REUP: RETRIEVAL-AUGMENTED UNDERSTANDING AND PREDICTION OF HUMAN BEHAVIOUR FROM SPARSE EGOCENTRIC SIGNALS

Anonymous authors
Paper under double-blind review

ABSTRACT

An assistant such as smart glasses needs a holistic picture of its wearer. It must know what the wearer does, where the wearer is, how the body moves and what comes next. Existing systems run large models on continuous egocentric video or train one model per question. We show that two cheap signals suffice: the head pose the glasses already track and one image. We present **ReUP**, a **Re**trieval-augmented model for **U**nderstanding and **P**redicting human behaviour. Its memory holds fully instrumented walkthroughs, where every moment carries all the answers. Multi-modal contrastive learning gives its head-motion key rich semantics from the body and narration of each moment. Behaviour-guided contrastive learning draws together moments that share a place, an activity or a future. At query time, head motion and image each retrieve similar moments. For each task, a learned scorer picks the useful ones from both, and a light decoder turns them into an answer. The model blends what only each cue offers: the place from the image, the activity and body from the motion. With fewer than 2M parameters, one model recognises the current and next action, estimates and forecasts the body pose and localises the wearer. In unseen buildings and on an unseen dataset, it is on average 3.5% better than the strongest baseline of each metric. Retrieving recorded human behaviour gives a lightweight model a prior for holistic and effective understanding and prediction, from minimal input and in unseen scenes.

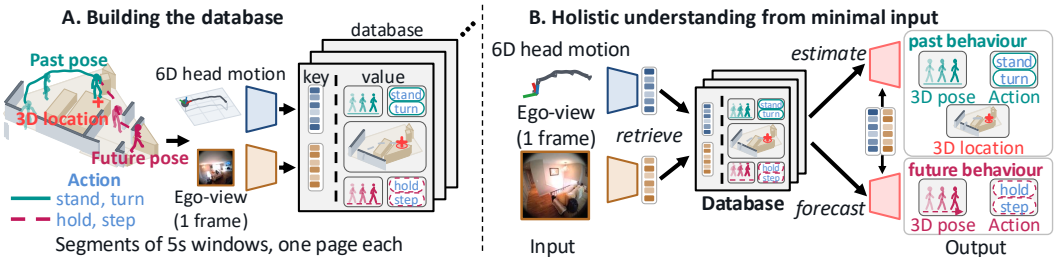


Figure 1: **Overview of ReUP.** (A) Each five-second window becomes a database page: keys from its head motion and one frame, values from past pose, action, 3D location, and future pose, action. (B) A query's keys retrieve pages, which are decoded into past and future behaviour and 3D location.

1 INTRODUCTION

An assistant that lives with a person, from smart glasses to robots, needs a running model of them (Grauman et al., 2022; 2024; Ma et al., 2024; Yang et al., 2025). It must know what the person is doing, how their body moves now and next, and where they are. These are facets of one situation, so one model can answer them more cheaply and consistently.

Two families of systems answer these questions today; neither fits an always-on wearable. Language models over egocentric video and motion (Lin et al., 2022; Zhao et al., 2023; Ye et al., 2025; Hong et al., 2025; Bae et al., 2026; Yang et al., 2025; Li et al., 2025a) are holistic but need continuous

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.

video and backbones far too large for the device. Dedicated models are light but answer one question each: body estimators conditioned on the head (Li et al., 2023; Yi et al., 2025; Jung & Kim, 2026; Tran et al., 2025; Jiang et al., 2022; Du et al., 2023), forecasters (Yuan & Kitani, 2019; Escobar et al., 2025; Patel et al., 2025), activity recognisers (Tan et al., 2023; Huang et al., 2026; Xue et al., 2026) and localisers (Herath et al., 2022; Zhang et al., 2025a). Together they cost more and can disagree. Each also stores its knowledge in its weights, so it needs retraining for a new building or dataset, or degrades there.

We ask how little input holistic understanding needs, and how to keep it working in new places. Smart glasses already track the head pose for rendering (Engel et al., 2023). This signal is cheap and always on, and it is informative about activity (Tsutsui et al., 2021; Tan et al., 2023; Huang et al., 2026), body (Li et al., 2023; Yi et al., 2025) and place (Herath et al., 2022). We add a single image. The difficulty is ambiguity: one walk down a corridor looks like every other. What a trajectory means also lives in modalities that are absent at test time.

Our premise is that a minimal input does not have to *describe* the situation; it only has to *address* it. Suppose a building is walked once with full instrumentation. Every moment of that walk then carries all its answers, recorded together. A query that finds the moments it resembles can read every answer from them. Where the wearer is can only be read from a memory of the same building. How people act and move, however, is shared across buildings. A memory of other buildings, or of another dataset, thus still answers the behaviour questions. Because the knowledge lives in the memory, as in retrieval-augmented language models (Khandelwal et al., 2020; Guu et al., 2020; Lewis et al., 2020; Borgeaud et al., 2022), the model only learns to find and correct recalled answers. It stays small, and a new building is added by indexing a walk through it, without a gradient step.

We build this idea as ReUP (Figure 1), a retrieve-then-edit model (Hashimoto et al., 2018). First, a head-pose encoder learns from privileged modalities that exist only in the recordings (Vapnik & Vashist, 2009; Lopez-Paz et al., 2016). It is bound to the synchronised body and narration (Radford et al., 2021; Girdhar et al., 2023) and organised by place, activity and near future (Khosla et al., 2020). As a result, windows lie close only if their situations are alike. Every window is then indexed under this motion key and under the embedding of one frame. The frame says where the wearer is (Arandjelović et al., 2016) but little about the body. At query time, both keys retrieve similar moments, and each task weighs them with its own learned balance. Each task then reads a base answer and refines it with a light decoder that starts at zero. All tasks are trained together.

This gives one small model the strengths that separate experts have alone. We test every method on the same input in a known building, an unseen building and an unseen dataset (§3.3). ReUP matches or beats the expert on most metrics and is on average 3.8% better than the strongest of them. The database makes this hold in new scenes. There, retrieving from other buildings improves the behaviour tasks and forecast tasks up to 3.6% over the same model without memory.

In summary, ReUP achieves the following:

- *Efficiency*: fewer than 2M learned parameters answer all five questions, two orders of magnitude fewer than the language models;
- *Holistic effectiveness*: one memory and one retrieval serve every task, on average 3.5% better than the strongest baseline of each metric in unseen buildings and on an unseen dataset;
- *Generalisability*: a database of other buildings, or of another dataset, improves the behaviour tasks in unseen scenes.

2 Related work

Egocentric understanding. Egocentric perception has been driven by first-person video collections (Damen et al., 2018; Grauman et al., 2022; 2024) and by Aria recordings (Engel et al., 2023) with motion capture and narration (Ma et al., 2024) or a digital twin of the scene (Pan et al., 2023). Video models recognise and anticipate actions (Furnari & Farinella, 2019; Girdhar & Grauman, 2021; Tai et al., 2026), learn from narrated clips (Lin et al., 2022; Zhao et al., 2023) and estimate, forecast or generate the body (Yuan & Kitani, 2019; Li et al., 2023; Patel et al., 2025; Xiu et al., 2026), and language models over video and motion (Ye et al., 2025; Hong et al., 2025; Bae et al.,

2026; Yang et al., 2025; Li et al., 2025a) answer many questions at the price of a continuous video stream and a large backbone. Head motion can replace most of the video (Tan et al., 2023; Tsutsui et al., 2021), but only for activity. ReUP keeps the holistic scope (Figure 2) with the economy of head motion plus one frame through a frozen image encoder (Radford et al., 2021).

Understanding from sparse egocentric signals. The body has been recovered from six IMUs (Huang et al., 2018; Yi et al., 2021), from the three points of a VR headset and controllers (Jiang et al., 2022; Winkler et al., 2022; Du et al., 2023; Castillo et al., 2023; Jiang et al., 2024; Dai et al., 2024; Barquero et al., 2025), and from the head alone (Li et al., 2023; Yi et al., 2025; Tran et al., 2025; Jung & Kim, 2026). The same signals support body forecasting (Escobar et al., 2025), where zero-velocity persistence (Martinez et al., 2017) is a strong reference, activity recognition (Moon et al., 2023; Xu et al., 2025b; Zhang et al., 2025b; Huang et al., 2026; Xue et al., 2026) and inertial localisation (Herath et al., 2020; Liu et al., 2020; Herath et al., 2022; Zhang et al., 2025a), each as a separate model with one output, the localisers bound to one building. ReUP instead uses the head trajectory as a key into instrumented walkthroughs: one retrieval serves all five outputs, and, as in motion graphs and motion matching (Kovar et al., 2002; Holden et al., 2020), bodies are retrieved and corrected rather than generated.

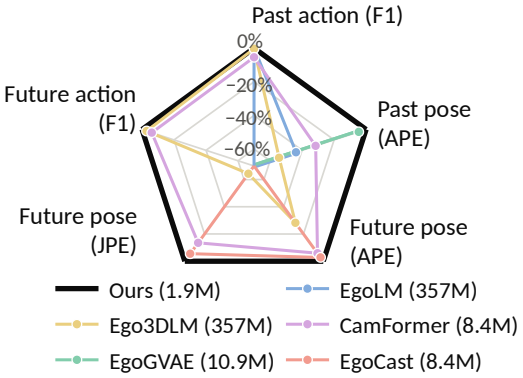


Figure 2: **ReUP against the baselines** on the five metrics of Setup B (unseen building). With one model, ReUP is best or on par on every metric. The radius is the difference from ReUP (black pentagon) in percent, outward is better, and a metric a baseline does not produce sits at the centre. In the legend: learned parameters producing the reported outputs, without the frozen image encoder.

Retrieval-augmented models. Retrieval-augmented models keep knowledge in an external memory instead of their weights, in language (Khandelwal et al., 2020; Guu et al., 2020; Lewis et al., 2020; Borgeaud et al., 2022; Hashimoto et al., 2018), long-tail recognition (Long et al., 2022; Iscen et al., 2023), image synthesis (Blattmann et al., 2022), vision-language tasks (Chen et al., 2022; Hu et al., 2023; Yasunaga et al., 2023), motion synthesis (Zhang et al., 2023; Xu et al., 2025a), robot policies (Memmel et al., 2025; Pfeifer et al., 2026), forecasting (Ning et al., 2025), egocentric video (Xu et al., 2024; Yang et al., 2025), and place recognition and localisation (Arandjelović et al., 2016; Tzachor et al., 2025; Chen et al., 2026; Barroso-Laguna et al., 2026). ReUP applies this motivation to human behaviour: it understands and predicts a person in a new scene from a database of how others behaved in similar moments, and a new building is added by recording, not retraining. As the memory holds what the query lacks (body, narration, place), the key is learned contrastively (van den Oord et al., 2018; Khosla et al., 2020; Radford et al., 2021; Girdhar et al., 2023; Tevet et al., 2022; Petrovich et al., 2023; Li et al., 2025b; Kinfu & Vidal, 2025) with privileged information (Vapnik & Vashist, 2009; Lopez-Paz et al., 2016) to accurately retrieve necessary information for prediction in new scenes.

3 Method

3.1 Problem definition

A sample is a five-second window of a recording. Its input is the headset's 6-DoF pose over 100 steps, $\mathbf{H} = \{(\mathbf{p}_t, \mathbf{q}_t)\}_{t=0}^{99}$, from the glasses' SLAM (Engel et al., 2023), and one RGB frame I from the end of the window. Φ removes the window-mean position and heading, $\mathbf{h} = \Phi(\mathbf{H}) \in \mathbb{R}^{12\times100}$, so no absolute position or heading reaches any model. Concretely, Φ subtracts the window-mean head position, rotates the sequence about gravity so that its mean heading faces $+x$, and describes each step by its position, velocity and the first two columns of its rotation. A model

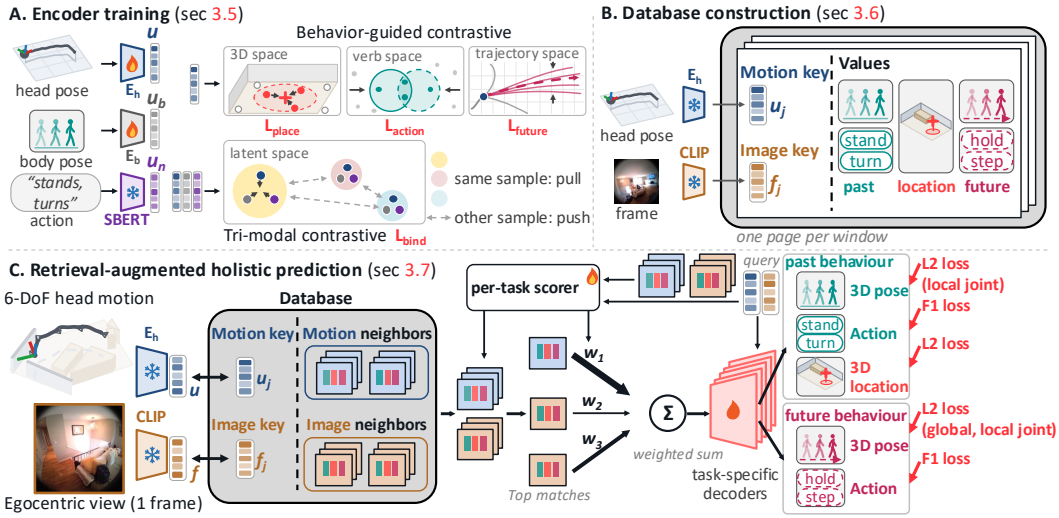


Figure 3: **ReUP. (A)** Contrastive training binds a window's head-motion key to its body, its narration and windows sharing its place, verbs or future. **(B)** Each window is a page with two keys (motion, CLIP frame) and three values (past behaviour, location, future behaviour). **(C)** A query's keys retrieve neighbours; per-task scorers weigh them and decoders edit the weighted answers.

answers five questions: *understanding* and *action prediction* score 44 verbs for the current window and for the next narrated activity, $\hat{\mathbf{a}}, \hat{\mathbf{a}}^F \in \mathbb{R}^{44}$; *tracking* and *forecasting* estimate the body as 22 SMPL joints (Loper et al., 2015) over the window and the next five seconds, $\hat{\mathbf{B}}, \hat{\mathbf{B}}^F \in \mathbb{R}^{100\times22\times3}$, and *localisation* gives the last head position in a known building's coordinates, $\hat{\mathbf{x}} \in \mathbb{R}^3$.

3.2 Dataset

Nymeria (Ma et al., 2024) provides training data, memories and in-dataset queries: everyday activities in twelve buildings, recorded with Aria glasses (Engel et al., 2023), a motion-capture suit and timestamped narrations, cut into five-second windows at 20 Hz. Verb labels are matched in a window's own narration, and for action prediction in those starting in the next five seconds. ADT (Pan et al., 2023) is the out-of-dataset test: one apartment with a different rig and skeleton, mapped once to SMPL-22 for all methods (Appendix A.1), with verb pseudo-labels from captions.

3.3 Evaluation protocol

We evaluate in three regimes; three Nymeria buildings are held out for evaluation and nine are used for training. In *Setup A (known building)*, each evaluation building's recordings are split into a *database* half, which forms the memory (baselines may also train on it), and a *query* half for evaluation. *Setup B (unseen building)* queries them against a memory of the nine training buildings, and *Setup C (unseen dataset)* trains on and remembers all twelve buildings and queries ADT. All methods read the same canonicalised input, use only their regime's allowed data and are scored on the same windows. Understanding and action prediction are scored by top-four micro-F1; tracking by APE, the head-relative joint error; forecasting by APE and by JPE, the joint error in the last head frame; and localisation, in Setup A only, by the ℓ_2 error at the last step (Appendix A).

3.4 Overview

ReUP answers the five questions by recall (Figure 3). A walk through a building with full instrumentation records, at every moment, the body, the activity, the place and what came next. ReUP stores these moments as memory pages keyed by head motion and one frame (§3.6); a query retrieves the moments it resembles, and each task reads its answer from them and corrects it with a small decoder (§3.7). Learning is spent on the key (§3.5) and on these corrections, while the memory supplies what the window cannot show, and a new building is added by indexing, without training.

3.5 HEAD-POSE ENCODER

A retrieval key is judged by its neighbourhoods, not by what can be decoded from it. It should bring together windows in which the body moves alike, the wearer does the same thing, in the same place, and continues the same way; these facets exist only at indexing time, so we use them as privileged information (Vapnik & Vashist, 2009) that shapes the key in training only.

The head encoder E_h is a temporal convolutional network on $\mathbf{h}$ followed by a projector, and gives the unit-norm key $\mathbf{u} \in \mathbb{R}^{128}$. During training only, a body encoder E_b of the same form embeds the window's body as $\mathbf{u}_b$, and a projector on a frozen sentence embedding (Reimers & Gurevych, 2019) embeds its narration as $\mathbf{u}_n$. Two kinds of contrastive terms shape the key, always as constraints on which windows lie close, never as targets regressed from it:

$$\begin{aligned} \mathcal{L}_{\text{enc}} &= \mathcal{L}_{\text{bind}} + \lambda_p \, \mathcal{L}_{\text{place}} + \lambda_a \, \mathcal{L}_{\text{action}} + \lambda_f \, \mathcal{L}_{\text{future}}, \\ \mathcal{L}_{\text{bind}} &= \lambda_{hb} \, \ell(\mathbf{u}, \mathbf{u}_b) + \lambda_{hn} \, \ell(\mathbf{u}, \mathbf{u}_n) + \lambda_{bn} \, \ell(\mathbf{u}_b, \mathbf{u}_n). \end{aligned} \tag{1}$$

The *tri-modal contrastive* term $\mathcal{L}_{\text{bind}}$ pulls the three embeddings of a window together and away from those of other windows, with ℓ the symmetric InfoNCE loss (van den Oord et al., 2018) and fixed weights λ (Appendix B.3). It places a head trajectory where its kinematics and intent would lie (Girdhar et al., 2023): the rich modalities only organise the space of the poorest one. The *behaviour-guided contrastive* terms instead pull together *different* windows whose behaviour agrees, in where it happens, what is done and how it continues: $\mathcal{L}_{\text{place}}$, $\mathcal{L}_{\text{action}}$ and $\mathcal{L}_{\text{future}}$ are supervised contrastive losses (Khosla et al., 2020) on $\mathbf{u}$ over a batch $\mathcal{B}$ with temperature τ that differ only in their positive sets $P(i)$,

$$\mathcal{L}_{\text{sup}} = -\frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \frac{1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(\mathbf{u}_i^\top \mathbf{u}_p / \tau)}{\sum_{k \in A(i)} \exp(\mathbf{u}_i^\top \mathbf{u}_k / \tau)}. \tag{2}$$

For $\mathcal{L}_{\text{place}}$, $P(i)$ holds the windows of the same building whose end positions lie within a radius ρ, and only windows of that building compete in $A(i)$. For $\mathcal{L}_{\text{action}}$, it holds the windows whose verb sets overlap with a Jaccard index of at least η. $\mathcal{L}_{\text{future}}$ is the retrieval counterpart of forecasting: instead of asking the key to predict the next five seconds, it asks its neighbours to share them, and $P(i)$ holds the ν windows of the batch whose future head paths, each in its own window's head frame, are closest. A second projector on the frozen trunk, trained with the head–narration and action terms, gives a semantic embedding $\mathbf{u}_s$ that only describes the query to the decoder.

3.6 DATABASE CONSTRUCTION

Every window of the recordings that a regime allows becomes one page of the memory $\mathcal{M}$. A page j holds two keys, the motion key $\mathbf{u}_j$ of its head pose and the CLIP ViT-L/14 embedding $\mathbf{f}_j$ of its frame (Radford et al., 2021), and the values of all five questions, recorded together: its past behaviour (the body $\mathbf{B}_j$ and the verbs $\mathbf{a}_j$), its location $\mathbf{x}_j$, and its future behaviour (the body $\mathbf{B}_j^F$ and the verbs $\mathbf{a}_j^F$ of the next five seconds). Bodies are stored in the page's own observed head frame, so that they can be moved onto any query, and query recordings never enter the memory. The two keys are complementary: motion finds the same movement anywhere but not where, and the image finds the same place but says little about the body.

3.7 RETRIEVAL-AUGMENTED HOLISTIC PREDICTION

Retrieve. A query is encoded into $\mathbf{u}$ and $\mathbf{f} = \text{CLIP}(I)$, and each key retrieves its ξ nearest pages by cosine similarity. The candidate set U is their union: every candidate can be judged by both keys.

Score per task. Each task t blends the place evidence of the image with the movement evidence of the motion in its own distribution over U,

$$\mathbf{w}_t(j) = \operatorname*{softmax}_{j \in U} \Big(\kappa \, \gamma_t \, s_u(j) + \kappa \, \sigma_u \, \beta_t \, z_f(j) \Big), \qquad \gamma_t = \omega^{\tanh \alpha_t}, \quad (\alpha_t, \beta_t) = W_t \, \mathbf{c}, \tag{3}$$

where $s_u(j)$ is the motion similarity of candidate j, $z_f(j)$ its image similarity standardised over U, σ_u the spread of s_u over U, and κ and ω fix the scale and the range of the sharpness. Both

Method	Unified?	Params	Setup A: known building						Setup B: unseen building				
			Pose (mm)↓			Action↑		Loc. (m)↓	Pose (mm)↓			Action↑	
			APE	JPE	APE	F1	F1	L2	APE	JPE	APE	F1	F1
NiLoc*	✗	2.03M	–	–	–	–	–	1.790	–	–	–	–	–
EAIL (+point cloud)	✗	0.61M	–	–	–	–	–	2.752	–	–	–	–	–
EgoEgo*	✗	4.31M	141.7	–	–	–	–	–	136.0	–	–	–	–
EgoAllo*	✗	50.8M	179.7	–	–	–	–	–	159.0	–	–	–	–
EgoGVAE*	✗	10.9M	**128.5**	–	–	–	–	–	127.1	–	–	–	–
EgoCast*	✗	8.38M	–	410.7	178.4	–	–	–	–	404.8	173.1	–	–
EgoLM	○	357M	184.8	–	–	0.522	–	–	174.8	–	–	**0.523**	–
Ego3DLM, no scene	○	357M	190.9	562.9	218.3	**0.526**	0.461	–	187.7	630.3	216.0	0.519	0.455
CamFormer* (our readout)	✓	8.37M	165.4	447.4	183.6	0.424	0.345	1.794	159.9	435.5	178.3	0.493	0.440
Ours, retrieval only	✓	1.92M	229.1	665.0	269.5	0.404	0.324	3.276	206.0	597.9	247.5	0.463	0.389
Ours	✓	1.92M	**128.5**	**400.0**	**176.2**	0.520	**0.473**	**1.415**	**121.6**	**383.7**	**168.5**	0.521	**0.469**
% vs. best baseline			+0.0%	+2.6%	+1.2%	-1.2%	+2.6%	+20.9%	+4.3%	+5.2%	+2.7%	-0.4%	+3.0%

Table 1: **Results on Nymeria in a known building (A) and an unseen building (B).** Teal metrics score the input window (past), crimson ones the next 5 s (future). Best in **bold**, second best underlined, a tie for the best both; the last row gives our relative difference to the best baseline (green: better or equal, grey: worse). Params: learned parameters that produce the reported outputs, the frozen image encoder not counted; NiLoc trains one model per building (the largest is listed). *: re-implemented under our protocol to take the 6-DoF head-pose sequence and one egocentric RGB frame as input; the original method takes a different input.

similarities pass through a low-rank adapter shared by all tasks, so the task losses jointly reshape the metric within U. The coefficients come from a zero-initialised map W_t of a context $\mathbf{c}$ describing the query and what each key returned, so each task decides per query how sharply to concentrate and how much to trust the image; at initialisation the motion key alone ranks. Since the weights are trained by each task's own loss, the scorer learns which neighbours are *useful* for that answer rather than merely similar to the query.

Read and edit. Each task keeps the ζ highest-weight candidates $\mathcal{N}_t$ of $\mathbf{w}_t$ that carry its value $\mathbf{y}_{t,j}$ (page j's answer to task t), renormalises their weights to $\bar{\mathbf{w}}_t$, and answers

$$\hat{\mathbf{y}}_t = \mathrm{base}_t\big(\mathcal{N}_t, \bar{\mathbf{w}}_t\big) + \mathrm{Dec}_t\big(\mathrm{base}_t,\ \{\mathbf{y}_{t,j}\}_{j\in\mathcal{N}_t},\ \bar{\mathbf{w}}_t,\ \mathbf{c}\big). \tag{4}$$

The base is read without parameters and follows the kind of answer. Where the retrieved answers scatter around one answer that the input constrains — verbs, positions and the current body — it is their weighted mean; where averaging would blend distinct outcomes — the future body, which nothing in the input observes — it is the single highest-weight future, moved onto the body the model has just estimated. Dec_t is one architecture for all tasks: per output slot (a time step of a body, or a single vector), the base token cross-attends to the ζ neighbours' raw values, weights and ranks, a small transformer mixes the slots, and a zero-initialised layer emits the correction; the two body decoders also read the query's own head features (Appendix B.1).

Training. The encoder, the semantic projector and an image probe used by the verb readouts are trained on the regime's training recordings and frozen. The retrieval stage (adapter, scoring maps, context and decoders) is then trained with the training windows as both queries and memory, a query's own recording never retrievable. To reproduce both Nymeria regimes, each query retrieves, independently per epoch, either from other recordings of its own building, sized like a Setup A memory, or from the other buildings only, where relative positions are zeroed and the localisation loss is off. Each task has one loss on its final output: the ℓ_2 error for localisation, a batch-level soft micro-F1 for understanding and action prediction, the head-relative joint error for tracking, and the joint error plus its head-relative form for forecasting. Two auxiliary losses, on the occupancy of the floor around the wearer and on the surrounding furniture, are optimised as well, and their outputs are not evaluated (Appendix B.2). A straight-through estimator makes the single-page forecast base differentiable in its weights, so every loss trains how its task retrieves and, through the shared context and adapter, how the others do. The losses are normalised by their first-batch values and summed, the two body losses with weight λ_{body}; the optimisation settings are listed in Appendix B.3.

Method	Pose (mm)↓ APE	JPE	APE	Action↑ F1	F1
EgoEgo*	169.8	–	–	–	–
EgoAllo*	194.7	–	–	–	–
EgoGVAE*	157.8	–	–	–	–
EgoCast*	–	512.5	177.3	–	–
EgoLM	181.6	–	–	0.365	–
Ego3DLM, no scene	194.1	597.6	224.7	0.387	0.386
CamFormer*	172.9	**456.5**	181.8	0.372	0.355
Ours, retrieval only	211.9	608.0	243.7	0.312	0.300
Ours	**140.0**	467.5	**168.9**	**0.394**	**0.404**
% vs. best baseline	+11.3%	-2.4%	+4.7%	+1.9%	+4.6%

Table 2: **Results on ADT, an unseen dataset (C).** Colours, marks and * as in Table 1.

Tri-modal	Behaviour-guided			Past		Future		
				Action	Motion	Action	Motion	
$\mathcal{L}_{\text{bind}}$	$\mathcal{L}_{\text{place}}$	$\mathcal{L}_{\text{action}}$	$\mathcal{L}_{\text{future}}$	F1↑	APE↓	F1↑	JPE↓	APE↓
✓				0.374	234.6	0.300	706.3	279.0
✓			✓	0.374	235.0	0.296	671.5	276.7
✓	✓	✓		0.392	229.4	0.320	702.1	273.1
✓	✓	✓	✓	**0.404**	**229.1**	**0.324**	**665.0**	**269.5**

Table 3: **Encoder objective** (Setup A, retrieval only). Each row trains the head-pose key with the checked terms of Eq. 1 and reads its nearest neighbours, without the image or the decoder. JPE and APE in mm.

Inference. At test time only E_h, CLIP and the retrieval stage run: one pass against the regime's memory gives all five answers, and the memory can be swapped without retraining.

4 Experiments

We compare ReUP with task-specific experts, for localisation (NiLoc (Herath et al., 2022), EAIL (Zhang et al., 2025a)), body tracking (EgoEgo (Li et al., 2023), EgoAllo (Yi et al., 2025), EgoGVAE (Jung & Kim, 2026)), forecasting (EgoCast (Escobar et al., 2025)) and activity (CamFormer (Xue et al., 2026)), and with the language models EgoLM (Hong et al., 2025) and Ego3DLM (Bae et al., 2026), which answer several questions. All are re-trained under our protocol on the same input, the head pose and the last frame (*: the original input differs), and Appendix A.3 describes each and our changes; "Ours, retrieval only" is ReUP without use of decoders.

4.1 Quantitative results

Known and unseen buildings. Table 1 compares all methods in a known building (A) and in unseen buildings (B). In the known building, one model is best or tied on five of the six metrics: it localises with a fifth less error than NiLoc, forecasts better than EgoCast, ties the best tracker, and predicts the next activity better than the language models, which stay slightly ahead only on the current activity. In unseen buildings the gap widens: averaged over the behaviour metrics, ReUP leads the strongest expert of each by 3.0%, against 1.0% in the known building, with the largest gains on tracking and forecasting. A plausible reading is that the experts must generalise from what their weights absorbed in the training buildings, whereas ReUP retrieves, for each query, recorded moments whose behaviour resembles it. In this regime the memory holds only other buildings, so what it contributes is behaviour that recurs across places — how people walk, sit or reach — rather than the place itself, and such behaviour appears to transfer more readily than learned weights do. The retrieval-only row separates the two parts of ReUP: retrieval already gives plausible answers, and the decoders, which also read the query's own head motion, remove roughly two fifths of the remaining body error by fitting the recalled bodies to this wearer's movement.

Unseen dataset. Table 2 carries the comparison to ADT, recorded with a different rig in a different apartment. The advantage holds across datasets: ReUP is ahead on four of the five metrics, by 4.0% on average and by 11% on tracking, and only CamFormer with a forecasting head stays slightly ahead on the forecast in the ADT frame. A memory of Nymeria walkthroughs thus still supplies useful behaviour for people in another dataset, although the rig and skeleton differ.

4.2 Qualitative results

Figure 4 follows one window through all five questions: the wearer walks into a kitchen and keeps walking. From a single frame and the head's path, ReUP names the current activity as well as the language model, predicts more of what comes next, and tracks the body more closely than the best tracker. Its forecast heads the way the wearer actually walks and ends near the true end point,

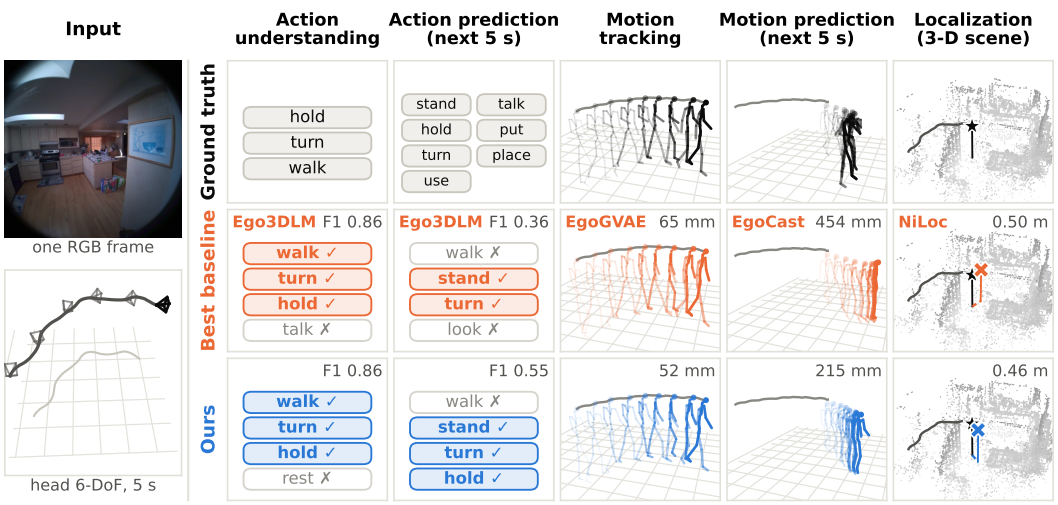


Figure 4: **One input, every answer** (Setup A). The input (one RGB frame and 5 s of head 6-DoF pose), the ground truth, the best baseline of each task (Table 1) and ReUP; ✓ marks correct verbs, and values are this window's F1 or error.

	Past		Future			Place
	Action	Motion	Action	Motion		Loc.
Model	F1↑	APE↓	F1↑	JPE↓	APE↓	L2↓
Setup A: known building (Nymeria)						
Ours	**0.520**	**128.5**	**0.473**	**400.0**	**176.2**	**1.415**
motion only	0.463	147.9	0.430	430.1	193.8	3.394
image only	0.487	170.6	0.468	498.5	194.8	1.680
Setup B: unseen building (Nymeria)						
Ours	**0.521**	**121.6**	**0.469**	**383.7**	**168.5**	–
motion only	0.481	138.8	0.425	408.4	184.2	–
image only	0.469	163.3	0.458	495.3	186.9	–

Table 4: **Input ablation.** Each row is trained and evaluated with the input shown. JPE and APE in mm, L2 in m; best of each block in bold.

	Past		Future			Place
	Action	Motion	Action	Motion		Loc.
Model	F1↑	APE↓	F1↑	JPE↓	APE↓	L2↓
Setup A: known building (Nymeria)						
Ours	**0.520**	**128.5**	**0.473**	**400.0**	**176.2**	**1.415**
single task	0.502	129.7	0.465	401.3	176.3	1.435
Setup B: unseen building (Nymeria)						
Ours	**0.521**	**121.6**	**0.469**	**383.7**	**168.5**	–
single task	0.508	122.0	0.454	387.2	168.8	–

Table 5: **Task ablation.** *Single task*: each column comes from a decoder trained on that task alone. JPE and APE in mm, L2 in m (Setup A only); best of each block in bold.

whereas EgoCast veers off, and it places the wearer slightly more precisely than NiLoc. All five answers are read from the same retrieved moments, which keeps them coherent in this example: the predicted walk continues from the tracked posture towards the estimated place. More examples in Appendix D show where the advantage comes from: when the head moves a lot, the activity scores tie and ReUP leads on the body and its forecast; when the head barely moves, the body errors are close and ReUP leads on the activity, which the moments retrieved by the frame help reveal.

4.3 Ablation Studies

The encoder objective. Table 3 evaluates the neighbourhoods of the motion key alone, before any decoder is trained. Tri-modal contrastive learning alone, CamFormer's objective, already gives a key whose neighbours carry behaviour, and each behaviour-guided term makes them more informative: the future term, which asks neighbours to continue alike, lowers the forecast error, and the place and action terms raise the action F1 and lower the tracking error. With the tri-modal and all behaviour-guided terms, retrieval alone is best on every column, which suggests that organising the key by where, what and what next lets the sparse head pose index behaviour rather than raw motion.

Both keys are needed. Table 4 retrains ReUP with one key only. Without the image, the localisation error more than doubles and action and body degrade; without the head motion, the body errors grow most, by about a third on tracking and a quarter on the forecast, and localisation also worsens.

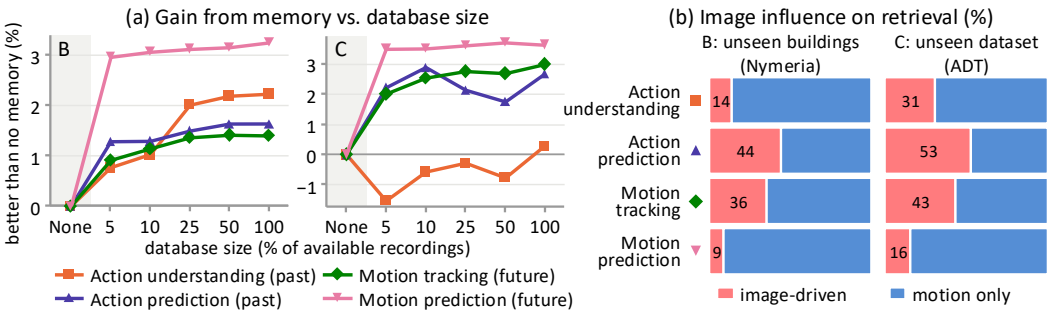


Figure 5: **Memory in unseen scenes** (Setups B and C). **(a)** Relative improvement over the same model without memory (*None*) as the share of indexed database recordings grows; points below 100% average three random subsets. **(b)** Per task, the share of the ζ highest-weight rows of $\mathbf{w}_t$ that change when the image term is removed ($\beta_t = 0$).

The place comes mainly from the image and the body from the head motion, and each key helps the other's tasks: the image narrows the candidate activities and the motion separates look-alike places.

The database in unseen scenes. Figure 5 examines how the memory contributes in unseen scenes. In (a), the model is evaluated with growing shares of its database indexed, against the same model without memory. The gain grows with the database on most metrics: with all of it, retrieval improves the behaviour metrics by 1.4–3.3% in unseen buildings and up to 3.6% on the forecast in the unseen dataset, where only the current activity barely changes; indexing as little as 1% of the recordings, however, hurts the current activity, since too few moments resemble the query. In (b), the image key changes a different share of the moments each task reads. Action prediction relies on it most (44% in B, 53% in C), plausibly because what the wearer looks at, a door, a sink, a chair, hints at what they will do next. Tracking follows (36% and 43%), consistent with the scene constraining the posture, whereas the current activity and the forecast lean on the motion key, which already carries how the wearer moves. All shares are larger on the unseen dataset, consistent with the model turning to the image when the head motion is less familiar. Each key thus contributes information of its own, and each task draws on them in its own proportion.

The tasks help one another. Table 5 compares ReUP with separate decoders, each trained on one task. Trained together, ReUP is better on every metric in both settings, most clearly on the current and next activity. The tasks share the context and the retrieval metric, so each one's supervision shapes what the others retrieve; the activity tasks gain most, as they benefit from neighbours that also agree in place and in movement, which the other tasks' losses favour. Removing one task at a time (Appendix C.2) shows the same: without the localisation loss, activity recognition drops in both Nymeria regimes, so learning where the wearer is also sharpens activity recognition. Appendix B.2, C.3 and C.1 report the auxiliary losses, a small image encoder and the computational cost.

5 Conclusion

We have shown that the head pose a pair of smart glasses already tracks, together with a single image, is enough for a holistic picture of the wearer when it is used to recall rather than to regress. ReUP indexes instrumented walkthroughs once and answers what the wearer does, how their body moves now and next, and where they are by retrieving similar moments and correcting their answers. Its key learns rich semantics from the poorest signal by contrastive learning, bound to the body and the narration recorded alongside it and shaped by priors on where behaviour happens, what it is and how it continues. With fewer than 2M parameters it matches or beats task-specific experts and language models, and its database helps even in unseen buildings and datasets. Code will be released.

AI USE STATEMENT

We used generative AI tools as a coding and writing assistant: for implementing the baseline re-implementations and the evaluation harness, for drafting and copy-editing prose, and for auditing the evaluation protocol for ground-truth leakage (§3.3). All experimental designs, hypotheses and conclusions are the authors'. All AI-assisted code was reviewed and its outputs independently verified by the authors; every number reported in this paper was produced by executing the released code. One set of labels was produced with generative AI: Aria Digital Twin has no narrations, so its activity labels are derived from captions that a vision–language model (Gemini) generated from keyframes; we report every activity result on that dataset as a pseudo-label result. Generative AI was not used to generate or alter any other experimental data. We take full responsibility for the content of this work.

ETHICS STATEMENT

This work uses two existing, publicly released egocentric datasets, Nymeria (Ma et al., 2024) and Aria Digital Twin (Pan et al., 2023), whose participants consented to capture and release under the datasets' own protocols; we collected no new human data and ran no new human-subject procedures. We use the head-pose trajectories, one egocentric frame per five-second window, the SMPL body joints and the narrations distributed with those datasets, and we release no participant-identifying material.

The privacy implications point in two directions, and we state both. Positively, the method needs no continuous video: it consumes a six-degree-of-freedom pose signal that the device's SLAM already produces and a single frame every five seconds, so a device can understand its wearer while processing far fewer images of the people and rooms the wearer looks at. Negatively, the same result shows that a head-pose trace — even one stripped of absolute position and heading, as every input is here — carries recoverable information about the wearer's activity, body posture and, together with a memory of the building, their location in it. Pose telemetry should therefore be treated as personal data rather than as harmless metadata, and the memory of walkthroughs is itself sensitive: it is a record of how particular people move through a particular home, and it should stay on the device.

REPRODUCIBILITY STATEMENT

The problem, the data, the splits and the three evaluation regimes are defined in §3.1–3.3; the encoder, the retrieval, the readouts, all loss weights and the training settings in §3.5–3.7. Every reported number was produced by executing the released code under a single declared configuration per method, and the isolation rules of §3.3 are enforced by checks in that code. The supplementary material includes the result files and a script that regenerates every table cell from them.

AUTHOR CONTRIBUTIONS

Omitted for anonymous review.

ACKNOWLEDGMENTS

Omitted for anonymous review.

REFERENCES

Relja Arandjelović, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *CVPR*, 2016.

Yujin Bae, Jaewoo Jeong, Hyeonseong Kim, and Kuk-Jin Yoon. Ego-human motion prediction with 3d-aware llm. *arXiv preprint arXiv:2607.07001*, 2026.

German Barquero, Nadine Bertsch, Manojkumar Marramreddy, Carlos Chacón, Filippo Arcadu, et al. From sparse signal to smooth motion: Real-time motion generation with rolling prediction models. In *CVPR*, 2025.

Axel Barroso-Laguna, Tommaso Cavallari, Victor Adrian Prisacariu, and Eric Brachmann. A scene is worth a thousand features: Feed-forward camera localization from a collection of image features. In *ICLR*, 2026.

Andreas Blattmann, Robin Rombach, Kaan Oktay, Jonas Müller, and Björn Ommer. Retrieval-augmented diffusion models. In *NeurIPS*, 2022.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. Improving language models by retrieving from trillions of tokens. In *ICML*, 2022.

Angela Castillo, Maria Escobar, Guillaume Jeanneret, Albert Pumarola, Pablo Arbeláez, Ali Thabet, and Artsiom Sanakoyeu. BoDiffusion: Diffusing sparse observations for full-body human motion synthesis. In *ICCV Workshops*, 2023.

Shunpeng Chen, Changwei Wang, Rongtao Xu, Xingtian Pei, Yukun Song, Jinzhou Lin, Wenhao Xu, Jingyi Zhang, Li Guo, and Shibiao Xu. SAGE: Spatial-visual adaptive graph exploration for efficient visual place recognition. In *ICLR*, 2026.

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William Cohen. MuRAG: Multimodal retrieval-augmented generator for open question answering over images and text. In *EMNLP*, 2022.

Peng Dai, Yang Zhang, Tao Liu, Zhen Fan, Tianyuan Du, Zhuo Su, Xiaozheng Zheng, and Zeming Li. HMD-Poser: On-device real-time human motion tracking from scalable sparse observations. In *CVPR*, 2024.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The EPIC-KITCHENS dataset. In *ECCV*, 2018.

Yuming Du, Robin Kips, Albert Pumarola, Sebastian Starke, Ali Thabet, and Artsiom Sanakoyeu. Avatars grow legs: Generating smooth human motion from sparse tracking inputs with diffusion model. In *CVPR*, 2023.

Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, et al. Project aria: A new tool for egocentric multi-modal AI research. *arXiv:2308.13561*, 2023.

Maria Escobar, Juanita Puentes, Cristhian Forigua, Jordi Pont-Tuset, Kevis-Kokitsi Maninis, and Pablo Arbeláez. Egocast: Forecasting egocentric human pose in the wild. In *WACV*, 2025.

Antonino Furnari and Giovanni Maria Farinella. What would you expect? Anticipating egocentric actions with rolling-unrolling LSTMs and modality attention. In *ICCV*, 2019.

Rohit Girdhar and Kristen Grauman. Anticipative video transformer. In *ICCV*, 2021.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *CVPR*, 2023.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *CVPR*, 2022.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, et al. Ego-Exo4D: Understanding skilled human activity from first- and third-person perspectives. In *CVPR*, 2024.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. Retrieval augmented language model pre-training. In *ICML*, 2020.

Tatsunori B. Hashimoto, Kelvin Guu, Yonatan Oren, and Percy Liang. A retrieve-and-edit framework for predicting structured outputs. In *NeurIPS*, 2018.

Sachini Herath, Hang Yan, and Yasutaka Furukawa. RoNIN: Robust neural inertial navigation in the wild: Benchmark, evaluations, & new methods. In *ICRA*, 2020.

Sachini Herath, David Caruso, Chen Liu, Yufan Chen, and Yasutaka Furukawa. Neural inertial localization. In *CVPR*, 2022.

Daniel Holden, Oussama Kanoun, Maksym Perepichka, and Tiberiu Popa. Learned motion matching. *ACM Transactions on Graphics (TOG)*, 39(4), 2020.

Fangzhou Hong, Vladimir Guzov, Hyo Jin Kim, Yuting Ye, Richard Newcombe, Ziwei Liu, and Lingni Ma. EgoLM: Multi-modal language model of egocentric motions. In *CVPR*, 2025.

Ziniu Hu, Ahmet Iscen, Chen Sun, Zirui Wang, Kai-Wei Chang, Yizhou Sun, Cordelia Schmid, David A. Ross, and Alireza Fathi. REVEAL: Retrieval-augmented visual-language pre-training with multi-source multimodal knowledge memory. In *CVPR*, 2023.

Chung-Ta Huang, Leopold Das, Jeffrey Zhou, Faizaan Siddique, Julia Seungjoo Baek, Serena Liu, et al. Beyond motion primitives: Behavioral activity recognition from head-mounted IMU. *arXiv:2605.27464*, 2026.

Yinghao Huang, Manuel Kaufmann, Emre Aksan, Michael J. Black, Otmar Hilliges, and Gerard Pons-Moll. Deep inertial poser: Learning to reconstruct human pose from sparse inertial measurements in real time. *ACM Transactions on Graphics (TOG)*, 37(6), 2018.

Ahmet Iscen, Alireza Fathi, and Cordelia Schmid. Improving image recognition by retrieving from web-scale image-text data. In *CVPR*, 2023.

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 *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 *ECCV*, 2024.

Jaehun Jung and Wonjun Kim. EgoGVAE: Ego-body mesh reconstruction via guided variational autoencoder. In *ECCV*, 2026.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Generalization through memorization: Nearest neighbor language models. In *ICLR*, 2020.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *NeurIPS*, 2020.

Kaleab A. Kinfu and Rene Vidal. MotionBind: Multi-modal human motion alignment for retrieval, recognition, and generation. In *NeurIPS*, 2025.

Lucas Kovar, Michael Gleicher, and Frédéric Pighin. Motion graphs. *ACM Transactions on Graphics (TOG)*, 21(3):473–482, 2002.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *NeurIPS*, 2020.

Gen Li, Yutong Chen, Yiqian Wu, Kaifeng Zhao, Marc Pollefeys, and Siyu Tang. EgoM2P: Egocentric multimodal multitask pretraining. In *ICCV*, 2025a.

Jiaman Li, Karen Liu, and Jiajun Wu. Ego-body pose estimation via ego-head pose estimation. In *CVPR*, 2023.

Zhe Li, Weihao Yuan, Yisheng He, Lingteng Qiu, Shenhao Zhu, et al. LaMP: Language-motion pretraining for motion generation, retrieval, and captioning. In *ICLR*, 2025b.

Kevin Qinghong Lin, Alex Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Zhongcong Xu, Difei Gao, Rong-Cheng Tu, Wenzhe Zhao, Weijie Kong, Chengfei Cai, Hongfa Wang, Dima Damen, Bernard Ghanem, Wei Liu, and Mike Zheng Shou. Egocentric video-language pretraining. In *NeurIPS*, 2022.

Wenxin Liu, David Caruso, Eddy Ilg, Jing Dong, Anastasios I. Mourikis, Kostas Daniilidis, Vijay Kumar, and Jakob Engel. TLIO: Tight learned inertial odometry. *IEEE Robotics and Automation Letters*, 5(4):5653–5660, 2020.

Alexander Long, Wei Yin, Thalaiyasingam Ajanthan, Vu Nguyen, Pulak Purkait, Ravi Garg, Alan Blair, Chunhua Shen, and Anton van den Hengel. Retrieval augmented classification for long-tail visual recognition. In *CVPR*, 2022.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics (TOG)*, 34(6), 2015.

David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. In *ICLR*, 2016.

Lingni Ma, Yuting Ye, Fangzhou Hong, Vladimir Guzov, Yifeng Jiang, Rowan Postyeni, Luis Pesqueira, Alexander Gamino, Vijay Baiyya, Hyo Jin Kim, Kevin Bailey, David Soriano Fosas, C. Karen Liu, Ziwei Liu, Jakob Engel, Renzo De Nardi, and Richard Newcombe. Nymeria: A massive collection of multimodal egocentric daily motion in the wild. In *ECCV*, 2024.

Julieta Martinez, Michael J. Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *CVPR*, 2017.

Marius Memmel, Jacob Berg, Bingqing Chen, Abhishek Gupta, and Jonathan Francis. STRAP: Robot sub-trajectory retrieval for augmented policy learning. In *ICLR*, 2025.

Seungwhan Moon, Andrea Madotto, Zhaojiang Lin, Aparajita Saraf, Amy Bearman, and Babak Damavandi. IMU2CLIP: Language-grounded motion sensor translation with multimodal contrastive learning. In *Findings of EMNLP*, 2023.

Kanghui Ning, Zijie Pan, Yu Liu, Yushan Jiang, James Zhang, et al. TS-RAG: Retrieval-augmented generation based time series foundation models are stronger zero-shot forecaster. In *NeurIPS*, 2025.

Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar Parkhi, Richard Newcombe, and Yuheng Carl Ren. Aria digital twin: A new benchmark dataset for egocentric 3d machine perception. In *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 *ICCV*, 2025.

Mathis Petrovich, Michael J. Black, and Gül Varol. TMR: Text-to-motion retrieval using contrastive 3D human motion synthesis. In *ICCV*, 2023.

Quinn Pfeifer, Ethan Pronovost, Paarth Shah, Khimya Khetarpal, Siddhartha Srinivasa, and Abhishek Gupta. Difference-aware retrieval policies for imitation learning. In *ICLR*, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP*, 2019.

Tsung-Ming Tai, Sofia Casarin, Andrea Pilzer, Werner Nutt, and Oswald Lanz. Action-guided attention for video action anticipation. In *ICLR*, 2026.

Shuhan Tan, Tushar Nagarajan, and Kristen Grauman. EgoDistill: Egocentric head motion distillation for efficient video understanding. In *NeurIPS*, 2023.

Guy Tevet, Brian Gordon, Amir Hertz, Amit H. Bermano, and Daniel Cohen-Or. MotionCLIP: Exposing human motion generation to CLIP space. In *ECCV*, 2022.

Minh Tran, Hongda Mao, Qingshuang Chen, and Yelin Kim. Head2Body: Body pose generation from multi-sensory head-mounted inputs. In *ICCV*, 2025.

Satoshi Tsutsui, Ruta Desai, and Karl Ridgeway. How you move your head tells what you do: Self-supervised video representation learning with egocentric cameras and IMU sensors. In *ICCV Workshop on Egocentric Perception, Interaction and Computing (EPIC)*, 2021.

Issar Tzachor, Boaz Lerner, Matan Levy, Michael Green, Tal Berkovitz Shalev, et al. EffoVPR: Effective foundation model utilization for visual place recognition. In *ICLR*, 2025.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv:1807.03748*, 2018.

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

Alexander Winkler, Jungdam Won, and Yuting Ye. QuestSim: Human motion tracking from sparse sensors with simulated avatars. In *SIGGRAPH Asia*, 2022.

Jingqiao Xiu, Fangzhou Hong, Yicong Li, Mengze Li, Wentao Wang, Sirui Han, Liang Pan, and Ziwei Liu. EgoTwin: Dreaming body and view in first person. In *ICLR*, 2026.

Haidong Xu, Guangwei Xu, Zhedong Zheng, Xiatian Zhu, Wei Ji, et al. VimoRAG: Video-based retrieval-augmented 3D motion generation for motion language models. In *NeurIPS*, 2025a.

Jilan Xu, Yifei Huang, Junlin Hou, Guo Chen, Yuejie Zhang, Rui Feng, and Weidi Xie. Retrieval-augmented egocentric video captioning. In *CVPR*, 2024.

Maxwell Xu, Jaya Narain, Gregory Darnell, Haraldur Hallgrimsson, Hyewon Jeong, et al. RelCon: Relative contrastive learning for a motion foundation model for wearable data. In *ICLR*, 2025b.

Zihui Xue, Kristen Grauman, Dima Damen, Andrew Zisserman, and Tengda Han. Seeing without pixels: Perception from camera trajectories. In *CVPR*, 2026.

Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, et al. EgoLife: Towards egocentric life assistant. In *CVPR*, 2025.

Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Richard James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. Retrieval-augmented multimodal language modeling. In *ICML*, 2023.

Hanrong Ye, Haotian Zhang, Erik Daxberger, Lin Chen, Zongyu Lin, et al. MMEgo: Towards building egocentric multimodal LLMs for video QA. In *ICLR*, 2025.

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 *CVPR*, 2025.

Xinyu Yi, Yuxiao Zhou, and Feng Xu. TransPose: Real-time 3D human translation and pose estimation with six inertial sensors. *ACM Transactions on Graphics (TOG)*, 40(4), 2021.

Ye Yuan and Kris Kitani. Ego-pose estimation and forecasting as real-time PD control. In *ICCV*, 2019.

Mingfang Zhang, Ryo Yonetani, Yifei Huang, Liangyang Ouyang, Ruicong Liu, and Yoichi Sato. Egocentric action-aware inertial localization in point clouds with vision-language guidance. In *ICCV*, 2025a.

Mingyuan Zhang, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li, Lei Yang, and Ziwei Liu. ReMoDiffuse: Retrieval-augmented motion diffusion model. In *ICCV*, 2023.

Yuwei Zhang, Kumar Ayush, Siyuan Qiao, A. Ali Heydari, Girish Narayanswamy, et al. SensorLM: Learning the language of wearable sensors. In *NeurIPS*, 2025b.

Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning video representations from large language models. In *CVPR*, 2023.

Contents of the appendix

A **Experimental setup** . . . 15
A.1 Dataset parsing . . . 15
A.2 Setups and protocol . . . 16
A.3 Baselines: re-implementation and changes . . . 17
B **Method details** . . . 18
B.1 Task-specific decoding . . . 18
B.2 Auxiliary losses . . . 19
B.3 Hyperparameters . . . 19
C **More ablations** . . . 20
C.1 Computational cost . . . 20
C.2 Leave one task out . . . 20
C.3 A lightweight image encoder . . . 21
D **More qualitative results** . . . 22
E **Limitations** . . . 22

A Experimental setup

This section gives the details that §3.2 and §3.3 summarise.

A.1 Dataset parsing

Streams. The input head pose is the glasses' closed-loop SLAM trajectory, sampled at the 100 steps of the window at 20 Hz; nothing from the motion capture, the narration or the scene annotation enters the input of any method. The body targets are the motion-capture solve as 22 SMPL joints (Loper et al., 2015) at the same 20 Hz steps: the current body over the window and the future body over the five seconds after it. Memory rows and training windows carry the same streams with their labels: body, verbs, future verbs and the end position.

Nymeria. A window is one narration window: its input starts at the narration's t_0 and spans 100 steps. Windows whose head input is incomplete are dropped from every task and every method. Action-prediction labels are the union of the verb labels of the windows of the same recording that start in $[t_0{+}5, t_0{+}10)$ s; windows with no such window are excluded from that task and counted. Consecutive narration windows start about 5 s apart, so the directly following window almost always starts just before $t_0{+}5$ s and falls outside the span: the label is in practice the narration of the window that starts about 10 s after t_0 and describes roughly $[t_0{+}10, t_0{+}15)$ s. Each recording's body clock is re-registered to the head stream by cross-correlating their velocities; tracking and forecasting are scored only where the body is complete over all 100 steps. One recording whose body solve does not follow its wearer at any time offset is excluded from the body tasks and kept for all others. Localisation targets and every memory position are expressed in one per-recording re-registration whose consensus is built only from recordings the regime allows; query recordings are registered to that consensus but never contribute to it.

ADT. ADT has 72 recordings in one apartment, with a 23-joint skeleton and captions instead of narrations; its 8 rig-calibration recordings are not scored. Each chunk's 10-Hz poses are interpolated to the same 100-step input, the head input comes from the glasses' own SLAM, never from the ground-truth trajectory, and the prediction targets are those of the next chunk. The skeleton mapping (Appendix A.1) leaves a rig mismatch of about 90 mm per joint, the same for all methods.

The image. Every preview video stores each frame's device-clock capture time, which is how I is chosen. The conventional mapping through a per-recording video offset assumes that video frame 0

coincides with the first SLAM pose, which does not hold on any recording of either dataset and pairs each pose with an image almost a second older.

ADT: mapping a foreign rig. ADT ground truth uses a 23-joint skeleton; our memory and every baseline's output are SMPL-22. The joint correspondence and a single body scale are fitted once, only on the 8 rig-calibration recordings, saved, and shared unchanged by every method. The anatomical-left chain of the ADT rig is resolved geometrically rather than by name. The scaled mapping is applied about the head joint, and the scorer aligns each recording's SLAM frame to the ground-truth frame with a 4-DoF (yaw and translation) rigid transform; models never see ground-truth device poses. The mean canonical skeletons still differ by about 90 mm per joint after mapping; this offset is identical for every method and sets a floor under all Setup C body errors.

A.2 Setups and protocol

Canonicalisation. All recordings of a building share one world frame, so an absolute position or heading in the input would carry the localisation answer. Every method, ours and every baseline, therefore canonicalises each window using only that window's *own observed* head. Φ subtracts the window-mean position, rotates by the negated window-mean heading ψ_{mean}, and emits a 12-channel sequence (gravity-aligned axes, forward direction, and velocity). Outputs are mapped back with the same observed pose, and every body quantity is expressed in the observed step-99 head frame c (origin at $\mathbf{p}_{99}$, x horizontally forward, z up).

Where localisation is scored. Localisation asks for coordinates in the query's own building, which exist only when the memory holds recordings of that building. The model produces all five outputs in every regime; Setups B and C report localisation as not applicable rather than as a large error.

Regimes and isolation. Setup A models a building that has been walked once with full instrumentation; its split into a database half and a query half is fixed once, by recording. Table 6 summarises the three regimes. Training may use any five-second span of a regime's allowed recordings; the

Table 6: **The three regimes.** In Setup A the baselines may also train on the evaluation buildings' database recordings.

	A. known building	B. unseen building	C. unseen dataset
ReUP trains on	9 training buildings	the model of A	all 12 buildings
baselines train on	the same + the eval. buildings' database	9 training buildings	all 12 buildings
memory	that building's database	9 training buildings	12 buildings
queries	that building's queries	3 eval. buildings	ADT
localisation	scored	–	–

memory and every scored window are protocol windows. The isolation covers encoders, projectors, probes, text embeddings, calibrations, normalisation statistics and priors, for every method.

Metrics. Every method is scored on its single output per window (best-of-1). Understanding and action prediction are scored by the micro-F1 of the four highest-scoring verbs of each window; four is the mean number of verbs of a training window, and no threshold is tuned. Tracking is scored by APE: in every frame, the head joint of the prediction and that of the ground truth are each subtracted from their own joints, and the error is the mean over the 21 other joints and the frames. Forecasting is scored by APE over the future frames and by JPE, the mean error of all 22 joints with prediction and ground truth in the observed head frame at the last input step, so that JPE also counts where the body goes. Localisation is scored by the distance between the estimated and the true head position at the last input step, summarised by the median within each building. Nymeria scores are averaged within each evaluation building and then over the three buildings; ADT scores are averaged over chunks.

A.3 BASELINES: RE-IMPLEMENTATION AND CHANGES

Every baseline is retrained under the protocol of §3.3; no published number or checkpoint is reused as a result. Each receives the input of ReUP, the 6-DoF head pose of the window and one egocentric frame, and nothing else.

Common changes. Every method takes the 100-step head pose of the window, canonicalised by the window's own observed head, and maps its outputs back with the same observed pose. Code paths that read ground truth at inference are removed: a floor height or pelvis centring taken from the ground-truth body, a head joint taken from motion capture in place of the glasses' pose, and calibrations fitted on evaluation data. Every fitted quantity, from normalisation statistics and calibrations to motion tokenisers, is fitted on the regime's training recordings only. The frame enters every baseline marked * as the same CLIP feature ReUP reads, through a single linear projection initialised at zero and added to the model's own hidden representation; at initialisation each such model therefore computes exactly its original output. Each baseline keeps its published architecture and recipe; training uses no evaluation data, and the last checkpoint is reported, or the one where a training-loss stopping rule ends training.

NiLoc. NiLoc (Herath et al., 2022) localises a person in a known building from inertial motion alone: a convolutional front end and a transformer turn the horizontal velocity sequence into a posterior over the cells of a building map. We feed it the head's horizontal displacements, rotated into the window's observed heading, train one model per evaluation building on its database recordings with a 1 m grid, and read out the most probable cell at the last step.

EAIL. EAIL (Zhang et al., 2025a) localises egocentric inertial motion in a point cloud of the building, aligning motion and image features with point-cloud cells under vision–language guidance. As no code is released, we re-implement its cross-modal cell classifier with a PointNet++ encoder of 1 m cells; its point cloud is the SLAM point cloud of one allowed recording per building, the only scene input any method receives. The query reads all 100 head steps and the same frame as ReUP, and the most probable cell gives the position at the last step.

CamFormer. CamFormer (Xue et al., 2026) learns what a camera trajectory reveals by aligning a transformer encoding of the trajectory with a frozen text embedding of the narration. We port it faithfully and train it per regime; to answer all five questions, its embedding replaces our motion key in our untrained retrieval readout, whose weighted neighbour positions and verbs give localisation and activity, while tracking and forecasting heads are trained on its frozen embedding ("our readout" in the tables).

EgoEgo. EgoEgo (Li et al., 2023) estimates the head pose from egocentric video and then the full body from the head with a conditional diffusion model. We keep its second stage, a head-conditioned diffusion model over the body joints, feed it the headset pose instead of an estimated head, and retrain it on 100-step windows.

EgoAllo. EgoAllo (Yi et al., 2025) samples the body from the head trajectory with a transformer diffusion model, in a frame invariant to the head's position and heading. We fine-tune its public checkpoint, trained on AMASS only, on the regime's training recordings for at most 25 epochs, stopping when the epoch-mean training loss has not improved by 1% for five epochs. Its floor height, taken from the ground-truth body, is removed, its camera-to-head calibration is refitted on training data, and the head joint's target rotation is that of the glasses composed with this calibration; with the original target, a head-orientation mismatch leaves an error floor that fine-tuning does not remove.

EgoGVAE. EgoGVAE (Jung & Kim, 2026) is a guided variational autoencoder: an autoencoder of full-body motion guides a head-to-motion network, which then samples the body in one step. It is fine-tuned from its public checkpoint, trained on AMASS only, exactly as EgoAllo, with the same calibration and head target.

EgoCast. EgoCast (Escobar et al., 2025) forecasts the body from the past head trajectory and video with two transformers, one estimating the current body and one continuing it. We port it with the frame in place of its video stream, map the 100 input steps to the next 100, and anchor the forecast at the observed head of the last step.

EgoLM and Ego3DLM. EgoLM (Hong et al., 2025) is a language model over egocentric video and motion sensors that tokenises motion with a vector-quantised autoencoder and answers in text and in motion tokens; Ego3DLM (Bae et al., 2026) adds awareness of the 3D scene and decodes past and future pose and narration together. We train both from GPT-2 medium with their instruction-tuning pipeline, without any scene input and with the head as the only motion sensor; the motion tokeniser is retrained on 20 Hz SMPL-22 motion of the training recordings, and the frame enters as the pipeline's frame embedding. They emit the current body and, for Ego3DLM, the next five seconds as motion tokens, and a linear head on the language model's last hidden state scores the 44 verbs.

The lightweight setting. For Appendix C.3, every method that reads the frame is re-run with the same small image encoder in place of CLIP, with its recipe unchanged.

B METHOD DETAILS

B.1 TASK-SPECIFIC DECODING

Every task follows the rule of §3.7: a parameter-free base read off the task's ζ highest-weight candidates, with weights $\bar{\mathbf{w}}_t$ renormalised over the candidates that carry the payload, plus a zero-initialised residual from one decoder architecture. Table 7 lists, for every task, the payload each memory row stores, how the base is formed and the slots the decoder corrects.

Table 7: **Payloads and readouts.** Each memory row stores the two keys and these five payloads. The base is formed from the task's ζ highest-weight candidates, with no parameters; the decoder adds a zero-initialised correction on the listed slots. Future: over $[t_0{+}5, t_0{+}10)$ s. $\mathbf{h}$: the query's own canonicalised head features.

task	payload	base	decoder slots
understanding	44 verbs	weighted vote	1
action prediction	44 future verbs	weighted vote	1
localisation	step-99 position	weighted mean	1
tracking	body, $100 \times 22 \times 3$	weighted mean	100 steps, $+\,\mathbf{h}$ per step
forecasting	future body	top row, re-anchored	100 steps, $+\,\mathbf{h}$ as prefix

Understanding and action prediction. The base is the weighted vote $\bar{\mathbf{v}} = \sum_{j \in \mathcal{N}_t} \bar{\mathbf{w}}_t(j)\, \mathbf{a}_j$ of the candidates' 44-dimensional verb vectors, which is a per-verb probability. It enters the verb logit together with two cues from the query's frame,

$$\boldsymbol{\ell} = \mathbf{g}_v \odot \mathrm{logit}(\bar{\mathbf{v}}) + \mathbf{g}_\chi \odot \boldsymbol{\chi} + \mathbf{g}_\pi \odot \boldsymbol{\pi} + \Delta, \tag{5}$$

where $\boldsymbol{\chi}$ are CLIP's zero-shot scores of the frame against one text prompt per verb, $\boldsymbol{\pi}$ the logits of the image probe below, $\mathbf{g}_v, \mathbf{g}_\chi, \mathbf{g}_\pi \in \mathbb{R}^{44}$ per-verb gains initialised to 1, 0 and 0, and Δ the decoder's correction, computed from the vote, the ζ candidates' own verb vectors, their weights and the context. At initialisation $\boldsymbol{\ell} = \mathrm{logit}(\bar{\mathbf{v}})$, the retrieval vote itself. The four verbs with the highest scores are the answer. Action prediction uses the same form with the candidates' future-verb vectors and its own scalars, weights and decoder.

Image probe. The probe is a two-layer network ($768 \to 512 \to 44$, with dropout) on the frame's CLIP embedding, trained with binary cross-entropy to predict the window's verbs on the training frames of the regime, before the decoder and independently of the encoder, and frozen afterwards. The probe used at query time is trained on all training buildings. While the decoder trains, every training window instead receives the prediction of a probe trained without its building (three folds of the training buildings), so the decoder never learns to trust a probe that has already seen the window; the decoder learns only how much to trust the probe for each verb, through $\mathbf{g}_\pi$.

Localisation. The base is the weighted mean of the candidates' step-99 positions in building coordinates. The decoder receives the candidates' positions *relative* to that mean and emits a three-dimensional correction, so no absolute position of the query enters and translating the memory translates the answer by the same amount. It is scored only when the memory holds the query's building (Setup A).

Tracking. The base is the weighted mean of the candidates' body sequences, each expressed in its own window's frame (the observed window-mean heading, origin at the observed step-99 head) and thereby carried into the query's. The decoder works on the 100 time steps, and every step's token also carries the query's own head features at that step. The output is rotated into the query's observed step-99 frame for scoring.

Forecasting. The base is the future of the single highest-weight candidate $j^\star$, re-anchored on the body ReUP has just tracked,

$$\mathrm{base}_{\mathrm{fcst}} = \hat{\mathbf{B}}_{99} + \left(\mathbf{B}^{F}_{j^\star} - \mathbf{B}_{j^\star,99}\right), \tag{6}$$

so the neighbour contributes what it did next rather than where it was, and no ground-truth quantity of the query enters. The ten candidates are read by the decoder anchored the same way. The decoder works on the 100 future steps, with the query's 100 steps of head features as prefix tokens. The top row is read with a straight-through estimator: its value is exactly the top row, and its gradient with respect to $\bar{\mathbf{w}}_t$ is that of the weighted mean, so the forecast loss trains the forecast's own scoring coefficients.

Why the current body is averaged and the future body is not. The current body is constrained by the observed head path: the candidates are samples scattered around one answer, and their mean is a better starting point than any one of them. The future is not observed, and different candidates continue in different directions; their mean is a path that no candidate took.

B.2 Auxiliary losses

ReUP is trained with two auxiliary losses on the static furniture around the wearer, whose outputs are not evaluated. Their targets come from the furniture annotations of the training recordings, in the frame of the observed head at the last input step: the (at most 16) nearest static objects within 3 m of the head, each with its centre, extent and category, and an occupancy grid of 24×24 cells of 0.25 m around the head, a cell being occupied when it lies in the floor footprint of one of these objects. Walls, floors, ceilings, doors, windows, wall art and objects larger than 4 m are excluded. Each page stores both, from its own recording's annotations, as two further values, and two further decoders of the architecture of §3.7, each with its own scoring distribution, read them from the candidates. The occupancy base is the weighted mean of the grids of the 32 highest-weight candidates, and its decoder corrects the grid as one vector. The box base starts from the boxes of the highest-weight candidate, matches the boxes of the other nine candidates to them by the distance of their centres on the floor, and replaces each box by the weighted mean of its matches, its category by a weighted vote; the decoder corrects every box and gives each a logit for whether it exists. The occupancy loss is the binary cross-entropy of the grid plus a soft Dice loss. The box loss matches the predicted boxes to the annotated ones (Hungarian) and penalises the ℓ_1 distance of their centres, half that of their extents and, with weight 0.2, the cross-entropy of their categories; a binary cross-entropy teaches each existence logit whether an annotated box lies within 0.5 m on the floor. Both losses are normalised like the task losses and added to their sum with weight one (§3.7). They reach the reported outputs only through what all tasks share, the adapter and the context: no reported output reads their decoders, whose parameters are therefore not counted in Table 1. Retrained without them, ReUP is slightly worse on most reported metrics and unchanged on the others at the printed precision (Table 8).

B.3 Hyperparameters

All values are fixed before training and shared by every regime; none is chosen on evaluation data (Table 9).

		Pose (mm)↓			Action↑		Loc. (m)↓
Regime	Model	APE	JPE	APE	F1	F1	L2
Known building (A)	ReUP	**128.5**	**400.0**	**176.2**	**0.520**	**0.473**	**1.415**
	without auxiliary losses	128.7	400.5	**176.2**	0.515	0.467	1.483
Unseen building (B)	ReUP	**121.6**	**383.7**	**168.5**	**0.521**	**0.469**	–
	without auxiliary losses	122.0	386.5	168.6	0.516	0.464	–
Unseen dataset (C)	ReUP	**140.0**	**467.5**	**168.9**	**0.394**	**0.404**	–
	without auxiliary losses	**140.0**	468.5	169.5	0.389	**0.404**	–

Table 8: **Auxiliary losses.** ReUP with and without its two auxiliary losses (Appendix B.2). Colours as in Table 1; the better value of each pair in bold.

Table 9: **Hyperparameters** of ReUP, with the symbols of §3. †Scoring maps, adapter and per-verb scalars.

stage	hyperparameter	value
encoder	binding weights λ_{hb}, λ_{hn}, λ_{bn}	0.5, 0.3, 0.3
	place, action and future weights λ_p, λ_a, λ_f	1.5, 0.3, 1.0
	temperature τ (place; others)	0.1; 0.07
	place radius ρ	1 m
	action Jaccard threshold η	0.5
	future positives per window ν	4
	key dimension; training epochs	128; 45
retrieval	candidates per key ξ	32
	candidates read per task ζ	10
	score scale κ; sharpness range ω	20; 10
	adapter rank (motion; image)	32; 64
training	body-loss weight λ_{body}	2
	auxiliary loss weight (Appendix B.2)	1
	optimiser; weight decay	AdamW; 10^{-3}
	peak learning rate (gates†)	3×10^{-4} (3×10^{-3})
	schedule	warm-up 2 ep., cosine to 5%
	gradient clipping; batch size; epochs	1.0; 64; 40
	weight averaging (EMA decay); seed	0.999; 0

C More ablations

C.1 Computational cost

Table 10 compares the cost of one five-second window. ReUP answers all five questions with one forward pass and the fewest learned parameters of any method but EAIL, which answers one. The diffusion-based body trackers EgoEgo and EgoAllo and the two language models need roughly two to three orders of magnitude more computation per window, each for one or a few of the outputs. The price of retrieval is a memory of the indexed walk, kept on the GPU, and a latency above that of the smallest single-task models.

C.2 Leave one task out

Table 11 retrains the decoder on all tasks but one. Removing a task's loss mainly degrades that task, as expected from task-specific outputs on a shared decoder. The exceptions show how the tasks interact. Without the localisation loss, activity recognition drops in both Nymeria regimes, so learning where the wearer is also helps to recognise what they do. Removing tracking or forecasting also lowers current-activity recognition slightly, in every regime. In the other direction, removing the understanding loss improves the forecast in Nymeria, a trade-off between the two tasks within the shared context and scoring.

Method	Tasks	Params	GFLOP	Latency (ms)
NiLoc*	1	2.03M	0.186	2.9
EAIL (+point cloud)	1	0.61M	0.0114	3.1
EgoEgo*	1	4.31M	27.1	41.2
EgoAllo*	1	50.8M	329	1017.4
EgoGVAE*	1	10.9M	2.17	78.7
EgoCast*	1	8.38M	1.37	2.6
EgoLM	2	357M	110†	309.1†
Ego3DLM, no scene	4	357M	110†	309.1†
CamFormer* (our readout)	5	8.37M	0.189	5.5
Ours	5	1.92M	0.411	22.2

Table 10: **Computational cost** per five-second window on one RTX A6000 at batch size 1: tasks answered, learned parameters, GFLOP and latency, without the shared frozen image encoder (11.3 ms, 162 GFLOP). ReUP includes retrieval over the largest Setup A memory, which takes under 100 MiB of GPU memory. †Measured with a GPT-2 medium model of the same size. * as in Table 1.

	Past		Future			Place
	Action	Motion	Action	Motion		Loc.
Model	F1↑	APE↓	F1↑	JPE↓	APE↓	L2↓
Setup A: known building (Nymeria)						
Ours	**0.520**	128.5	0.473	400.0	176.2	1.415
w/o localisation	0.496	**128.1**	0.464	401.0	175.6	3.185
w/o understanding	0.399	128.6	**0.474**	**393.5**	175.5	1.458
w/o action prediction	0.517	**128.1**	0.337	398.1	**175.4**	1.427
w/o tracking	0.511	193.8	0.473	399.5	177.1	1.392
w/o forecasting	0.510	128.5	0.470	611.7	231.5	**1.386**
Setup B: unseen building (Nymeria)						
Ours	**0.521**	**121.6**	0.469	383.7	168.5	–
w/o localisation	0.512	**121.6**	0.453	386.2	168.8	–
w/o understanding	0.459	**121.6**	**0.470**	**382.0**	**167.7**	–
w/o action prediction	0.517	121.9	0.391	383.2	168.4	–
w/o tracking	0.516	164.0	0.466	384.2	169.2	–
w/o forecasting	0.516	122.5	0.466	567.2	221.3	–
Setup C: unseen dataset (ADT)						
Ours	**0.394**	140.0	0.404	467.5	**168.9**	–
w/o localisation	0.389	141.2	0.400	**465.2**	170.5	–
w/o understanding	0.308	**139.6**	**0.410**	468.4	169.1	–
w/o action prediction	**0.394**	140.5	0.300	469.9	169.4	–
w/o tracking	0.385	165.6	0.394	472.9	169.1	–
w/o forecasting	0.390	139.9	0.402	563.6	215.1	–

Table 11: **Leave one task out.** Each *w/o* row is trained on all tasks but the one named. JPE and APE in mm, L2 in m (Setup A only); best of each block in bold.

C.3 A LIGHTWEIGHT IMAGE ENCODER

The image key and the image context of ReUP come from CLIP ViT-L/14, by far the largest part of the model. Table 12 replaces it with a small convolutional image encoder, distilled into CLIP's image-embedding space on the frames of the training recordings only, and gives the same encoder to every method that reads the frame. With the small encoder, ReUP is less accurate than with CLIP on every metric. Among the methods that read the same small encoder, including the two language models, it has the lowest pose error on the input window and over the next five seconds, the highest current and future activity scores in both Nymeria regimes, and the lowest localisation error in the known building (Table 12). Its future joint error also stays below that of EgoCast reading the full CLIP feature (Table 1).

Method	Unified?	Params	Setup A: known building						Setup B: unseen building				
			Pose (mm)↓			Action↑		Loc. (m)↓	Pose (mm)↓			Action↑	
			APE	JPE	APE	F1	F1	L2	APE	JPE	APE	F1	F1
NiLoc*	✗	2.03M	–	–	–	–	–	2.032	–	–	–	–	–
EgoEgo*	✗	4.31M	157.4	–	–	–	–	–	153.8	–	–	–	–
EgoAllo*	✗	50.8M	171.2	–	–	–	–	–	164.2	–	–	–	–
EgoGVAE*	✗	10.9M	138.9	–	–	–	–	–	132.9	–	–	–	–
EgoCast*	✗	8.38M	–	424.0	192.1	–	–	–	–	416.1	184.1	–	–
EgoLM	○	357M	190.5	–	–	0.489	–	–	181.8	–	–	0.483	–
Ego3DLM, no scene	○	357M	205.4	590.5	224.3	0.486	0.429	–	198.2	697.6	221.2	0.484	0.417
CamFormer* (our readout)	✓	8.37M	180.0	467.2	200.1	0.404	0.330	2.066	174.6	450.5	191.9	0.444	0.389
Ours	✓	1.92M	**137.1**	**409.2**	**184.2**	**0.490**	**0.444**	**1.510**	**129.1**	**394.2**	**175.3**	**0.496**	**0.437**
% vs. best baseline			+1.3%	+3.5%	+4.1%	+0.2%	+3.4%	+25.7%	+2.8%	+5.3%	+4.8%	+2.5%	+4.9%

Table 12: **Lightweight setting.** Every method reads the frame through a small image encoder distilled from CLIP ViT-L/14 on the training frames. Colours, marks, Params and * as in Table 1.

D More qualitative results

Figure 6 adds three query windows of the known building to Figure 4, laid out the same way: the input, the ground truth, the best baseline of each task and ReUP. Two windows move a lot and one hardly moves, and ReUP leads in both kinds, for different reasons.

Large motion. In (a) the wearer walks through the room, and in (b) walks and turns. Here the activity is easy to name from the head path alone: ReUP and the language model return the same verbs, with the same score. The difference lies in the motion. ReUP tracks the body at least as closely as the best body tracker, and its forecast follows the direction the wearer actually takes, with a clearly lower error than EgoCast, whose forecast stays close to where the wearer stands. Retrieval supplies moments in which a similar walk continued, so the forecast inherits a plausible continuation rather than extrapolating the last steps.

Small motion. In (c) the wearer sits and plays a board game, and the head barely moves. The motion is now easy for every method, and the body errors are close. The difference lies in the semantics: with almost no motion to go on, the language model names only part of the activity and anticipates little of what follows, whereas ReUP recognises every narrated verb, both now and in the next five seconds. The frame retrieves recorded moments at the same table, and their narrations carry the activity that a still head cannot reveal.

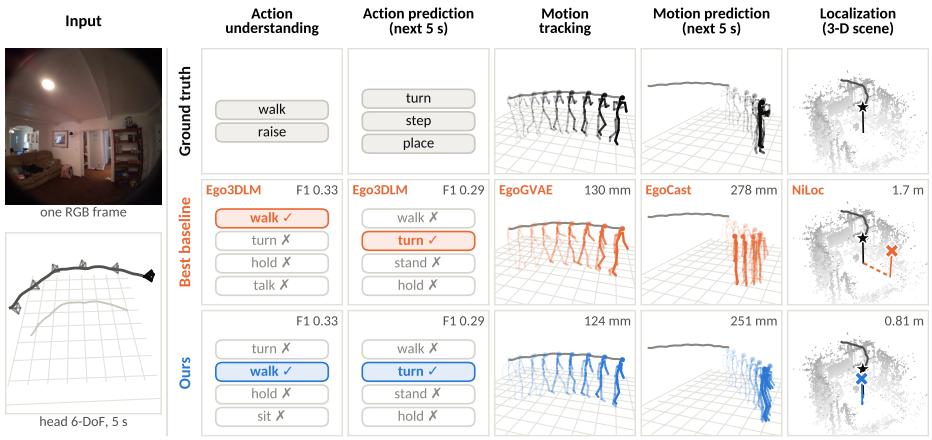
Across the three windows, ReUP also places the wearer closer to the true position than NiLoc.

E Limitations

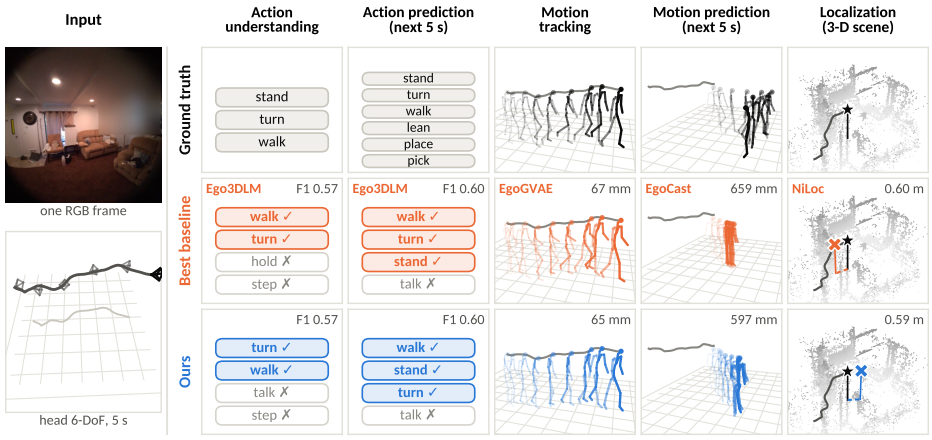
Localisation needs a memory of the same building. ReUP places the wearer by reading the positions of retrieved moments, so it can do so only when the database holds recordings of the building the wearer is in. In an unseen building or dataset the database still helps every other task, but no position can be read from it, and we therefore report localisation in the known building only.

Activity labels on the unseen dataset are pseudo-labels. ADT has no narrations, so its current and future activity labels are derived from captions with the same verb vocabulary. They are the same for every method and support a comparison between methods, but they are noisier than the narrated labels of Nymeria and the absolute scores on them should be read with that in mind.

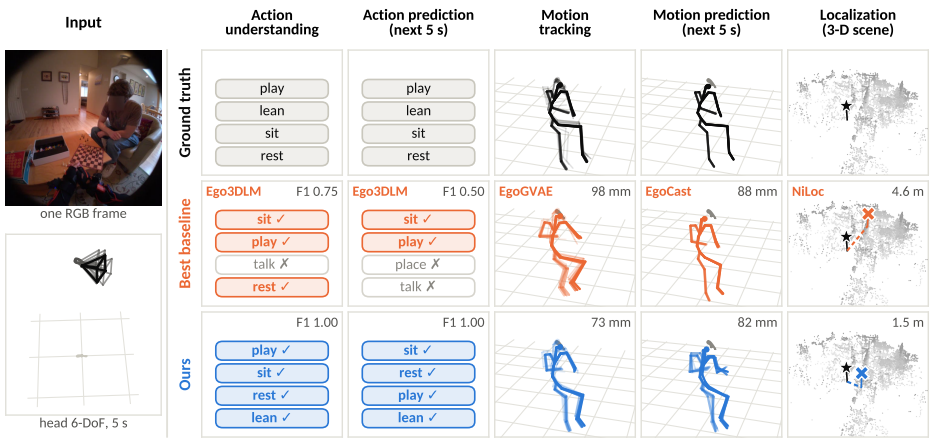
Answers are bounded by the memory. ReUP recalls answers and only learns corrections to them, so a behaviour that no recorded moment resembles, such as a rare activity or movement, can only be approximated by its nearest neighbours. Its answers therefore follow the coverage of the memory: indexing a small share of the recordings even hurts the current activity (Figure 5a), and the forecast,



(a) Large motion: walking.



(b) Large motion: walking and turning.



(c) Small motion: seated, playing a board game.

Figure 6: **More qualitative results** (Setup A), laid out as Figure 4: (a, b) large head motion, (c) small head motion.

read from the single highest-weight future, commits to one continuation rather than describing the spread of possible ones. A larger memory, or a decoder free to go beyond the retrieved answers, could relax this bound, at the cost of more recordings or more parameters.