EGOAHEAD: EGOCENTRIC HUMAN MOTION UNDERSTANDING WITH JOINT VIDEO–MOTION PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Understanding and anticipating human motion from egocentric observations is a key task for proactive AR/XR and embodied assistants. Existing egocentric behavior understanding models are typically trained only to describe the current motion, while methods that forecast future motion rely on either hints from the future recording or precomputed 3D environment scene features, limiting their applicability in real-world settings. To address this gap, we introduce EgoAhead, a unified multimodal model trained jointly for motion description and future video and pose prediction. Extending from an unified backbone Lance, EgoAhead allows multimodal generation of future video frames and poses to co-supervise a shared representation alongside language-based motion description during training. At inference, the model can output language descriptions without generating video or pose tokens, preserving inference efficiency. On the Nymeria benchmark, EgoAhead outperforms baseline models on both current-motion and future-motion descriptions despite using a smaller backbone. Our findings demonstrate that multimodal generation supervision provides an effective way to strengthen egocentric behavior understanding for wearable AI assistants.

1 INTRODUCTION

Understanding and anticipating human motion from egocentric observations is essential for proactive extended reality (XR) and embodied AI assistants. Rather than waiting for explicit user prompts, an effective assistant must infer ongoing activity and forecast immediate downstream actions. While recognizing current behavior enables context-aware support and automated logging, short-term motion anticipation allows the system to act preemptively by cueing sequential task steps, warning of imminent hazards, or pre-fetching relevant information. Achieving this requires jointly decoding user intent from visual cues alongside other wearable sensing modalities such as body kinematics and inertial measurement units (IMUs). With recent progress in multimodal large language models (MLLMs) and video generative models, there is growing interest in adapting these models to understand and predict egocentric human behavior.

For understanding current behavior, EgoLM (Hong et al., 2025) tokenize body motion and train a language model to narrate it from video and sparse sensors. Ego4o (Wang et al., 2025) introduces a part-aware motion code representation and incorporates IMU signals into a pretrained MLLM for current-motion narration. A contemporaneous work Ego3DLM (Bae et al., 2026) can output next-motion narration, though it requires precomputed semantic 3D scene features, limiting its applicability.

Concurrently, several works predict user's future behavior through video generative models. PEVA (Bai et al., 2025) predicts user's future egocentric video with past video and future body pose references. EgoTwin (Xiu et al., 2026) predicts future egocentric video and body motion based on current video, pose and future motion's text narration. EgoControl (Pallotta et al., 2026) predicts future egocentric video with past video and future poses. Although these methods achieve future video prediction, they all rely on some form of future-behavior reference, such as future poses or future-motion descriptions, which would not generally be available to a proactive wearable assistant at inference time. Moreover, these methods do not output a textual narration of future behavior.

Is it possible to leverage both the benefits of MLLMs and video generative models to learn a better representation that leads to better egocentric human behavior understanding for both current and

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.

Table 1: **Comparison of inputs and outputs.** Dashes denote inputs or outputs not supported. Future motion denotes a future body pose trajectory.

	Observed inputs		Additional conditioning			Outputs			
Method	Video	Pose/ IMU	3D scene	Future motion	Future narration	Current narration	Future narration	Future motion	Future video
EgoLM	✓	✓	–	–	–	✓	–	–	–
Ego4o	✓	✓	–	–	–	✓	–	–	–
Ego3DLM	✓	✓	✓	–	–	✓	✓	✓	–
EgoTwin	✓	✓	–	–	✓	–	–	✓	✓
PEVA	✓	–	–	✓	–	–	–	–	✓
EgoControl	✓	–	–	✓	–	–	–	–	✓
EgoAhead	✓	✓	–	–	–	✓	✓	✓	✓

future motions? In this work, we present EgoAhead, an egocentric behavior understanding model build on a Unified Multimodal Model (UMM) backbone Lance (Fu et al., 2026), jointly trained with video-motion prediction in addition to the language narration task. Importantly, while joint training with pose and video generation help improve narration, they remain optional at inference, enabling efficient deployment for behavior understanding. Table 1 summarizes the inputs and outputs of EgoAhead and existing methods.

We evaluate our approach on the Nymeria benchmark (Ma et al., 2024; DeTone et al., 2026), a standard large-scale dataset for multimodal egocentric motion understanding captured by Project Aria smart glasses (Engel et al., 2023). Despite using a lightweight 3B backbone, EgoAhead outperforms baseline models on both narration tasks. We also show that both joint video-motion prediction SFT and multimodal GRPO post-training further improves next-motion narrations.

In summary, our key contributions are:

- We introduce EgoAhead, a unified framework for egocentric motion understanding that learns current- and next-motion narration jointly with future body-motion and egocentric-video prediction. Future trajectories provide complementary physical and visual supervision during training, while narration requires only observed wearable inputs at inference.
- We extend a pretrained UMM to wearable motion by integrating structured pose and IMU representations through sensor-extended MaPE, together with flow-based future-motion and video prediction. Training combines joint predictive supervision with multimodal GRPO while preserving direct narration inference without future generation.
- On Nymeria, EgoAhead with a 3B language backbone achieves the strongest reported results among the compared methods on both current- and next-motion narration. Ablations and representation analyses further show complementary roles for predictive supervision and GRPO post-training.

2 Related Work and Background

2.1 Egocentric Motion Understanding and Anticipation

Prior motion–language modeling largely focuses on mapping isolated 3D skeleton sequences to text. Earlier formulations address bidirectional motion–text translation (Guo et al., 2022), while recent frameworks cast motion understanding and synthesis as unified sequence modeling within pretrained language models (Jiang et al., 2023; Zhou et al., 2024). However, these methods operate predominantly on isolated third-person kinematic trajectories, different from the egocentric visual context and sensor streams inherent to wearable devices.

Nymeria provides 300 hours of in-the-wild egocentric recordings with synchronized full-body motion capture and language narrations (Ma et al., 2024; DeTone et al., 2026). Leveraging these rich multimodal resources, EgoLM (Hong et al., 2025) represents body motion with discrete tokens and

combines it with egocentric video and wearable signals for narration, while Ego4o (Wang et al., 2025) introduces part-aware motion tokens and links motion reconstruction with language understanding. Together, these works demonstrate the value of combining visual and wearable motion signals to capture behavioral cues that egocentric vision alone could miss.

For anticipation, AntGPT and PALM use language-based reasoning over observed context to predict future action-category sequences (Zhao et al., 2024; Kim et al., 2024). Ego3DLM (Bae et al., 2026) jointly models past and future pose trajectories together with textual descriptions, but requires precomputed semantic 3D scene features at inference time. EgoAhead instead uses future motion and video prediction as auxiliary supervision within a unified architecture, improving motion narration from observed wearable inputs alone.

2.2 Unified Multimodal Understanding and Generation

Unified multimodal models (UMMs) combine understanding and generation within a shared architecture. These approaches unify language modeling with image or video generation, as well as additional modalities. We further discuss representative UMM architectures and their applications in Appendix A.2. LANCE (Fu et al., 2026) uses interleaved multimodal context with decoupled understanding and generation pathways, providing the UMM backbone used in EgoAhead. However, existing unified motion–language models primarily model text and body motion, while general UMMs do not explicitly incorporate the wearable pose and inertial streams central to egocentric behavior understanding.

2.3 Predictive Supervision for Egocentric Representations

Predictive objectives provide supervision beyond language annotations. As shown by Wu et al. (2026), shared task knowledge can support understanding–generation transfer. Further discussion of predictive supervision in visual representation learning, embodied control, and video reasoning appears in Appendix A.3.

Egocentric generative models explicitly couple body kinematics with visual dynamics, but typically operate in controlled synthesis settings. PEVA (Bai et al., 2025) and EgoControl (Pallotta et al., 2026) generate video from supplied future body trajectories, while EgoTwin (Xiu et al., 2026) co-generates motion and video from an initial state and an intended-action description. These controlled configurations synthesize specified behaviors instead of anticipating the wearer's next action. EgoAhead brings predictive supervision to egocentric behavior understanding by jointly forecasting future body motion and egocentric video from observed inputs alone. Together, these targets provide complementary supervision for body dynamics and their visual consequences, strengthening current- and next-motion narration.

3 Methods

EgoAhead jointly trains current- and next-motion narration with future body-motion and egocentric video prediction. The two future targets provide complementary supervision: pose captures body dynamics, while video captures changes in viewpoint and interactions with the environment. We combine this predictive training with structured wearable encoding and multimodal GRPO post-training.

3.1 Overview

Let t_0 denote the start time of the current-motion annotation and the endpoint of the observed context. The inputs $O = (X, P, I)$ comprise egocentric video, body pose, and inertial measurement unit (IMU) signals available through t_0. A text task prompt q selects current- or next-motion narration, generated using the same language head. The current reference y^- is the text of the raw annotation starting at t_0; the next reference y^+ is the text of the earliest raw annotation starting strictly after t_0 in the same recording.

In addition to narration, EgoAhead predicts future motion and video. The recorded future targets P^+ and X^+ cover a fixed prediction window, nominally $[t_0, t_0 + 2\,\mathrm{s}]$, at matching recorded times-

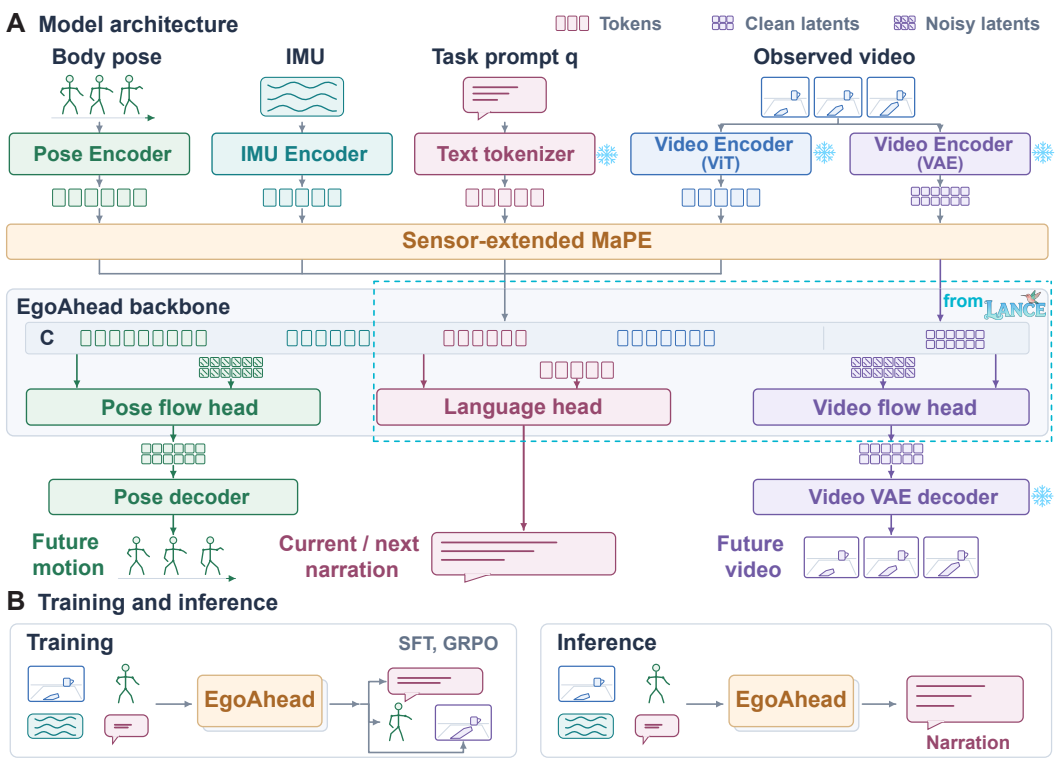

Figure 1: **EgoAhead architecture and training–inference workflow.** (a) Pose, IMU, text, and video representations enter the LANCE backbone with sensor-extended MaPE. The observed context C contains pose, IMU, prompt, and ViT tokens. Video prediction additionally includes VAE tokens from the observed clip. The hatched tokens entering the pose and video flow heads denote noisy future latents. Flow predictions yield estimates or samples of future latents for decoding. ❄ indicates modules frozen during training. (b) Training supervises narration, future motion, and future video; narration inference bypasses future generation.

tamps. Each flow pathway predicts its corresponding future modality from observed inputs. The next narration is selected by annotation order, whereas the motion and video targets cover a fixed prediction window. The selection of y^+ is therefore independent of that window's endpoint.

EgoAhead extends LANCE's unified architecture and video-generation framework (Fu et al., 2026) with sensor encoding, sensor-extended MaPE, and future-pose prediction. Figure 1A summarizes the architecture. Pose and IMU encoders, a text tokenizer, and a pretrained ViT produce tokens from the observed inputs and task prompt. Sensor-extended MaPE assigns rotary coordinates to these tokens, and we denote the resulting observed conditioning sequence by C. Video prediction additionally includes VAE tokens from the observed clip. The language head generates narration, while the flow heads predict latents decoded into future motion and video. Future-prediction losses supervise latent velocities and decoded joint motion, alongside narration supervision (Section 3.3). These losses train the shared representation used for narration. Figure 1B shows the training and inference workflow; narration inference bypasses the video VAE and future sampling.

3.2 Multimodal Encoding

Video and text. We follow LANCE's visual pathways for understanding and generation (Fu et al., 2026). The pretrained ViT encodes the observed frames as semantic visual tokens. The text tokenizer and embedding layer represent the task prompt. For video prediction, a frozen video VAE is also used to encode the observed clip as additional conditioning tokens.

Pose and inertial features. We first train a temporal pose codec with residual vector quantization (RVQ) to reconstruct body motion, then freeze it during EgoAhead training. The codec compresses the observed pose sequence into pose latent frames. The RVQ codebooks are shared across all pose latent frames, yielding one discrete code index per codebook at each latent frame. A separate trainable embedding table maps each code index to a backbone-width input token. Its entries are learned during supervised joint training, while the RVQ codebook vectors remain fixed. We add codebook-level and modality embeddings before layer normalization and dropout. An MLP embeds the inertial features. We average adjacent inertial samples to match the observed pose latent frames. Frame sampling, feature dimensions, and codec settings are provided in Appendix B.1.

Sensor-extended MaPE. Modality-aware rotary positional encoding (MaPE) assigns each visual token three rotary coordinates corresponding to time and its two-dimensional image location (Fu et al., 2026). We extend the same coordinate format to pose and IMU tokens. For pose latent-frame index k and codebook level ℓ, we assign

$$\begin{aligned} \rho^{p}_{k,\ell} &= (b_t + k,\ b_h + \ell,\ b_w), \\ \rho^{i}_{k} &= (b_t + k,\ b_h + L,\ b_w), \end{aligned} \tag{1}$$

where L is the number of pose codebooks. Pose and IMU tokens do not have a 2D image location, so the coordinates take on a different role: the first indexes the pose latent frame, the second distinguishes the pose codebooks and the IMU stream, and the third is kept fixed. The offset (b_t, b_h, b_w) anchors the sensor-token coordinates to the coordinate system used by the observed visual tokens. Pose and IMU tokens from the same latent frame therefore share the same temporal coordinate, aligning the sensor streams with the video in time rather than with individual image patches.

3.3 Joint Narration and Future Prediction

Narration supervision. Each training example contains observed inputs $\mathcal{O}$, a text task prompt q, and the corresponding reference narration $y = (y_1, \ldots, y_{|y|})$. Let $\mathcal{D}_{\text{text}}$ denote the narration training set. The language head is trained with teacher-forced cross-entropy, $\mathcal{L}_{\text{text}}$, defined in Appendix B.2. The task prompt indicates whether the requested output is current- or next-motion narration, and y provides the corresponding reference.

Future pose and video prediction. We model future body motion and egocentric video in continuous latent spaces using conditional flow matching (Lipman et al., 2023). The recorded future targets are encoded separately from the observed context C:

$$z_0^p = E_p(P^+), \qquad z_0^v = E_v(X^+). \tag{2}$$

Here, E_v is the frozen video VAE encoder. For pose, E_p forms a continuous latent by summing the selected RVQ codebook vectors at each pose latent frame and normalizing them with training-set statistics. The encoded future trajectories serve as supervision targets and are not included in the observed conditioning C.

For modality $m \in \{p, v\}$, given Gaussian noise $\epsilon_m \sim \mathcal{N}(0, I)$ and flow noise level $\tau \sim p_m(\tau)$, we construct

$$\begin{aligned} z_\tau^m &= (1-\tau) z_0^m + \tau \epsilon_m, \\ \mathcal{L}_m &= \mathbb{E}\left[\frac{1}{d_m} \left\| v_\theta^m(z_\tau^m, \tau, C) - (\epsilon_m - z_0^m) \right\|_2^2 \right], \end{aligned} \tag{3}$$

where v_θ^m predicts the flow velocity and d_m is the number of valid scalar elements in the corresponding target latent. Here, τ denotes the flow noise level, rather than physical recording time. The predicted flow field gives an estimate of the noise-free target latent:

$$\widehat{z}_0^m = z_\tau^m - \tau v_\theta^m(z_\tau^m, \tau, C), \qquad m \in \{p, v\}. \tag{4}$$

The subscript 0 denotes the noise-free endpoint of the flow, distinct from the observation time t_0.

For pose, we also minimize errors in the decoded joint trajectory. Let D_p denote the frozen pose decoder, including inverse normalization and joint-coordinate extraction. The decoded-motion loss $\mathcal{L}_{\text{dec}}$ compares joint positions, velocities, and the final predicted pose with the codec reconstruction of the recorded target. The decoder parameters remain fixed, while gradients with respect to

$\widehat{z}_0^p$ propagate through D_p to the pose-prediction pathway. The noise-level sampling schedule and expanded decoded-motion loss are given in Appendix B.3. At generation time, each flow pathway starts from Gaussian noise and integrates from $\tau = 1$ to $\tau = 0$ before decoding the resulting latent into future motion or video.

Information flow. All tasks condition on the observed multimodal inputs and task prompt. Video prediction additionally receives VAE tokens from the observed clip. During training, noisy future-target tokens enter their respective prediction branches. Attention masks prevent observed-context and narration tokens from attending to these future-target tokens, while prediction tokens can attend to the observed context. Future-prediction losses therefore supervise the shared representation without exposing future observations to the language pathway. At inference, both current- and next-motion narrations are generated directly from observed inputs.

3.4 Two-Stage Training

Training proceeds in two stages. Stage 1 jointly optimizes narration and future pose/video prediction from the same observed inputs. Stage 2 applies multimodal GRPO to sampled narration, pose, and video outputs using modality-specific rewards. Figure 1B summarizes this workflow.

Stage 1: supervised joint training. We jointly optimize narration, future pose and video prediction, and decoded-motion supervision:

$$\mathcal{L}_{\mathrm{SFT}} = \lambda_t \mathcal{L}_{\mathrm{text}} + \lambda_p \mathcal{L}_p + \lambda_v \mathcal{L}_v + \lambda_d \mathcal{L}_{\mathrm{dec}}. \tag{5}$$

Each term is normalized over its valid elements before weighting. We train the pose-token embeddings, IMU encoder, and future-pose prediction pathway together with low-rank updates to the LANCE backbone (Hu et al., 2022), while the pretrained visual encoders and pose/video codecs remain frozen. Because the prediction branches attend to shared trainable backbone states, their gradients update the same LoRA parameters used by the narration pathway.

Stage 2: multimodal GRPO. For the same observed example and task prompt, the rollout policy samples a group of G candidates for each output modality. Narration is scored using BERTScore F1 and SBERT. Pose is scored using negative decoded position, velocity, and endpoint errors, while video is scored using negative latent-prediction MSE. Current-motion narrations are compared with current-motion references and next-motion narrations with next-motion references; pose and video are evaluated against their fixed-horizon future targets.

We center each component within the group and normalize it using a fixed scale calibrated from training examples. We average the two standardized narration rewards and the three standardized pose rewards. Each modality therefore retains its own credit signal while jointly updating the shared model. We optimize a clipped policy objective with KL regularization toward the frozen stage-1 reference policy. The modality-specific advantages, likelihood ratios, and full GRPO objective are given in Appendix B.4.

3.5 Narration-Only Inference

At inference, the task instruction specifies whether the model generates current- or next-motion narration from the observed video, pose, and IMU context. The language head decodes autoregressively without running the video VAE or sampling future pose or video latents. Narration therefore incurs no future-generation step at inference.

4 Experiments

4.1 Experimental Setup

Dataset setup. We use Nymeria (Ma et al., 2024; DeTone et al., 2026), with 119,146 training windows. Each window contains 25 observed RGB frames together with five seconds of pose and inertial history ending at the same t_0. Narration metrics are evaluated separately on three held-out test subsets, with details described in Appendix C.3.

Table 2: **Current- and next-motion narration.** EgoLM and Ego3DLM results are from their original papers since public implementations are not available. Our evaluations report means across three held-out test subsets. All scores are multiplied by 100; BERTScore denotes F1. Variation across the three held-out test subsets is reported in Table 3.

Model	BLEU-1	BLEU-4	ROUGE-L	SBERT	BERTScore
Current motion					
EgoLM	30.02	6.49	26.42	59.00	19.40
Ego4o	35.24	13.00	37.31	72.04	27.51
Ego3DLM	39.66	11.07	31.95	64.58	–
Qwen3.5-4B	28.30	10.06	34.67	65.87	22.25
EgoAhead	**39.95**	**14.55**	**38.76**	**75.70**	**30.06**
Next motion					
Ego3DLM	38.39	10.39	30.04	62.06	–
Qwen3.5-4B	29.30	9.79	33.56	65.05	20.10
EgoAhead	**42.96**	**14.66**	**36.67**	**74.41**	**27.25**

EgoAhead model variants. All variants use the same observed video, pose, IMU, and task prompts and are initialized with LANCE (Fu et al., 2026) weights. During SFT, all models are trained on both current- and next-motion narration. *Narration-only* uses only the narration objective; *+video* additionally predicts future video; and *+video+pose* further adds future-pose prediction and decoded-motion supervision. Full EgoAhead *(+video+pose+GRPO)* applies multimodal GRPO post-training to the *+video+pose* model. These variants isolate the effects of predictive supervision and GRPO post-training under the same narration tasks.

Baselines. We compare EgoAhead with existing egocentric motion-narration methods, including EgoLM (Hong et al., 2025), Ego4o (Wang et al., 2025), and Ego3DLM (Bae et al., 2026), as well as Qwen3.5-4B (Qwen Team, 2026) as a general MLLM baseline. Qwen3.5-4B uses only vision and language inputs, without pose or IMU signals.

4.2 EgoAhead outperforms existing methods on motion narration

Current-motion narration. EgoAhead achieves the best performance across all current-motion narration metrics in Table 2. Among the baselines, Ego4o is the strongest overall, ranking second on four of the five metrics. EgoAhead improves over Ego4o by 4.71 BLEU-1, 3.66 SBERT, and 2.55 BERTScore F1 points, despite using a smaller 3B language backbone compared with Ego4o's LLaVA-7B (Wang et al., 2025). This shows that EgoAhead can achieve stronger egocentric motion narration with a more compact backbone.

Next-motion narration. We compare EgoAhead with Ego3DLM, which supports future motion narration, and with Qwen3.5-4B trained for the next-motion task. EgoAhead achieves the best performance across all reported metrics, showing stronger agreement with the reference next-motion descriptions.

4.3 Predictive Supervision and GRPO Improve Next-Motion Narration

Future video prediction strengthens narration. Table 3 isolates the contribution of predictive video supervision to motion narration. Adding future-video supervision improves all five next-motion metrics over Narration-only. Such generative video supervision on current-motion narration shows smaller improvement. This difference is consistent with the tasks: Current narration can draw directly from the observed motion and scene; next-motion narration must infer a continuation. Predicting the upcoming egocentric view therefore provides useful information for describing what the wearer is likely to do next.

Pose supervision changes the balance of narration metrics. Adding pose prediction raises next-motion SBERT from 70.98 to 72.43 and BERTScore recall from 17.31 to 17.45 (Appendix Table 4), indicating higher sentence-level semantic similarity and reference coverage. Before GRPO

Table 3: **Effect of joint predictive supervision.** Video and pose denote additional prediction objectives; GRPO denotes subsequent multimodal GRPO post-training. Entries are mean ± sample standard deviation across three held-out test subsets, multiplied by 100. Bold marks the best mean within each task.

Training objective	BLEU-1	BLEU-4	ROUGE-L	SBERT	BERTScore F1
Current motion					
Narration-only	46.45 ± 0.10	16.78 ± 0.05	38.70 ± 0.03	75.39 ± 0.08	29.87 ± 0.03
+video	47.04 ± 0.09	$\mathbf{16.93 \pm 0.04}$	38.45 ± 0.02	75.25 ± 0.07	29.56 ± 0.02
+video+pose	$\mathbf{47.10 \pm 0.07}$	16.74 ± 0.01	37.38 ± 0.08	$\mathbf{75.91 \pm 0.08}$	27.94 ± 0.06
Full EgoAhead (+video+pose+GRPO)	39.95 ± 0.13	14.55 ± 0.03	$\mathbf{38.76 \pm 0.01}$	75.70 ± 0.06	$\mathbf{30.06 \pm 0.05}$
Next motion					
Narration-only	35.95 ± 0.18	12.07 ± 0.06	34.91 ± 0.08	70.14 ± 0.05	24.63 ± 0.06
+video	38.45 ± 0.25	13.80 ± 0.10	35.41 ± 0.09	70.98 ± 0.01	27.13 ± 0.09
+video+pose	36.09 ± 0.08	12.33 ± 0.02	34.11 ± 0.01	72.43 ± 0.03	21.55 ± 0.05
Full EgoAhead (+video+pose+GRPO)	$\mathbf{42.96 \pm 0.20}$	$\mathbf{14.66 \pm 0.04}$	$\mathbf{36.67 \pm 0.10}$	$\mathbf{74.41 \pm 0.09}$	$\mathbf{27.25 \pm 0.23}$

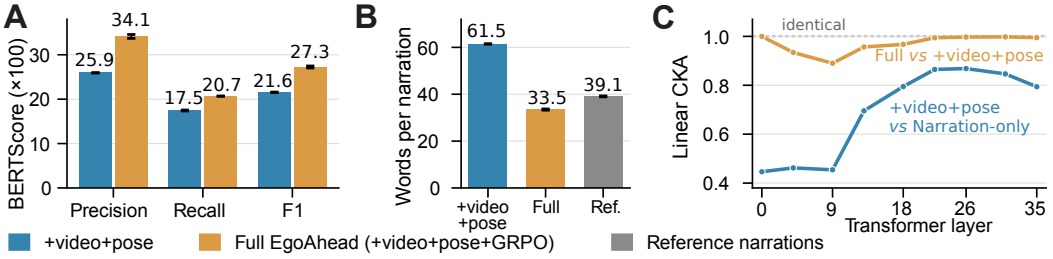

Figure 2: **How predictive supervision and GRPO post-training shape narration.** (a) Next-motion BERTScore precision, recall, and F1 for *+video+pose* and Full EgoAhead. (b) Average next-motion narration length for *+video+pose*, Full EgoAhead, and the reference narrations; GRPO reduces excess verbosity and moves the generated length closer to the reference distribution. (c) Linear CKA of final-prompt-token representations across transformer layers for *+video+pose* versus Narration-only and Full EgoAhead versus *+video+pose*. Predictive supervision produces substantially larger representation changes than GRPO.

post-training, however, the *+video* model remains stronger on BLEU, ROUGE-L, and BERTScore F1. This difference is consistent with the metrics: SBERT compares sentence-level embeddings, whereas BLEU, ROUGE-L, and BERTScore are more sensitive to lexical or token-level matching. Because LANCE is pretrained primarily for language and visual modeling, adding pose prediction introduces a new motion space, and the resulting metric pattern is consistent with a trade-off between representing the future continuation and matching the reference wording.

Multimodal GRPO post-training refines narration. Relative to *+video+pose*, Full EgoAhead improves all five next-motion metrics, including the unrewarded BLEU-1, BLEU-4, and ROUGE-L metrics. While BERTScore F1 and SBERT are used as narration rewards during GRPO, the gains on lexical metrics indicate that the improvement extends beyond the reward functions themselves. As shown in Section 4.4, GRPO largely preserves the representation learned through predictive supervision while producing more concise and reference-aligned narration.

4.4 How Predictive Supervision and GRPO Shape Narration

The ablations above suggest complementary roles for predictive supervision and multimodal GRPO post-training. We examine these effects through the generated narrations and the representations formed inside the language backbone.

GRPO produces more concise, reference-aligned narration. Relative to *+video+pose*, Full EgoAhead increases next-motion BERTScore precision from 25.93 to 34.13, recall from 17.45 to 20.68, and F1 from 21.55 to 27.25 (Figure 2A). At the same time, average narration length decreases from 61.46 to 33.45 words, compared with 39.09 words for the references (Figure 2B). Post-training therefore improves reference agreement while substantially reducing verbosity.

Predictive supervision changes the backbone representation far more than GRPO. Figure 2C compares final-prompt-token representations across transformer layers using linear CKA, where higher values indicate more similar representation geometry. Comparing *+video+pose* with *Narration-only*, CKA is approximately 0.45 through layer 9 and rises to 0.85–0.87 in the late layers. In contrast, Full EgoAhead remains much closer to *+video+pose*: its minimum CKA is 0.89 and similarity is at least 0.99 from layer 22 onward. This pattern is consistent with predictive supervision substantially reshaping the multimodal representation, while GRPO largely preserves that representation as it improves narration. See Appendix C.5 for additional analysis.

4.5 Future Motion and Video Prediction

Although future motion and video prediction only serve as auxiliary supervision for EgoAhead's narration tasks, in Appendix C.6 we also evaluate these predictions on held-out test data to ensure they learn meaningful future visual and motion structure from observed inputs. For motion, the full EgoAhead model predicts upcoming body trajectories directly from observed inputs, obtaining 276.56 mm MPJPE over held-out test subsets. For video, future-video supervision reduces MSE from 0.3014 to 0.0600 and increases PSNR from 5.26 to 12.43 dB; adding pose supervision further improves performance and Full EgoAhead retains comparable video quality after GRPO post-training. A qualitative prediction example is provided in Appendix C.7.

5 Discussions

We present EgoAhead, a unified multimodal model trained jointly for motion narration and future video and pose prediction. On Nymeria, EgoAhead achieves the strongest results among the compared methods on both current- and next-motion narration with a compact 3B backbone. Our ablations further show complementary benefits from predictive supervision and multimodal GRPO post-training.

Limitations. EgoAhead has several limitations. First, although the future video and pose pathways learn meaningful predictions, their generation quality can be further improved. Since these pathways mainly serve as auxiliary supervision, stronger motion or video generative models may provide richer training signals for narration. Second, EgoAhead uses a relatively compact 3B language backbone, within the scale of recent smartphone-oriented models, but deployment on more resource-constrained wearable devices would still benefit from more efficient visual encoding, model compression, and lower-precision inference.

Future work. A natural extension is to incorporate additional wearable modalities into egocentric motion understanding. Audio, gaze and other physiological signals could provide complementary cues about the environment, user attention, and physical state beyond vision and body motion. These modalities could provide richer context for understanding and anticipating wearer behavior. Another promising direction is personalized adaptation at test time. The current model processes each observation using fixed model parameters and a local context window, while a deployed wearable assistant could continuously accumulate information about an individual user. This could be achieved through test-time adaptation or training that updates the model from new observations, or through an agentic memory mechanism that stores and retrieves relevant past activities, interactions, and behavioral patterns without modifying model parameters. Both approaches could capture recurring routines and user-specific behavior over longer time scales, enabling increasingly personalized motion understanding and anticipation.

AI USE STATEMENT

We used generative AI tools to assist with research ideation and execution, including brainstorming research directions and executing experiments. We also used these tools to support literature retrieval and discovery, and to aid paper writing. We did not use generative AI to generate synthetic datasets, formulate or prove mathematical claims, or replace the authors' interpretation of experimental results. Relevant literature and citations suggested through AI-assisted retrieval were checked against the original sources, and all AI-assisted research and writing were reviewed and verified by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses the publicly released Nymeria dataset and does not involve new human-subject data collection. Nymeria contains egocentric recordings of daily human behavior and therefore raises important privacy considerations. According to the dataset documentation, consent for recording and data use was obtained from participants and property owners, the data were collected and stored in an anonymized manner, and faces and license plates were blurred before further processing and analysis. We use only the publicly released, de-identified data and do not attempt to identify or re-identify participants. The video, pose, and inertial data used in this work are released for non-commercial use under the CC BY-NC 4.0 license (Ma et al., 2024), and our use follows these terms.

More broadly, egocentric behavior understanding and anticipation can reveal sensitive information about a user's activities, routines, and environment. Deployment in wearable assistants could therefore introduce risks of unwanted monitoring or behavioral profiling. Future applications should prioritize privacy-preserving sensing, data minimization, transparent user consent and control, and careful evaluation across users and environments.

REFERENCES

Yujin Bae, Jaewoo Jeong, Hyeonseong Kim, and Kuk-Jin Yoon. Ego-human motion prediction with 3d-aware llm. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2026.

Yutong Bai, Danny Tran, Amir Bar, Yann LeCun, Trevor Darrell, and Jitendra Malik. Whole-body conditioned egocentric video prediction. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 164375–164418. Curran Associates, Inc., 2025. doi: 10.52202/085713-5482. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/f0552f14388d95b19740dee809f5cad1-Paper-Conference.pdf.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mido Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=QaCCuDfBk2. Featured Certification.

Junhao Cheng, Liang Hou, Xin Tao, and Jing Liao. Video-as-answer: Predict and generate next video event with joint-grpo. *arXiv preprint arXiv:2511.16669*, 2025.

Daniel DeTone, Federica Bogo, Eric-Tuan Le, Duncan Frost, Julian Straub, Yawar Siddiqui, Yuting Ye, Jakob Engel, Richard Newcombe, and Lingni Ma. Nymeriaplus: Enriching nymeria dataset with additional annotations and data, 2026. URL https://arxiv.org/abs/2603.18496.

Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brighid Meredith, Cheng Peng, Chris Sweeney, Cole Wilson, Dan Barnes, Daniel DeTone, David Caruso, Derek Valleroy, Dinesh Ginjupalli, Duncan Frost, Edward Miller, Elias Mueggler, Evgeniy Oleinik, Fan Zhang, Guruprasad Somasundaram, Gustavo Solaira, Harry Lanaras, Henry Howard-Jenkins, Huixuan Tang, Hyo Jin Kim, Jaime Rivera, Ji Luo, Jing Dong, Julian Straub, Kevin Bailey, Kevin Eckenhoff, Lingni Ma, Luis Pesqueira, Mark Schwesinger, Maurizio Monge, Nan Yang, Nick Charron, Nikhil Raina, Omkar Parkhi, Peter Borschowa, Pierre Moulon, Prince Gupta, Raul Mur-Artal, Robbie Pennington,

Sachin Kulkarni, Sagar Miglani, Santosh Gondi, Saransh Solanki, Sean Diener, Shangyi Cheng, Simon Green, Steve Saarinen, Suvam Patra, Tassos Mourikis, Thomas Whelan, Tripti Singh, Vasileios Balntas, Vijay Baiyya, Wilson Dreewes, Xiaqing Pan, Yang Lou, Yipu Zhao, Yusuf Mansour, Yuyang Zou, Zhaoyang Lv, Zijian Wang, Mingfei Yan, Carl Ren, Renzo De Nardi, and Richard Newcombe. Project aria: A new tool for egocentric multi-modal ai research, 2023. URL https://arxiv.org/abs/2308.13561.

Fengyi Fu, Mengqi Huang, Shaojin Wu, Yunsheng Jiang, Yufei Huo, Hao Li, Yinghang Song, Fei Ding, Jianzhu Guo, Qian He, Zheren Fu, Zhendong Mao, and Yongdong Zhang. Lance: Unified multimodal modeling by multi-task synergy, 2026. URL https://arxiv.org/abs/2605.18678.

Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. Tm2t: Stochastic andnbsp;tokenized modeling fornbsp;thenbsp;reciprocal generation ofnbsp;3d human motions andnbsp;texts. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV*, pp. 580–597, Berlin, Heidelberg, 2022. Springer-Verlag. ISBN 978-3-031-19832-8. doi: 10.1007/978-3-031-19833-5_34. URL https://doi.org/10.1007/978-3-031-19833-5_34.

Zhongbin Guo, Jiahao Xie, Dongling Xiao, Qianle Wang, Ruiqi Lu, Xiaomin He, Wanxuan Sun, and Cheng Yang. Generation as auxiliary supervision: Enhancing visual understanding at zero inference overhead via decoupled embedding prediction, 2026. URL https://arxiv.org/abs/2608.12209.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5344–5354, 2025. doi: 10.1109/CVPR52734.2025.00503.

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*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 20067–20079. Curran Associates, Inc., 2023. doi: 10.52202/075280-0880. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/3fbf0c1ea0716c03dea93bb6be78dd6f-Paper-Conference.pdf.

Tianxiang Jiang, Linquan Wu, Sheng Xia, Songze Li, Ziang Yan, Haoyu Yang, Yu Qiao, and Yi Wang. Imagine before you predict: Interleaved latent visual reasoning for video event prediction. *arXiv preprint arXiv:2606.05769*, 2026.

Sanghwan Kim, Daoji Huang, Yongqin Xian, Otmar Hilliges, Luc Van Gool, and Xi Wang. Palm: Predicting actions through language models. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXXXII*, pp. 140–158, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-73006-1. doi: 10.1007/978-3-031-73007-8_9. URL https://doi.org/10.1007/978-3-031-73007-8_9.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

Shuang Li, Yihuai Gao, Dorsa Sadigh, and Shuran Song. Unified Video Action Model. In *Proceedings of Robotics: Science and Systems*, LosAngeles, CA, USA, June 2025. doi: 10.15607/RSS.2025.XXI.074.

Zekun Li, Sizhe An, Chengcheng Tang, Chuan Guo, Ivan Shugurov, Linguang Zhang, Amy Zhao, Srinath Sridhar, Lingling Tao, and Abhay Mittal. Llamo: Scaling pretrained language models for unified motion understanding and generation with continuous autoregressive tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

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 *the 18th European Conference on Computer Vision (ECCV)*, 2024. URL https://arxiv.org/abs/2406.09905.

Enrico Pallotta, Sina Mokhtarzadeh Azar, Lars Doorenbos, Serdar Ozsoy, Umar Iqbal, and Juergen Gall. Egocontrol: Controllable egocentric video generation via 3d full-body poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4269–4279, June 2026.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin (eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040/.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410/.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Zihan Su, Hongyang Wei, Kangrui Cen, Yong Wang, Guanhua Chen, Chun Yuan, and Xiangxiang Chu. Generation enhances understanding in unified multimodal models via multi-representation generation. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=uYaNE2Cys2.

Haonan Wang, Hongfu Liu, Xiangyan Liu, Chao Du, Kenji Kawaguchi, Ye Wang, and Tianyu Pang. Fostering video reasoning via next-event prediction. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=8nUgzuvskm.

Jian Wang, Rishabh Dabral, Diogo Luvizon, Zhe Cao, Lingjie Liu, Thabo Beeler, and Christian Theobalt. Ego4o: Egocentric Human Motion Capture and Understanding from Multi-Modal Input . In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22668–22679, Los Alamitos, CA, USA, June 2025. IEEE Computer Society. doi: 10.1109/CVPR52734.2025.02111. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52734.2025.02111.

Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, and Ping Luo. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12966–12977, June 2025.

Penghao Wu, Haiwen Diao, Weichen Fan, Lewei Lu, Dahua Lin, and Ziwei Liu. Uncovering understanding-generation synergy in native unified multimodal models: From representation, task to system, 2026. URL https://arxiv.org/abs/2609.01607.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=o6Ynz6OIQ6.

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 *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=QFJkvv3zMi.

Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, et al. Dancegrpo: Unleashing grpo on visual generation. *arXiv preprint arXiv:2505.07818*, 2025.

Tianyuan Yuan, Zibin Dong, Yicheng Liu, and Hang Zhao. Fast-wam: Do world action models need test-time future imagination? *arXiv preprint arXiv:2603.16666*, 2026. URL https://arxiv.org/abs/2603.16666.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.

Qi Zhao, Shijie Wang, Ce Zhang, Changcheng Fu, Minh Quan Do, Nakul Agarwal, Kwonjoon Lee, and Chen Sun. Antgpt: Can large language models help long-term action anticipation from videos? In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 56677–56697, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/faa8be9311811ba7c36fa1ceec13b862-Paper-Conference.pdf.

Chunting Zhou, LILI YU, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=SI2hI0frk6.

Zixiang Zhou, Yu Wan, and Baoyuan Wang. Avatargpt: All-in-one framework for motion understanding planning generation and beyond. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1357–1366, June 2024.

Bingfan Zhu, Biao Jiang, Sunyi Wang, SHIXIANG TANG, Tao Chen, Linjie Luo, Youyi Zheng, and Xin Chen. MotionGPT3: Human motion as a second modality. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Ha075JDMZR.

A Additional Related Work

A.1 Motion–Language Modeling

Earlier formulations address bidirectional motion–text translation (Guo et al., 2022), while recent frameworks cast motion understanding and synthesis as unified sequence modeling within pretrained language models (Jiang et al., 2023; Zhou et al., 2024). For instance, MotionGPT (Jiang et al., 2023) treats 3D motion tokens as an additional modality for captioning and temporal generation, and AvatarGPT (Zhou et al., 2024) instruction-tunes an LLM across understanding and planning tasks.

A.2 Unified Multimodal Architectures

Transfusion (Zhou et al., 2025) couples autoregressive language modeling with continuous visual denoising, Show-o (Xie et al., 2025) combines language modeling with discrete diffusion, and Janus (Wu et al., 2025) separates visual pathways for understanding and generation to reduce representational interference. More recently, unified modeling has extended beyond images to human motion. MotionGPT3 (Zhu et al., 2026) uses coupled language and motion streams with diffusion-based motion generation, while LLaMo (Li et al., 2026) integrates continuous motion generation into an autoregressive language model through flow matching. These methods connect language with body motion, but do not jointly model how that motion changes the wearer's first-person visual experience.

A.3 Predictive Supervision Across Tasks

In visual representation learning, V-JEPA (Bardes et al., 2024) predicts masked spatiotemporal features, while UniMRG (Su et al., 2026) and GAS (Guo et al., 2026) use auxiliary generation objectives to improve visual understanding without additional inference computation. Similar principles appear in embodied control: UVA (Li et al., 2025) and Fast-WAM (Yuan et al., 2026) co-train on future video and action trajectories, allowing downstream policies to benefit from predictive dynamics without the latency of runtime video generation.

In video reasoning, forecasting downstream events has emerged as an effective supervisory signal for learning temporal dynamics. NEP (Wang et al., 2026) forecasts textual summaries of future events from past frames, VANS (Cheng et al., 2025) couples vision–language reasoning with video diffusion via reinforcement learning, and Future-L1 (Jiang et al., 2026) interleaves latent visual spans within language reasoning traces. EgoAhead shares similar ideas while also introducing body motion prediction and egocentric perception.

B Additional Method Details

B.1 Sampling and Encoding Settings

Video sampling and future alignment. The observed video is a single 25-frame clip covering approximately two seconds before t_0. All 25 frames are encoded by the pretrained ViT for every task. For future-video prediction, the same clip is also encoded by the frozen video VAE. The future video target contains 25 recorded frames spanning the nominal interval $[t_0, t_0 + 2\,\mathrm{s}]$. Their timestamps lie on the 15-fps camera clock, giving an effective sampling rate of approximately 12 fps. The future pose target uses exactly the same recorded timestamps.

Pose codec and observed inputs. The pose codec uses temporal convolutions with one temporal downsampling stage (`down_t` $= 1$) and therefore supports variable-length sequences. It is pretrained on 50-frame pose sequences sampled at 10 Hz, which are encoded into 25 latent frames. The observed pose input uses this 50-frame format over the preceding five-second context. Each pose frame contains 69 features: 22 pelvis-relative 3D joint positions and the pelvis displacement from the preceding frame. Six 512-entry RVQ codebooks are shared across latent frames, yielding six code indices per latent frame.

Future pose encoding. Future trajectories are encoded independently rather than cropped from a five-second sequence. Each 25-frame future target is passed directly through the frozen variable-length codec, producing 12 latent frames. Decoding these latents yields 24 pose frames, which are linearly interpolated to the original 25-frame grid for supervision and evaluation. The frozen codec reconstructs the future windows with 30.08 mm MPJPE, indicating that the substantially larger forecasting error primarily reflects uncertainty in predicting future motion rather than pose-codec reconstruction error.

Inertial features and observed coordinates. A two-layer MLP embeds the nine-dimensional inertial features: three acceleration channels and a six-dimensional orientation representation. For observed conditioning, we average adjacent inertial samples to match the 25 observed pose latent frames, giving six pose tokens and one IMU token per observed latent frame. For these observed conditioning tokens, the coordinates in Equation (1) use $k = 0, \ldots, 24$, $\ell = 0, \ldots, L-1$, and $L = 6$.

B.2 Narration Objective

The language head is trained with teacher-forced cross-entropy:

$$\mathcal{L}_{\text{text}} = -\mathbb{E}_{(\mathcal{O},q,y)\sim\mathcal{D}_{\text{text}}} \left[\frac{1}{|y|} \sum_{j=1}^{|y|} \log p_\theta(y_j \mid y_{<j}, \mathcal{O}, q) \right], \tag{6}$$

where j indexes narration tokens, $|y|$ is the target length, and $y_{<j}$ denotes the preceding tokens.

B.3 Flow Sampling and Decoded Motion Supervision

Noise-level sampling. We sample τ from a shifted logistic-normal distribution: for $a \sim \mathcal{N}(0,1)$ and $u = \text{sigmoid}(a)$,

$$\tau = \frac{s_m u}{1 + (s_m - 1)u}, \tag{7}$$

with modality-specific shift $s_m > 0$.

Decoded motion loss. We supervise decoded pose trajectories using a fixed decoding pipeline D_p. After inverse latent normalization, the frozen codec decodes each sequence of 12 pose latent frames into 24 frames. We linearly interpolate both predicted and reference reconstructions to a common 25-frame index grid and extract the 22 joint coordinates. We compute

$$\widehat{J} = D_p(\widehat{z}_0^p), \qquad \overline{J} = D_p(z_0^p), \tag{8}$$

where $\widehat{J}, \overline{J} \in \mathbb{R}^{F \times N_j \times 3}$ are the predicted joint trajectory and the codec reconstruction of the recorded target, respectively. We optimize

$$\begin{aligned} \mathcal{L}_{\text{dec}} = \alpha_{\text{pos}} &\|\widehat{J} - \overline{J}\|_{2,\text{avg}} \\ &+ \alpha_{\text{vel}} \|\Delta\widehat{J} - \Delta\overline{J}\|_{2,\text{avg}} \\ &+ \alpha_{\text{end}} \|\widehat{J}_F - \overline{J}_F\|_{2,\text{avg}}. \end{aligned} \tag{9}$$

Here, $F = 25$ and $N_j = 22$. Velocities here denote frame-to-frame joint differences scaled by the nominal frame rate, $\Delta J_f = 12(J_f - J_{f-1})$, applied identically to the predicted and reference trajectories. The norms average joint-wise Euclidean distances over the 25 frames for position, 24 intervals for temporal differences, and the final frame for the endpoint term. For supervised joint training, we set $(\alpha_{\text{pos}}, \alpha_{\text{vel}}, \alpha_{\text{end}}) = (1.0, 0.5, 0.25)$. The decoder parameters remain fixed, while gradients with respect to $\widehat{z}_0^p$ propagate through D_p to the pose-prediction pathway.

B.4 Multimodal GRPO Objective

For the same observed example and task prompt, the rollout policy samples a group of G candidates for each output modality; we use $G = 4$ in our experiments. Let $i = 1, \ldots, G$ index candidates and $m \in \{n, p, v\}$ denote narration, pose, and video.

For reward component $r \in \mathcal{K}_m$, let r_{ir}^m be the score of candidate i and μ_r^m its mean within the candidate group. We center each component within the group and normalize it using a fixed scale σ_r^m calibrated from training examples. The modality-specific advantage is

$$A_i^m = \sum_{r \in \mathcal{K}_m} \omega_r^m \frac{r_{ir}^m - \mu_r^m}{\sigma_r^m}, \qquad \omega_r^m \geq 0, \qquad \sum_{r \in \mathcal{K}_m} \omega_r^m = 1. \tag{10}$$

We average the two standardized narration rewards and the three standardized pose rewards. Each modality therefore retains its own credit signal while jointly updating the shared model.

For narration, each generated token is treated as one policy action. For pose and video, each stochastic latent transition in the flow sampler is treated as one policy action. Let $u_{i,s}^m$ denote the sampled token or transition at step s, and $h_{i,s}^m$ its preceding token or sampling history. We define

$$R_{i,s}^m(\theta) = \frac{\pi_\theta^m \left(u_{i,s}^m \mid h_{i,s}^m, \mathcal{O}, q\right)}{\pi_{\text{old}}^m \left(u_{i,s}^m \mid h_{i,s}^m, \mathcal{O}, q\right)}, \tag{11}$$

where narration uses token probabilities and pose/video use Gaussian transition densities from stochastic flow sampling (Xue et al., 2025).

Writing $\ell_\delta(R, A) = \min\{RA, \text{clip}(R, 1-\delta, 1+\delta)A\}$, we optimize

$$\mathcal{J}_{\text{GRPO}} = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G} \sum_{m \in \{n,p,v\}} \frac{\eta_m}{Z_m} \sum_{s \in \mathcal{S}_i^m} \ell_\delta\left(R_{i,s}^m, A_i^m\right)\right] - \beta \mathcal{D}_{\text{KL}}, \tag{12}$$

where η_m balances the three output branches. For narration, $\mathcal{S}_i^n$ contains the generated token positions and $Z_n = 128$ is the fixed normalization used during training. For pose and video, one stochastic transition is sampled for each update, so $Z_p = Z_v = 1$. The KL term regularizes each branch toward the frozen stage-1 reference policy. The rollout policy π_{old} is the policy that generated the current candidate group.

C Additional Experimental Details

C.1 Narration Target Construction

Each example is anchored at the start time t_0 of an atomic-action annotation, whose text provides the current-motion reference y^-. The next-motion reference y^+ is the text of the earliest annotation starting strictly after t_0 in the same recording. Observed pose and inertial inputs cover $[t_0 - 5\,\text{s}, t_0]$, while aligned future pose and video targets span the nominal interval $[t_0, t_0 + 2\,\text{s}]$.

C.2 Training Settings

Supervised fine-tuning. All EgoAhead variants are initialized from LANCE (Fu et al., 2026) and use both current- and next-motion narration targets. We use rank-64 LoRA (Hu et al., 2022) with scaling 128, dropout 0.05, a learning rate of 2×10^{-5}, and 500 warmup steps. The variants differ in whether future-video and future-pose objectives supplement narration. The joint motion–video variant also uses decoded pose losses, as specified in Section 3.3.

Multimodal post-training. We run 612 GRPO updates with a learning rate of 10^{-6}, each paired with current- and next-motion prompts. Each update samples four candidates per output modality for the same observed example. Narration sampling uses temperature 0.9 and a maximum of 128 new tokens. Pose and video use 12 stochastic flow steps with noise strength 0.3, with one transition sampled per update for each flow pathway. The policy-ratio clipping parameter is 0.2 and the reference-KL coefficient is 0.01. Rewards and modality-specific advantages follow Section 3.4 (Shao et al., 2024; Xue et al., 2025). We update the backbone LoRA weights and future-pose interface while keeping the observed-input encoders and pose/video codecs frozen. Both SFT and post-training use current- and next-motion references.

C.3 Evaluation Metrics

Each narration task is evaluated on three heldout test subsets, each containing 29,438 examples. For EgoLM (Hong et al., 2025) and Ego3DLM (Bae et al., 2026), there are publicly available implementations and evaluation splits, so we report the scores published in their original papers for comparison. For EgoAhead, Ego4o and Qwen3.5-4B, we used our own splits. Narration metrics compare a generated description with its reference: BLEU and ROUGE-L measure lexical agreement, while SBERT and BERTScore use pretrained text representations. All narration scores are multiplied by 100. Higher narration scores, PSNR, and global SSIM are better; lower motion error and video MSE are better.

BLEU-1 and BLEU-4. BLEU uses modified n-gram precision, which clips each generated n-gram count by its count in the reference (Papineni et al., 2002). Counts are accumulated across the evaluation corpus. BLEU-1 uses unigram precision; BLEU-4 combines precisions for one- through four-word sequences using their geometric mean. Both include a brevity penalty for descriptions that are shorter than their references. Writing p_n for corpus-level modified precision, c for total candidate length, and r for total reference length, the standard form is

$$\mathrm{BLEU}_N = \exp\left(\min\left(0, 1 - \frac{r}{c}\right)\right) \exp\left(\frac{1}{N}\sum_{n=1}^{N} \log p_n\right), \qquad N \in \{1, 4\}. \tag{13}$$

BLEU-1 reflects word-level agreement, while BLEU-4 places additional weight on matching phrases and local word order. Differences in wording can therefore change BLEU even when descriptions are semantically close.

ROUGE-L. ROUGE-L compares the longest common subsequence between a generated narration and its reference (Lin, 2004). A subsequence preserves word order but allows intervening words. Its precision and recall normalize the matched length by the generated and reference lengths, respectively, and are combined into a score. ROUGE-L therefore captures ordered word overlap without requiring every matched word to belong to a contiguous phrase. It complements the shorter-range phrase matching measured by BLEU-4.

SBERT similarity. Sentence-BERT represents each description with a sentence embedding and compares embeddings using cosine similarity (Reimers & Gurevych, 2019). We use all-mpnet-base-v2. For a generated narration $\widehat{y}$, its reference y, and sentence encoder e, the score is

$$\mathrm{SBERT}(\widehat{y}, y) = \frac{e(\widehat{y})^{\mathsf{T}} e(y)}{\|e(\widehat{y})\|_2 \|e(y)\|_2}. \tag{14}$$

The score measures whole-description similarity and can credit paraphrases that have limited exact word overlap. It complements metrics that examine individual words or token matches.

BERTScore precision, recall, and F1. BERTScore matches contextual token embeddings between the generated and reference descriptions (Zhang et al., 2020). Precision averages how well each generated token matches a reference token; recall averages how well each reference token is covered by the generated description. F1 combines the two. Precision thus characterizes reference-aligned expression, while recall characterizes reference coverage.

We use RoBERTa-large with inverse document frequency (IDF) weights estimated from the reference corpus and official baseline rescaling. IDF gives more weight to tokens that occur in fewer reference narrations. Before rescaling, the per-example F1 is the harmonic mean of that example's precision and recall. The official transformation rescales each component using its corresponding baseline:

$$\widetilde{s} = \frac{s - b_s}{1 - b_s}, \tag{15}$$

where s is the raw precision, recall, or F1 and b_s is its baseline. We aggregate the resulting per-example scores. Accordingly, the reported mean F1 is not obtained by taking the harmonic mean of the reported mean precision and recall. Together, the components distinguish closely matched expression from coverage of the reference description.

Table 4: **BERTScore component analysis.** Entries are mean ± sample standard deviation across held-out test subsets and are multiplied by 100. F1 is aggregated from per-example scores.

	Current motion			Next motion		
Model	Precision	Recall	F1	Precision	Recall	F1
Narration-only	36.87 ± 0.09	$\mathbf{23.20 \pm 0.03}$	29.87 ± 0.03	33.23 ± 0.02	16.51 ± 0.11	24.63 ± 0.06
+video	36.25 ± 0.06	23.18 ± 0.07	29.56 ± 0.02	$\mathbf{37.50 \pm 0.05}$	17.31 ± 0.13	27.13 ± 0.09
+video+pose	33.65 ± 0.15	22.44 ± 0.04	27.94 ± 0.06	25.93 ± 0.05	17.45 ± 0.11	21.55 ± 0.05
Full EgoAhead (+video+pose+GRPO)	$\mathbf{38.45 \pm 0.12}$	22.00 ± 0.08	$\mathbf{30.06 \pm 0.05}$	34.13 ± 0.44	$\mathbf{20.68 \pm 0.04}$	$\mathbf{27.25 \pm 0.23}$

Motion prediction. Pelvis-relative mean per-joint position error (MPJPE) averages the Euclidean distance between predicted and recorded 3D joints at matching timestamps. Each pose is expressed relative to its own pelvis. For F evaluated frames and J joints, let $\widehat{p}_{f,j}$ and $p_{f,j}$ denote these relative coordinates. We compute

$$\mathrm{MPJPE} = \frac{1}{FJ} \sum_{f=1}^{F} \sum_{j=1}^{J} \|\widehat{p}_{f,j} - p_{f,j}\|_2. \tag{16}$$

The reported error is in millimeters. Pelvis-relative evaluation measures body configuration without global root translation; we apply no scale fitting or Procrustes alignment. The forecasting evaluation uses recorded joints, distinct from the codec-reconstructed targets used by the decoded training loss.

Video prediction. MSE averages squared differences between corresponding predicted and reference pixel values. PSNR expresses reconstruction fidelity on a logarithmic scale. For a pixel data range L,

$$\mathrm{PSNR} = 10 \log_{10} \left(\frac{L^2}{\mathrm{MSE}} \right). \tag{17}$$

Lower MSE and higher PSNR indicate closer pixel-level agreement. Because PSNR is nonlinear, averaging individual PSNR scores is different from transforming an already averaged MSE.

Structural similarity compares image luminance, contrast, and correlation (Wang et al., 2004). Our global SSIM uses whole-frame means, variances, and covariance in this comparison. It summarizes overall structural agreement, while MSE and PSNR summarize pixel fidelity. We retain the term *global SSIM* to distinguish this implementation from spatially averaged local-window SSIM. All video measurements compare predictions with the recorded continuation at matching timestamps and aligned image resolution.

C.4 BERTScore Component Analysis

Table 4 reports BERTScore precision, recall, and F1 for the EgoAhead training variants.

C.5 Representation Similarity Analysis

Protocol. We compare pairs of model variants with linear centered kernel alignment (CKA) (Kornblith et al., 2019). Both variants receive the same 4,000 held-out windows drawn from 117 recordings with the same task prompt, and we collect hidden states of the 36-layer LANCE backbone at layers 0, 4, 9, 13, 18, 22, 26, 31, and 35 for the final prompt token and at layers 0–31 for pooled sensor tokens (layer 0 denotes the output of the first block). For the final prompt token we take the hidden state at the last prompt position, the final conditioning token before narration begins. For pooled pose and pooled IMU we mean-pool hidden states over the observed pose or IMU token positions. Each representation gives an $N \times d$ matrix per variant, and we report $\mathrm{CKA}(X_l, Y_l)$ with a linear kernel after column centering. As a chance reference, CKA between X_l and a row-shuffled Y_l is below 0.006 at every layer. To test whether similarity is driven by recording identity, we also subtract each recording's mean feature vector from both variants before computing CKA; the resulting values follow the same layer profile.

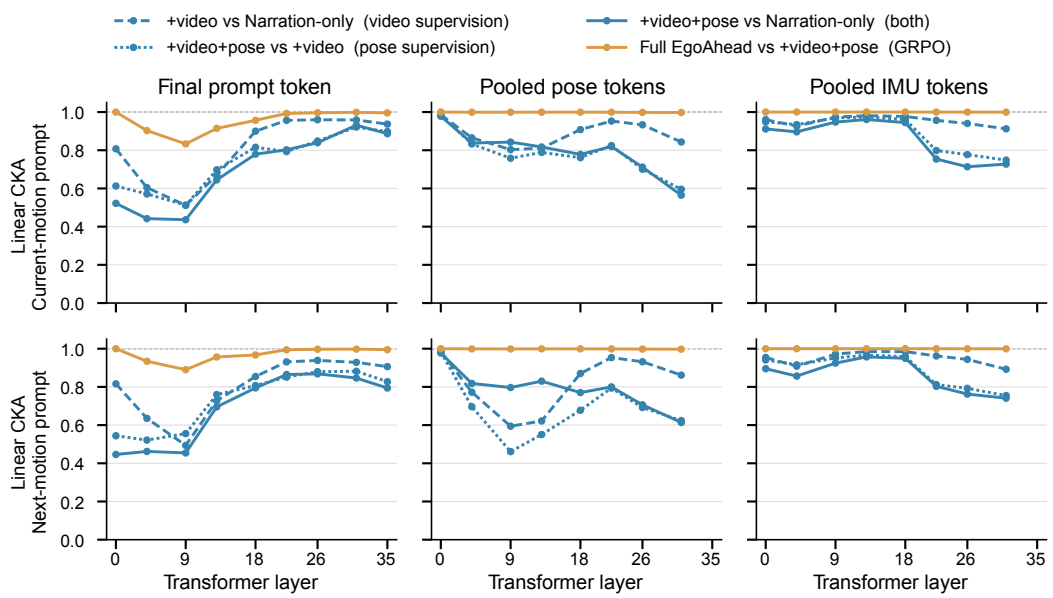

Figure 3: **Per-layer linear CKA for all training steps.** Rows: current-motion and next-motion prompt. Columns: final prompt token, pooled pose tokens, pooled IMU tokens. Blue lines compare supervised variants (dashed: video supervision only; dotted: pose supervision added to *+video*; solid: both, as in the main text); orange compares Full EgoAhead with *+video+pose*. Values are computed on the same 4,000 held-out windows per comparison; a shuffled control gives < 0.01. Pooled sensor tokens are reported for layers 0–31.

Results. Figure 3 extends the main-text analysis across both task prompts, intermediate supervision stages, and pooled sensor-token representations. The current- and next-motion prompts produce nearly identical final-token CKA profiles, indicating that the observed layer-wise pattern is not specific to one narration query. Adding video supervision changes the intermediate representation, and adding pose supervision on top of the video-supervised model produces a further change; the jointly supervised model shows the largest departure from Narration-only in these layers.

The sensor-token analysis reveals a related pattern. Adding pose supervision produces substantial changes in pooled pose-token representations in the later evaluated layers, whereas Full EgoAhead and *+video+pose* retain CKA of 0.99–1.00 for pooled pose and IMU representations. Together with the recording-centered control, these results support the main-text conclusion that predictive supervision produces substantially larger representational changes than GRPO post-training.

C.6 Future Motion and Video Prediction

Future motion prediction. Full EgoAhead achieves 276.56 mm pelvis-relative MPJPE, slightly improving over 276.74 mm for *+video+pose*. On average displacement error (ADE) over the first 2 s of the predicted trajectory, EgoAhead obtains 361.2 mm, while Ego3DLM reports 205.9 mm. As future-motion prediction in EgoAhead primarily serves as auxiliary supervision for narration, stronger dedicated motion forecasting remains an important direction for future work.

Additionally, flow model also provides a lower-error point prediction before completing the stochastic sampling trajectory. On two 128-window held-out test subsets, this clean-latent estimate obtains 232.6 and 227.2 mm MPJPE, compared with 278.6 and 272.0 mm for the corresponding completed samples. Thus, when a point forecast is sufficient, EgoAhead can produce a more accurate prediction without running the full flow sampler.

Table 5: **Future-video prediction on heldout test data.** Lower MSE and higher PSNR/global SSIM are better.

Model	MSE	PSNR (dB)	Global SSIM
Narration-only	0.3014	5.26	−0.108
+video	0.0600	12.43	0.342
+video+pose	**0.0589**	**12.53**	**0.355**
Full EgoAhead	0.0599	12.46	0.354

Future video prediction. We evaluate video prediction on heldout test data. Each target contains 25 future frames. Generation uses 30 sampling steps, noise seed 2026, and guidance scale 1, and the generated frames are aligned to the reference resolution before scoring.

Adding future-video supervision produces a large improvement over the otherwise matched Narration-only model: MSE decreases from 0.3014 to 0.0600, PSNR increases from 5.26 to 12.43 dB, and global SSIM increases from -0.108 to 0.342. Adding future-pose supervision slightly improves all three metrics, with *+video+pose* reaching 0.0589 MSE, 12.53 dB PSNR, and 0.355 global SSIM. Full EgoAhead retains similar future-video quality after GRPO post-training, obtaining 0.0599 MSE, 12.46 dB PSNR, and 0.354 global SSIM. These results show that future-video supervision learns meaningful visual dynamics, pose prediction provides additional complementary supervision, and GRPO post-training largely preserves the learned video forecasting capability.

C.7 Qualitative Example of Future Prediction

Figure 4 shows future-video predictions and corresponding next-motion narrations for one held-out test example. Without future-video supervision, *Narration-only* produces a visual continuation inconsistent with the observed bedroom scene, with a PSNR of 7.63 dB. Adding future-video supervision substantially improves this example: *+video* preserves the bedroom environment and coarse egocentric structure, increasing PSNR to 13.55 dB. Adding pose supervision further increases PSNR to 14.36 dB, while Full EgoAhead obtains 14.40 dB after GRPO post-training. For this sequence, the later training stages therefore preserve the scene-consistent future-video prediction established by video supervision. This qualitative behavior is consistent with the aggregate results in Table 5, where future-video supervision provides the largest improvement, pose supervision further improves the video metrics, and Full EgoAhead retains comparable future-video quality after GRPO post-training.

The corresponding next-motion narrations are shown in Table 6. In this example, *Narration-only* correctly describes placing the clothing on the bed but adds an unsupported continuation in which the wearer walks toward the door and hallway. The *+video* variant focuses on the relevant folding action but refers to the clothing as a blanket. The *+video+pose* variant captures the clothing placement but adds an unsupported continuation around the bed. Full EgoAhead captures both folding the clothes and placing them on the bed without introducing an additional action absent from the reference. Thus, for this example, post-training produces a more concise narration that more closely matches the annotated future behavior. This example complements the corpus-level narration results in the main paper, rather than serving as independent evidence for a general performance trend.

Table 6: **Next-motion narrations for the qualitative example in Figure 4.**

Model	Next-motion narration
Reference	C is standing in the bedroom while holding the clothes with both hands, then lays the clothes on the bed as she slightly bends toward the bed and folds the clothes with both hands.
Narration-only	C is standing in the bedroom while holding a piece of clothing with both hands, then she puts the piece of clothing on the bed with both hands. C then turns to her right, walks towards the door, then turns to her left and walks towards the hallway.
+video	C is standing in the bedroom while folding the blanket with both hands.
+video+pose	C is standing in the bedroom while holding a piece of clothing with both of her hands, and then she puts the piece of clothing on the bed with both of her hands. C then turns to her left and walks towards the other side of the bed.
Full EgoAhead	**C is standing in the bedroom while folding the clothes with both of her hands, and then she puts the folded clothes on the bed with both of her hands.**

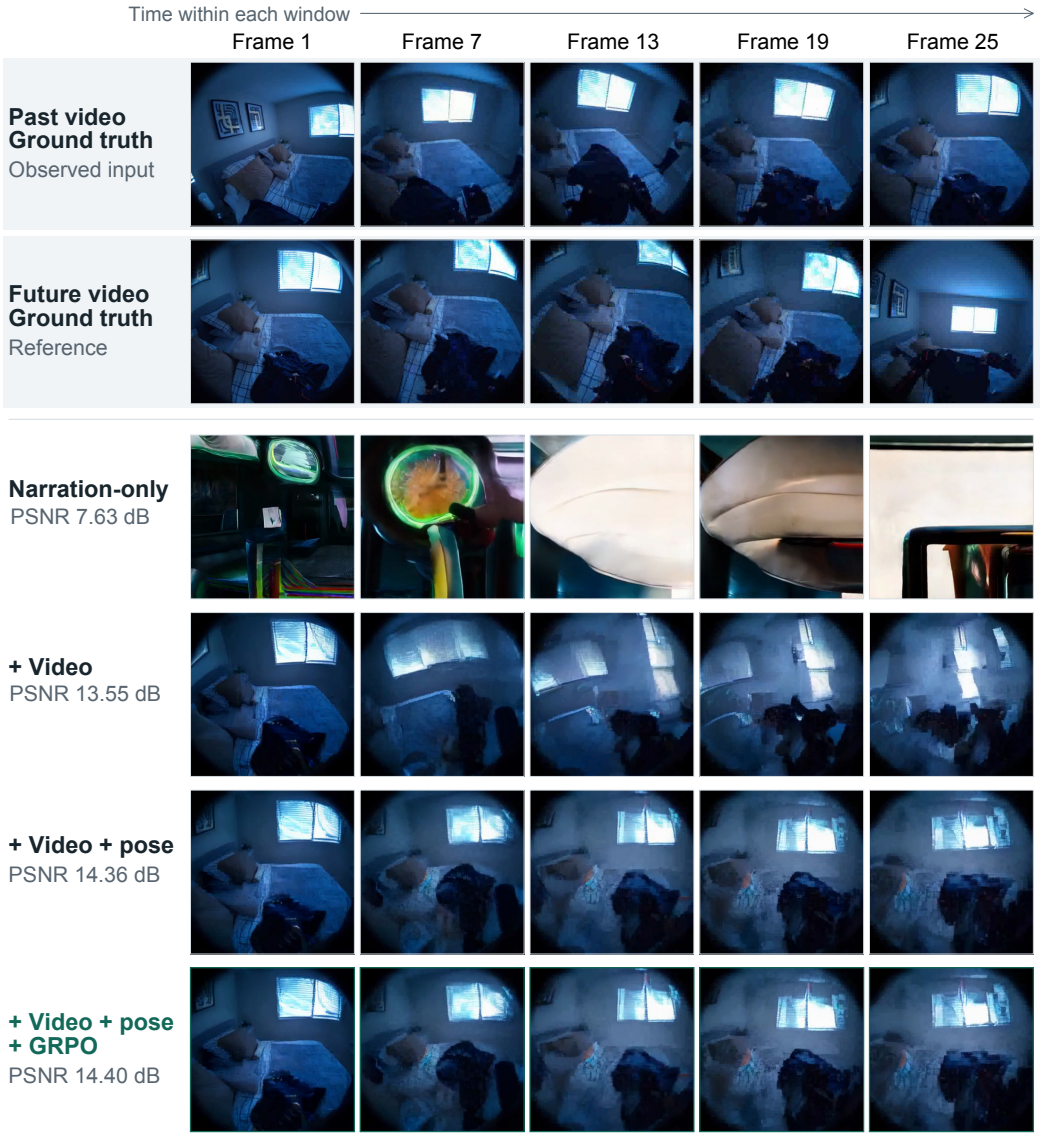

Figure 4: **Qualitative future-video prediction.** Five frames are shown from the observed past and future windows, with the future reference and predictions evaluated at matching frame indices. The corresponding PSNR for each predicted sequence is shown on the left. Future-video supervision substantially improves scene-consistent prediction, while pose supervision and GRPO post-training preserve and further refine the predicted continuation.