2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

SHAF: Small Language Model Integrated with Motion Modality for Multimodal Interaction

Aamir Ahmad Ansari*, Nguyen Tan Viet Tuyen*, and Sarvapali D Ramchurn

Abstract— **Multimodal interaction plays a vital role in human–AI interaction, enabling robots or AI agents to interpret human input from multiple sensory channels and respond through diverse communication modalities. This paper introduces SHAF, an LLM-based multimodal model capable of handling text, image, and human motion as both input and output modalities across different multi-turn conversational settings. In SHAF, vector quantization is employed to convert images and human motion into an aligned set of tokens, followed by pre-training and instruction fine-tuning of a small Large Language Model (LLM) on our newly created SHAF dataset. Experimental results demonstrate that SHAF achieves competitive performance in text-to-motion and motion-to-text tasks in comparison to relevant works, while handling an additional modality and supporting a broader range of tasks. This research contributes an LLM-based multimodal approach, with the aim of fostering deeper exploration of human motion modality in LLMs within the context of HRI and related domains.**

I. Introduction

Seamless interaction between humans and robots or AI assistants requires systems to dynamically accept human data collected through different modalities to infer human intentions and provide appropriate assistance. At the same time, these systems should be able to respond through diverse output modalities depending on particular interaction contexts. Beyond verbal communication, incorporating modalities such as images, video, and non-verbal gestures enables more natural and context-aware engagement in human–robot interaction (HRI) [1].

Recent advances in Large Language Models (LLMs), particularly Multimodal Large Language Models (MLLMs), have opened new opportunities for addressing the topic of multimodal interaction in HRI through their powerful capabilities in reasoning and generation across different modalities. While substantial progress has been made in integrating common modalities such as images, video, and audio in recent years [2], the incorporation of human motion, an essential modality for understanding and synthesizing human interaction, has not been sufficiently explored in the community. This paper aims to contribute to this research gap by introducing SHAF, an LLM-based model capable of handling text, image, and motion as both input and output modalities across different multi-turn conversational settings, thereby simulating practical multimodal interaction between humans and AI systems. SHAF employs a small language model as its backbone, with the goal of providing a lightweight, cost-effective, and time-efficient multimodal solution that paves the way for more adaptable human–AI interactions applicable to both virtual AI agents and socially assistive robots.

The main contributions of this paper are two-fold. First, we introduce SHAF, an LLM-based multimodal approach that unifies text, image, and motion within a shared representation space, thereby enabling cross-modality alignment. This unified framework improves both the interpretation of multimodal input and the generation of multimodal output. Empirical results demonstrate that SHAF achieves competitive performance compared to relevant models specialized in text-to-motion and motion-to-text tasks, while additionally supporting an extra modality and a broader set of tasks. To the best of our knowledge, SHAF is among the first vision–language–motion models capable of handling a wide spectrum of conversational tasks. Second, we present the SHAF dataset, a multi-turn multimodal dataset centered on human motion in daily activities across different settings. The experimental results reported in Section V provide the first benchmark for this dataset. SHAF dataset is expected to serve as a valuable resource for the community, fostering deeper integration of the motion modality into LLMs within the context of HRI and related research domains. Further details are available at: `https://tvtn.github.io/SHAF.github.io/`.

II. Background

A. Multimodal Human-Robot Interaction

Multimodal interaction plays a vital role in human–robot systems, enabling them to interpret human input from multiple sensory channels and respond through diverse communication modalities, thereby enhancing the overall interaction experience. Beyond verbal communication, modalities such as images, video, and non-verbal gestures foster more natural and context-sensitive engagement between humans and robots [1] or virtual AI agents [3]. In particular, human motion is an essential modality for understanding human intentions [4], [5] and for generating human-inspired motion [6], [7], ultimately improving the ability of agents to interact with humans.

Recent studies have emphasized the role of multimodal HRI with a focus on human motion data [8], [9]. The framework introduced in [8] enables robots to learn from human non-verbal gestures and reconfigure their own body motions accordingly. However, this work overlooks the integration of textual information from human interaction, thereby limiting robots' adaptability across broader HRI contexts. In [9], the authors proposed the VITA, a multimodal HRI system

* Equal contributions.

The authors are with the school of Electronics and Computer Science, University of Southampton, UK. Corresponding author: tuyen.nguyen@soton.ac.uk

This work was supported by the Engineering and Physical Sciences Research Council [grant number EP/Y009800/1], through funding from Responsible Ai UK.

designed to adapt to coaches' multimodal behaviors and deliver coaching exercises to promote mental well-being via a social robot equipped with verbal and non-verbal gesture channels. While VITA leverages an LLM for processing verbal input, human non-verbal signals are interfaced with the LLM through prompt engineering rather than being encoded as signal tokens. This design approach may restrict cross-modal learning between human signal inputs and text tokens in a shared embedding space, thereby constraining the system's capacity to comprehensively interpret behavioral input and to generate coherent robot behaviors. To overcome these limitations, this paper introduces an LLM-based model that unifies text, image, and motion within a shared representation space, enabling cross-modality alignment. This unified framework aims to pave the way for more adaptable and context-aware human–AI interactions, applicable to both virtual agents and social robots.

B. Multimodal LLMs

MLLMs extend LLMs to represent and generate modalities such as text, images, audio, and video by aligning modality features within a shared embedding space. This cross-modal alignment enables richer reasoning and more versatile interaction compared with text-only models. For example, an MLLM can process an image alongside its textual description [10], generate detailed captions, or answer queries requiring both visual and linguistic grounding. Similarly, audio [11] and video [12] integration enhances temporal and perceptual understanding, such as affect recognition in speech or summarizing motion-rich visual scenes. Frameworks like AnyGPT [5] further expand the LLMs beyond dual modalities by unifying text, audio, video, and image modalities in a shared embedding space, towards enhancing the model capabilities in reasoning and generation across different modalities.

Recent efforts have incorporated human motion to enrich human–AI interaction by enabling systems to comprehend and generate human motion. Models such as MotionGPT [13], TM2T [14], and MotionChain [15] discretize continuous human 3D pose sequences into motion tokens, creating a motion vocabulary analogous to words. This allows unified modeling of text-to-motion, motion-to-text, and multimodal conversational tasks. Embedding motion tokens in the same space as text tokens enables LLMs to generate diverse actions from natural language, describe perceived movement, and perform motion-driven conversations. However, MotionGPT and TM2T are limited to text–motion and motion-motion alignment, overlooking the contribution of vision modality. While MotionChain represents a recent effort to integrate vision alongside motion into a vision–motion–language model, its capabilities remain restricted to multimodal input rather than fully multimodal output. Specifically, MotionChain can generate text and motion conditioned on text, image, or motion input, but it cannot produce image outputs. This limitation prevents it from supporting the broader spectrum of multi-turn conversations that are critical for advancing practical human–AI interaction. In contrast, our proposed approach addresses these shortcomings by enabling users to interact with SHAF across a much wider range of conversational tasks, including motion-to-image, image reasoning, motion reasoning, text-to-image, and beyond.

To our best knowledge, there are currently no publicly available multi-turn multimodal datasets centered on human motion that cover a wide spectrum of conversational scenarios. The HumanML3D [16], while serving as a large-scale dataset that bridges the gap between human motion and textual input, does not incorporate the vision modality. The MotionChain dataset [15] is only partially available to the research community, as the current version lacks associated vision data. In addition, the current release is restricted to single-turn conversations, which limits its usability for training multi-turn multimodal tasks. To overcome these limitations, we created the SHAF dataset, a multi-turn multimodal conversational dataset centered on human daily activities for training the proposed SHAF model in the supervised fine-tuning stage.

III. SHAF MODEL

A key idea behind our SHAF model is to extend text-only large language models with additional modalities commonly used in human interaction, including images and human motions, by learning them as a foreign language. As language models natively operate on discrete text tokens, introducing a new modality in the form of discrete tokens enables seamless integration with minimal modifications to the model's architecture.

A. Overview

Fig. 1 presents an overview of the SHAF model architecture. SHAF is a small Large Language Model (LLM) that acts as the backbone of the framework, enabling joint reasoning over text, image, and human motion. Each modality is transformed into a sequence of discrete tokens using a modality-specific tokenizer. To mark modality boundaries and ensure modality-aware token generation, special tokens indicating the start and end of each modality are appended accordingly. These tokens, along with text tokens, are concatenated into a unified sequence and passed to the LLM backbone, where cross-modal dependencies are modeled through attention mechanisms. The LLM then autoregressively generates output tokens, which may correspond to text, motion, or image representations depending on particular interaction contexts. These generated tokens are subsequently decoded back into their respective modalities using dedicated de-tokenizers. In this way, the SHAF architecture supports a wide range of multimodal conversations, such as text-to-motion, motion-to-image, image-to-text, text-to-image, image-to-motion, motion reasoning, motion-to-text, and image reasoning. The overall architecture of SHAF has 3 core components: the tokenizers/quantizers, the LLM backbone, and the de-tokenizer(often a part of the quantizer). These components are described in the following subsections in detail.

B. Image Quantization

To extend the capabilities of our LLM, it is important to convert images from the continuous domain to a discrete and more manageable domain. SHAF utilizes the SEED tokenizer[17] to convert an image into a small sequence of

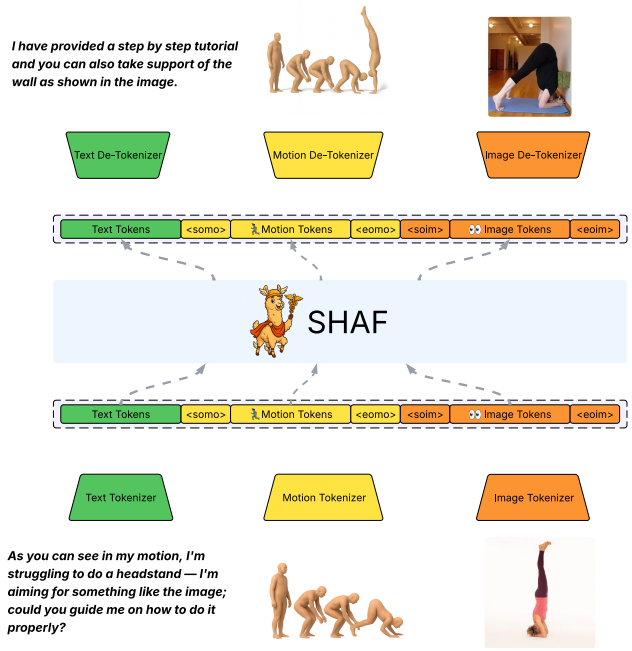

Fig. 1. An overview of the SHAF model. All modalities are converted into discrete tokens, upon which the LLM performs multimodal comprehension and generation tasks in an autoregressive fashion.

discrete tokens. Its design follows two foundational rules: (i) the tokens must be produced with 1D causal dependency (left-to-right), and (ii) they must carry high-level semantics information as do the words in language. The tokenizer has five parts: a ViT image encoder, a causal Q-Former, a vector-quantized (VQ) codebook, a small Transformer decoder, and an MLP that maps codes into a diffusion-model conditioning space. The ViT and UNet follow BLIP-2 and unCLIP Stable Diffusion, respectively [18]. The ViT outputs a 16×16 feature grid. This feature grid is then processed by a Q-Former (adopted from BLIP-2 [18]), and it converts these semantic features into 32 learnable queries that produce 32 causal embeddings. These embedding vectors are quantized via nearest-neighbor lookup in the VQ codebook to yield 32 visual codes. An MLP then maps the codes to conditioning embeddings for de-tokenization, improving realism and semantic fidelity. The codebook contains 8192 unique vectors; we therefore add 8192 new tokens to the LLM vocabulary. Compared to raw pixels, this discrete representation of images substantially reduces storage and computation while preserving semantic information.

C. Motion Quantization

To integrate motion modality into our LLM, the model requires a discrete representation of the 3D human motions. We adapt the motion quantizer from [14], which is based on the VQ-VAE framework. Unlike prior approaches that use quantized motion representation directly as sequences for LLMs, SHAF incorporates these discrete motion codes into the input string sequence, and feed them to the LLM as part of the standard text input during both the pre-training and supervised fine-tuning stages.

Specifically, given a motion sequence $m \in \mathbb{R}^{T \times D_p}$, where T is the number of frames and D_p is the pose dimension, several 1D convolutions are used to encode it into a set of latent vectors $\hat{b} \in \mathbb{R}^{t \times d}$ via an encoder $E(\cdot)$, i.e., $\hat{b} = E(m)$. These vectors are converted to a discrete set of tokens using a learnable codebook $B = \{b_k\}_{k=1}^{K} \subset \mathbb{R}^d$ of K embedding vectors. Each latent vector $\hat{b}_i$ is quantized to its nearest codebook entry:

$$b_q = Q(\hat{b}) := \left(\arg\min_{b_k \in B} \|\hat{b}_i - b_k\| \right) \in \mathbb{R}^{t \times d}. \quad (1)$$

The quantized sequence b_q is then decoded back to a reconstructed pose sequence $\hat{m} = D(b_q) = D(Q(E(m)))$. The training objective includes reconstruction and commitment losses as below:

$$\mathcal{L}_{vq} = \|\hat{m} - m\|_1 + \|\text{sg}[E(m)] - b_q\|_2^2 + \beta \|E(m) - \text{sg}[b_q]\|_2^2, \quad (2)$$

where $\text{sg}[\cdot]$ denotes the stop-gradient operator. In our study, we utilise the sequence b_q to represent the motion in the LLM. 1024 new motion tokens are added to the vocabulary, as it is the size of the codebook.

D. Text Quantization

Text tokenization is a critical component of SHAF model. It not only converts natural language into numerical tokens interpretable by the model, but also manages tokens from other modalities. Specifically, motion and image tokens are appended to the existing vocabulary so that they are not treated as general words, and no further preprocessing is applied to these tokens. We employ the Llama 3.2 tokenizer, which contains over 128k unique tokens.

E. LLM Backbone

The Large Language Model (LLM) functions as the central backbone of our pipeline, enabling a unified representation of text, image, and human motion. Each modality, originally residing in a continuous domain, is discretized into a finite sequence of codes using modality-specific quantizers [14], [17]. These codes, together with their corresponding textual captions, are then arranged into a single sequence. To ensure a clear separation between different modalities, special tokens, for motion `<somo>`(start of motion), `<eomo>`(end of motion), and for images `<soim>`(start of image), `<eoim>`, are appended. This forms an explicit boundary between the modalities.

During pre-training, this sequence may represent either a text-to-modality generation task or a modality-to-text generation task. The LLM first consumes the input tokens, which are embedded and propagated through multiple layers of attention and feed-forward networks. The attention modules facilitate two critical functions: (i) capturing dependencies between text tokens and modality tokens, and (ii) modeling intra-modality relationships among non-text tokens. At each decoding step, the model autoregressively predicts the next token, which can belong to the text vocabulary or the modality-specific codebook. The training loss is computed between the predicted tokens and the ground truth sequence, and the model parameters are updated accordingly.

In the supervised fine-tuning (SFT) stage, the principle of tokenization and modality discretization remains consistent, but the structure of the input sequence differs from pre-training. Instead of simple caption–modality pairs, the training data consists of multi-turn, multimodal conversations. Each conversational turn is represented as a sequence that begins with a start-of-turn token and ends with an end-of-turn token. The LLM is promoted to learn to generate a response based on the user instructions by masking the user turns in the conversation and only computing loss on the assistant turns that are actually predicted by the LLM.

Overall, the LLM backbone is not only responsible for aligning various modalities within a common tokenized framework but also for orchestrating their interactions through attention-based mechanisms and autoregressive generation. Its ability to seamlessly transition from caption–modality mappings in pre-training to multi-turn, multimodal dialogues in fine-tuning underscores its role as the decisive component that enables coherent cross-modal understanding and response generation across our proposed approach.

F. Training Strategy

The training pipeline consists of two stages: pre-training and supervised fine-tuning. In this first stage of the training process, SHAF extends the capabilities of the Llama 3.2 model, with 3 billion parameters. The core idea in this stage is to learn the alignment between text and the new modalities. We use prompt-based pre-training, where the corresponding text captions of the motion and images are utilized in a modality-to-text or text-to-modality task. This technique embeds rich cross-modal information into the tokens along with learning the modality-specific information. Given a multimodal sequence $x = \{x_1, \ldots, x_T\}$, we optimize the negative log-likelihood as illustrated in Eq. 3 for encouraging intra-modal and inter-modal dependencies. To avoid token overlap, modality identifiers with start/end markers are added to the quantized image and motion codes as illustrated in Fig. 1.

$$\mathcal{L}_{\text{NLL}} = -\sum_{t=1}^{T} \log p_\theta(x_t \mid x_{<t}) \tag{3}$$

The second stage of the training, supervised fine-tuning, improves the model's capabilities to follow human instructions more closely. The model receives multimodal prompts (instruction I) and generates modality-specific responses $A = \{a_1, \ldots, a_m\}$. The loss is applied only to SHAF outputs as shown in Eq. 4. This way of computing the loss function and updating the model enforces semantic alignment across modalities, as well as allows the model to follow human instructions more accurately.

$$\mathcal{L}_{\text{instr}} = -\sum_{j=1}^{m} \log p_\theta(a_j \mid x_{<t_j}), \tag{4}$$

IV. SHAF Datasets

The SHAF dataset is employed in our supervised fine-tuning stage with the aim of enhancing the model's ability to engage in multi-turn conversations based on user inputs. This stage also enables the model to learn relationships between human motion and image modalities. SHAF is designed to support daily user interactions across diverse contexts, such as well-being practices, instructions, and beyond, where the model should be able to follow multi-turn conversations and generate appropriate responses aligned with user requests. To this end, the SHAF dataset includes a set of tasks that cover a wide range of multi-turn conversational settings:

- Motion to Image Alignment: Given a human motion represented by 3D joint coordinates as input in a multi-turn conversation, the model is expected to generate images in response.
- Image to Motion Alignment: Given an image as input in a multi-turn conversation, the model is expected to generate human motion in response.
- Motion Reasoning: Provided a human motion as an initial input, the user can indulge in multiple rounds of questioning with the model.
- Image Reasoning: Provided an image as an initial input, the user can indulge in multiple rounds of questioning with the model.

To construct the dataset, the motion-to-image alignment task was developed using motion data collected from the HumanML3D dataset. The OpenAI API was utilized to prompt GPT-4o [19] with specific instructions for rephrasing the caption of each motion and for generating five distinct scenarios corresponding to the caption. For each scenario, a conversation was subsequently generated, resulting in five different conversations for each motion-text pair. Within each conversation, motion sequences were placed in the user's turns, whereas images were assigned to the assistant's turns (SHAF). Additional follow-up questions and answers were incorporated by the LLM, resulting in a multi-turn conversational structure. When constructing these conversations, we also ensured that the user's input explicitly requested the assistant to generate an image conditioned on the input motion. Furthermore, instructions were provided to the model to include image prompts in the assistant turns so that the images could later be generated using GenAI models. Similarly, we develop conversational dataset for image-to-motion alignment and motion reasoning tasks centered around the motion data available from the HumanML3D dataset. During the development of the image-to-motion alignment task samples, an LLM was instructed to keep images within the user instructions and motion data in the assistant's response. A mix of single-turn and two-turn conversations was retained so that the model could learn from diverse instruction structures. In terms of the motion reasoning task dataset, it provides longer contextual information to the model during both training and inference, with a minimum length of three turns for each conversation. Here, the LLM was asked to add some general questions about human motion potentially asked by the user, and the assistant provided clarity on them. Finally, for image reasoning, a subset of the LLaVA Instruct 150k dataset [20] was employed. At least 100 samples from each dataset were reserved as the evaluation/test set. Overall, the total number of samples generated for the SHAF dataset is reported in Table I.

TABLE I
NUMBER OF SAMPLES IN EACH CATEGORY OF THE SHAF DATASET.

Task	# of Samples
Motion to Image	2,300
Image to Motion	3,668
Motion Reasoning	4,828
Image Reasoning	3,500
Total	14296

V. EXPERIMENT & RESULTS

A. Experiment Setup

In addition to our SHAF dataset, HumanML3D dataset[16] was also utilized for training text and motion alignment during the pre-training (Pre) and for multimodal alignment during the supervised fine-tuning (SFT) stage. HumanML3D contains 3D human motion clips with corresponding textual descriptions. HumanML3D is designed for training models that bridge the gap between text and human motion. The dataset contains 14,616 text-motion pairs, sampled at 20 frames per second. On the other hand, for training text and image alignment in the pre-training stage, we used the text-to-image-2M [21] dataset that has 2M unique pairs of synthetically generated images from various GenAI models. HumanML3D and text-to-image-2M form the basis of the pre-training stage of the SHAF model. They are essential to learn the embeddings of the newly added codes for image and motion into the model. The SHAF datasets introduced in Section IV were employed in the supervised fine-tuning stage. We adopt Llama 3.2 (3B parameters) [22] as our LLM backbone due to its strong performance and relatively small parameter size. The hyperparameters for training model, including the learning rate of $1e^{-6}$, the maximum gradient norm of 1, and the effective batch size of 128, are chosen through our empirical experiments.

B. Evaluation Metrics

For evaluating the quality of generated motion and image, we use **Frechet Inception Distance (FID)** [23] that measures how close the generated motions are to real motions by comparing feature distributions. Generation Diversity is measured using **Diversity** [23] metric, which looks at how varied the generated motions are by comparing the variance between input and output features. The alignment between the text input and motion output is evaluated using **R-Precision (R@k)**[14]. This metric measures how well a motion matches the input text by checking if the correct motion is among the top k results. For evaluating text and image alignment, we utilize **Clip Score**[24], which computes the distance between the texts and images in a shared embedding space. To evaluate the text captions or reasoning outputs of our model, we adopt BLEU, ROUGE, CIDEr, and BertScore, as they are commonly used metrics in this domain:

- **BLEU** [25]: Compares n-gram overlaps between generated captions and reference captions. Higher is better.
- **ROUGE** [26]: Focuses on recall-based overlap between generated and reference captions. Higher is better.
- **CIDEr** [27]: Checks how well the generated captions agree with multiple human references. Higher is better.
- **BertScore** [28]: Uses BERT embeddings to measure semantic similarity between generated and reference captions. Higher is better.

C. Results and Discussion

The experimental results reported in Table II and Table III evaluate the effectiveness of our proposed multi-turn multimodal approach against relevant models specialized in text-to-motion and motion-to-text tasks. Meanwhile, Table IV and Table V present the model's performance on motion-to-image, image-to-motion, image-to-text, text-to-image, motion reasoning, and image reasoning tasks, establishing the first benchmark on our newly created SHAF dataset.

Text to Motion Table II presents experimental results on the text-to-motion task on the testing set of the HumanML3D dataset. The experimental results displayed in Table II demonstrate that the proposed SHAF model achieves a good balance between motion quality, accuracy in retrieving, and diversity, although both SHAF(Pre) and SHAF(SFT) have lower R-Precision values compared to state-of-the-art models like MotionGPT [13]. The lower R-precision scores could be because the model ignores the text conditioning and is more focused on generating motion closer to the ground truth distribution. This points towards a weaker alignment between text input and motion output. In terms of FID score, SHAF significantly outperforms all other models, as seen in the Pre-trained version, which attains the FID score of 0.0003. This result indicates that SHAF produces motions significantly more in line with real data distributions, even when SHAF handles more than one modality compared to others, and covers a higher number of tasks. In terms of motion diversity, SHAF (SFT) achieves a Diversity score of 9.036, remaining competitive though slightly below other transformer-based models such as T2M-GPT (9.761) and MotionGPT (SFT) (9.528).

Overall, the experimental results suggest that SHAF delivers performance comparable to models specialized for this task, particularly in FID and Diversity, while additionally supporting multiple tasks beyond text-to-motion and incorporating vision as an extra modality.

Motion to Text The experimental results on the testing set of the HumanML3D dataset are reported in Table III. SHAF model attains competitive performance across the evaluation metrics. While scores based on exact lexical overlap, such as BLEU and ROUGE, are modest (BLEU@1 of 35.0 and BLEU@4 of 1.7 for SHAF(Pre) and BLEU@1 of 32.8 and BLEU@4 of 1.5 for SHAF(SFT)), these measures primarily reflect the quality of surface-level similarity of reference captions during inference and are not always indicative of semantic alignment between the predictions and the references. Notably, SHAF surpasses state-of-the-art models on metrics that evaluate semantic similarity, with a CIDEr score of 42.9 and a BertScore of 88.2. These empirical results demonstrate that SHAF can generate captions that capture the intended meaning of motions with high fidelity, yielding descriptions that align more closely with human interpretability, even when lexical forms diverge from the references.

The results presented in Table III also indicate that SHAF (SFT) achieves slightly lower performance than SHAF (Pre)

TABLE II

COMPARISON OF MODELS ON TEXT-TO-MOTION TASK. → MEANS THE CLOSER TO REAL MOTIONS THE BETTER.

Methods	R-Precision↑			FID↓	Diversity→
	Top1	Top2	Top3		
Real	$0.511^{\pm.003}$	$0.703^{\pm.003}$	$0.797^{\pm.002}$	$0.002^{\pm.000}$	$9.503^{\pm.065}$
TM2T [14]	$0.424^{\pm.003}$	$0.618^{\pm.003}$	$0.729^{\pm.002}$	$1.501^{\pm.017}$	$8.589^{\pm.076}$
T2M [7]	$0.457^{\pm.002}$	$0.639^{\pm.003}$	$0.740^{\pm.003}$	$1.067^{\pm.002}$	$9.188^{\pm.002}$
MotionDiffuse [6]	$0.491^{\pm.001}$	$\mathbf{0.681}^{\pm.001}$	$\mathbf{0.782}^{\pm.001}$	$0.630^{\pm.001}$	$9.410^{\pm.049}$
MDM [29]	$0.320^{\pm.005}$	$0.498^{\pm.004}$	$0.611^{\pm.007}$	$0.544^{\pm.044}$	$9.559^{\pm.086}$
MLD [30]	$0.481^{\pm.003}$	$0.673^{\pm.003}$	$0.772^{\pm.002}$	$0.473^{\pm.013}$	$9.724^{\pm.082}$
T2M-GPT [31]	$0.491^{\pm.003}$	$0.680^{\pm.003}$	$0.775^{\pm.002}$	$0.116^{\pm.004}$	$9.761^{\pm.082}$
MotionGPT (Pre) [13]	$0.435^{\pm.003}$	$0.607^{\pm.002}$	$0.700^{\pm.002}$	$0.160^{\pm.008}$	$9.411^{\pm.081}$
MotionGPT (SFT) [13]	$\mathbf{0.492}^{\pm.003}$	$0.681^{\pm.003}$	$0.778^{\pm.002}$	$0.232^{\pm.008}$	$\mathbf{9.528}^{\pm.071}$
SHAF (Pre)	$0.063^{\pm.04}$	$0.123^{\pm.048}$	$0.167^{\pm.055}$	$\mathbf{0.0003}^{\pm.00004}$	$8.920^{\pm.420}$
SHAF (SFT)	$0.064^{\pm.03}$	$0.126^{\pm.053}$	$0.173^{\pm.060}$	$0.0003^{\pm.00003}$	$9.036^{\pm.198}$

TABLE III

COMPARISON OF MODELS ON MOTION-TO-TEXT TASKS.

Methods	Bleu@1↑	Bleu@4↑	Rouge↑	CIDEr↑	BertScore↑
TM2T [14]	**48.9**	7.00	**38.1**	16.8	32.2
MotionGPT (SFT) [13]	48.2	**12.47**	37.4	29.2	32.4
SHAF (Pre)	35.0	1.7	33.7	**42.9**	**88.2**
SHAF (SFT)	32.8	1.5	30.6	36.14	87.6

TABLE IV

PERFORMANCE OF SHAF ON MOTION-TO-IMAGE AND IMAGE-TO-MOTION TASKS.

Methods	FID↓	Diversity↑	Euclidean Distance↓
	Motion-to-Image		
SHAF (SFT)	0.131	0.534	0.4556
	Image-to-Motion		
SHAF (SFT)	10.602	8.666	5.345

TABLE V

PERFORMANCE OF SHAF ON IMAGE-TO-TEXT, MOTION REASONING, AND IMAGE REASONING TASKS.

Methods	Bleu@1↑	Bleu@4↑	Rouge↑	CIDEr↑	BertScore↑
	Image-to-Text				
SHAF (SFT)	26.2	0.5	28.9	37.7	88.0
	Motion Reasoning				
SHAF (SFT)	42.2	7.2	43.1	32.2	91.4
	Image Reasoning				
SHAF (SFT)	54.0	9.5	57.4	27.25	93.1

across evaluation metrics. One possible explanation is that SHAF (SFT) is trained to balance multiple modalities and adapt to a wider range of tasks, whereas SHAF (Pre) is more specialized.

Motion to Image The experimental results in Table IV show that SHAF (SFT) achieves a low FID score of 0.131 and a low Euclidean distance score, providing evidence that the images generated by SHAF (SFT) are highly similar to real images in both visual quality and distributional properties. These results also imply that the output of the model is determined not just by sharpness and clarity but by statistical equivalence to real images as well. Indeed, the low Diversity score suggests that the images are closely bound together inside the embedding space of the encoder. One reason for this is the presence of a human, almost in every image, with slight changes in context.

Image to Motion Table IV presents the performance of SHAF (SFT) in image-to-motion generation task. An FID score of 10.602 indicates that the generated motions exhibit some similarity to real image distributions, though the margin for improvement remains. The Euclidean distance score of 5.345 further reflects a moderate alignment between the generated motions and ground-truth sequences, underscoring the model's capacity to capture essential motion characteristics. Indeed, the Diversity score of 8.666 demonstrates the model's effectiveness in producing varied outputs, thereby reducing redundancy across samples. SHAF provides reasonable performance, besides having only 2300 samples for training this task (the lowest among tasks in the SHAF dataset).

Text to Image and Image to Text: The text-to-image task was evaluated using the Clip Score. Within this task, SHAF(SFT) achieves a Clip score of 30. Given that the prompts in the test subset [21] were quite detailed, the model can comprehend multiple concepts and generate relevant images from the provided text. For the image-to-text task, the results are reported in Table V. The BLEU scores indicate that exact word-for-word reproduction of the input sequence is not emphasized, but the model attains superior results on semantic-oriented metrics, achieving a CIDEr score of 37.7 and a BertScore value of 88.0. These outcomes highlight the model's capability in capturing the intrinsic content of images and in producing captions that are contextually coherent and semantically faithful.

Motion Reasoning SHAF obtains high performance accuracy in multi-turn motion reasoning tasks as observed in

Table V. The generated responses are consistently relevant, coherent, and well-aligned with the expected outputs. Beyond close reproduction of the input, a high degree of semantic understanding is demonstrated, reflecting the model's notable capability to capture contextual details across diverse dialogue interactions.

Image Reasoning Table V presents the results in the image reasoning task. The model achieves a Bleu@1 score of 54.0 and a Bleu@4 score of 9.5, reflecting its capability to identify a high percentage of unigram overlaps between reference captions. However, the model's capability to generate longer and semantically coherent n-gram sequences seems to be limited. The Rouge score of 57.4 indicates that SHAF (SFT) exhibits good competence in producing outputs marked by high lexical overlap with reference texts. The model also achieves a CIDEr score of 27.25, reflecting a moderate degree of compatibility with human-written captions. Particularly, SHAF (SFT) attains a high BertScore of 93.1, reflecting a strong semantic similarity between generated content and references despite the occurrence of instances where exact lexical correspondences are less common.

D. Example Outputs

This section presents examples generated by SHAF. Fig. 2 and 3 illustrate conversational scenarios, where green boxes denote user inputs and red boxes indicate system-generated outputs. These examples demonstrate SHAF's capability to manage multi-turn conversations involving different input/output modalities across different contexts.

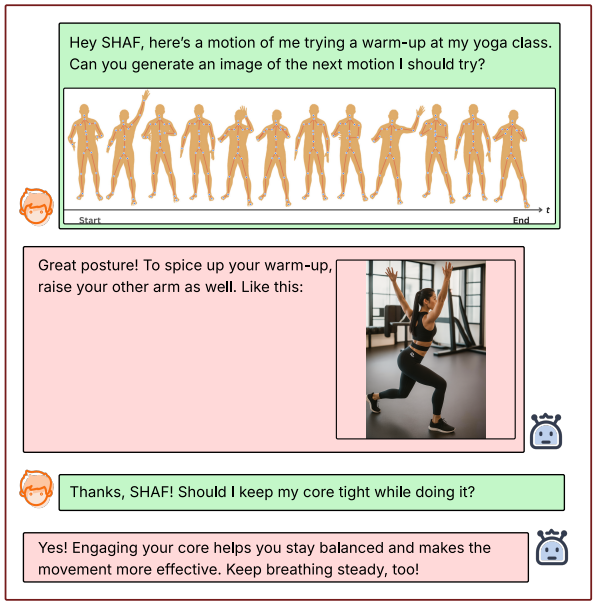

Fig. 2. A multi-turn conversation between the user and SHAF in the motion-to-image task.

As a proof of concept, Fig. 4 illustrates how the SHAF model can be implemented in an HRI setting to support humans in their well-being practices. The motion represented by a skeleton and the image displayed on the robot tablet were generated in response to human verbal input, *"Could you show me how to throw a ball with force?"*. Throughout a

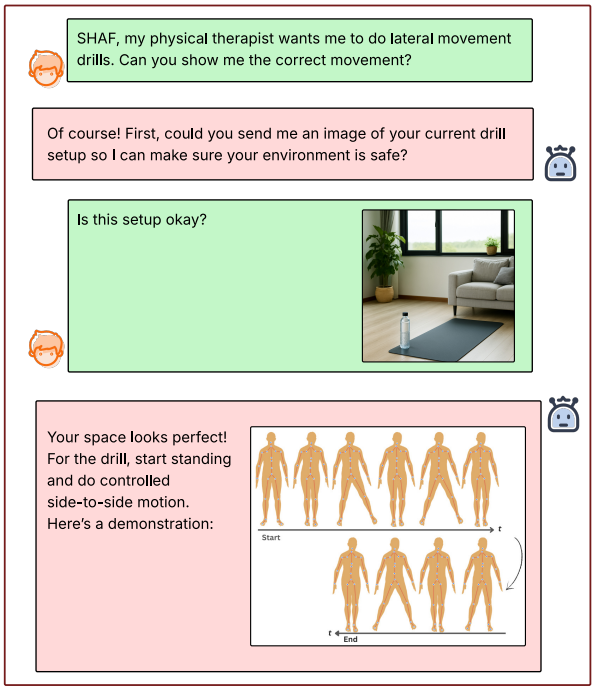

Fig. 3. A multi-turn conversation between the user and SHAF in the image-to-motion task.

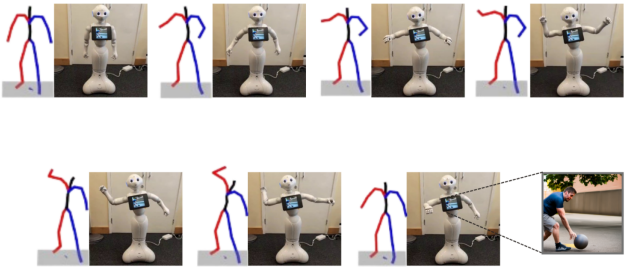

Fig. 4. The robot responds to a human input with a generated motion and a generated image of throwing a ball with force with the right hand. The generated human motion was converted into the robot's motion. The generated image is displayed on the robot's tablet while it performs the motion.

mapping module introduced in [8], generated motion was transferred into the robot motion space, being the robot gesture. Although the Pepper robot is limited to imitating the generated upper body motion, the SHAF model can be flexibly deployed on other humanoid robot platforms, virtual assistants, or game agents to support users in their well-being exercises and other contexts in daily human-AI interaction scenarios. While a comprehensive subjective evaluation in real-world HRI settings remains future work, this demonstration illustrates the feasibility of integrating SHAF into such pipelines.

VI. Conclusions

This paper introduces SHAF, a small multimodal large language model designed to unify text, image, and human motion within a single framework. By representing image and motion modalities as discrete token sequences through vector quantization, SHAF extends a text-only large language model to support multimodal conversations without architectural modifications of the LLM, thereby maintaining

efficiency for low-resource platforms. SHAF was pre-trained and fine-tuned on a variety of multimodal tasks, including text-to-motion, motion-to-text, text-to-image, and image-to-text. To further enhance interaction, we introduced several new tasks and developed dedicated datasets to train SHAF in such multi-turn multimodal settings. Experimental results demonstrate SHAF's ability to generate realistic motions with low FID, semantically accurate captions with strong CIDEr and BERTScore values, and diverse outputs compared to prior models. In addition, our SHAF dataset facilitates alignment between motion and image modalities.

While SHAF has demonstrated its potentials, the empirical results also highlight several areas for further investigation. For instance, SHAF places less emphasis on text conditioning during text-to-motion tasks. Furthermore, the performance of SHAF is constrained by the lossy nature of image and motion quantization techniques. Consequentially, our future works will explore improved quantization methods, cost functions to enhance overall performance, larger and more diverse multimodal datasets, and scaling SHAF to even smaller resource-efficient LLM models to further enhance its applicability for virtual AI agents and socially assistive robots.

REFERENCES

[1] H. Su, W. Qi, J. Chen, C. Yang, J. Sandoval, and M. A. Laribi, "Recent advancements in multimodal human–robot interaction," *Frontiers in Neurorobotics*, vol. 17, p. 1084000, 2023.

[2] J. Wang, H. Jiang, Y. Liu, C. Ma, X. Zhang, Y. Pan, M. Liu, P. Gu, S. Xia, W. Li, Y. Zhang, Z. Wu, Z. Liu, T. Zhong, B. Ge, T. Zhang, N. Qiang, X. Hu, X. Jiang, X. Zhang, W. Zhang, D. Shen, T. Liu, and S. Zhang, "A comprehensive review of multimodal large language models: Performance and challenges across different tasks." *CoRR*, vol. abs/2408.01319, 2024. [Online]. Available: http://dblp.uni-trier.de/db/journals/corr/corr2408.html#abs-2408-01319

[3] C. Oertel, G. Castellano, M. Chetouani, J. Nasir, M. Obaid, C. Pelachaud, and C. Peters, "Engagement in human-agent interaction: An overview," *Frontiers in Robotics and AI*, vol. 7, p. 92, 2020.

[4] Y. Goutsu and T. Inamura, "Linguistic descriptions of human motion with generative adversarial seq2seq learning," in *2021 IEEE International conference on robotics and automation (ICRA)*. IEEE, 2021, pp. 4281–4287.

[5] J. Zhan, J. Dai, J. Ye, Y. Zhou, D. Zhang, Z. Liu, X. Zhang, R. Yuan, G. Zhang, L. Li, H. Yan, J. Fu, T. Gui, T. Sun, Y.-G. Jiang, and X. Qiu, "AnyGPT: Unified multimodal LLM with discrete sequence modeling," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 9637–9662. [Online]. Available: https://aclanthology.org/2024.acl-long.521/

[6] M. Zhang, Z. Cai, L. Pan, F. Hong, X. Guo, L. Yang, and Z. Liu, "Motiondiffuse: Text-driven human motion generation with diffusion model," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 6, pp. 4115–4128, 2024.

[7] C. Guo, S. Zou, X. Zuo, S. Wang, W. Ji, X. Li, and L. Cheng, "Generating diverse and natural 3d human motions from text," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 5142–5151.

[8] N. T. V. Tuyen and O. Celiktutan, "It takes two, not one: context-aware nonverbal behaviour generation in dyadic interactions," *Advanced Robotics*, vol. 37, no. 24, pp. 1552–1565, 2023.

[9] M. Spitale, M. Axelsson, and H. Gunes, "Vita: A multi-modal llm-based system for longitudinal, autonomous and adaptive robotic mental well-being coaching," *ACM Transactions on Human-Robot Interaction*, vol. 14, no. 2, pp. 1–28, 2025.

[10] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36. Curran Associates, Inc., 2023, pp. 34 892–34 916. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/6dcf277ea32ce3288914faf369fe6de0-Paper-Conference.pdf

[11] M. Yang, S.-J. Chen, J. Xie, and J. Hansen, "Bridging the modality gap: Softly discretizing audio representation for llm-based automatic speech recognition," 2025. [Online]. Available: https://arxiv.org/abs/2506.05706

[12] M. Maaz, H. Rasheed, S. Khan, and F. Khan, "Videogpt+: Integrating image and video encoders for enhanced video understanding," 2024. [Online]. Available: https://arxiv.org/abs/2406.09418

[13] B. Jiang, X. Chen, W. Liu, J. Yu, G. Yu, and T. Chen, "Motiongpt: Human motion as a foreign language," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[14] C. Guo, X. Zuo, S. Wang, and L. Cheng, "Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts," in *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV*. Springer-Verlag, 2022, p. 580–597.

[15] B. Jiang, X. Chen, C. Zhang, F. Yin, Z. Li, G. Yu, and J. Fan, "Motionchain: Conversational motion controllers via multimodal prompts," in *European Conference on Computer Vision*. Springer, 2024, pp. 54–74.

[16] C. Guo, S. Zou, X. Zuo, S. Wang, W. Ji, X. Li, and L. Cheng, "Generating diverse and natural 3d human motions from text," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 5152–5161.

[17] Y. Ge, Y. Ge, Z. Zeng, X. Wang, and Y. Shan, "Planting a seed of vision in large language model," 2023. [Online]. Available: https://arxiv.org/abs/2307.08041

[18] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," 2023. [Online]. Available: https://arxiv.org/abs/2301.12597

[19] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford, *et al.*, "Gpt-4o system card," *arXiv preprint arXiv:2410.21276*, 2024.

[20] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *NeurIPS*, 2023.

[21] zk, "text-to-image-2m (revision e64fca4)," 2024. [Online]. Available: https://huggingface.co/datasets/jackyhate/text-to-image-2M

[22] Meta AI. (2024, Sept. 25) Llama 3.2: Revolutionizing edge ai and vision with open, customizable models. Accessed: 2025-09-03. [Online]. Available: https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/

[23] C. Guo, X. Zuo, S. Wang, S. Zou, Q. Sun, A. Deng, M. Gong, and L. Cheng, "Action2motion: Conditioned generation of 3d human motions," in *Proceedings of the 28th ACM International Conference on Multimedia*, ser. MM '20. ACM, Oct. 2020, p. 2021–2029. [Online]. Available: http://dx.doi.org/10.1145/3394171.3413635

[24] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, and Y. Choi, "Clipscore: A reference-free evaluation metric for image captioning," 2022. [Online]. Available: https://arxiv.org/abs/2104.08718

[25] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, P. Isabelle, E. Charniak, and D. Lin, Eds. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, July 2002, pp. 311–318. [Online]. Available: https://aclanthology.org/P02-1040/

[26] Y. Liu, A. Medlar, and D. Glowacka, "Rogue: A system for exploratory search of gans," in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 3278–3282. [Online]. Available: https://doi.org/10.1145/3477495.3531675

[27] R. Vedantam, C. L. Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 4566–4575.

[28] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," 2020. [Online]. Available: https://arxiv.org/abs/1904.09675

[29] G. Tevet, S. Raab, B. Gordon, Y. Shafir, D. Cohen-or, and A. H. Bermano, "Human motion diffusion model," in *The Eleventh International Conference on Learning Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=SJ1kSyO2jwu

[30] X. Chen, B. Jiang, W. Liu, Z. Huang, B. Fu, T. Chen, and G. Yu, "Executing your commands via motion diffusion in latent space," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 000–18 010.

[31] J. Zhang, Y. Zhang, X. Cun, Y. Zhang, H. Zhao, H. Lu, X. Shen, and S. Ying, "Generating human motion from textual descriptions with discrete representations," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 14 730–14 740.