Guided Attention for Interpretable Motion Captioning

Karim Radouane, Julien Lagarde, Sylvie Ranwez, Andon Tchechmedjiev

Guided Attention for Interpretable Motion Captioning

Karim Radouane[1]
karimradouane39@gmail.com
Julien Lagarde[2]
julien.lagarde@umontpellier.fr
Sylvie Ranwez[1]
sylvie.ranwez@mines-ales.fr
Andon Tchechmedjiev[1]
andon.tchechmedjiev@mines-ales.fr

[1] EuroMov Digital Health in Motion, University of Montpellier, IMT Mines Ales, Ales, France
[2] EuroMov Digital Health in Motion, University of Montpellier, IMT Mines Ales, Montpellier, France

Abstract

Diverse and extensive work has recently been conducted on text-conditioned human motion generation. However, progress in the reverse direction, motion captioning, has seen less comparable advancement. In this paper, we introduce a novel architecture design that enhances text generation quality by emphasizing interpretability through spatio-temporal and adaptive attention mechanisms. To encourage human-like reasoning, we propose methods for guiding attention during training, emphasizing relevant skeleton areas over time and distinguishing motion-related words. We discuss and quantify our model's interpretability using relevant histograms and density distributions. Furthermore, we leverage interpretability to derive fine-grained information about human motion, including action localization, body part identification, and the distinction of motion-related words. Finally, we discuss the transferability of our approaches to other tasks. Our experiments demonstrate that attention guidance leads to interpretable captioning while enhancing performance compared to higher parameter-count, non-interpretable state-of-the-art systems. The code is available at: https://github.com/rd20karim/M2T-Interpretable.

1 Introduction

Motion-to-language datasets such as KIT-ML [11] have garnered significant interest in motion-language applications. The motion captioning task is closely related to video captioning. However, human pose representation reduces the amount of data that needs to be processed and helps the model focus on the most important aspects of human motion, enabling more effective descriptions of human activities. In this context, the motion captioning task aims to generate natural language descriptions from sequences of human poses. Compared to the significant work done in vision-based captioning, which has seen different interpretable approaches identifying zones in images or videos that most contribute to the captions [15, 17], interpretability has been relatively less emphasized in motion captioning methods [5, 12].

Nonetheless, an interpretable model holds significant importance in ensuring model reliability, offering explainable predictions for users, understanding model limitations. In this paper, taking inspiration from captioning approaches in vision, we devise a novel interpretable motion captioning system incorporating spatio-temporal and adaptive attention mechanisms. Moreover, the attention is guided to better match the human perception. To the best of our knowledge, this is the first interpretable system for motion captioning at both spatial and temporal levels. We demonstrate the performance of our interpretable captioning approach on available benchmarks: KIT Motion-Language Dataset [11] and HumanML3D [4], using common metrics, in alignment with current best practices for this task. Our contributions are summarized as follows:

- We propose an interpretable architecture design that offers a transparent reasoning process, mimicking human-like attention perception and analysis, in contrast to black box approaches.
- Novel formulation of adaptive gating mechanism, along with spatio-temporal attention in the context of human motion captioning.
- We propose methodologies for adaptive and spatial attention supervision, aligned with our human skeleton partitioning method, which divides the body into six parts. This partitioning integrates separated local and global motion representations, aiming to enhance interpretability.
- We conduct extensive evaluations and analysis of our model's interpretability, involving qualitative assessments through attention maps and quantitative analyses utilizing specific proposed histograms and density distributions. Moreover, we demonstrate the capacity to leverage resulting model interpretability for action localization, body part identification, and distinguishing motion-words.

2 Related Work

Motion Captioning. The first approach on the KIT-ML dataset [11] was introduced by [12] using a bidirectional LSTM. Later systems mainly focused on motion generation [2, 7, 10], but motion captioning has seen a resurgence with the introduction of HumanML3D [4]. This dataset was firstly used for motion captioning by [5], which proposes learning motion tokens using VQ-VAE that are then mapped to word tokens through a Transformer [16]. The results of this approach was not high specifically on KIT-ML (BLEU@4 =18.4%). Then, [13] slightly improved text generation results using a combination of Multilayer Perceptron (MLP) and Gated Recurrent Unit (GRU). Multitask learning was introduced in MotionGPT [6], but the disparity in tasks prevents fair comparisons. However, this strategy negatively impacted motion captioning, yielding a low BLEU@4 score of 12.47% on HumanML3D and no reported results on the KIT-ML dataset.

Adaptive attention. Attending to the input (*e.g.,* image) for the generation of non-visual words can be misleading and degrading to the performance of attention networks. To alleviate this problem, [9] propose a formulation for a learnable gate variable β. The variable β is learned to choose either to rely on the image features or only on the context of language

generation through the visual sentinel vector. For motion captioning, this is particularly relevant as only specific words ("walks", "throw", etc.) need to access motion input information during prediction time, in contrast to non-motion words ("a", "the", etc.).

Guided attention. Attention mechanisms can focus on incorrect areas of the input or on regions with a strong bias that aren't particularly meaningful for human interpretation. To mitigate these limitations [8] propose attention supervision, a technique aimed at improving the performance and accuracy of image captioning models. This approach leads to more relevant attention maps, thereby enhancing interpretability. In the context of video captioning, spatial guiding of attention has also been shown to improve captioning performance [13].

3 Methods

We first present the general model architecture for our captioning approach (Section 3.1), followed by more in-depth presentations of our formulations for spatial and adaptive attention, as well as our attention guidance methodology (Section 3.2).

3.1 Architecture design for motion captioning

Our model, summarized in Figure 1, is composed of an encoder block, a spatio-temporal attention block and a text generation/decoder block incorporating an adaptive attention mechanism.

Let $\mathbf{X} \in \mathbb{R}^{T_x \times J \times D}$ be the input sequence of motion features of T_x time steps, where J is the number of joints in the skeleton and D is the number of spatial dimensions. We note by X_k the 3D joints positions and V_k their corresponding velocities at frame time k.

Skeleton partitioning. We group the joints in 6 body-parts: Left Arm, Right Arm, Torso, Left Leg, Right Leg, Root. We convert the global coordinates to root-relative coordinates, except for the root itself, which describes the global trajectory of the motion. X_{ik} denotes the group of joints of part i for every frame k as described in Figure 1.

Encoder. Each of the six body parts is embedded by two linear layers followed by *tanh* activations, as illustrated in Figure 1. Each linear layer (FC) encode positions X_{ik} and velocities V_{ik} separately. The final embedding P_{ik} for a given part i and frame k is the concatenation of the position and velocity embeddings. We note by P the frame-level motion features of all human body parts. $P \in \mathbb{R}^{T_x \times a \times h_{enc}}$ where h_{enc} the dimension of the final output encoder and $a = 6$ is the number of body parts $(P = Enc(X))$.

Decoder. We adopt a two-LSTM decoder configuration, a *Bottom LSTM* for learning attention weights and language context and a *Top LSTM* for final word generation based on the relevant information extracted from language and motion. We note by $\mathbf{y} = (y_1, \ldots, y_{T_y})$, $y_i \in \mathbb{R}^{K_y}$ the sequence of words describing the motion. Let $h_t \in \mathbb{R}^{h_{dec}}$ be the decoder hidden state of the bottom LSTM for a word w_t in the sequence and $\bar{h}_t$ for the Top LSTM. We note by K_y the size of the target vocabulary. T_x and T_y are respectively the length of the motion sequence and the length of its description. The decoder *Dec* is used to predict the next word y_t given the adaptive context vector described by $\bar{c}_t$ and the previous word y_{t-1} and the bottom hidden state h_t.

$$p(y_t \mid \{y_1, \cdots, y_{t-1}\}, \bar{c}_t) = Dec(y_{t-1}, h_t, \bar{c}_t) \tag{1}$$

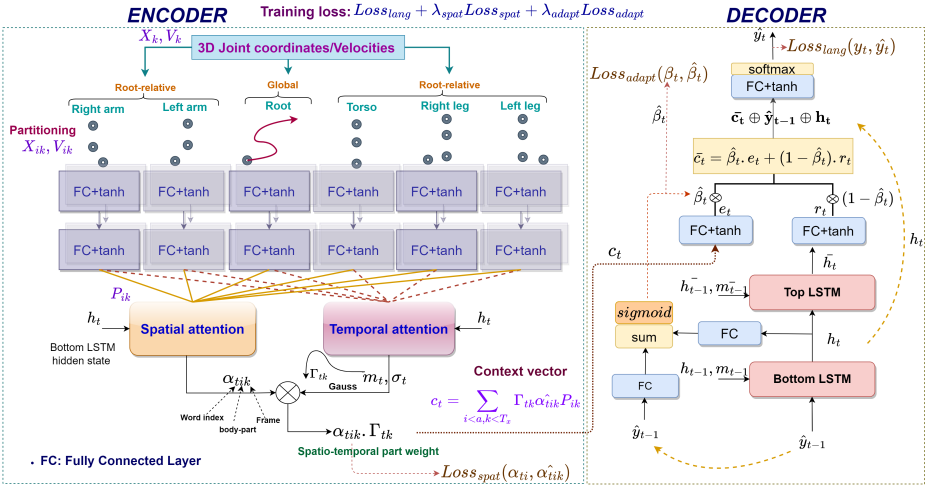


Figure 1: The encoder branch encodes frame-wise part-based motion representations from joint positions (X_{ik}) and velocities (V_{ik}), while the decoder branch takes as input (previous token y_{t-1}, previous state (h_{t-1}, m_{t-1})) and estimates the relative importance ($\hat{\beta}_t$ gate) of motion information to consider for word prediction $\hat{y}_t$. Spatial ($\hat{\alpha_{tik}}$) and temporal attention (Γ_{tk}) are computed from encoded part embeddings P_{ik} and h_t. The spatio-temporal weights are used to compute the context vector c_t which is then passed to the decoder adaptive gate. $Loss_{lang}$ the cross entropy between predicted, and target words is the main loss. We propose to guide spatial and adaptive attention with $Loss_{spat}$ and $Loss_{adapt}$.

The context vector c_t is computed by a spatial-temporal attention mechanism, where temporal attention determines **when** to focus attention, and spatial attention determines **where** to focus in the body part graph. In the following, we note by $P^* \in \mathbb{R}^{h_{enc} \times a \times T_x}$ the permutation of $P \in \mathbb{R}^{T_x \times a \times h_{enc}}$.

Temporal attention. The temporal weights are computed from extracted motion features P^* and the current decoder hidden state h_t.

$$z_t = w_h^T \tanh(W_p P^* + ep(W_h h_t)) \tag{2}$$

$$\gamma_t = \text{softmax}(z_t) \tag{3}$$

Here $W_p \in \mathbb{R}^{d \times h_{enc}}$,$W_h \in \mathbb{R}^{d \times h_{dec}}$ and $w_h \in \mathbb{R}^{d \times 1}$ are learnable parameter, ep is an expansion operator mapping to $d \times a \times T_x$, and a the number of body parts. Moreover, γ_t is the temporal attention weights for the word generated at time t. With the above formulation, we often have discontinuities in the attention maps, yet such discontinuities are undesired, as the action happens continuously in a given frame range. The distribution of attention weights for a *particular motion word* can be modelled as a Gaussian distribution with a learnable mean and standard deviation. The mean m_t and standard deviation σ_t are computed from the previous temporal attention weights γ_{tk}, which are replaced by Γ_{tk} during training in this case (See Figure 1). Intuitively, the mean m_t will approximately represent the center time of action duration described by a motion word w_t, and the spread of the distribution approximately corresponds to the duration of the action.

$$\Gamma_{tk} = \exp\left(-\frac{(k-m_t)^2}{2{\sigma_t}^2}\right) \tag{4}$$

Spatial attention. Spatial weights are computed for each body part (Torso, left/right arm, left/right leg) as follows:

$$s_t = w_s^T \tanh\left(W_{p_s} P^* + ep(W_{h_s} h_t)\right) \tag{5}$$

$$\alpha_t = \text{softmax}\left(s_t\right) \tag{6}$$

Here $s_t \in \mathbb{R}^a$. The learnable parameters are $W_{p_s} \in \mathbb{R}^{d \times h_{enc}}$,$W_{h_s} \in \mathbb{R}^{d \times h_{dec}}$ and $w_s \in \mathbb{R}^{d \times 1}$. We note by $\alpha_{t,m,k}$ the spatial attention score for part m of the skeleton graph at frame k for the word generated at time t. Thus, explicitly $\alpha_t = [\alpha_{t,1,1}, \alpha_{t,1,2}, \cdots, \alpha_{t,a,T_x}]$.

Adaptive attention. Non-motion words, particularly grammatical words, do not carry any information about the movement. Consequently, we propose to learn a gating variable $\hat{\beta}_t$ to decide the proportion to which to use language context over motion features.

$$\hat{\beta}_t = sigmoid(W_b^h.h_t + W_e.(E\hat{y}_{t-1})) \tag{7}$$

Where $W_b^h \in \mathbb{R}^{1 \times h_{dec}}, W_e \in \mathbb{R}^{1 \times d_{emb}}$ are learnable matrices. $E \in \mathbb{R}^{d_{emb} \times K_y}$ refers to embedding matrix of target words. The gating variable depends on the hidden state, which encodes residual information about generated words up to the time step t, as well as on the embedding of the previous word, as detailed in Equation (7).

Context vector. The context vector is derived by weighting the motion features with spatial and temporal attention weights, and averaging across the frame-time dimension 8.

$$c_t = \sum_{k=1}^{T_x} \sum_{i=1}^{a} \Gamma_{tk} \alpha_{tik} P_{ik} \tag{8}$$

The motion c_t and language information $\bar{h}_t$ are embedded into the same space through an linear layer with *tanh* activation (for bounded values in [-1,1]), giving $\mathbf{e_t}$ and $\mathbf{r_t}$ respectively.

Adaptive context vector. Given by Equation (9). When $\hat{\beta}_t = 1$ the model uses full motion information and when $\hat{\beta}_t$ is close to 0 the model relies more on language structure.

$$\bar{c}_t = \hat{\beta}_t.e_t + (1 - \hat{\beta}_t).r_t \tag{9}$$

Finally, the probability outputs are computed as in Equation (10), similarly to previous work on video captioning [14], except we include the bottom hidden state. This ensures that the language information of previously generated words is always present, which is important for correct syntax, even for motion words (e.g. jogs, jogging...).

$$p(\hat{y}_t \mid \hat{y}_{1:t-1}, \hat{c}_t) = softmax(\tanh\left(W_f.concat([\hat{c}_t; \hat{y}_{t-1}; h_t])\right)) \tag{10}$$

3.2 Spatial and adaptive attention supervision

To our knowledge, simultaneous supervision of attention mechanisms with an adaptive gate and spatial attention has never been applied to captioning tasks, particularly motion captioning. Below, we provide a formal definition of how the losses for attention supervision are formulated.

Language loss. The standard loss for motion-to-text generation is defined as the cross entropy between the target and predicted words:

$$Loss_{lang} = -\sum_{t=0}^{T_y-1} y_t \cdot \log(\hat{y}_t) \tag{11}$$

Adaptive attention loss. To build a ground truth for adaptive attention, we define mapping rules to distinguish between motion words, action verbs and qualifying adjectives (*e.g.,* walk, circle, slowly) from non-motion words (*e.g.,* of, person). We assign $\beta_t = \mathbf{1}$ for motion words and $\beta_t = \mathbf{0}$ for non-motion words (See Supp.C).

$$Loss_{adapt} = -\sum_{t=0}^{T_y-1} \beta_t \log(\hat{\beta}_t) + (1-\beta_t)\log(1-\hat{\beta}_t) \tag{12}$$

Spatial attention loss. The predicted attention score is $\hat{\alpha_{tik}}$ for a given word w_t and part i of the source motion at the frame k. The loss is formulated in Equation (13), where N_y is a normalization factor that count the number of supervised words for a given target description y (See Supp.C for attention guidance strategy).

$$Loss_{spat} = -\frac{1}{N_y}\sum_{i,t,k} \alpha_{ti}\log(\hat{\alpha_{ti}}) + (1-\alpha_{tik})\log(1-\hat{\alpha_{tik}}) \tag{13}$$

Global loss. To define the global loss, we add the loss terms for spatial attention $loss_{spat}$, adaptive attention gate $loss_{adapt}$ guidance, respectively weighted by $\lambda_{spat}, \lambda_{adapt}$, to control their contributions.

$$Loss = loss_{lang} + \lambda_{spat}.loss_{spat} + \lambda_{adapt}.loss_{adapt} \tag{14}$$

4 Experiments

We consider the commonly used benchmarks KIT-ML [10] and the HumanML3D (HML3D) [4] (Dataset statistics in Supp.B). We conduct ablation studies on both datasets to determine the impact of adaptive and guided attention, followed by a detailed analysis of our model's interpretability.

Ablation Study. We configure a search space for $(\lambda_{spat}, \lambda_{adapt})$ and run the search using WandB [1]. Table 1 quantifies the impact of attention guidance. Due to space constraints, more results can be found in Supp.D, and additional detailed analysis regarding the effectiveness of our architecture components can be found in Supp.E).

Hyperparameters. For both KIT-ML and HumanML3D datasets, we set respectively the word embedding size and decoder hidden size to ($d_{emb} = 64, h_{dec} = 128$) and ($d_{emb} = 128$, $h_{dec} = 256$), respectively. Additionally, the output dimension of each fully connected layer FC_i is 128 for layer 1 and 64 for layer 2 in KIT-ML, and 256 for layer 1 and 128 for layer 2 in HumanML3D. After concatenation, we obtain 128 and 256 joint-velocity features per frame respectively for KIT-ML and HML3D.

Dataset	λ_{spat}	λ_{adapt}	BLEU@1	BLEU@4	CIDEr	ROUGE_L	BERTScore
KIT-ML	0	0	57.3	23.6	109.9	57.8	41.1
	0	3	56.3	22.5	108.4	56.5	39.8
	2	**3**	**58.4**	**24.4**	**112.1**	**58.3**	**41.2**
HML3D	0	0	69.3	24.0	58.8	54.8	38.7
	0	**3**	**69.9**	**25.0**	61.6	**55.3**	**40.3**
	2	3	69.2	24.4	**61.7**	55.0	**40.3**

Table 1: Results for different supervision modes, where $\lambda_{spat} = \lambda_{adapt} = 0$ represents the case without any attention guidance for comparison. The gate (adapt) and spatial (spat) supervision, perform well when used together on KIT-ML (small). For HumanML3D adaptive attention was always beneficial, but guided spatial attention slightly degraded exact matching scores (BLEU@4, ROUGE) compared to only adaptive attention (Detailed experimented values in Supp.D). The impact is more significant on the interpretability aspect Section 4.2.

Dataset	Model	BLEU@1	BLEU@4	ROUGE-L	CIDEr	BERTScore
KIT-ML	SeqGAN [5]	3.12	5.20	32.4	29.5	2.20
	TM2T [5]	46.7	18.4	44.2	79.5	23.0
	MLP+GRU [13]	56.8	**25.4**	**58.8**	**125.7**	**42.1**
	Ours-[spat+adapt](2,3)	**58.4**	24.7	57.8	106.2	41.3
	***Ours-[spat+adapt](2,3)**	**58.4**	24.4	58.3	112.1	41.2
HML3D	SeqGAN [5]	47.8	13.5	39.2	50.2	23.4
	TM2T [5]	61.7	22.3	49.2	**72.5**	37.8
	MLP+GRU[13]	67.0	23.4	53.8	53.7	37.2
	Ours-[adapt](0,3)	67.9	**25.5**	54.7	64.6	**43.2**
	***Ours-[adapt](0,3)**	**69.9**	25.0	**55.3**	61.6	40.3

Table 2: Text generation performance, assessed with beam size 2 as in [5], while * indicate a greedy search. Our model performs better than Transformer-based (TM2T) method on both datasets and on HumanML3D compared to MLP+GRU.

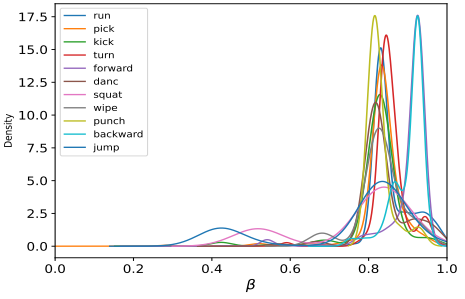
4.1 Evaluation and discussion

Table 2 presents the comparison to SOTA systems. Our approach performs significantly better than other state-of-the-art approaches without beam search on HML3D, including the Transformer TM2T. For KIT-ML dataset, MLP+GRU is slightly better than our approach in terms of NLP metrics. However, in terms of interpretability, our approach provides more information on the body parts involved in an action compared to MLP+GRU, which lacks spatial and adaptive attention. Therefore, in their case, the motion representation doesn't consider the skeleton graph structure and is always utilized for generating non-motion words that don't require motion information, which may lead to biased learning.

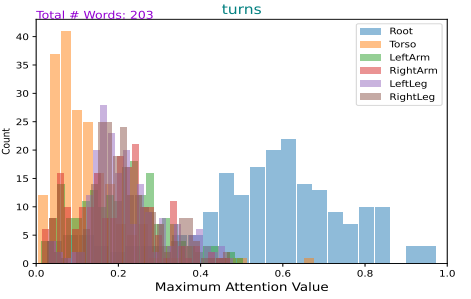
4.2 Interpretability analysis

In our context, interpretability is measured by the ability to establish a correspondence between learned attention mechanisms and human attention perception. In this section, we discuss the interpretability of learned attentions and how we can leverage interpretabiliy as illustrated in Figure 6. To demonstrate the role of each of the context vectors c_t and LSTMs hidden states ($\bar{h}_t$,h_t), we fix the $\hat{\beta}$ value at 1 and show a representative examples compared to adaptive gate in Table 3. Further analysis in Supp.E.

Spatial / Adaptive attention impact. When training a model without guiding adaptive



(a) With gate supervision, motion information is correctly used frequently for motion-words generation.



(b) Attention is frequently focused on relevant parts: e.g. on Root (global trajectory) for word "turns".

Figure 2: $\hat{\beta}$ test set density distribution for a few motion words stems on HumanML3D and the temporal maximum body-parts attention histogram for word "turn".

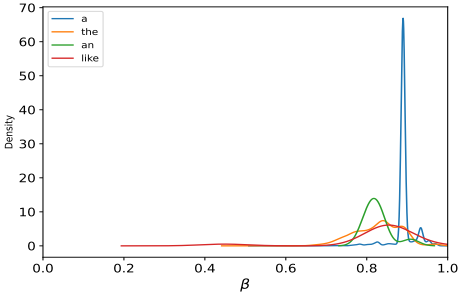
$\hat{\beta} = 1$	**Adaptive $\hat{\beta}$**	**Reference**
walks forward and sits down <eos>	a person walks forward turns around and sits down and gets back up and walk back <eos>	man walks forwards stops turns around and sits then gets up and walks back <eos>
jumping up and down in place <eos>	a person jumps up and down multiple times <eos>	someone jumps twice and looks down at the ground <eos>
punching boxing and moving hands around <eos>	a person is boxing with both hands <eos>	a person standing up is making boxing motions with their left and right arms <eos>

Table 3: Comparison of the prediction when setting $\hat{\beta} = 1$ and adaptive on HML3D-(0,3) using human motion samples involving different actions.

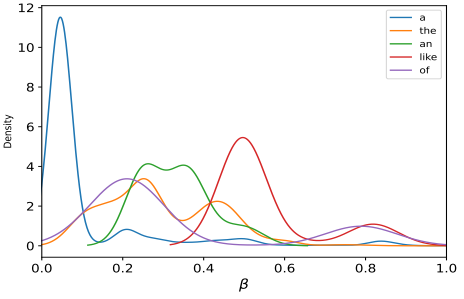
attention, we observe that $\hat{\beta}$ gate values frequently takes higher values for non-motion words (a:0.9, the:0.8) as illustrated in Figure 3a. This behavior degrades performance, as seen in Table 1 for both datasets. However, when we introduce adaptive gate supervision (cf. Figure 3b), the model more frequently assigns less weight $\hat{\beta}$ to non-motion words and begins to learn how to make decisions automatically when to use the context vector, as illustrated also in Figure title 7b, while guided spatial attention enhances the learned attention maps.

Body part identification. We can illustrate the effectiveness of our architecture in learning a correct body part association through spatio-temporal attention by viewing the density distribution for maximum attention across time per each body part for some motion words as illustrated in Figures 4 and 7 (Diverse examples in Supp.F).

Action localization. Another aspect that emerges from temporal Gaussian attention weights is action localization. The architecture shows ability to identify motion onset without temporal supervision. We can derive the action onset from spatio-temporal attention maps, as illustrated in Figure 5 where we also show their actual onset time.

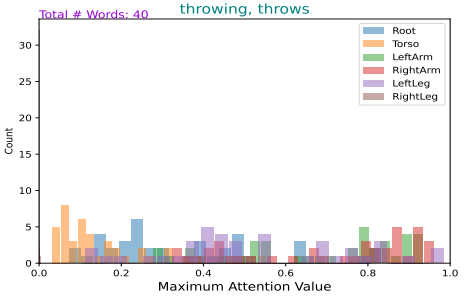


(a) Without gate supervision, decoder uses frequently motion information even for non-motion words (β frequently high).

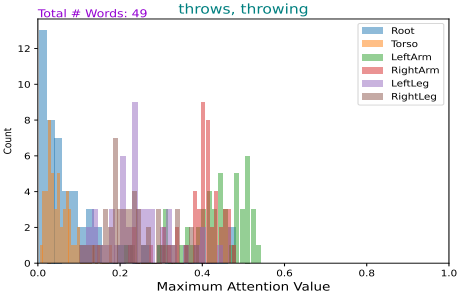


(b) With gate supervision, the decoder uses correctly more language context for non-motion words (β frequently small).

Figure 3: $\hat{\beta}$ density distribution over test set for some non-motion words (stemmed) on HumanML3D.



(a) Without spatial supervision, attention incorrectly focused on legs rather than arms for "throw" motion in some cases (left leg).



(b) With spatial supervision, spatial attention is always maximal on relevant part, for this example on the arms.

Figure 4: Effect of spatial supervision on HumanML3D across the entire test set for a given motion word (e.g. *throw*) (# Refer to number of the given motion words).

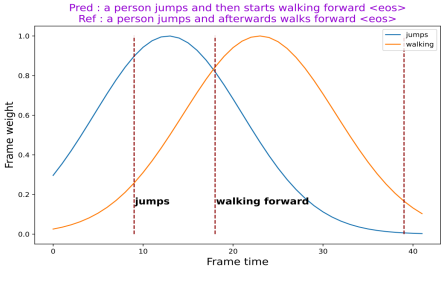


Figure 5: Temporal gaussian window displayed for different motion words given a prediction on KIT-ML.

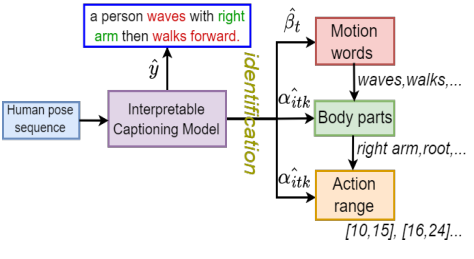
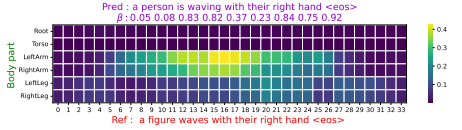
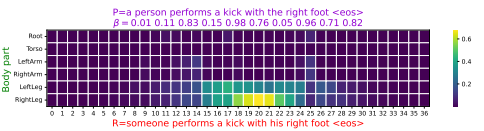


Figure 6: Interpretability use towards fine-grained captioning, based on spatial, temporal and adaptive attention scores.



(a) HML3D-(2,3), word *waving* in range $[4, 27]$.



(b) KIT-(2,3), word *kick* in range $[16, 26]$

Figure 7: Spatio-temporal attention maps for some words, with the color scale indicating attention score intensity per frame per body part. The model focalize correctly on relevant parts ((a).arms, (b).legs) at precise action timing and $\hat{\beta}$ values are semantically variable depending on the nature of predicted words as illustrated by the predictions in figures title (other examples in Supp.F

Transfer to adjacent tasks. Similar tasks can benefit from the proposed methodologies. In the context of skeleton based action recognition and localization, our proposed motion encoder and skeleton partitioning could be used to build an interpretable model. In a continuous stream, action segmentation tasks could be also cast as sequence to sequence learning, thus attention weights could be used to infer the action start/end times as an unsupervised learning. If the action time is available, these annotations could serve to supervise the spread of temporal weights, further enhancing the accuracy of action localization and spatio-temporal attention maps. Given an image, for each visual word in the caption, our spatial supervision could be transformed into maximizing the attention weights on relevant objects. Finally, the interpretability could be evaluated using the proposed density function for adaptive attention and histograms for attention distribution on spatial locations in other captioning context.

5 Conclusion

We have introduced guided attention with adaptive gate for motion captioning. After evaluating the influence of different weighting schemes for the main loss terms, we have found that our approach leads to interpretable captioning while improving performance. Interpretability is very important to consider when designing an architecture, it's gives insights on model capability to perform a true reasoning. This ensures the ability of generalizing instead of memorizing. The proposed model addressed the two challenges, given an interpretable result with accurate semantic captions. The model and proposed methodology can be transposed to other captioning tasks, such as supervision of spatial attention weights in action recognition tasks.

Acknowldgements

This work is supported by the Occitanie Region of France (Grant ALDOCT-001100 20007383) and the European Union's HORIZON Research and Innovation Programme (Grant 101120657, Project ENFIELD).

References

[1] Lukas Biewald. Experiment tracking with weights and biases, 2020. Software available from wandb.com.

[2] Anindita Ghosh, Noshaba Cheema, Cennet Oguz, Christian Theobalt, and Philipp Slusallek. Synthesis of compositional animations from textual descriptions. In *International Conference on Computer Vision (ICCV)*, pages 1376–1386, 2021.

[3] Yusuke Goutsu and Tetsunari Inamura. Linguistic descriptions of human motion with generative adversarial seq2seq learning. In *Proceedings of the International Conference on Robotics and Automation (ICRA)*, pages 4281–4287, 2021.

[4] Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5152–5161, June 2022.

[5] Chuan Guo, Xinxin Zuo, Sen Wang, and Li Cheng. Tm2t: Stochastic and tokenized modeling for the reciprocal generation of 3d human motions and texts. In *European Conference on Computer Vision (ECCV)*, 2022.

[6] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems*, 36, 2024.

[7] Angela S. Lin, Lemeng Wuk, Rodolfo Corona, Kevin Tai, Qixing Huang, and Raymond J. Mooney. Generating animated videos of human activities from natural language descriptions. In *Proceedings of the Visually Grounded Interaction and Language Workshop at NeurIPS 2018*, December 2018.

[8] Chenxi Liu, Junhua Mao, Fei Sha, and Alan Yuille. Attention correctness in neural image captioning. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 4176–4182. AAAI Press, 2017.

[9] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3242–3250, 2017.

[10] Mathis Petrovich, Michael J. Black, and Gül Varol. TEMOS: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision (ECCV)*, 2022.

[11] Matthias Plappert, Christian Mandery, and Tamim Asfour. The KIT motion-language dataset. *Big Data*, 4(4):236–252, dec 2016.

[12] Matthias Plappert, Christian Mandery, and Tamim Asfour. Learning a bidirectional mapping between human whole-body motion and natural language using deep recurrent neural networks. *Robotics and Autonomous Systems*, 109:13–26, 2017.

[13] Karim Radouane, Andon Tchechmedjiev, Julien Lagarde, and Sylvie Ranwez. Motion2language, unsupervised learning of synchronized semantic motion segmentation. *Neural Computing and Applications*, 36(8):4401–4420, 2023.

[14] Jingkuan Song, Lianli Gao, Zhao Guo, Wu Liu, Dongxiang Zhang, and Heng Tao Shen. Hierarchical lstm with adjusted temporal attention for video captioning. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, page 2737–2743. AAAI Press, 2017. ISBN 9780999241103.

[15] Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. From show to tell: A survey on deep learning-based image captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1): 539–559, 2023.

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[17] Xinyu Xiao, Lingfeng Wang, Bin Fan, Shinming Xiang, and Chunhong Pan. Guiding the flowing of semantics: Interpretable video captioning via POS tag. In *Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2068–2077. Association for Computational Linguistics, 2019.

[18] Youngjae Yu, Jongwook Choi, Yeonhwa Kim, Kyung Yoo, Sang-Hun Lee, and Gunhee Kim. Supervising neural attention models for video captioning by human gaze data. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6119–6127, 2017.

Supplementary

This supplementary provides more details on the method implementation, and more visualization for global evaluation of interpretability. Furthermore, we discuss the effectiveness of architecture design. All following analysis are conducted on the test set. For illustration, visual animations are included in the github repository[1]. The transparency level of the gold box represents the temporal attention variation for each predicted *motion word* selected based on *adaptive attention*. We note that grammatical errors mainly stem from the datasets themselves, which contain valid action descriptions but sometimes with incorrect language structure.

A Motivation

Our approach is **focusing on interpretability** while ameliorating motion captioning performance. This comes with additional challenging question on accurate methods for interpretability evaluations. To address this question, a first attempt is to draw multiple visualizations. However, for a global evaluation on test set, this become infeasible. To overcome this limitation, in addition, a simple solution, yet effective, is to display histogram and density

[1] https://github.com/rd20karim/M2T-Interpretable

distributions for attention weights across all test set instead of just sample wise visualizations.

The architectural design is primarily intended to be interpretable, allowing for the explanation of learned spatial, temporal, and adaptive attention weights. Designing an efficient architecture while maintaining interpretability can be very challenging, but has several advantages beyond focusing solely on increasing accuracy metrics. In addition to ensuring a reliable model, we can leverage the interpretability provided by attention mechanisms to extract other semantic motion information: action localization, body part and motion word identification. Let's recall the main novel contributions of our paper in this context:

- Interpretable architecture design.
- Supervision of adaptive and spatial attention.
- Effective tools for global interpretability evaluation.

Consequently, regarding each contribution aspect, we will show the concrete effectiveness of associated theoretical formulations.

B Datasets

We use the two commonly used benchmarks KIT-MLD and Human ML3D with the following statistical details:

Subset	Number	Train	Test	Val.
KIT-ML-aug	motions	4886	830	300
	samples	10408	1660	636
HML3D-aug	motions	22068	4160	1386
	samples	66734	12558	4186

Table 4: Data splits, for KIT and Human ML3D after augmentation (aug).

C Ground-truth generation for supervision

Predefined dictionary. We manually define a dictionary based on representative words in the dataset describing different motion characteristics. Intentionally the dictionary doesn't cover all datasets actions with their synonyms, we want the model to be able to generalize to remaining unsupervised words for their spatial and gate attention. We will see later that the model effectively converges for this intended behavior.

During training, the words in Table 5, and targets words, are stemmed to find correspondence for spatial weight supervision.

Spatial attention supervision. The ground truth spatial attention weights α_{ti} are generated based on the predefined dictionary and it's same for all frames, the temporal attention is the responsible for temporal filtering.

Category	Words	Body part
Trajectory	circle, circuit, clockwise, anticlockwise, forward, backward	Root
Local motion	open, waves, wipe, throw, punch, pick, boxing, clean, swipe, catch, handstand, draw	Arms
	kick, stomp, lift, kneel, squat, squad, stand, stumble, rotate	Legs
	bend, bow	Torso
Connection words	is, the, of, his, her, its, on, their	-
Subject	a, person, human, man	-

Table 5: Predefined dictionary for both datasets.

Adaptive attention supervision. The ground truth β_t is generated based on the Part Of Speech (POS) tagging.

D Hyperparameters selection

We run experiments for different values of $(\lambda_{spat}, \lambda_{adapt})$. The quantitative results are reported in Table 6.

Dataset	λ_{spat}	λ_{adapt}	BLEU@1	BLEU@4	CIDEr	ROUGE$_L$	BERTScore
KIT-ML	0	0	57.3	23.6	109.9	57.8	41.1
	0	3	56.3	22.5	108.4	56.5	39.8
	1	3	57.6	23.5	102.6	57.2	40.1
	2	3	**58.4**	**24.4**	**112.1**	**58.3**	**41.2**
	3	5	57.6	23.7	105.7	57.5	40.9
	5	5	56.5	22.0	99.4	56.8	39.9
HML3D	0	0	69.3	24.0	58.8	54.8	38.7
	0	3	**69.9**	**25.0**	61.6	**55.3**	**40.3**
	0.1	3	69.5	23.8	58.7	55.0	38.9
	0.25	3	68.7	23.8	59.7	54.7	39.3
	0.5	3	68.8	23.8	60.0	55.0	38.6
	1	3	68.7	23.7	58.2	54.6	39.0
	2	3	69.2	24.4	**61.7**	55.0	40.3
	3	3	68.3	23.2	56.5	54.5	37.1

Table 6: Spat+adapt supervision impact w.r.t each corresponding weights.

E Architecture compounds effectiveness

We aim in the following visualizations to demonstrate the global effectiveness of architecture design of each compound :

- Functionality of gating mechanism.
- Impact of Part based motion encoding.
- Spatio-temporal attention blocks.

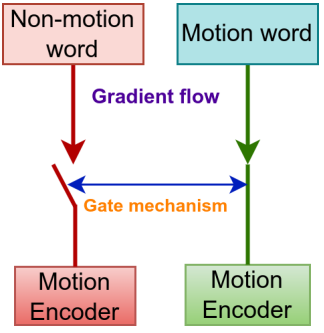
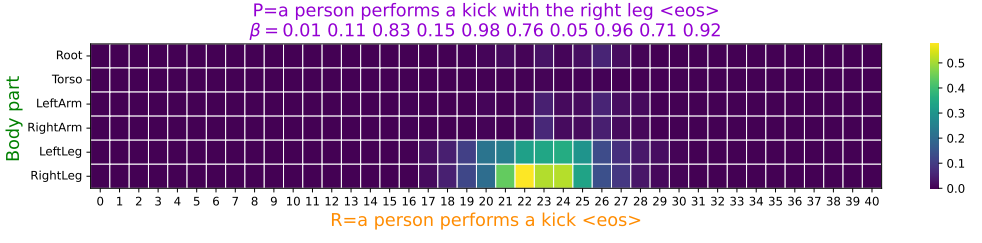


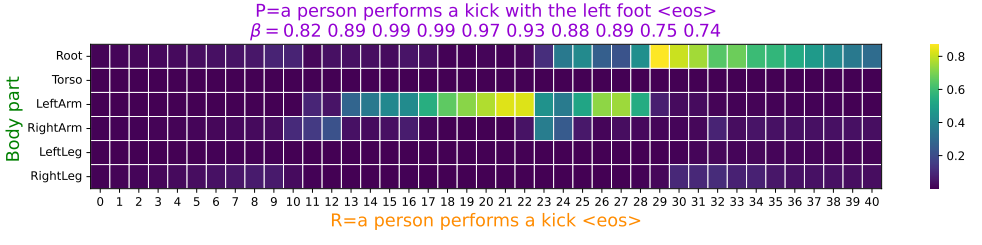
Figure 8: Illustration of our gating mechanism during training. This mechanism prevent the decoder from attending to motion for non-motion word. Consequently the motion encoder is prevented from receiving important gradients updates for non motion words.

Gating mechanism. The gate variable β allows the model to use or not the motion information given the word time step. To visualize this internal process of switching between motion and language, we display predictions for the best model on KIT-ML (results on HumanML3D were shown in the paper). As we see in the following Table, the context vector ($\beta = 1$) is successfully used for all motion characteristics: *action*, *speed*, *body parts*, *trajectory*, *direction*.... Particularly, we note that the end token <eos> is also motion related, as outputting this word depends on the end of the relevant human motion range.

Spatial+adapt attention supervision [KIT-ML]. We show comparison of Spatio-temporal attention maps and text generated between the case of supervision and w/o supervision:



(a) With supervision KIT-(2,3) (action range [19,28]/right kick).

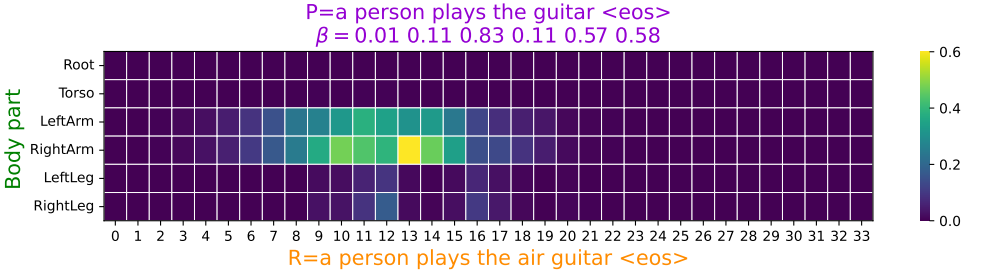


(b) Without supervision KIT-(0,0) (action range [19,27]/right kick).

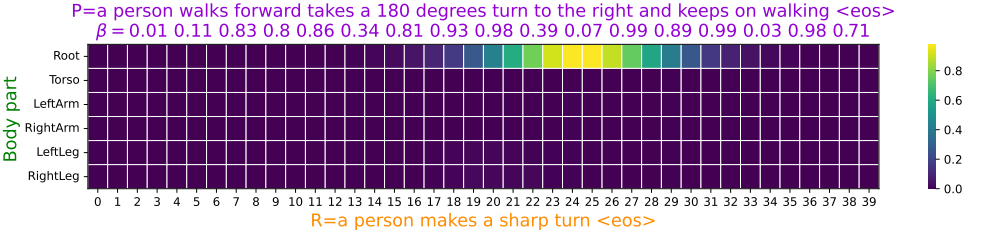
As we see in the case of supervision (Fig.9a) the part were correctly identified and perfectly localized in the range $[20, 26]$ with corresponding manually identified range $[19, 28]$ and small β values are associated with non-motion words. Without supervision (Fig.9b),

the model focuses on irrelevant part and consequently the range of action was not precisely localized. Additionally the β values are high for all kind of words.

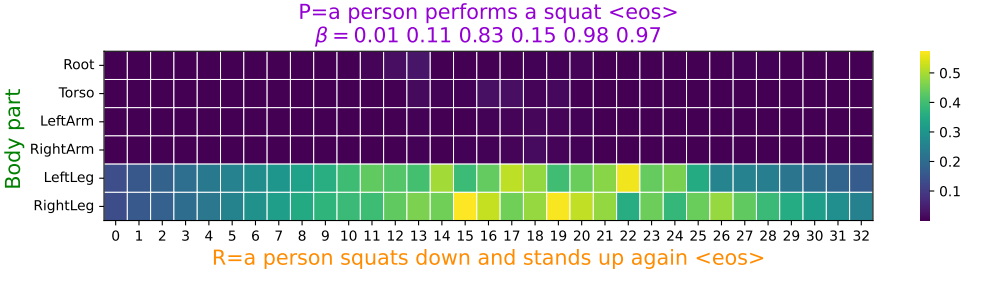
We visualize more samples (Fig.10) with Spatial+adapt supervision. Temporal range is mentioned for comparison, even if action localization wasn't the main focus in captioning task, the model was able to learn implicitly a temporal location through the temporal Gaussian attention mechanism.



(a) Play (action range [10,20]).



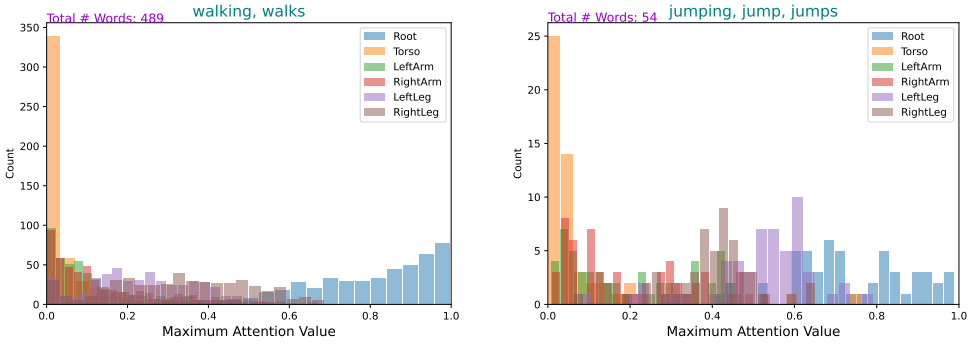
(b) Turn (action range [22,27]).



(c) Squat (action range [10,28]).

Figure 10: Spatio-temporal attention for different motion words on KIT-ML.

Trajectory and global motion. The attention was supervised only for words describing trajectory, but the model generalize successfully to motion words highly depending on global trajectory. This result on maximum attention distributed toward the *Root* body part, as we see in Figure 11.



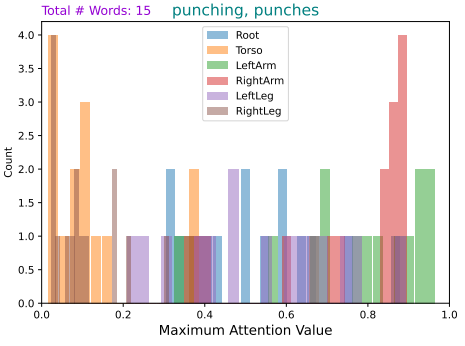
(a) Walk (b) Jump

Figure 11: [KIT-(2,3)]: Body part distribution (spat+adapt).

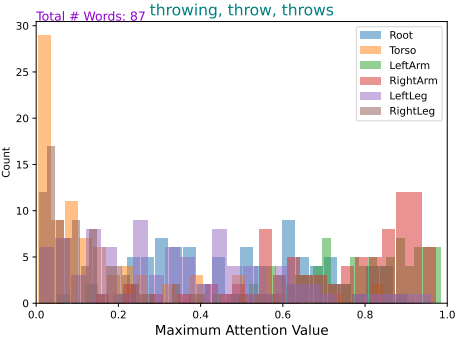
F Part based encoding & spatio-temporal attention

As mentioned in the paper, our architecture design could be sufficient in learning a correct spatial attention maps using larger dataset with rich semantic descriptions. For demonstration, we will use the model with no spatial supervision, to show that part based encoding and spatio-temporal can work solely and correctly together for focusing on relevant body parts w.r.t to the associated generated motion word. To this purpose, we propose to display the histogram distribution of temporal maximum attention weights for each body part over all test set and given a different motion words. This allows for an effective global evaluation of interpretability over all test set.

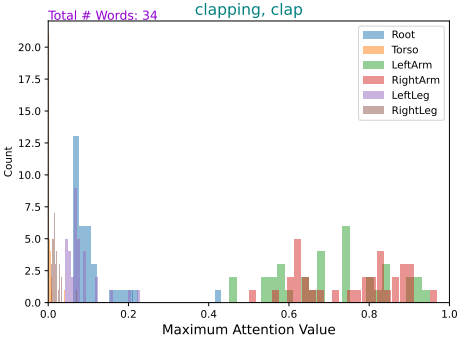
Histograms. In the following, we display the body parts histogram distribution across the test set for different motion words on the model with ***no spatial supervision*** as demonstration for the effectiveness in finding relevant parts to focus on using our interpretable architecture design that includes part-based encoding along with spatio-temporal attention. This is only in the case of the larger dataset HumanML3D. The KIT-ML small dataset still requires spatial supervision to help the architecture focusing on relevant part, as the vocabulary and its size are limited. As demonstrated in all following Figures, depending on the motion word, arms-based/legs-based actions, and particularly some motions with an emphasis on Torso body part.



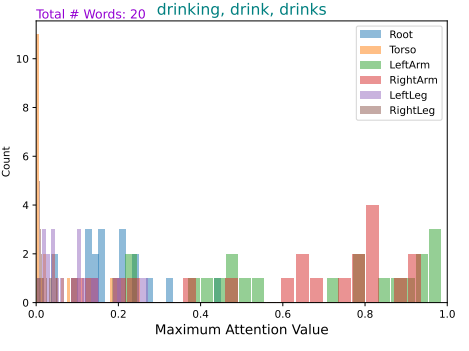
Punch



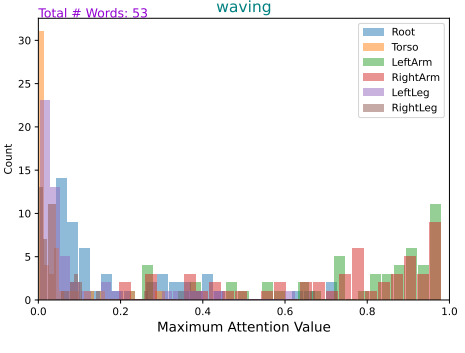
Throw



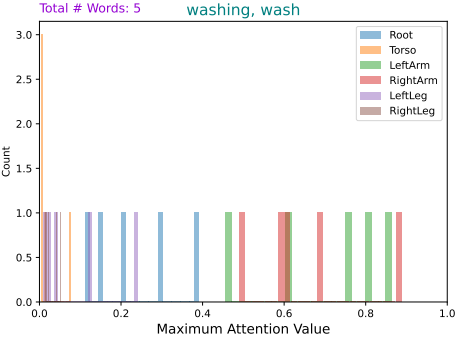
Clap



Drink

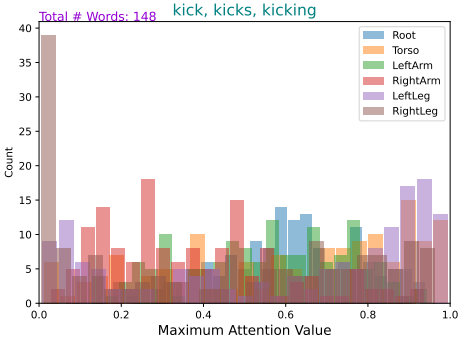


Wave

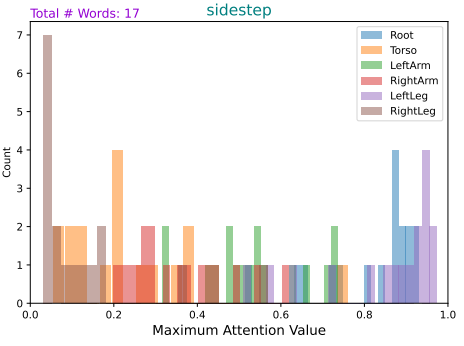


Wash

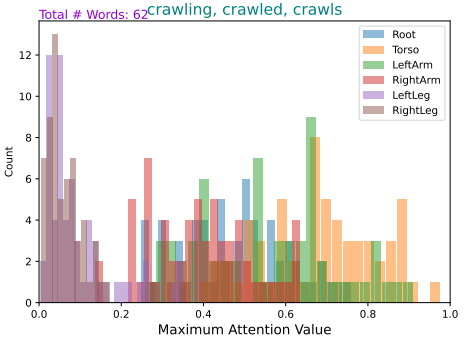
Figure 12: Histogram generated on the HML3D with the config (0,3).



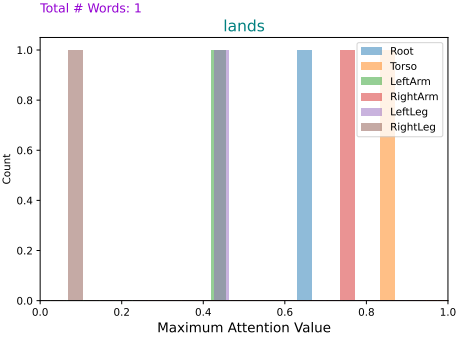
Kick



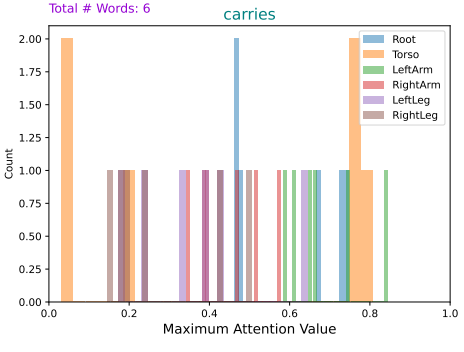
Sidestep



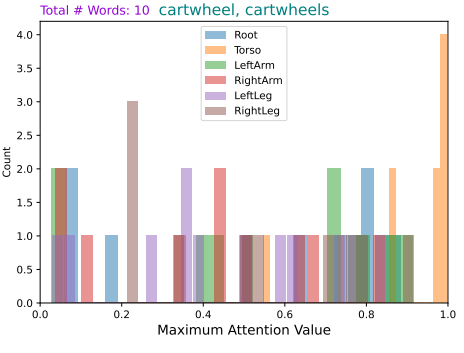
Crawl



Land

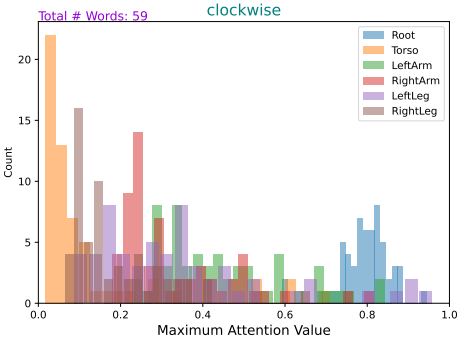


Carries

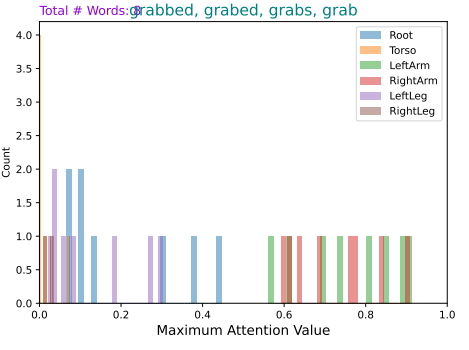


Cartwheel

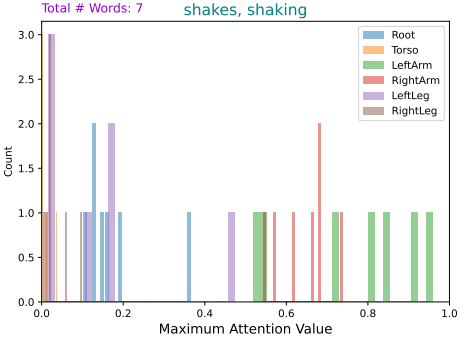
Figure 13: Histogram generated on HML3D with the config (0,3).



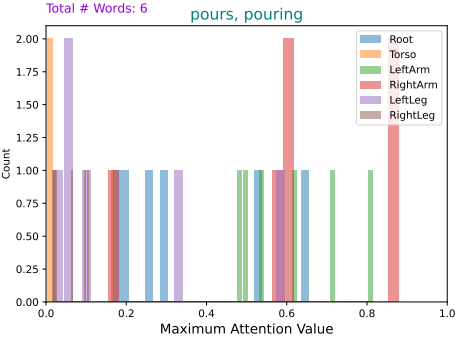
Clockwise



Grab



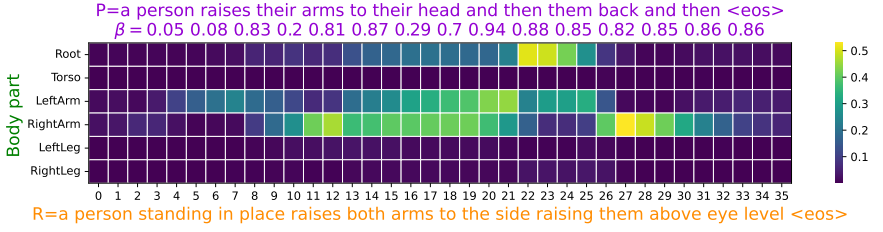
Shake



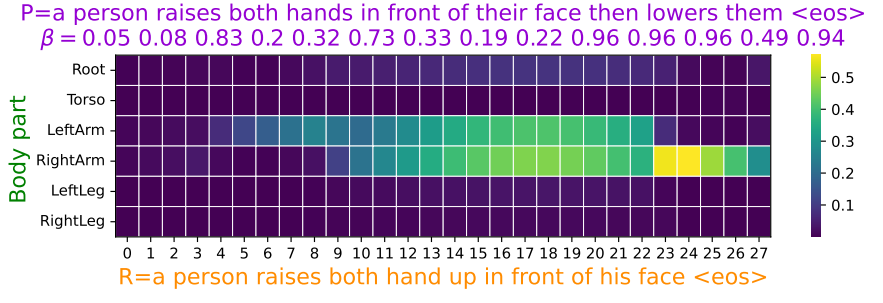
Pour

Figure 14: Histogram generated on HML3D with the config (0,3).

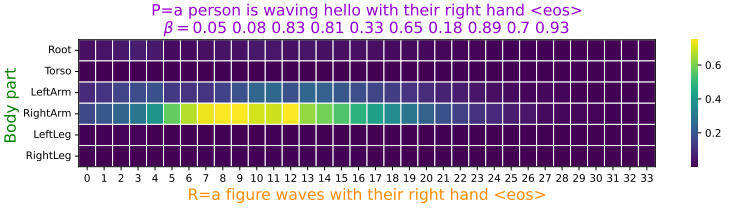
Spatio-temporal attention maps. In this part, we display attention maps for some interesting words for HML3D (0-3) /adapt. In the case of the model without spatial supervision, we have found that the model performs a correct attention focus. When an action is performed using right leg/arm, the model focuses correctly on the corresponding parts. Moreover, for actions performed with both arms/legs, the model focus on both parts. For all cases, body part words (left/right/both) are always accurately identified into the generated text. These observations are common across different representative samples (from different actions).



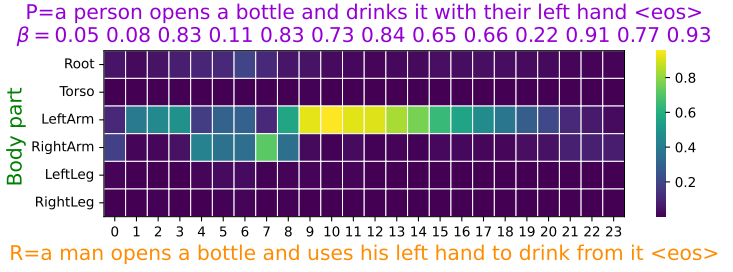
Raises.



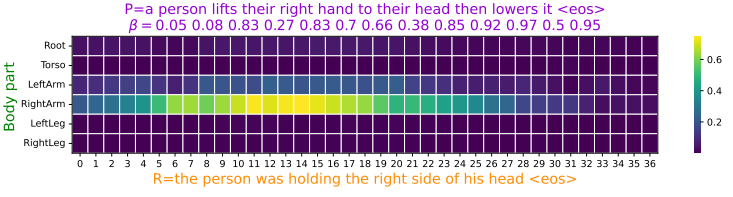
Lowers.



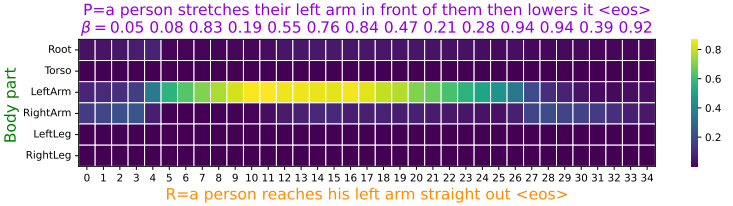
Waving.



Opens.



Lifts.



Stretches.