., 2023) have become a promising route to full-body motion capture from head-worn devices such as smart glasses. A single masked generator can be trained across heterogeneous datasets and queried for any subset of modalities, and it produces plausible articulated motion even when the body is almost never visible to the camera. To share one generator across datasets recorded with different world conventions, such models canonicalize each modality in its own first-frame coordinate system. For the camera stream this means that the first head pose is set to the identity.

Canonicalization is convenient for learning but has a subtle consequence: the absolute direction of gravity and the height of the head above the floor are no longer observable to the model. World-frame body pose, however, is defined relative to exactly these quantities. A person looking down at a table and a person bending over produce similar canonicalized camera motion; a tall person standing and a shorter person on tiptoe produce the same head trajectory up to a vertical offset. The generator must therefore guess quantities that the sensors actually measure.

We call these quantities the physical gauge of the scene and argue that an egocentric world model should maintain them as an explicit state rather than discard them. This paper builds on our EgoLite line of work, which frames egocentric body estimation as prediction inside a world model, and introduces EgoLite-G (Figure 1). Our contributions are:

- Diagnosis. We identify gauge removal as a structural property of first-frame canonicalization in token world models, characterize the removed coordinates (Propo-

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.

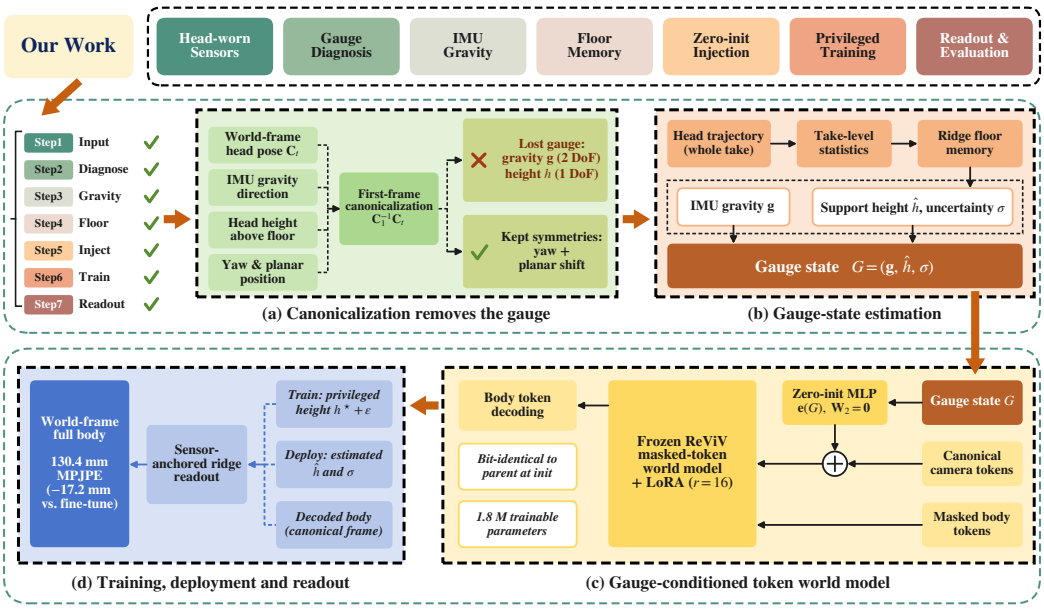


Figure 1: Overview of EgoLite-G. (a) First-frame canonicalization maps the gravity-aligned world into canonical camera tokens and thereby removes the physical gauge—the gravity direction (2 DoF) and the height above the support surface (1 DoF)—keeping only the true symmetries of world-frame pose (yaw and planar shift). (b) EgoLite-G estimates the gauge state $G = (\mathbf{g}, \hat{h}, \sigma)$ from IMU gravity and a persistent take-level floor memory computed from the head trajectory alone. (c) The gauge enters a frozen masked-token world model (ReViV with LoRA) through a zero-initialized embedding added to the camera tokens, so the adapted model is bit-identical to its parent at initialization. (d) Training uses privileged, noise-perturbed heights, deployment uses the legal estimates, and a sensor-anchored readout places the decoded body in the world frame.

sition 1), and quantify their cost with controlled interventions: gauge shuffling, pitch-stratified analysis and gauge-accuracy sweeps.

- Gauge-state world model. EgoLite-G estimates a gauge state $G = (\mathbf{g}, h, \sigma)$ from signals head-worn devices already provide—IMU gravity and a persistent take-level support height inferred from the head trajectory—and injects it into a frozen masked-token generator through zero-initialized embeddings. Privileged noisy-gauge training and an uncertainty-aware variant let the same network be trained with ground-truth heights and deployed with estimates.
- Evidence under a sealed protocol. On the official EE4D-Motion validation split, evaluated once after all predictions were frozen and hashed, EgoLite-G reduces world MPJPE by 17.2 mm over a budget-matched fine-tuned ReViV, improves 91% of participants, and outperforms EgoAllo (Yi et al., 2025) when both methods receive the same floor estimate. All pre-registered primary contrasts pass with Holm-corrected $p < 2 \times 10^{-4}$.

2 Related Work

Egocentric body estimation from head-worn sensors. Early work estimated the wearer's pose from downward-facing fisheye cameras; with head-mounted SLAM devices the problem shifted to inferring the full body from head motion. EgoEgo (Li et al., 2023) decomposes the task into head-pose estimation and head-conditioned body generation, and UniEgoMotion (Patel et al., 2025) unifies reconstruction, forecasting and generation with a head-centric motion representation. EgoAllo (Yi et al., 2025) conditions a diffusion model on head poses

expressed relative to a floor plane fitted to SLAM points and shows that invariant conditioning improves accuracy. ReViV (Lyu et al., 2026) unifies body, hands, camera, gaze and depth in a masked generative egocentric transformer. EgoLite-G is complementary to these approaches: it keeps the token world model and its canonicalization—which are what make multi-dataset training possible—and restores the gauge that canonicalization removes as an explicit, estimated state.

Sparse-sensor avatars. A parallel line of work reconstructs full-body motion from the head and hand trackers of VR headsets, e.g. AvatarPoser (Jiang et al., 2022) and EgoPoser (Jiang et al., 2024), and physics-based controllers regularize egocentric estimates with dynamics (Luo et al., 2021). These methods typically express inputs in a gravity-aligned, floor-relative frame by construction; EgoLite-G brings the same physical grounding to canonicalized token world models, where it is lost by design, and estimates the floor rather than assuming it.

Joint scene and self reconstruction. RESELF (Guan et al., 2026) jointly recovers scene geometry and wearer motion from egocentric video and refines the camera with closed-loop kinematic feedback. Scene-aware capture systems such as HPS (Guzov et al., 2021) localize the wearer in pre-scanned scenes. Our gauge state can be read as a minimal scene description—gravity and support height—that is estimated from sensor streams alone, requires no scan, and transfers across datasets and generators.

Masked multimodal token models and adaptation. 4M (Mizrahi et al., 2023) and its successors train any-to-any generators over tokenized modalities. Masked generative decoding (Chang et al., 2022) and classifier-free guidance (Ho & Salimans, 2022) make them efficient and controllable samplers, and parameter-efficient adaptation with LoRA (Hu et al., 2022) is the standard way to specialize them. We show that zero-initialized conditioning is a safe way to add physically meaningful inputs to a released generator: at initialization the adapted model reproduces the parent bit-for-bit, so every change in behaviour is attributable to learned use of the gauge.

3 Method

3.1 Setting and the gauge state

A head-worn device provides, for a window of T=60 frames, camera poses $\mathbf{C}_{1:T} \in SE(3)$ in a gravity-aligned world frame. The world model receives canonicalized camera tokens $\mathbf{c} = \mathrm{Tok}(\mathbf{C}_1^{-1}\mathbf{C}_{1:T})$ and generates body tokens that are decoded into a canonical body $\mathbf{B}^c$ and placed in the world as $\mathbf{B} = \mathbf{C}_1\mathbf{B}^c$.

Proposition 1 (Gauge removed by canonicalization). The canonical tokens $\mathbf{c}$ are invariant to every rigid transform $\mathbf{A} \in SE(3)$ of the world, $\mathbf{C}_t \mapsto \mathbf{A}\mathbf{C}_t$. World-frame body pose relative to the ground is invariant only to the subgroup of rotations about the gravity axis and horizontal translations (4 DoF). The quotient $SE(3)/(SO(2) \ltimes \mathbb{R}^2)$ is parameterized by the gravity direction in the first camera frame, $\mathbf{g} = \mathbf{R}_1^\top (0, 0, -1)^\top \in S^2$ (2 DoF), and the height of the head above the support surface, $h = z_1^{\mathrm{cpf}} - z_{\mathrm{floor}} \in \mathbb{R}$ (1 DoF). These three coordinates are therefore unobservable from $\mathbf{c}$, yet observable from the device.

Sketch. $\mathbf{C}_1^{-1}\mathbf{C}_t$ is unchanged by left multiplication with any $\mathbf{A}$. Rotations that tilt the gravity axis change $\mathbf{g}$, and vertical translations change h, while both leave $\mathbf{c}$ untouched; the remaining yaw and horizontal translations leave $\mathbf{g}$, h and the body relative to the ground unchanged. □

We collect the removed coordinates, together with an uncertainty σ of the height estimate, in the gauge state $G = (\mathbf{g}, h, \sigma)$.

Table 1: Floor-memory estimators (mean absolute floor error on calibration takes, cm). The ridge model is selected on calibration participants and used throughout.

Estimator (sensor trajectory only)	Floor MAE (cm)
Population-height prior (head p95 − mean height)	16.1
Constant training-mean floor	14.1
Gradient-boosted trees on take statistics	8.7
Ridge on take statistics (ours)	7.8

3.2 Estimating the gauge from sensors

Gravity. $\mathbf{g}$ is observed directly: the SLAM world frame of the device is gravity-aligned through its IMU, so $\mathbf{g}$ follows from the first camera rotation at no extra cost.

Persistent floor memory. The support height is a slowly varying property of the world, so EgoLite-G maintains it as a persistent state at the level of a take rather than re-estimating it in every window. A ridge regressor f_θ maps trajectory statistics $\phi(\mathbf{C}_{\text{take}})$ pooled over the take—head-height percentiles and spreads, gait oscillation amplitude and dominant frequency, stride, walking fraction, speed and head-pitch percentiles—to the floor height,

$$\hat{z}_{\text{floor}} = f_\theta\big(\phi(\mathbf{C}_{\text{take}})\big), \qquad \hat{h} = z_1^{\text{cpf}} - \hat{z}_{\text{floor}}. \tag{1}$$

The regressor is fitted on training participants only, its regularization is selected on calibration participants, and it uses no body information at test time. Figure 3b shows that it halves the floor error of a population-height prior (22.3 cm $\rightarrow$ 10.5 cm on held-out participants) and raises the correlation between estimated and true window height from 0.33 to 0.85. Table 1 compares candidate estimators on the calibration participants, on which the ridge model was selected.

3.3 Injecting the gauge into the world model

The gauge is embedded by a two-layer MLP and added to every camera-token embedding of the encoder,

$$\tilde{\mathbf{x}}_i^{\text{cam}} = \mathbf{x}_i^{\text{cam}} + \mathbf{e}(G), \qquad \mathbf{e}(G) = \mathbf{W}_2\,\text{GELU}(\mathbf{W}_1 G + \mathbf{b}_1) + \mathbf{b}_2, \tag{2}$$

with $\mathbf{W}_2 = \mathbf{0}$ and $\mathbf{b}_2 = \mathbf{0}$ at initialization, in the spirit of zero convolutions (Zhang et al., 2023). Because $\mathbf{e} \equiv \mathbf{0}$ at initialization, the adapted model reproduces the released parent exactly. We adapt the frozen world model with LoRA (r=16, α=32) on the encoder and decoder attention projections; together with the gauge MLP this amounts to 1.8 M trainable parameters. Training minimizes the native masked cross-entropy on body tokens $\mathbf{b}$,

$$\mathcal{L} = \mathbb{E}\Big[-\textstyle\sum_{j\in\mathcal{M}} \log p_\psi\big(b_j \mid \mathbf{b}_{\setminus\mathcal{M}}, \tilde{\mathbf{x}}^{\text{cam}}\big)\Big], \tag{3}$$

with condition dropout (p=0.1) that removes the camera tokens—and hence the gauge—jointly, consistent with classifier-free guidance at inference.

Privileged gauge training. During training the true floor of the training participants is available; at deployment only the estimate is. This is an instance of learning using privileged information (Vapnik & Vashist, 2009). We therefore train with the true height perturbed by Gaussian noise whose scale matches the calibration error of the floor memory, $h^{\text{train}} = h^\star + \epsilon$, $\epsilon \sim \mathcal{N}(0, \sigma_{\text{cal}}^2)$, and deploy with $\hat{h}$. This privileged-information scheme exposes the generator to exactly the kind of height error it will face at test time.

Uncertainty-aware gauge. The σ-aware variant samples a per-example noise level $\sigma \sim \text{LogUniform}(\sigma_{\min}, \sigma_{\max})$, perturbs $h^\star$ accordingly and feeds σ as an additional gauge coordinate. A single network thus learns to trust precise heights and to discount imprecise ones; at deployment σ is set to the calibration error of whichever floor source is used.

Table 2: Participant-disjoint splits of EE4D-Motion.

Split	Participants	Windows
Adaptation	240	99,913
Calibration	48	20,016
Development	51	19,369
Official validation	99	44,046

Table 3: Implementation details.

Backbone	ReViV MGET, 12+12 layers, d=768
Adaptation	LoRA r=16, α=32 + gauge MLP
Trainable params	1.8 M
Optimization	AdamW, lr 10^{-3}, cosine, 10k steps
Batch	64 windows (2×32)
Body tokens	30×7, VQ codebook 1024
Decoding	ROAR, 3 steps (selected on calib)
Compute	≈1 GPU-hour per run (24 GB)

Table 4: Official EE4D-Motion validation (99 unseen participants, scored once). Lower is better; EgoLite-G rows are means over three training seeds. †Uses a label-derived floor (oracle reference).

Method	World MPJPE	Pelvis-rel	Sim3-PA	Bone err. (%)	Tilt err. (°)
ReViV (released) (Lyu et al., 2026)	149.75	126.04	84.21	9.97	6.05
EgoAllo, prior floor (Yi et al., 2025)	150.10	125.90	84.65	9.65	5.44
ReViV fine-tuned (budget-matched)	147.59	124.31	81.99	10.48	5.75
EgoAllo + our floor estimate	136.02	111.98	77.70	9.02	5.41
EgoLite-G, gravity only	142.16	119.72	80.14	9.81	5.49
EgoLite-G, gravity + height	130.43	108.08	75.19	9.02	5.25
EgoAllo, label floor†	129.90	110.45	78.79	9.25	5.58

Sensor-anchored readout. All methods, including every baseline, share the same light readout that expresses the generated body in the per-frame yaw-aligned camera frame and corrects it by ridge regression fitted on calibration participants (R2); a small residual MLP (R2+MLP) is used for the strongest baseline variants. The native, readout-free output (R0) is reported as well.

4 Experiments

4.1 Protocol

Data. We use EE4D-Motion (Patel et al., 2025), which provides fitted SMPL-X bodies for Ego-Exo4D (Grauman et al., 2024) recordings, with 2-s windows of 10-Hz Aria MPS head trajectories interpolated to 30 Hz. Table 2 lists the participant-disjoint splits. The official validation split was sealed until every prediction had been generated and its hash recorded; it was then scored exactly once.

Metrics and statistics. World MPJPE is the primary metric; we also report pelvis-relative error, Sim(3)-aligned error (Sim3-PA), bone-length error and torso-tilt error. Differences are paired over participants with a participant-cluster bootstrap (10,000 resamples) and Holm correction over the pre-registered primary contrasts. Trained arms are averaged over three training seeds.

Baselines. (i) ReViV (Lyu et al., 2026) released weights with sampling tuned on calibration participants; (ii) ReViV fine-tuned with exactly our budget, data and LoRA configuration but a constant gauge (budget-matched control); (iii) EgoAllo (Yi et al., 2025) (official checkpoint) with a population-prior floor; (iv) EgoAllo supplied with our take-level floor estimate, the strongest fair baseline; and (v) EgoAllo with the label-derived floor as an oracle reference. Every method receives the same readout capacity.

4.2 Main results on the official validation set

Tables 4 and 5 and Figure 2 summarize the official validation results. Restoring gravity alone reduces world MPJPE by 5.4 mm relative to the budget-matched fine-tune, which receives identical data, capacity and training steps; this isolates the effect of the gauge from the effect of adaptation. Adding the height coordinate yields a further 11.7 mm, for a total

Table 5: Pre-registered and key contrasts on the official validation set (world MPJPE, mm; 95% participant-cluster bootstrap CI). The four pre-registered primary contrasts (⋆) all pass with Holm-corrected $p < 2 \times 10^{-4}$.

Contrast	Δ (mm)	95% CI
⋆ Gravity only − budget-matched ReViV	−5.43	[−7.10, −3.95]
⋆ Gravity only − strongest pre-registered baseline	−7.59	[−9.50, −5.83]
⋆ Gravity + height − budget-matched ReViV	−17.16	[−20.10, −14.09]
⋆ Gravity + height − strongest pre-registered baseline	−19.32	[−22.81, −15.99]
Gravity + height − gravity only	−11.73	[−14.59, −8.83]
Gravity + height − EgoAllo with the same floor	−5.59	[−8.04, −3.22]
Gravity + height − EgoAllo with prior floor	−19.67	[−23.73, −15.89]

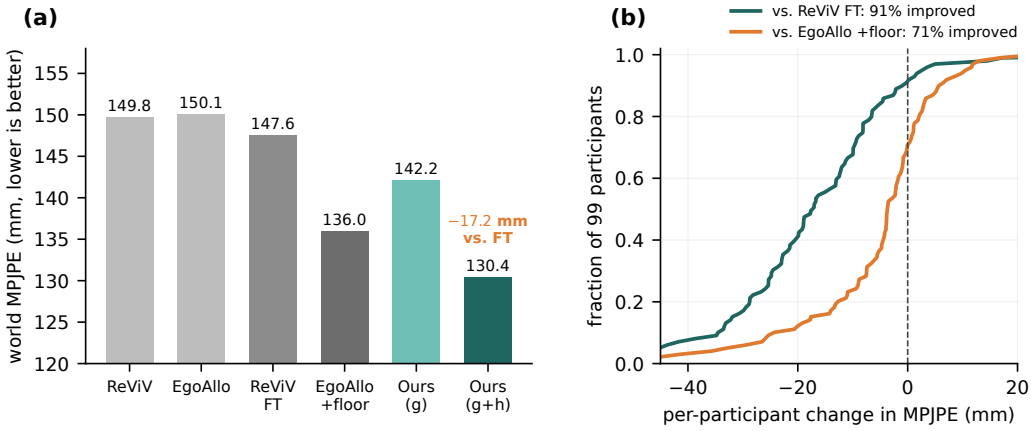


Figure 2: Official validation. (a) World MPJPE of all methods (FT: budget-matched fine-tune; +floor: EgoAllo with our floor estimate). (b) Cumulative distribution of the per-participant change in MPJPE of EgoLite-G (gravity + height): it improves 91% of the 99 unseen participants over the budget-matched ReViV and 71% over EgoAllo given the same floor.

reduction of 17.2 mm (11.6%). Improvements are consistent across metrics: pelvis-relative error drops by 16.2 mm, Sim(3)-aligned error by 6.8 mm, bone-length error from 10.5% to 9.0% and torso-tilt error from 5.75° to 5.25°, showing that the gauge improves the articulated pose and not only its global placement. The improvement is broad rather than driven by a few participants: 91% of the unseen participants improve over the budget-matched model and 93% over the released ReViV (Figure 2b).

When EgoAllo is given the same take-level floor estimate—which by itself is a useful contribution of our floor memory, improving EgoAllo by 14.1 mm—EgoLite-G remains 5.6 mm better, and it matches EgoAllo supplied with a label-derived floor (+0.5 mm, CI [−3.7, 4.5]). Results are stable across training seeds (Table 6): every seed of EgoLite-G is better than every seed of the budget-matched control.

The gain is not produced by the readout. Table 7 compares the native output (R0) with the readout (R2). The gauge gains are even larger without the readout: EgoLite-G lowers native world MPJPE from 184.2 mm to 154.2 mm (30.0 mm), and it is 43.9 mm below native EgoAllo and 87.7 mm below the released ReViV.

4.3 The gain is caused by the gauge

Interventions on gravity. We compare the true-gravity model with two controls of identical capacity: a constant gauge and a gauge whose gravity is randomly permuted across windows

Table 6: Per-seed official validation results (world MPJPE, mm, readout R2).

Method	Seed 1	Seed 2	Seed 3	Mean
ReViV fine-tuned (budget-matched)	147.03	148.11	147.62	147.59
EgoLite-G, gravity only	140.71	144.07	141.70	142.16
EgoLite-G, gravity + height	132.76	129.13	129.40	130.43

Table 7: Native (R0) versus sensor-anchored readout (R2) on the official validation set (world MPJPE / pelvis-relative, mm). The gauge gain holds without any readout.

	Native output (R0)		Readout (R2)	
Method	World	Pelvis-rel	World	Pelvis-rel
ReViV (released)	241.95	213.53	156.45	132.42
EgoAllo, prior floor	198.10	168.00	154.73	130.90
ReViV fine-tuned (budget-matched)	184.22	142.85	147.59	124.31
EgoLite-G, gravity only	166.66	137.15	142.16	119.72
EgoLite-G, gravity + height	154.23	127.29	130.43	108.08

during training. On the development set, the true-gravity model is 6.3 mm and 7.5 mm better than these controls, and permuting gravity at inference increases its error by 15.9 mm while changing 86.4% of the generated tokens—the model actively uses $\mathbf{g}$. Figure 3a shows a clear dose–response: the gain grows monotonically with the deviation of the head pitch from the training mean, reaching 11.1–11.7 mm in the top quartile, exactly where canonicalized camera motion is most ambiguous about posture.

Replication. Table 8 shows that the gravity effect replicates on an independent re-draw of the participant split and across training seeds.

Ablation of the gauge state. Table 9 adds the gauge components one at a time on the development set. Gravity alone removes 6.3 mm; adding a height coordinate whose values are shuffled across windows gives no benefit, whereas the estimated height yields up to a further 10.5 mm. With an exact height the model reaches 111.4 mm, which quantifies how much of the gauge is still to be gained from better floor estimation.

Gauge accuracy sweep. Figure 4a perturbs the height gauge with take-level Gaussian errors of 0–12 cm. Body error decreases monotonically as the gauge becomes more accurate, reaching a 23.4 mm gain over the gravity-only model with an exact gauge and retaining 94% of it at 3 cm. The σ-aware variant exploits accurate gauges: at 0, 3 and 6 cm it improves over fixed-noise training by 4.4, 4.8 and 3.2 mm (all CIs below zero). The sweep identifies precise support-height estimation, e.g. from SLAM point clouds, as the main lever for further gains.

Persistent world state. Figure 4b estimates the floor from the current 2-s window only, from the causal history up to the window, or from the whole take. Floor errors decrease from 12.1 to 10.8 to 9.1 cm and body errors decrease accordingly; the take-level state improves over the window-level estimate by 2.8 mm (CI $[-5.70, -0.03]$). This supports treating the gauge as a persistent property of the world model rather than as a per-window quantity.

5 Discussion

Why a gauge state? The gravity and height coordinates are not new sensor channels; they are already present in the device pose but are removed by the canonicalization that makes multi-dataset token models possible. Treating them as a separate, explicit state has three advantages. First, it preserves the canonical token interface, so the released generator, its tokenizers and its training mixture remain untouched. Second, it makes the physical assumptions of the model inspectable and controllable: the gauge can be shuffled, perturbed or replaced by a better estimate at test time, which is what enables the causal

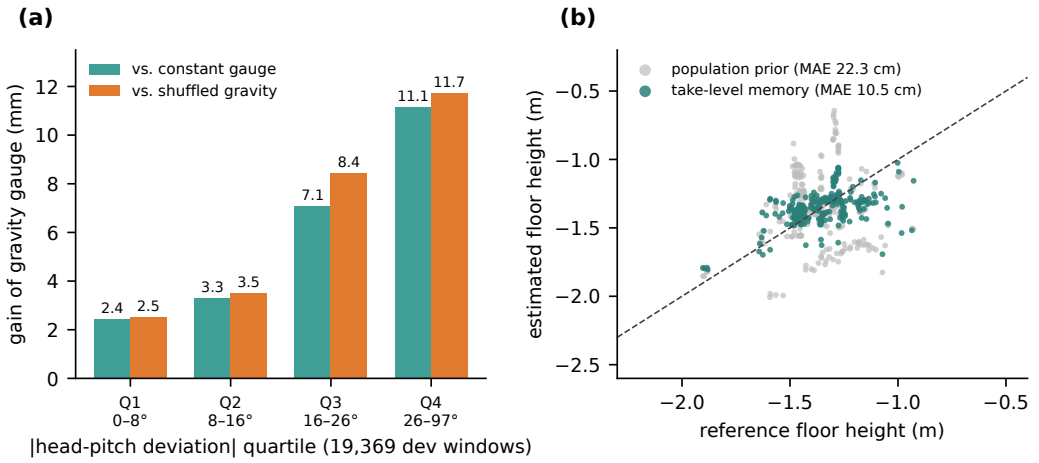


Figure 3: Mechanism. (a) Gain of the gravity gauge on the 19,369 development windows, stratified by the deviation of the first-frame head pitch from the training mean: the gain grows from 2.4 mm to 11.1 mm (vs. constant gauge) and from 2.5 mm to 11.7 mm (vs. shuffled gravity). (b) Take-level floor memory versus a population-height prior on the 249 development takes.

Table 8: Replication of the gravity gauge (development participants, world MPJPE, mm; 95% CI).

Setting	Contrast	Δ	95% CI
Independent split	gravity − constant gauge	−2.87	[−5.72, −0.25]
Independent split	gravity − shuffled gravity	−6.06	[−8.71, −3.58]
Independent split	gravity − released ReViV (R2+MLP)	−8.12	[−11.20, −5.28]
Three training seeds	gravity − constant gauge	−5.16	[−8.44, −2.23]

analyses of Section 4.3. Third, it separates what the generator needs (a gauge) from how it is obtained (IMU, trajectory statistics or, in the future, scene geometry), so improvements in gauge estimation translate directly into body accuracy, as quantified by the gauge-accuracy sweep.

Zero-initialized conditioning. Because $\mathbf{e}(G) \equiv \mathbf{0}$ at initialization, the adapted model reproduces the released generator bit-for-bit on every evaluation window before training, which we verified for the parent, the true-gauge arm and the shuffled-gauge arm. Every difference between arms is therefore caused by learned use of the gauge rather than by a changed initialization. The embedding adds only about 50 k parameters to the LoRA budget, and the whole adaptation fits in about one GPU-hour.

Privileged heights as training signal. The ablation in Table 9 shows that the value of the height coordinate comes from combining a legal estimator at test time with privileged supervision at training time. Supervising the gauge with ground-truth heights of the training participants teaches the generator how height relates to posture; the persistent floor memory then provides a good enough estimate at deployment. The σ-aware variant extends this idea by letting one network adapt to the reliability of whichever floor source is available.

Broader applicability. Nothing in EgoLite-G is specific to a particular generator: any egocentric model that canonicalizes its inputs by the first pose loses the same three coordinates and can recover them with the same zero-initialized embedding. The gauge-accuracy sweep further suggests that coupling the gauge state with scene geometry from SLAM point clouds, available on current smart-glasses platforms (Engel et al., 2023; Pan et al., 2023; Ma et al., 2024), is a direct path to further improvement.

Table 9: Ablation of the gauge state (development set, world MPJPE with readout R2, mm; single training seed).

Gauge input	World MPJPE	Δ vs. constant
Constant gauge (budget-matched control)	141.14	—
Gravity **g**	134.85	−6.29
g + height, shuffled across windows	136.52	−4.62
g + estimated height, noisy privileged training	127.02	−14.12
g + estimated height, σ-aware	125.91	−15.23
g + estimated height, clean privileged training	124.36	**−16.78**
g + exact height (reference)	111.44	−29.70

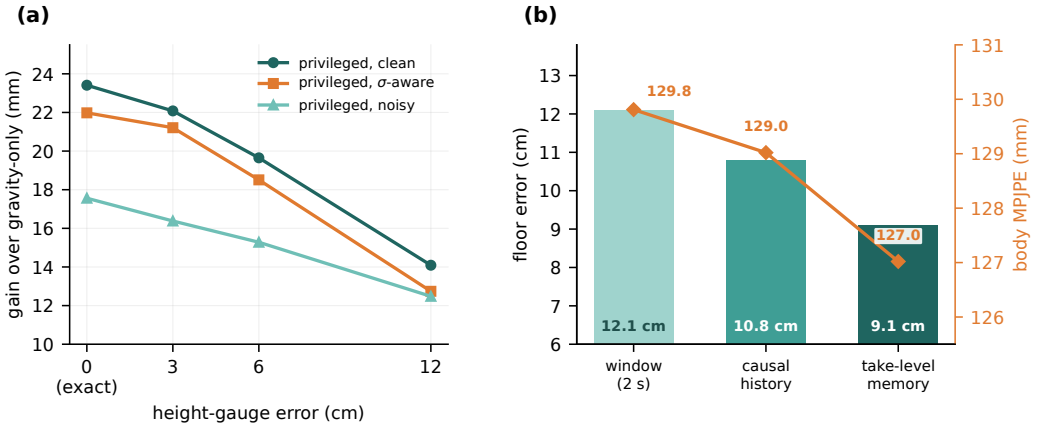


Figure 4: Gauge accuracy and persistence (development set). (a) Gain over the gravity-only model as the height gauge is perturbed with take-level Gaussian errors: body error decreases monotonically with gauge accuracy, and the σ-aware gauge exploits accurate heights. (b) Estimating the floor from the current window, from the causal history or from the whole take: a more persistent world state yields a more accurate floor (bars) and a lower body error (line).

6 Conclusion

First-frame canonicalization lets egocentric token world models share one generator across datasets, but it removes the physical gauge on which world-frame body pose depends. EgoLite-G restores gravity and support height as an explicit, persistent state estimated from head-worn sensors and injects it through zero-initialized embeddings with privileged training. On the official EE4D-Motion validation set, evaluated once under a sealed protocol, this lowers world MPJPE by 17.2 mm over a budget-matched fine-tune, improves 91% of unseen participants, and outperforms EgoAllo given the same floor estimate. Controlled interventions attribute the gain to the gauge, and the gauge-accuracy sweep points to scene-level floor estimation as a direct path to further improvement. Because the gauge enters through a zero-initialized embedding, the approach applies to any canonicalized egocentric generator.

A Statements

Ethics statement. This work uses only publicly released research datasets (EE4D-Motion annotations of Ego-Exo4D recordings) under their licenses; all participants consented to data collection by the dataset providers, and no new human data were collected. We report results at the participant level and never attempt to identify individuals. Egocentric body estimation can support assistive technology, rehabilitation and AR/VR avatars; like any motion-capture technology it could also be misused for unconsented monitoring, and we encourage deployment only with the informed consent of the wearer.

Reproducibility statement. All splits are participant-disjoint and fixed by seeded hashing before training. The official validation split was sealed until all predictions were generated and their SHA-256 hashes recorded, and was then scored once. Section 4 and Table 3 list all hyperparameters (LoRA r=16, α=32, 10k steps, batch 64, learning rate 10^{-3}, condition dropout 0.1, ROAR decoding with three steps). The pre-registration documents, frozen prediction hashes, training configurations and evaluation scripts will be released. Each training run takes about one GPU-hour on a 24 GB GPU.

Use of large language models. Large language models were used as general-purpose assistants for code writing and debugging, for polishing the language of the manuscript, and for preparing figure scripts. All research ideas, experimental designs, results and claims were verified by the authors, who take full responsibility for the content of the paper.

References

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T. Freeman. MaskGIT: Masked generative image transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.

Jakob Engel et al. Project Aria: A new tool for egocentric multi-modal AI research. arXiv preprint arXiv:2308.13561, 2023.

Kristen Grauman et al. Ego-Exo4D: Understanding skilled human activity from first- and third-person perspectives. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024.

Kai Guan et al. Seeing the world and the self from egocentric video. arXiv preprint arXiv:2609.01276, 2026.

Vladimir Guzov, Aymen Mir, Torsten Sattler, and Gerard Pons-Moll. Human POSEitioning system (HPS): 3d human pose estimation and self-localization in large scenes from body-mounted sensors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598, 2022.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In International Conference on Learning Representations (ICLR), 2022.

Jiaxi Jiang, Paul Streli, Huajian Qiu, Andreas Fender, Larissa Laich, Patrick Snape, and Christian Holz. AvatarPoser: Articulated full-body pose tracking from sparse motion sensing. In European Conference on Computer Vision (ECCV), 2022.

Jiaxi Jiang, Paul Streli, Manuel Meier, and Christian Holz. EgoPoser: Robust real-time egocentric pose estimation from sparse and intermittent observations everywhere. In European Conference on Computer Vision (ECCV), 2024.

Jiaman Li, Karen Liu, and Jiajun Wu. Ego-body pose estimation via ego-head pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023.

Zhengyi Luo, Ryo Hachiuma, Ye Yuan, and Kris Kitani. Dynamics-regulated kinematic policy for egocentric pose estimation. In Advances in Neural Information Processing Systems (NeurIPS), 2021.

Xiaozhong Lyu, Gen Li, Zhiyin Qian, Xucong Zhang, Marc Pollefeys, and Siyu Tang. ReViV: Reconstructing the viewer and the view in 4d from monocular egocentric video. In European Conference on Computer Vision (ECCV), 2026. arXiv:2607.17790.

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. In European Conference on Computer Vision (ECCV), 2024.

David Mizrahi, Roman Bachmann, Oğuzhan Fatih Kar, Teresa Yeo, Mingfei Gao, Afshin Dehghan, and Amir Zamir. 4M: Massively multimodal masked modeling. In Advances in Neural Information Processing Systems (NeurIPS), 2023.

Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar Parkhi, Richard Newcombe, and Carl Yuheng Ren. Aria Digital Twin: A new benchmark dataset for egocentric 3d machine perception. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023.

Chaitanya Patel, Hiroki Nakamura, Yuta Kyuragi, Kazuki Kozuka, Juan Carlos Niebles, and Ehsan Adeli. UniEgoMotion: A unified model for egocentric motion reconstruction, forecasting, and generation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2025.

Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. Neural Networks, 22(5-6):544–557, 2009.

Brent Yi, Vickie Ye, Maya Zheng, Yunqi Li, Lea Müller, Georgios Pavlakos, Yi Ma, Jitendra Malik, and Angjoo Kanazawa. Estimating body and hand motion in an ego-sensed world. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023.