GazeInterpreter: Parsing Eye Gaze to Generate Eye-Body-Coordinated Narrations

Qing Chang[1*], **Zhiming Hu**[1,2*†]

[1]The Hong Kong University of Science and Technology (Guangzhou), China
[2]The Hong Kong University of Science and Technology, China
changqinghs@163.com, zhiminghu@hkust-gz.edu.cn

Abstract

Comprehensively interpreting human behavior is a core challenge in human-aware artificial intelligence. However, prior works typically focused on body behavior, neglecting the crucial role of eye gaze and its synergy with body motion. We present *GazeInterpreter* – a novel large language model-based (LLM-based) approach that parses eye gaze data to generate eye-body-coordinated narrations. Specifically, our method features 1) a *symbolic gaze parser* that translates raw gaze signals into symbolic gaze events; 2) a *hierarchical structure* that first uses an LLM to generate eye gaze narration at semantic level and then integrates gaze with body motion within the same observation window to produce integrated narration; and 3) a *self-correcting loop* that iteratively refines the *modality match*, *temporal coherence*, and *completeness* of the integrated narration. This hierarchical and iterative processing can effectively align physical values and semantic text in the temporal and spatial domains. We validated the effectiveness of our eye-body-coordinated narrations on the text-driven motion generation task in the large-scale Nymeria benchmark. Moreover, we report significant performance improvements for the sample downstream tasks of action anticipation and behavior summarization. Taken together, these results reveal the significant potential of parsing eye gaze to interpret human behavior and open up a new direction for human behavior understanding.

Code — zhiminghu.net/chang26_gazeinterpreter

Introduction

Comprehensively interpreting human behavior is a foundational challenge in human-aware artificial intelligence. A robust understanding of human behavior underpins many critical applications, including human motion generation (Zhang et al. 2022; Yan et al. 2024), human intention recognition (Hu et al. 2022; Belardinelli et al. 2022), proactive action anticipation (Hu et al. 2024c,a), and efficient behavior summarization (Zhang et al. 2025). With the recent success of large language models (LLMs), researchers have begun leveraging them to generate natural language explanations of human behavior, making notable progress in interpreting body motion (Jiang et al. 2023; Chen et al. 2024).

However, a crucial modality is largely overlooked in this new paradigm: human eye gaze. As a powerful non-verbal cue, eye gaze is not only a direct window into human intention (Hu et al. 2022; Belardinelli et al. 2022) but is also intrinsically correlated with body motion (Hu et al. 2024b; Sidenmark and Gellersen 2019). For instance, when a person intends to grasp a cup, their eyes typically fixate on it just before or during the arm's movement. Despite this, prior works (Kong and Fu 2022; Chang et al. 2025) have predominantly focused on interpreting body behavior in isolation, neglecting the potential information conveyed by eye gaze and its synergistic relationship with body motion. This omission results in a significant gap, leaving interpretations of human behavior incomplete and less robust.

To fill this gap, we present *GazeInterpreter* – a novel LLM-based framework that parses eye gaze data to generate comprehensive eye-body-coordinated narrations, integrating human potential intentions and fine-grained motion features. Specifically, we first introduce a *symbolic gaze parser* that converts raw eye gaze signals into symbolic gaze events that serve as input to the LLM. Then, a *hierarchical structure* composed of multiple LLMs generates eye gaze narrations from gaze events, and these narrations are semantically integrated with body motion narrations, producing eye-body-coordinated narrations. To ensure alignment between narrations and reality, we further present a *self-correcting loop* that iteratively refines the *continuity* of the gaze narrations and the *modality match*, *temporal coherence*, and *completeness* of the integrated narrations by an LLM-driven evaluation-feedback mechanism.

We extensively evaluate our method for text-driven motion generation on the large-scale, in-the-wild Nymeria benchmark (Ma et al. 2024). Experiments show that our eye-body-coordinated narrations lead to superior performance for generating motions. Furthermore, we validate the effectiveness of our narrations on representative downstream tasks of action anticipation and behavior summarization, demonstrating the advantages of our approach in comprehensive human behavior interpretation.

The contributions of our work are three-fold:

- We propose *GazeInterpreter* – a novel LLM-based framework for interpreting gaze behavior that features a

*These authors contributed equally.
†Corresponding author.

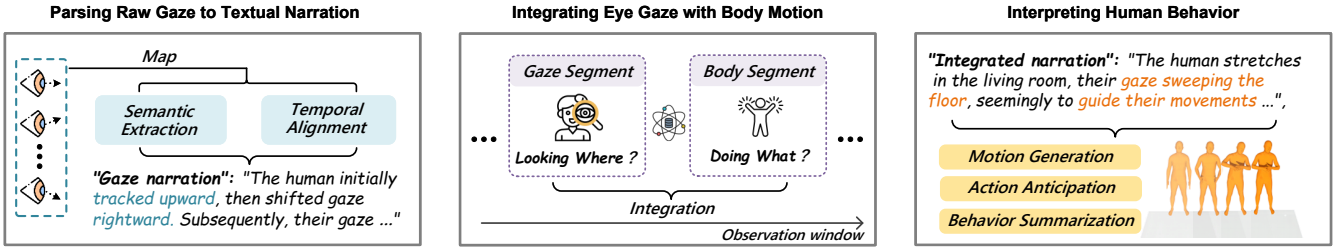

Figure 1: GazeInterpreter is a novel method for interpreting gaze behavior that first parses raw gaze signals to a textual narration (Left) and then integrates gaze with body motion to generate eye-body-coordinated narration (Middle). The integrated narration provides a superior basis for interpreting human behavior (Right).

symbolic gaze parser to convert raw gaze signal, a hierarchical structure to integrate gaze with body motion, and a self-correcting loop for refinement.

- We conduct extensive experiments on the large-scale Nymeria benchmark and demonstrate that our narrations can significantly improve performance in text-driven motion generation.
- We demonstrate the broad applicability of our method by significantly enhancing performance on the sample downstream tasks of action anticipation and behavior summarization.

Related Work

Human Behavior Interpreting

Comprehensively interpreting human behavior is a crucial topic in the areas of human-centered computing and human-aware artificial intelligence. Earlier works typically focused on rule-based techniques to interpret human behavior (Pons-Moll, Fleet, and Rosenhahn 2014; Delmas et al. 2022). For example, Pons-Moll et al. manually set rules on *joint distance*, *articulation angle*, and *relative position* to describe relationships between body parts (Pons-Moll, Fleet, and Rosenhahn 2014) while Delmas et al. defined a set of rules on the 3D keypoints to generate the description of full-body pose (Delmas et al. 2022). Recently, with the great success of LLMs, many researchers have started to interpret human behavior directly using LLMs (Jiang et al. 2023; Chen et al. 2024). Specifically, Jiang et al. proposed a uniform motion-language model to link human body motion with natural language (Jiang et al. 2023) while Chen et al. projected human body motion and video data into the linguistic space to generate better narrations of human body behavior (Chen et al. 2024). However, prior works mainly focused on interpreting human body behavior, neglecting the human eye gaze.

Eye Gaze Analysis

Analyzing human gaze behavior has been a popular topic in the area of vision research for decades (Yarbus 1967; Itti, Koch, and Niebur 1998; Chen, Jiang, and Zhao 2024). Prior works typically focused on hand-crafted statistical indicators of gaze behavior such as gaze velocity (Hu et al. 2019; Kothari et al. 2020), gaze distribution (Sitzmann et al. 2018; Hu et al. 2019; Hadnett-Hunter et al. 2019), saccade amplitude (Hu et al. 2019; Hadnett-Hunter et al. 2019; Hu et al. 2020), fixation number (Coutrot, Hsiao, and Chan 2018; Hu et al. 2022), fixation duration (Hadnett-Hunter et al. 2019; Kothari et al. 2020; Hu et al. 2022), fixation dispersion (Coutrot, Hsiao, and Chan 2018; Hu et al. 2022), and fixation clusters (Coutrot, Hsiao, and Chan 2018; Hu et al. 2021). However, these hand-crafted statistical indicators are cumbersome to compute, provide only limited information, and lack explainability. In stark contrast, in this work, we directly convert raw gaze signals into natural language narrations that can effectively improve the understanding of eye gaze behavior.

Eye-Body Coordination

Many researchers have investigated the coordination of human eye gaze and their body movements. Specifically, Kothari et al. discovered the coordinated patterns of eye and head movements during daily activities (Kothari et al. 2020) while Hu et al. examined virtual environments and revealed the eye-head coordination during free-viewing and task-oriented situations (Hu et al. 2019, 2020, 2021, 2025). Sidenmark et al. analyzed the gaze shift process in virtual reality and identified the coordination of eye, head, and torso movements (Sidenmark and Gellersen 2019). Hu et al. further revealed the strong correlation between eye gaze and human full-body movements in various daily activities (Hu et al. 2024b). Inspired by the close link between eye gaze and body movements, in this work, we integrate eye gaze narrations with body motion narrations to improve human behavior understanding.

Method

Problem Formulation

We define interpreting eye gaze behavior as the task of generating comprehensive eye-body-coordinated narrations from eye gaze signals and body motion narrations. For a given time segment i, we define the input as a multi-modal tuple $O_i = (S_i^{\mathrm{g}}, S_i^{\mathrm{m}})$, where $S_i^{\mathrm{g}} \in \mathbb{R}^{N_g \times 2}$ represents the raw, continuous gaze signal, composed of N_g samples of yaw and pitch coordinates. Concurrently, $S_i^{\mathrm{m}} = (m_1, \ldots, m_{N_m})$ is a sequence of N_m discrete atomic body motion narrations from a predefined set $\mathcal{M}$. Our objective is to learn a mapping $\mathcal{F} : (\mathbb{R}^{N_g \times 2}, \mathcal{M}^{N_m}) \rightarrow \mathcal{T}$, which generates a textual narration $\hat{T}_i \in \mathcal{T}$ that not only captures overt eye-body actions but also facilitates human behavior understanding.

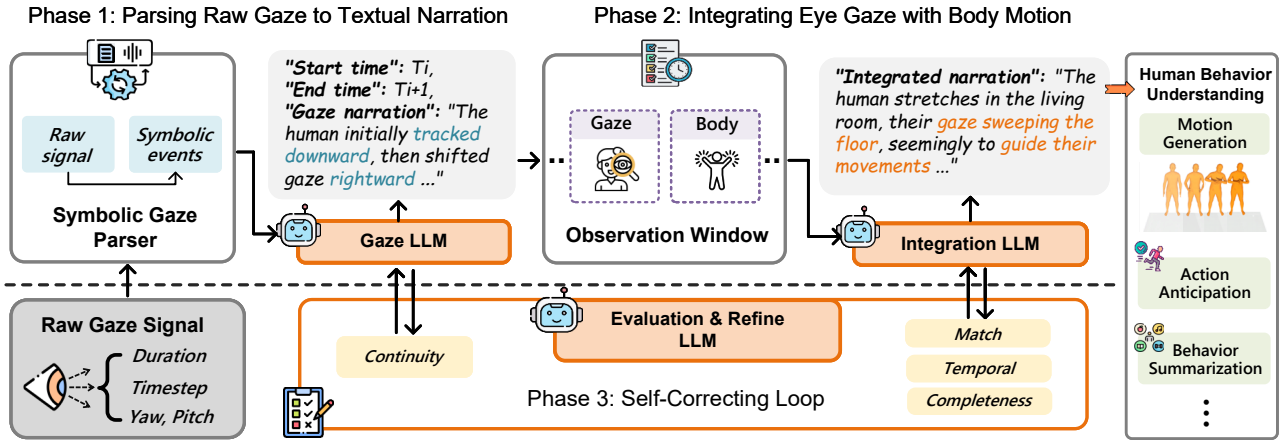

Figure 2: Architecture of GazeInterpreter. Our method first uses a symbolic gaze parser to convert raw gaze signals into symbolic events and then employs an LLM to generate textual narration (Phase 1), which is further integrated with body motion narration in an observation window to produce the eye-body-coordinated narration (Phase 2). A self-correcting loop is applied to iteratively refine both the gaze and integrated narration, ensuring feature alignment across different dimensions (Phase 3).

A fundamental challenge in this task is to bridge the semantic gap between low-level, numerical gaze sensor data and the high-level, abstract concepts needed for comprehensive behavior understanding. The core research question is: *How can we reliably transform noisy, continuous gaze signals S^g into a structured, semantic representation that facilitates faithful multi-modal reasoning?*

While recent approaches have explored direct numerical-to-text conversion using LLMs (Wang et al. 2024) or employed auxiliary techniques like contrastive learning (Chang and Tong 2024), such end-to-end methods remain susceptible to common pitfalls, including factual hallucination and a failure to robustly ground the generated text in the source signals (Xu et al. 2025). To address these limitations, we propose to first abstract low-level raw gaze data into an intermediate symbolic vocabulary of gaze events, providing a reliable basis for motion fusion. This numerical-to-symbolic conversion contributes to the generation of more robust behavioral descriptions.

Framework Overview

To address the formulated problem, we propose *GazeInterpreter* – a novel hierarchical coarse-to-fine LLM-based method that transforms raw low-level gaze signals into high-level, interpretable narrations through three phases: *parsing raw gaze to textual narration*, *integrating eye gaze with body motion*, and *self-correcting loop*, as depicted in Fig. 2. In the first phase, our method parses the raw gaze signal into a sequence of structured symbolic gaze events and then applies an LLM to generate texture narrations of these structured gaze events. Benefiting from the results of the first phase, we can align eye gaze narrations with atomic body motion narrations in the semantic space during the next phase and synthesize eye-body-coordinated narrations. In the third phase, our method applies a self-correcting loop to iteratively refine both the gaze narrations and the integrated narrations based on explicit quality criteria.

This decomposed, hierarchical coarse-to-fine architecture is central to our contribution. It ensures that the generated narrations are not only robustly grounded in observable sensor data but are also contextually holistic. By design, this approach overcomes the factual inconsistency and local incoherence that often plague monolithic, end-to-end models (Lee et al. 2023; Yan et al. 2023), leading to more faithful and useful representations of human behavior.

Phase 1: Parsing Raw Gaze to Textual Narration

The first phase of GazeInterpreter is *parsing raw gaze to textual narration*, designed to bridge the gap between raw, continuous sensor readings and a discrete, semantic representation. Directly interpreting noisy numerical data with LLMs poses significant grounding challenges, often leading to factual inconsistencies (Bommasani et al. 2021; Lee et al. 2023). To circumvent this, we adopt a two-step approach that decomposes the problem into deterministic symbolic parsing followed by conditioned language synthesis.

Symbolic Gaze Parser. This initial step is a deterministic module that extracts a vocabulary of fundamental gaze events from the raw signal. For a given segment i, the raw gaze signal $S_i^{\mathrm{g}} \in \mathbb{R}^{N_g \times 2}$ consists of a sequence of N_g timestamped coordinates, which we denote as $\{(t_j, y_j, p_j)\}_{j=1}^{N_g}$ where (y_j, p_j) corresponds to the yaw and pitch values. The parser maps this numerical sequence S_i^{g} into a structured sequence of M symbolic event primitives $E_i = (e_1, e_2, \ldots, e_M)$. Specifically, we first computed the instantaneous angular velocity ω_j for each point j:

$$\omega_j = \frac{\sqrt{(y_j - y_{j-1})^2 + (p_j - p_{j-1})^2}}{t_j - t_{j-1}}. \quad (1)$$

We then follow the established Identification-by-Velocity-Threshold (I-VT) algorithm (Salvucci and Goldberg 2000), which used a two-threshold classification scheme to segment the signal into a vocabulary of primitives,

Type	Dimesion	High	Low
Gaze	Continuity	5: Perfect gaze transitions with natural flow.	0: Contains abrupt, illogical, or disjointed event descriptions.
Integrated	Match	5: Mutually supportive integration of modalities.	0: Modalities are disconnected, redundant, or contradictory.
	Temporal	5: Clear, logical, chronological progression.	0: Lacks a discernible temporal structure or causal flow.
	Completeness	5: Fully includes all key elements and actions.	0: Essential information or key behavioral events are omitted.

Table 1: Scoring rubric for the multi-dimensional narration evaluation.

$\mathcal{V} = \{\text{Fixation, Saccade, SmoothPursuit}\}$, based on whether ω_j falls below a low threshold v_{low} or exceeds a high threshold v_{high}. Each resulting primitive $e_m \in E_i$ is a rich data object, encapsulating not only its class from $\mathcal{V}$ but also quantitative attributes (e.g., duration, amplitude, peak_velocity) and their corresponding qualitative descriptors (e.g., duration_label: "Brief"). This process abstracts the noisy, high-dimensional signal into a compact, machine-readable symbolic representation.

Symbolic-to-Text Synthesizer. The second step leverages an LLM to translate the sequence of gaze symbolic events E_i into a coherent textual narration, T_i^{g}. The module's objective is to generate this narration by modeling the conditional probability $P(T_i^{\text{g}}|E_i)$. We operationalized this by first serializing the sequence of event objects E_i into a descriptive string, which is then embedded within a carefully engineered few-shot prompt (Kojima et al. 2022). By conditioning the generative process on this structured symbolic input, we fundamentally change the nature of the task for the LLM. Instead of performing risky inference from raw numbers, the model is constrained to a factual translation task: converting a symbolic, verifiable account of behavior into fluent natural language. This ensures the resulting gaze narration is not only descriptive but also verifiably accurate with respect to the underlying physical measurements.

Phase 2: Integrating Eye Gaze with Body Motion

The first phase produces a continuous stream of gaze narrations that remain faithful to the raw signal but are fragmented and lack explanatory depth. Thus, the core challenge is moving from narration to holistic interpretation. Leveraging the strong coupling between gaze and body motion (Hu et al. 2024b), the second phase integrates gaze narration with atomic body narration to generate a unified, eye–body–coordinated description of human behavior.

Historical Context for Coherence. Human behavior exhibits temporal coherence (Wei et al. 2022). To model this dependency, we use a sliding observation window to aggregate the historical context $\mathcal{H}_i$. Formally, $\mathcal{H}_i$ adopts dictionary format, primarily consisting of two parts: (i) W segments of integrated narrations previously inferred by our pipeline and (ii) feedback content obtained from the last round of self-correcting (Phase 3). Additional scene metadata (e.g., location, focus, etc.) can be included when data is available. $\mathcal{H}_i$ is updated in each iteration, efficiently preserving the necessary context and providing a solid foundation for the model to gain insights into temporal coherence and causal relationships.

Algorithm 1: Self-Correcting Loop

Require: Initial narration $\hat{T}^{(0)}$, max iterations $K_{\max}$, score thresholds $\boldsymbol{\tau}$
Ensure: A refined, high-quality narration $\hat{T}^*$

$\hat{T}^* \leftarrow \hat{T}^{(0)}$
for $k = 0$ to $K_{\max} - 1$ **do**
  $(\mathbf{s}^{(k)}, \phi^{(k)}) \leftarrow \text{LLM}_{\text{eval}}(\hat{T}^{(k)})$
  if all components $s_j^{(k)} \geq \tau_j$ **then**
    return $\hat{T}^{(k)}$
  end if
  $\hat{T}^{(k+1)} \leftarrow \text{LLM}_{\text{refine}}(\hat{T}^{(k)}, \phi^{(k)})$
  $\hat{T}^* \leftarrow \hat{T}^{(k+1)}$
end for
return $\hat{T}^*$

Eye-Body-Coordinated Narration Synthesis. Then, we can integrate narrations of eye movements and body motions using LLM. For each segment i, the LLM conditions its generation on both the textual gaze narration T_i^{g} from the first phase of our pipeline and the corresponding sequence of atomic body motion narration S_i^{m}. The objective is to model the conditional probability $P(\hat{T}_i \mid T_i^{\text{g}}, S_i^{\text{m}}, \mathcal{H}_i)$, where $\mathcal{H}_i$ represents the historical context to ensure temporally coherent. We first carefully designed a structured prompt template $\Pi_{\text{integ}}(i)$ to fill in all the above information:

$$\Pi_{\text{integ}}(i) = [\texttt{CTX}: \mathcal{H}_i;\ \texttt{GAZE}: T_i^{\text{g}};\ \texttt{MOTION}: S_i^{\text{m}}]. \quad (2)$$

Here, special tokens such as `CTX:` and `GAZE:` act as explicit delimiters. This structured formatting is crucial as it guides the LLM to differentiate between historical context, observed gaze behavior, and concurrent body actions, thus facilitating more precise multi-modal reasoning. By conditioning on this structured input, the LLM's task extends beyond mere summarization to active reasoning. For instance, it learns to associate a gaze shift described in T_i^{g} with a "user is walking" description in S_i^{m} to infer the holistic action: "The user carefully scans the ground while walking."

The final integrated eye-body-coordinated narration $\hat{T}_i$ for the current segment is then generated as:

$$\hat{T}_i = \arg\max_{T \in \mathcal{T}} P(T \mid \Pi_{\text{integ}}(i)). \quad (3)$$

This mechanism ensures that the resulting narration is not only grounded in the current observation but also logically consistent with preceding actions.

Scene Type	Method	MM Dist↓	FID↓	Top-1↑	Top-2↑	Top-3↑	MM↑
Low-level	MotionGPT	$6.748^{\pm.098}$	$7.458^{\pm.322}$	$0.052^{\pm.002}$	$0.126^{\pm.003}$	$0.187^{\pm.004}$	$3.469^{\pm.051}$
	MotionGPT†	$\mathbf{6.406}^{\pm.053}$	$\mathbf{6.801}^{\pm.241}$	$\mathbf{0.102}^{\pm.002}$	$\mathbf{0.153}^{\pm.003}$	$\mathbf{0.214}^{\pm.003}$	$\mathbf{3.727}^{\pm.044}$
High-level	MotionGPT	$7.133^{\pm.033}$	$8.804^{\pm.142}$	$0.054^{\pm.002}$	$0.123^{\pm.002}$	$0.162^{\pm.005}$	$3.223^{\pm.024}$
	MotionGPT†	$\mathbf{6.862}^{\pm.025}$	$\mathbf{8.134}^{\pm.323}$	$\mathbf{0.062}^{\pm.001}$	$\mathbf{0.127}^{\pm.003}$	$\mathbf{0.193}^{\pm.004}$	$\mathbf{3.864}^{\pm.017}$
All	MotionGPT	$6.941^{\pm.056}$	$8.131^{\pm.304}$	$0.053^{\pm.003}$	$0.124^{\pm.003}$	$0.175^{\pm.004}$	$3.346^{\pm.036}$
	MotionGPT†	$\mathbf{6.634}^{\pm.042}$	$\mathbf{7.468}^{\pm.277}$	$\mathbf{0.082}^{\pm.004}$	$\mathbf{0.140}^{\pm.005}$	$\mathbf{0.204}^{\pm.005}$	$\mathbf{3.796}^{\pm.028}$

Table 2: Comparison of different input types in text-driven motion generation tasks. We fix the generation model weight and evaluate the effect of different text inputs. † indicates using our eye-body-coordinated narrations as input.

Type	Train Set	Simil.↑	BERT F1↑	ROUGE-L↑	Action F1↑
Low	Nymeria	0.525	0.869	0.241	0.281
	GazeInterpreter	**0.575**	**0.877**	**0.276**	**0.294**
High	Nymeria	0.393	0.866	0.163	0.171
	GazeInterpreter	**0.436**	**0.881**	**0.186**	**0.201**
All	Nymeria	0.459	0.868	0.202	0.226
	GazeInterpreter	**0.506**	**0.879**	**0.231**	**0.248**

Table 3: Comparison of baselines in action anticipation tasks. Simil. indicates Cosine Similarity.

Type	Train Set	Simil.↑	BERT F1↑	ROUGE-1↑	ROUGE-L↑
Low	Nymeria	0.564	0.851	0.228	0.174
	GazeInterpreter	**0.583**	**0.860**	**0.283**	**0.219**
High	Nymeria	0.395	0.820	0.165	0.126
	GazeInterpreter	**0.490**	**0.859**	**0.175**	**0.133**
All	Nymeria	0.480	0.836	0.197	0.150
	GazeInterpreter	**0.537**	**0.860**	**0.575**	**0.229**

Table 4: Comparison of baselines in behavior summarization tasks. Simil. indicates Cosine Similarity.

Phase 3: Self-Correcting Loop

The output of LLM may fail to emphasize the most salient behavioral cues and suffer from hallucinations (Li et al. 2024). We therefore introduce a final self-correcting loop. Unlike generic feedback mechanisms that apply superficial edits (Nacke et al. 2011), our loop iteratively evaluates and refines the narration across key quality dimensions, ensuring suitability for downstream tasks requiring a deep understanding of human behavior.

Multi-Dimensional Narration Evaluation. The self-correcting loop is driven by a specialized LLM_{eval}. Inspired by (Han et al. 2025), we tailor the criteria to the specific dimensions of each narration type. For gaze narration, the key dimension is Continuity, as natural gaze behavior unfolds as a continuous event stream (Yin et al. 2024). This ensures the narration reflects coherent transitions between actions rather than isolated events. For integrated narration, evaluation considers three factors: (i) *Modality Match*, assessing whether gaze and body actions are synergistic and contextually aligned; (ii) *Temporal Coherence*, ensuring the narration follows a logical chronological flow; and (iii) *Completeness*, verifying that all critical behavioral events are retained.

Based on these dimensions, the output of LLM_{eval} is twofold: a structured score vector $\mathbf{s}$ that quantifies quality, and a textual critique ϕ that provides targeted feedback for improvement. The specific scoring rubrics that guide this assessment are detailed in Table 1. This approach ensures the evaluation process is transparent, replicable, and provides fine-grained diagnostics for refinement.

Threshold-Governed Iterative Refinement. The refinement process, detailed in Algorithm 1, is structured as a collaborative loop between two specialized LLMs: LLM_{eval} and $\text{LLM}_{\text{refine}}$. The loop iterates until the narration's quality, measured by a score vector $\mathbf{s}$, meets or exceeds a predefined threshold vector $\boldsymbol{\tau}$. In each iteration k, LLM_{eval} first assesses the current narration $\hat{T}^{(k)}$ to generate both the scores $\mathbf{s}^{(k)}$ and a textual critique $\phi^{(k)}$. If these scores are insufficient, $\text{LLM}_{\text{refine}}$ then uses the original narration and the critique to produce an improved version $\hat{T}^{(k+1)}$.

This self-correcting loop iteratively filters out noise and redundancy based on set criteria, ensuring that narration correctly reflects facts and ultimately provides a comprehensive overview of human behavior.

Experiments and Results

Our experimental validation is twofold. We first assess the descriptive fidelity and effectiveness of our integrated narrations via a text-driven motion generation task. Subsequently, we demonstrate their utility for higher-level human behavior understanding on the downstream tasks of action anticipation and behavior summarization.

Implementation Details

Dataset. The Nymeria dataset (Ma et al. 2024) is currently the only public source of synchronized gaze and motion narrations. As the largest human motion dataset with 300 hours of daily activities, its scale allows robust tests. To facilitate a comprehensive analysis of our method's capabilities, we follow the EgoCHARM (Padmanabha et al. 2025) to categorize the 236 sequences with motion narration annotations into two subsets: *high-level* activities, which involve complex, goal-oriented interactions (e.g., housekeeping), and *low-level* activities, which consist of simple, atomic movements (e.g., walking). This division allows us to specifically test our approach's performance on tasks with varying degrees of behavioral complexity.

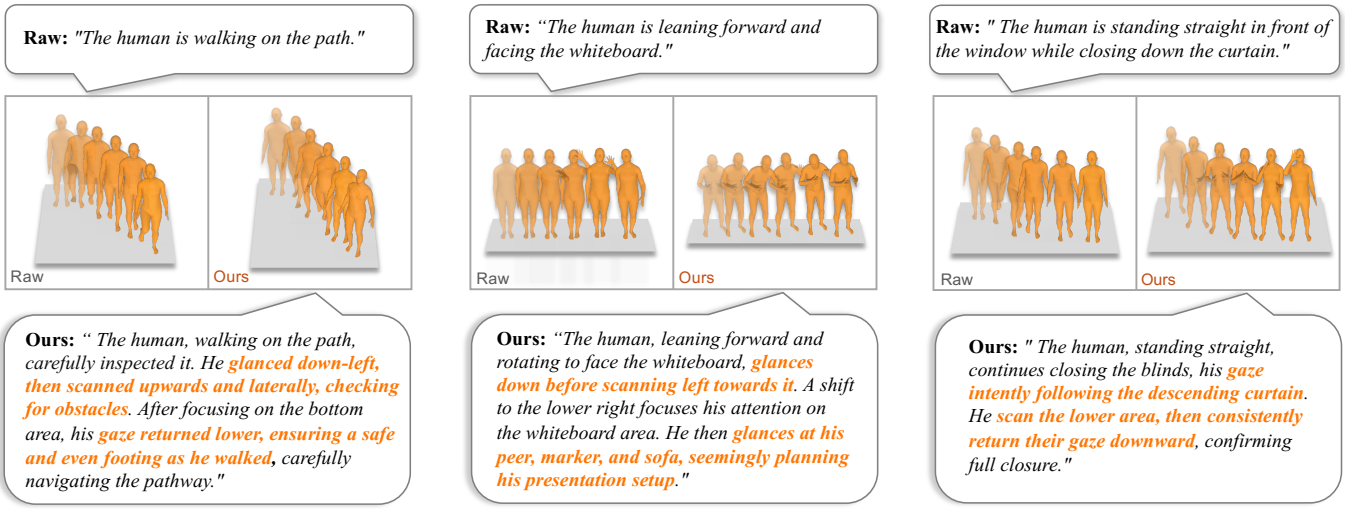

Figure 3: Qualitative comparison for text-driven motion generation on **our** integrated narrations versus **raw** atomic body narrations, including examples from both low-level scenes (e.g., walking) and high-level scenes (e.g., housework).

Implementation Specifics. All LLM-based modules in our framework utilize the Gemini-2.5-Flash model (Comanici et al. 2025), guided by a few-shot, in-context learning strategy (Comanici et al. 2025). For the *symbolic gaze parser*, we follow general guidelines (Salvucci and Goldberg 2000) to set the I-VT velocity thresholds to $v_{\text{low}} = 30°/s$ and $v_{\text{high}} = 100°/s$. For integrating gaze and body motion, the size of the sliding window is set to $W = 2$, and the *self-correcting loop* runs for a maximum of $K_{\text{max}} = 3$ iterations with a score threshold of $\tau = 4.5$. For more details about LLM prompts and parameters, please refer to our supplementary materials and source code.

Evaluation Metrics. We follow the same metrics as (Lin et al. 2023) to ensure a comprehensive evaluation in the text-driven motion. We assess the alignment between text and motion using Multimodal Distance (MM Dist), which measures the average feature space distance, and R-Precision (Top-k), which evaluates retrieval accuracy. The realism of the generated motions is measured by the Frechet Inception Distance (FID), which reflects their distributional similarity to real motions. Finally, we evaluate generation diversity using Multimodality (MM), which quantifies the variation among different motions produced from the same text prompt.

Main Experiments

Experiment Setup. Following Motion-X (Lin et al. 2023), we fix the weights of the motion generation model, using the atomic motion narration provided by Nymeria and the integrated narration obtained through our method to drive the model in generating motion sequences, and finally measuring the advantages of our narrations using a wide range of metrics. We choose the popular MotionGPT (Jiang et al. 2023) model as the motion generation baseline. Moreover, we follow the commonly used feature extractor (Guo et al. 2022) as the basis for calculating FID metrics.

Quantitative Results. When the MotionGPT baseline weights are fixed, Table 2 shows that narrations from GazeInterpreter significantly outperform the atomic body narrations provided by Nymeria across all metrics. Notably, this advantage is pronounced on low-level activities, where our method substantially reduces the FID from 7.458 to 6.801. This suggests that the factually-grounded details in our narrations provide a superior conditioning signal for generating precise, atomic motions. The performance gains persist for complex high-level scenes, where GazeInterpreter continues to yield lower distribution distances and higher matching scores. These findings validate that our eye-body-coordinated narration is effectively grounded in physical behavior and leads to higher-fidelity motion synthesis.

Qualitative Results. The qualitative results in Fig. 3 vividly demonstrate the superior fidelity of motions generated from our narrations compared to those from atomic body narrations of Nymeria. By capturing the subtle interplay between gaze and physical action, our eye-body-coordinated narration provides crucial contextual richness. This allows the synthesis model to generate actions that are not only more natural and detailed but also more aligned with potential human intentions, providing a solid foundation for a comprehensive explanation of human behavior.

Downstream Tasks Experiments

To demonstrate the advantage of our integrated narrations in comprehensively understanding human behavior, we evaluate their performance on two representative downstream tasks: action anticipation (Hu et al. 2024c,a,d) and behavior summarization (Zhang et al. 2025). For both tasks, we follow (Zhang et al. 2019) to employ a comprehensive suite of metrics to assess the generated text. Semantic fidelity is measured by Cosine Similarity and BERTScore F1, while structural coherence and lexical accuracy are evaluated using ROUGE (1/L) and a keyword-based Action F1 Score.

Module	MM Dist ↓	FID ↓	Top-1 ↑
w/o Structure	$8.135^{\pm.076}$	$9.124^{\pm.442}$	$0.059^{\pm.005}$
w/o Parser	$7.642^{\pm.060}$	$7.893^{\pm.459}$	$0.061^{\pm.005}$
w/o Loop	$7.425^{\pm.066}$	$7.831^{\pm.344}$	$0.063^{\pm.004}$
Ours	$\mathbf{6.634}^{\pm.042}$	$\mathbf{7.468}^{\pm.277}$	$\mathbf{0.082}^{\pm.004}$

Table 5: Ablation studies for Hierarchical Structure, Symbolic Gaze Parser, and Self-Correcting Loop.

Action Anticipation Results. To evaluate the predictive power of our narrations, we perform an action anticipation task where the goal is to generate a textual description of the next action based on the current context. We employ a frozen Gemini-2.5-Flash model as a zero-shot predictor, providing it with either the original Nymeria atomic motion narrations or our integrated narrations as input. As shown in Table 3, using the narrations from our method leads to a significant improvement in prediction performance across all metrics. In all cases, our method boosts the cosine similarity of predictions from 0.459 to 0.506 and the keyword-centric Action F1 score from 0.226 to 0.248. This striking result suggests that by systematically integrating multi-modal cues like gaze, our generated narration encapsulates more predictive, intent-rich information than the human-annotated text alone. This advantage holds for both simple *low-level* and complex *high-level* scenarios, demonstrating the potential for uncovering human intentions.

Behavior Summarization Results. This task requires generating a high-level behavior summary from a sequence of motion descriptions. We employed a frozen Gemini-2.5-Flash model, comparing the summaries it produces when conditioned on our integrated narrations versus those from the original Nymeria motion narrations. The results in Table 4 show that our narration serves as a superior conditioning signal, outperforming the baseline across all metrics. This performance gain is especially pronounced in complex *High-level* scenarios, where our method boosts the summary's cosine similarity from 0.395 to 0.490. This consistent success across motion generation, anticipation, and summarization tasks compellingly demonstrates that our integrated narrations provide a robust and versatile behavioral representation, effectively enhancing a wide range of human-aware applications.

Ablation Study

Overall Ablation Study. To validate the contribution of our framework's core components, we conduct a comprehensive ablation study, with results presented in Table 5. The removal of the *hierarchical structure* causes the most significant performance degradation across all metrics, highlighting the importance of integrating gaze semantic extraction into the motion hierarchy. Similarly, ablating the *symbolic gaze parser* by using raw signal inputs, or deactivating the *self-correcting loop*, both lead to a distinct decline in performance. These results validate that each component of the GazeInterpreter is essential and contributes effectively to generating high-fidelity behavioral narration.

Continuity	Match	Temporal	Completeness	Top-1↑	FID↓
				0.063	7.831
✓				0.069	7.722
✓	✓			0.072	7.644
✓	✓	✓		0.074	7.573
✓	✓	✓	✓	**0.082**	**7.468**

Table 6: Effects of each evaluation dimension.

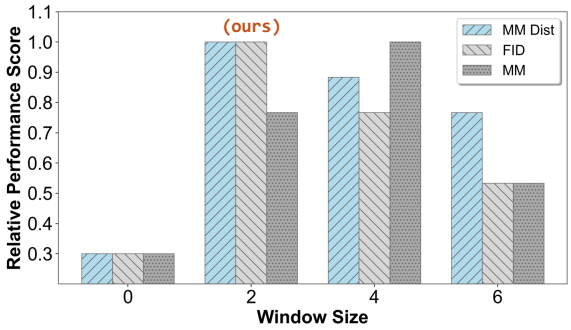

Figure 4: Effects of different observation window sizes.

Ablation Study on Quality Dimension. We also analyze the impact of each quality dimension within the self-correcting loop. As detailed in Table 6, the results reveal a clear additive effect: cumulatively introducing each of our designed criteria—*Continuity*, *Modality Match*, *Temporal Coherence*, and *Completeness*, progressively improves performance. This validates our multi-dimensional evaluation rubric, with each dimension providing an essential criterion for judging how accurately the narrations correspond to real-world behavioral features.

Ablation Study on Windows Size. We further analyze the effect of the sliding observation window and investigate how many preceding narration segments W should be incorporated into the historical context $\mathcal{H}_i$. As shown in Fig. 4, a window size of $W = 2$ consistently achieves the best balance between contextual richness and computational efficiency. Increasing W provides only marginal gains while noticeably increasing inference cost and occasionally introducing redundant or noisy historical cues.

Conclusion

In this work, we proposed a novel LLM-based method that features a symbolic gaze parser to convert raw gaze signal, a hierarchical framework to integrate gaze with body motion, and a self-correcting loop for refinement. Through extensive experiments on a large-scale benchmark, we showed the advantages of our eye-body-coordinated narrations in text-driven motion generation. We also demonstrated the effectiveness of our method for the sample downstream tasks of action anticipation and behavior summarization. As such, our work reveals the significant information content available in eye gaze for interpreting human behavior and guides future work on this promising direction.

References

Belardinelli, A.; Kondapally, A. R.; Ruiken, D.; Tanneberg, D.; and Watabe, T. 2022. Intention estimation from gaze and motion features for human-robot shared-control object manipulation. In *Proceedings of the 2022 IEEE International Conference on Intelligent Robots and Systems*, 9806–9813. IEEE.

Bommasani, R.; Hudson, D. A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M. S.; Bohg, J.; Bosselut, A.; Brunskill, E.; et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Chang, Q.; Dai, W.; Shuai, Z.; Yu, L.; and Yue, Y. 2025. Spatial-Temporal Perception with Causal Inference for Naturalistic Driving Action Recognition. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Chang, Q.; and Tong, Y. 2024. A hybrid global-local perception network for lane detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 981–989.

Chen, L.-H.; Lu, S.; Zeng, A.; Zhang, H.; Wang, B.; Zhang, R.; and Zhang, L. 2024. Motionllm: Understanding human behaviors from human motions and videos. *arXiv preprint arXiv:2405.20340*.

Chen, X.; Jiang, M.; and Zhao, Q. 2024. Gazexplain: Learning to predict natural language explanations of visual scanpaths. In *European Conference on Computer Vision*, 314–333. Springer.

Comanici, G.; Bieber, E.; Schaekermann, M.; Pasupat, I.; Sachdeva, N.; Dhillon, I.; Blistein, M.; Ram, O.; Zhang, D.; Rosen, E.; et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Coutrot, A.; Hsiao, J. H.; and Chan, A. B. 2018. Scanpath modeling and classification with hidden Markov models. *Behavior Research Methods*, 50(1): 362–379.

Delmas, G.; Weinzaepfel, P.; Lucas, T.; Moreno-Noguer, F.; and Rogez, G. 2022. Posescript: 3d human poses from natural language. In *European Conference on Computer Vision*, 346–362. Springer.

Guo, C.; Zou, S.; Zuo, X.; Wang, S.; Ji, W.; Li, X.; and Cheng, L. 2022. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5152–5161.

Hadnett-Hunter, J.; Nicolaou, G.; O'Neill, E.; and Proulx, M. 2019. The effect of task on visual attention in interactive virtual environments. *ACM Transactions on Applied Perception*, 16(3): 1–17.

Han, H.; Wu, X.; Liao, H.; Xu, Z.; Hu, Z.; Li, R.; Zhang, Y.; and Li, X. 2025. Atom: Aligning text-to-motion model at event-level with gpt-4vision reward. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 22746–22755.

Hu, Z.; Bulling, A.; Li, S.; and Wang, G. 2021. FixationNet: forecasting eye fixations in task-oriented virtual environments. *IEEE Transactions on Visualization and Computer Graphics*, 27(5): 2681–2690.

Hu, Z.; Bulling, A.; Li, S.; and Wang, G. 2022. EHTask: recognizing user tasks from eye and head movements in immersive virtual reality. *IEEE Transactions on Visualization and Computer Graphics*.

Hu, Z.; Haeufle, D.; Schmitt, S.; and Bulling, A. 2025. HOIGaze: Gaze Estimation During Hand-Object Interactions in Extended Reality Exploiting Eye-Hand-Head Coordination. In *Proceedings of the ACM Special Interest Group on Computer Graphics and Interactive Techniques*, 1–10.

Hu, Z.; Li, S.; Zhang, C.; Yi, K.; Wang, G.; and Manocha, D. 2020. DGaze: CNN-based gaze prediction in dynamic scenes. *IEEE Transactions on Visualization and Computer Graphics*, 26(5): 1902–1911.

Hu, Z.; Schmitt, S.; Haeufle, D.; and Bulling, A. 2024a. GazeMotion: Gaze-guided Human Motion Forecasting. In *Proceedings of the 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems*.

Hu, Z.; Xu, J.; Schmitt, S.; and Bulling, A. 2024b. Pose2Gaze: Eye-body Coordination during Daily Activities for Gaze Prediction from Full-body Poses. *IEEE Transactions on Visualization and Computer Graphics*.

Hu, Z.; Yin, Z.; Haeufle, D.; Schmitt, S.; and Bulling, A. 2024c. HOIMotion: Forecasting Human Motion During Human-Object Interactions Using Egocentric 3D Object Bounding Boxes. *IEEE Transactions on Visualization and Computer Graphics*.

Hu, Z.; Zhang, C.; Li, S.; Wang, G.; and Manocha, D. 2019. SGaze: a data-driven eye-head coordination model for real-time gaze prediction. *IEEE Transactions on Visualization and Computer Graphics*, 25(5): 2002–2010.

Hu, Z.; Zhang, G.; Yin, Z.; Haeufle, D.; Schmitt, S.; and Bulling, A. 2024d. HaHeAE: Learning Generalisable Joint Representations of Human Hand and Head Movements in Extended Reality. *arXiv preprint arXiv:2410.16430*.

Itti, L.; Koch, C.; and Niebur, E. 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11): 1254–1259.

Jiang, B.; Chen, X.; Liu, W.; Yu, J.; Yu, G.; and Chen, T. 2023. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems*, 36: 20067–20079.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213.

Kong, Y.; and Fu, Y. 2022. Human action recognition and prediction: A survey. *International Journal of Computer Vision*, 130(5): 1366–1401.

Kothari, R.; Yang, Z.; Kanan, C.; Bailey, R.; Pelz, J. B.; and Diaz, G. J. 2020. Gaze-in-wild: a dataset for studying eye and head coordination in everyday activities. *Scientific Reports*, 10(1): 1–18.

Lee, H.; Joo, S.; Kim, C.; Jang, J.; Kim, D.; On, K.-W.; and Seo, M. 2023. How well do large language models truly ground? *arXiv preprint arXiv:2311.09069*.

Li, X.; Wang, S.; Zeng, S.; Wu, Y.; and Yang, Y. 2024. A survey on LLM-based multi-agent systems: workflow, infrastructure, and challenges. *Vicinagearth*, 1(1): 9.

Lin, J.; Zeng, A.; Lu, S.; Cai, Y.; Zhang, R.; Wang, H.; and Zhang, L. 2023. Motion-x: A large-scale 3d expressive whole-body human motion dataset. *Advances in Neural Information Processing Systems*, 36: 25268–25280.

Ma, L.; Ye, Y.; Hong, F.; Guzov, V.; Jiang, Y.; Postyeni, R.; Pesqueira, L.; Gamino, A.; Baiyya, V.; Kim, H. J.; et al. 2024. Nymeria: A massive collection of multimodal egocentric daily motion in the wild. In *European Conference on Computer Vision*, 445–465. Springer.

Nacke, L. E.; Kalyn, M.; Lough, C.; and Mandryk, R. L. 2011. Biofeedback game design: using direct and indirect physiological control to enhance game interaction. In *Proceedings of the SIGCHI conference on human factors in computing systems*, 103–112.

Padmanabha, A.; Govindarajan, S.; Kim, H.; Ortiz, S.; Rajan, R.; Senkal, D.; and Kadetotad, S. 2025. EgoCHARM: Resource-Efficient Hierarchical Activity Recognition using an Egocentric IMU Sensor. *arXiv preprint arXiv:2504.17735*.

Pons-Moll, G.; Fleet, D. J.; and Rosenhahn, B. 2014. Posebits for monocular human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2337–2344.

Salvucci, D. D.; and Goldberg, J. H. 2000. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 ACM Symposium on Eye Tracking Research and Applications*, 71–78.

Sidenmark, L.; and Gellersen, H. 2019. Eye, head and torso coordination during gaze shifts in virtual reality. *ACM Transactions on Computer-Human Interaction*, 27(1): 1–40.

Sitzmann, V.; Serrano, A.; Pavel, A.; Agrawala, M.; Gutierrez, D.; Masia, B.; and Wetzstein, G. 2018. Saliency in VR: how do people explore virtual environments? *IEEE Transactions on Visualization and Computer Graphics*, 24(4): 1633–1642.

Wang, Z.; Liu, S.; Zhang, Z.; Ma, T.; Zhang, C.; and Ye, Y. 2024. Can LLMs Convert Graphs to Text-Attributed Graphs? *arXiv preprint arXiv:2412.10136*.

Wei, W.-L.; Lin, J.-C.; Liu, T.-L.; and Liao, H.-Y. M. 2022. Capturing humans in motion: Temporal-attentive 3D human pose and shape estimation from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13211–13220.

Xu, X.; Ni, C.; Guo, X.; Liu, S.; Wang, X.; Liu, K.; and Yang, X. 2025. Distinguishing llm-generated from human-written code by contrastive learning. *ACM Transactions on Software Engineering and Methodology*, 34(4): 1–31.

Yan, H.; Hu, Z.; Schmitt, S.; and Bulling, A. 2024. GazeMoDiff: Gaze-guided Diffusion Model for Stochastic Human Motion Prediction. In *Proceedings of the 2024 Pacific Conference on Computer Graphics and Applications*.

Yan, X.; Guo, J.; Lou, X.; Wang, J.; Zhang, H.; and Du, Y. 2023. An efficient end-to-end training approach for zero-shot human-AI coordination. *Advances in neural information processing systems*, 36: 2636–2658.

Yarbus, A. L. 1967. *Eye movements and vision*. Springer.

Yin, P.; Wang, J.; Zeng, G.; Xie, D.; and Zhu, J. 2024. Lg-gaze: Learning geometry-aware continuous prompts for language-guided gaze estimation. In *European Conference on Computer Vision*, 1–17. Springer.

Zhang, G.; Ahmed, M.; Hu, Z.; and Bulling, A. 2025. SummAct: Uncovering User Intentions Through Interactive Behaviour Summarisation. In *Proc. ACM CHI Conference on Human Factors in Computing Systems (CHI)*, 1–17.

Zhang, M.; Cai, Z.; Pan, L.; Hong, F.; Guo, X.; Yang, L.; and Liu, Z. 2022. Motiondiffuse: Text-driven human motion generation with diffusion model. *arXiv preprint arXiv:2208.15001*.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.