



HAL
open science

Transformer with Controlled Attention for Synchronous Motion Captioning

Karim Radouane, Sylvie Ranwez, Julien Lagarde, Andon Tchechmedjiev

► **To cite this version:**

Karim Radouane, Sylvie Ranwez, Julien Lagarde, Andon Tchechmedjiev. Transformer with Controlled Attention for Synchronous Motion Captioning. 2024. hal-04697946

HAL Id: hal-04697946

<https://hal.science/hal-04697946v1>

Preprint submitted on 14 Sep 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

Transformer with Controlled Attention for Synchronous Motion Captioning

Karim Radouane^{1,3} Sylvie Ranwez¹ Julien Lagarde² Andon Tchechmedjiev¹
EuroMov DHM, IMT Mines Ales¹, University of Montpellier², LIPN-University Sorbonne Paris Nord³

Abstract

In this paper, we address a challenging task, synchronous motion captioning, that aim to generate a language description synchronized with human motion sequences. This task pertains to numerous applications, such as aligned sign language transcription, unsupervised action segmentation and temporal grounding. Our method introduces mechanisms to control self- and cross-attention distributions of the Transformer, allowing interpretability and time-aligned text generation. We achieve this through masking strategies and structuring losses that push the model to maximize attention only on the most important frames contributing to the generation of a motion word. These constraints aim to prevent undesired mixing of information in attention maps and to provide a monotonic attention distribution across tokens. Thus, the cross attentions of tokens are used for progressive text generation in synchronization with human motion sequences. We demonstrate the superior performance of our approach through evaluation on the two available benchmark datasets, KIT-ML and HumanML3D. As visual evaluation is essential for this task, we provide a comprehensive set of animated visual illustrations in the code repository: <https://github.com/rd20karim/Synch-Transformer>.

1. Introduction

Motion-Language processing has garnered much interest in the computer vision community, where it has been revitalized along with increasing popularity of generative AI. In machine learning, captioning is the process of generating textual descriptions from a given input data, such as images or videos. The interest in captioning tasks stems from the need for a more efficient and effective way to understand and process visual data. Current approaches, mainly focus on often vision-based input, thus, typically relies on a combination of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) or more recently use the Transformers [23]. The aim is to produce detailed

and human-like captions that can be used in several applications such as image and video retrieval and understanding. While captioning tasks have primarily focused on images and videos, limited research has explored motion captioning or human skeleton-based captioning [9, 17].

This approach generates captions for human motion based on estimated or ground-truth poses. The human skeleton offers a concise and semantically rich representation of motion, enabling better understanding and description of human activities. This task involve associating human pose sequence with close textual descriptions. The past three years have seen the emergence of larger and better quality motion-language datasets and an effervescence of ever-improving offline language to motion systems [8]. Although such systems have been a significant focus of research [9, 15, 17], there has also been an interest in motion-to-language generation [7, 9], that has picked up steam with recent papers addressing synchronous motion to language generation [20].

The first motion captioning architecture [20] aiming to synchronize the generation of descriptions with human actions was based on a very simple pre-Transformer model (RNN) [5] and introduced extensions to the canonical attention mechanism. Their experiments were mainly conducted on the original version of KIT-ML [17] before augmentation [8]. While the performance exhibited was honorable, and outperformed previous offline generation systems, particularly on older and smaller datasets like KIT-ML, the emergence of larger datasets such as HumanML3D, calls for a transition to more modern architectures that have been proven to be more effective for language modelling [23].

In this paper, we propose an architecture design for synchronous motion captioning based on Transformer operations. We incorporate mechanisms to control self- and cross-attention distributions, combined with structuring losses to achieve both synchronous generation along better text quality generation. We also propose masking approaches to solve mixing information problems. Subsequently, we annotate a representative subset of the test set from HumanML3D containing a more diverse range of

compositional motions. This allows for an effective quantitative evaluation of the synchronization performance derived from learned attention under our proposed strategy for motion-language alignment control.

2. Related work

In recent years, numerous motion encoders have been proposed to address the challenges of motion and text generation. Excluding studies focusing on bidirectional mapping [9, 18], it is evident that the field of motion generation has witnessed significant advancements, with extensive research efforts dedicated to this task [4, 6, 8, 15, 25]. In contrast, progress in language generation from motion has been comparatively less substantial [7]. In this section, we will present the datasets used for both motion and language generation. Subsequently, we will discuss relevant work related to our study.

2.1. Motion-Language Datasets

The study of complex human movements and actions often requires the use of datasets based on motion-capture. One of the most widely used datasets is the KIT Motion Language Dataset (KIT-ML) [12]. The annotations describe the entirety of each movement, often in the form of single sentences. Recently, an updated version of the KIT-MLD dataset was introduced by augmentation [8], along with a much larger dataset, Human-ML3D. The Motion-Language datasets include recordings of various movement types (walking, running, waving, etc.), where the descriptions give fine-grained details specifying the body parts involved, the manner in which the motion is executed (e.g., speed).

2.2. Motion captioning approach

The motion captioning task is similar to Video Captioning, where the input is a sequence of human poses instead of images. Existing motion-captioning approaches were based on recurrent neural network encoder-decoder architectures, only transitioning to using Transformer-based architectures in recent years.

RNN-based design. A first model addressing the bidirectional generation task was proposed by [18], using the original KIT-MLD dataset. The motion sequence is initially encoded using a stack of bidirectional RNNs to obtain a context vector c . This context vector is further decoded by another stack of unidirectional RNNs into a sequence of text. A similar design was used for motion generation in the reverse direction.

Transformer-based design. More recently, both modes of generation have been addressed by [9]. The authors proposed a transformer-based architecture to handle the generation of both text and motion. This is achieved straightforwardly by representing motion as token sequences us-

ing a codebook obtained through pretraining a VQ-VAE. Their experiments were conducted on the more recent HumanML3D dataset [8] and augmented KIT-MLD. More recently, MotionGPT [10] involves multi-task learning: motion generation and captioning, among other tasks. The disparity in tasks prevents a fair comparison of results. However, its learning process adversely affected motion captioning, resulting in a notably low BLEU@4 score of 12.47% on HumanML3D and no reported results on KIT-ML dataset for motion captioning.

Synchronous Motion Captioning. This task aims to provide a captioning aligned with the motion sequence represented by the human poses in time. The model learns to output a synchronized description with motion, where motion words are generated at the time of the corresponding actions. We can find some analogies with dense aligned captioning [11]. But the alignment is performed at the phrase level instead of the word level, and, thus, it doesn't involve progressive word generation.

Motion primitives and description. Synchronizing motion and language involve implicitly to localize motion primitives and their part of description in the complete sentence. This process intersect with moment retrieval that was presented as use case of *text-to-motion retrieval* (TMR) task introduced by [16]. TMR model performs motion retrieval based on natural language descriptions, and shows qualitative results and initial possibilities to temporally localizing a natural language query in a long 3D motion sequence. On the other end, synchronized captioning approaches [20], involve automatic unsupervised alignment, enabling a simultaneous *progressive text generation* and *motion segmentation*.

3. Methods

We aim for a motion to language system generating text synchronously while being fed a movement sequence. Like in the approach by [20] who used a modified NMT architecture to enable synchronicity, we propose an evolution of the Transformer architecture to achieve the same objective. In this section, we describe our contributions by going over the main components of our approach. Figure 1 gives an overview on the proposed architecture design, on the left a higher level conceptual view of the interaction between the main components of the architecture, and on the right a more details schematic representation of a forward pass during inference.

3.1. Mixing Information in Transformer

In the context of Neural Machine Translation (NMT), the Transformer employs a Multi-Head Attention Mechanism to learn contextualized token representations. Within each encoder layer, an input token's representation is formed as an aggregated representations of input tokens with differ-

ent contributions (attention weights). This process results in context mixing [13]. Several studies have explored information mixing in the Transformer and its influence on predictions [21], aiming to improve the use of attention for interpretability. While this mixing is effective in learning contextual representations for machine translation, it becomes misleading for interpretability analysis. The information mixing process across heads and, or even layers makes it challenging to keep track of the most relevant information used to make predictions. The increase of the number of layers makes it all the more difficult to keep track of the attention flow [1] by using attention weights directly. Consequently, there are two sources of information mixing, the use of multiple Transformer Layers and the attention mechanism itself. We aim to utilize attention weights to identify the most pertinent frames that contribute to the prediction of an action word. Thus, we opt for working with a single Transformer layer. We make use of masking strategies to obtain direct information about motion time through attention, but also to construct latent *compact local motion representations*. A sequence of pose frames is thus transformed into a sequence of compact motion representation which then act akin to a dictionary to retrieve the most relevant frame given a motion word query. Additionally, introducing multiple layers in the Encoder results in an expansion of the receptive field for local motions at each layer, forming a global motion representation. Our objective is for each frame to receive information from a fixed-size window defining what is *local* in the motion. This setup enables us to extract precise motion localization from the attention weights without the undesirable mixing in the information source. To prevent these behaviors, we propose masking strategies incorporated in both self and cross attention mechanisms.

Our model is fully illustrated in Figure 1. We use only one layer in the Encoder and Decoder for the reasons elicited above 3.1. Moreover, including a higher number of Transformer layers isn’t documented to lead to better performance [9] independently from interpretability.

3.2. Masked Attention

Let’s first define the semantics of attention in the context of our task. The attention mechanism is based on the common concepts of Key-Query-Value, here: *Query* u_t : What is the most relevant local human motion information to use for the prediction of word w_t ?

Value v_i : Compact local motion representation around a frame i .

Key k_i : Relevant key representation to learn for a value v_i .

Information Interaction. As illustrated in Figure 1a, for a given query u_t , the goal of cross attention is to search

among the provided motion keys and to retrieve the most relevant motion values v_j , maximising $u_t^T \cdot k_j$ and used to predict the current word w_t .

Masking Strategies. To prevent this mixing in information with long range frame communication, we propose to apply a window centered on each frame i with a range of r so that the new representation becomes a compact local summary of temporal information carried by frames in the range $\Gamma_i = [i - r, i + r]$. This window attention was also applied in another context of long text generation [2], referred to as *sliding window*, but for different main reasons, such as computational efficiency. Masking is also incorporated in the cross-attention, as illustrated in Fig. 1c. In the following, we discuss the window definition for both cases in detail.

Self Attention Window Γ_1 . Self attention in its original form, as proposed by [23] lets each token attend to all other tokens. However, this results in uninformative attention weights for synchronous captioning. Here, we have a pose vector that represents the embedding of each motion frame. Using full self attention, leads to source information mixing and in turn leads to a global representation that encode information about all the actions in the sequence, while we need separate local information to localize each action involved in the human motion separately. Intuitively, without local masking, the representation of a frame i in the next output layer may contain information about different non contiguous frames. Therefore, when the cross-attention is maximized on the final representation of frame i , the attention weights cannot directly be used to access the most relevant set of frames used for the current word prediction. The precise frame source of information used for the predicted motion word is lost. Moreover, including long-distance isolated frames reduces the ability of the model to learn correct local information.

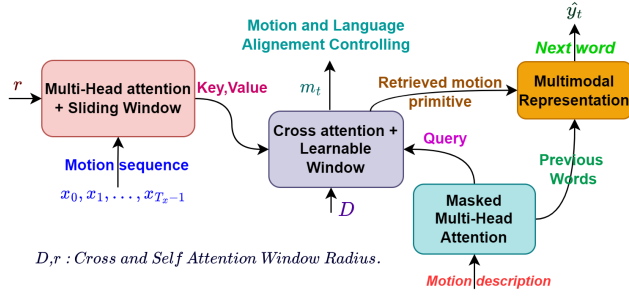
Cross Attention Window γ_t . We constrained attention scores to be around a learnable frame position m_t . This learnable value represents the center of the cross window search range $\gamma_t = [m_t - D, m_t + D]$.

Receptive Field. Regarding the receptive field, taking into account the two masking strategies, the query u_t at step t searches in the motion frames across a window of width $L = 2(D + r)$. This results in mask accumulation ranging over $[m_t - r - D, m_t + r + D]$, as illustrated in Fig. 1b.

3.3. Transformer Operations in Masking Context

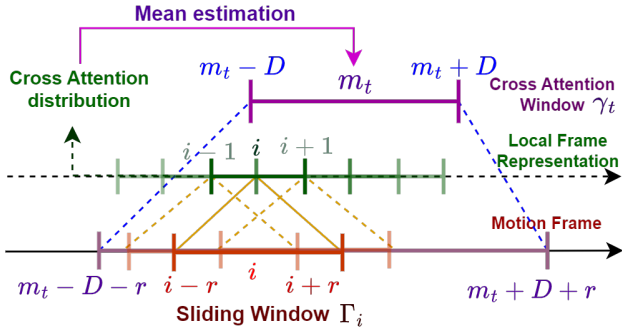
After introducing our masking strategies (Sec. 3.2), we will formulate the Transformer operations, taking into account cross- and self-attention masking.

Multi-Head Attention. Given a sequence of pose vectors $p_i \in \mathbb{R}^c$. The pose of each frame i is transformed into x_i by Eq. (1), where PE is the common positional encoding. Then, in each head $h \in \{1, \dots, H\}$ in the self-attention

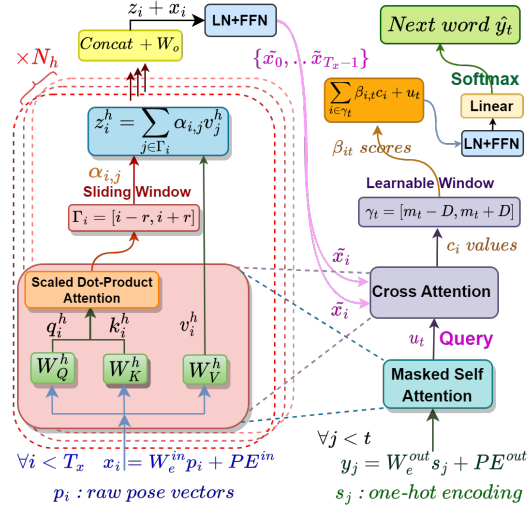


D, r : Cross and Self Attention Window Radius.

(a) Information interaction in our model: A motion primitive is retrieved based on the words query through the relevant key, using weighted sum of corresponding attention scores. The alignment motion-language is controlled through structuring losses.



(b) Receptive field of the decoder during generation spans the motion range $[m_t - D - r, m_t + D + r]$, where m_t is the motion-word alignment position estimated from the cross attention distributions.



(c) Our Transformer operations in inference phase: A static mask Γ_i is incorporated in the encoder side and learnable mask γ_t for the decoder, attention maps are controlled during training to allows the inference of synchronization time between motion and language in an unsupervised manner.

Figure 1. Overview of general proposed framework with relevant details.

block, x_i is transformed into a query q_i^h , a key k_i^h , and a value v_i^h .

$$x_i = (p_i W_e + b_e) + PE \quad (1)$$

$$q_i^h = x_i W_Q^h + b_Q^h \quad k_i^h = x_i W_K^h + b_K^h \quad v_i^h = x_i W_V^h + b_V^h \quad (2)$$

The context vector z_i^h for the i^{th} token of each attention head is then generated as a weighted sum over the transformed value vectors inside the sliding window Γ_i (Sec. 3.2).

$$z_i^h = \sum_{j \in \Gamma_i} \alpha_{i,j}^h v_j^h \quad (3)$$

where $\alpha_{i,j}^h$ is the attention weight assigned to the j^{th} frame, and computed using Eq. (4). We note that scores outside the window Γ_i are not considered in the soft-max operation (masked with $-\infty$).

$$\alpha_{i,j}^h = \frac{\exp(q_i^{hT} \cdot k_j^h / \sqrt{d})}{\sum_{j \in \Gamma_i} \exp(q_i^{hT} \cdot k_j^h / \sqrt{d})} \quad (4)$$

The context vector ($z_i \in \mathbb{R}^d$) aggregates information from each head through the W_O projection layer Eq. (5).

$$z_i = \text{CONCAT}(z_i^1, \dots, z_i^{N_h}) W_O \quad (5)$$

LN + FFN. Represent the mapping $f_{W_1^{in}, W_2^{in}} : (z_i, x_i) \mapsto \tilde{x}_i$ as defined in Equations (6) to (8).

$$\tilde{z}_i = \text{LN}(z_i + x_i) \quad (6)$$

$$\tilde{x}_i = \max(0, \tilde{z}_i W_1^{in} + b_1) W_2^{in} + b_2 \quad (7)$$

$$\tilde{\tilde{x}}_i = \text{LN}_{\text{FFN}}(\tilde{x}_i + \tilde{z}_i) \quad (8)$$

Where LN is the Layer Normalization, while Feed Forward operation (FF) is given by Eq. (7).

Compact Local Representation. Refers to the final motion encoding vector \tilde{x}_i (Eq. (8)). Intuitively, \tilde{x}_i captures local motion information centered on a frame i within Γ_i .

Cross Attention Weights. In our cross-attention formulation we only have one attention head, and attention scores are formulated as :

$$\beta_{i,t} = \frac{\exp(u_t^T \cdot k_i / \sqrt{d})}{\sum_{j \in \gamma_t} \exp(u_t^T \cdot k_j / \sqrt{d})} \quad (9)$$

Retrieved Motion Primitive. Refers to the local motion information selected as relevant for the prediction of next word \hat{y}_t , defined in Eq. (10).

$$r_t = \sum_{i \in \gamma_t} \beta_{i,t} c_i \quad (10)$$

Multimodal Representation. Denoted as g_t , quantifies information about: i) previous generated words up to time t given by u_t , and ii) local motion information c_j to consider for the prediction of the next word y_t . Where c_j is the value produced by the cross attention block for frame j (cf. Fig. 1c).

$$g_t = f_{W_1^{out}, W_2^{out}}(r_t, u_t) \quad (11)$$

3.4. Transformer with Controlled Attention

Learnable Cross Window Center. Given a language query u_t for a motion input. Let's A_t be the discrete random variable that associates each local motion representation around the frame i to its probability $p(A_t = i)$ of being the most relevant information contributing to the prediction of the current word w_t . Formally, we consider the learnable center window position m_t as the center of the A_t distribution (Eq. (12)), where T_x is the human motion length.

$$m_t = \mathbb{E}[A_t] = \sum_{i=0}^{T_x-1} i \cdot p(A_t = i) = \sum_{i=0}^{T_x-1} i \cdot \beta_{i,t} \quad (12)$$

Constraint on Alignment Position m_t . In order to obtain synchronous generation, inspired by [20], we include a constraint on m_t such that $m_{t-1} < m_t$ in the training loss. Although this constraint is language dependent and not universally true at the word level, it holds for motion words. For example, the words {"the", "a", "person"} are not related to the monotony of frame generation, but for action words like ("walk", "jump"), the succession 'walk' then 'jump' happens successively in time, as results the word describing these appear successively in the human description references. The words are generated progressively with human motion evolution. Synchronous motion captioning aims to associate every set of words in the sentence describing one action to the relevant set of frames based on m_t and the attention weights distribution of A_t .

Initial Alignment Position. Formally, this position is m_0 . To encourage the model to see the whole motion from its start, we push m_0 to be close to the *first* motion frame and become a reference for the next learnable attention mean $m_t, \forall t > 0$.

Motion and Language Alignment Control. The model attention distributions are forced to converge toward a solution that respects the constraint $m_{t-1} < m_t, \forall t > 0$ using the attention *structuring losses*:

$$\begin{aligned} Loss_0 &= m_0 / T_x \\ Loss_m &= \frac{1}{T_x} \sum_{t < T_x - 1} \max((m_t + m) - m_{t+1}, 0)^2 \end{aligned}$$

During training, the loss constraining monotonic positions $Loss_m$ will be only penalized when the constraint $m_t + margin \leq m_{t+1}$ is violated. We added a margin value to ensure that m_{t+1} is strictly superior to m_t which prevents the trivial case resulting in m_t been constant for all words. This enables the *attention controlling* for synchronous captioning. In all experiments, we set the margin value $m = 1$.

Training loss. We define the global loss that can be observed as two goals of supervision mode. First, a loss term, focusing on the direct language generation. Secondly, losses focusing on attention structuring.

$$Loss = Loss_{lang} + \lambda_0 Loss_0 + \lambda_m Loss_m \quad (13)$$

Where (λ_0, λ_m) are balancing coefficients, and the language loss (Eq. (14)) is defined as the standard text generation objective minimizing cross entropy between the target and predicted words.

$$Loss_{lang} = -\frac{1}{T_y} \sum_{j=1}^{T_y} y_j \log(\hat{y}_j) \quad (14)$$

Attention Heads N_h . While we use only one layer on both the encoder and decoder sides, multi-head attention is incorporated in both the Encoder and Decoder, except for the cross-attention which uses only one head. This choice is motivated by the necessity to capture information from different frames inside the sliding window. On the decoder side, we maintain a query that takes into consideration all previously generated words.

4. Quantitative and Qualitative Results

In our specific case, our objective extends beyond maximizing the BLEU score; we also aim to align each motion word w_t with the most accurate center time of action execution. Our goal is to infer alignment information using only cross attention weights. Thus, we need to evaluate quality of both text generation and synchronicity. Given an attention distribution over frames, effective localization of an action occurs when the mean of attention weights ideally matches the center time of the action, and the start and end frames are defined by the spread of attention distribution. We will first discuss NLP metrics, qualitative analysis then evaluate synchronization.

Dataset	D	r	BLEU@1	BLEU@4	CIDEr	ROUGE_L	BERTScore
HML3D	5	10	66.4	25.1	61.9	54.3	42.0
	10	10	68.7	26.6	68.0	55.6	44.3
	20	20	69.2	27.1	70.3	56.1	45.5
	∞	∞	68.9	26.5	69.0	56.0	45.0
KIT-ML	10	5	54.3	21.2	93.7	54.8	39.0
	10	10	59.0	26.4	117.8	58.1	43.5
	20	20	57.6	24.4	116.7	58.1	44.1
	∞	∞	58.8	26.5	132.3	58.7	45.8

Table 1. Controlled attention with different values for D and r . The masking approach helped improve the NLP metrics in case of HML3D. However, these parameters have a more significant effect on our main goal of motion-language synchronization as will be demonstrated in Table 3.

4.1. Ablation and Evaluation Study

We recall that our architecture incorporates a single encoder/decoder layer Transformer. More complex designs tend to yield less interpretable attention maps and are not directly controllable. However, interpretability and attention control are crucial for inferring synchronization between motion and language in unsupervised setting. Consequently in our context the ablation study concern only two aspects : i) Effect of motion and language alignment controlling (structuring losses) and ii) Effect of masking approach: learnable and sliding window (more analysis in Supp.B).

Hyperparameters of Attention Control. To enable attention control, we set $\lambda_m = 1000, \lambda_0 = 0.1$ and experiment with different values for window size, D for cross attention, and r for self-attention. Table 1 presents quantitative results for this hyperparameter search. First, we note by $D = \infty, r = \infty$ the case where full context length is used without self- and cross-attention masking. The hidden size d_m and the number of heads N_h are set respectively to 128 and 4 for HumanML3D and to 64 and 4 for KIT-ML. We note that higher values of D and r in some cases give better results in terms of text quality (cf. Tab. 1) but not in terms of synchronization between motion and language (cf.Tab. 3). Consequently, many alternative models can yield good or equivalent solutions in terms of text quality generation, but not all lead to good synchronization.

Comparison with SOTAs. Although our primary objective goes beyond merely enhancing the quality of the generated text, for comparison, we present the standard text generation metrics in Table 2 based solely on text quality generation. On KIT-ML, our model significantly outperforms the TM2T model which is also Transformer-based model but with 3 layers in the Encoder and Decoder. In contrast, our model employs only one layer with fewer parameters and does not utilize beam searching, while achieving synchronous captioning. In comparison to the

model MLP+GRU, we achieve significantly better results than SOTA both on KIT-ML (SOTA + 1.1% BLEU4, - 0.1% ROUGE, +6.6 CIDEr, +3.7% BERTScore), and on HumanML3D (SOTA + 3.7% BLEU4, +2.3% ROUGE, - 2.2 CIDEr, +8.3% BERTScore).

4.2. Qualitative analysis

In this part we discuss qualitative results at the level of attention maps and human motion sequences frozen in time.

Cross Attention Maps. Examples of compositional motions are shown in Figure 2 with corresponding motion ranges. The violet rectangles represents the position of maximum attention. Each word is generated at it’s corresponding position. In Fig. 2a, considering the motion words, the spread of attention for the phrase *walks up stairs* is in the range [17, 45], as compared to the manuel observation [10, 40]. The subject turns at Frame 45, where the predicted attention for the word *turn* is maximal at the frame 44. Similar analyses could be conducted on other samples (Figs. 2b and 2c). However, the evaluation remains subjective, specifically in terms of defining the start/end of each action. To address this limitation, the *Intersection over Prediction* (IoP) and *Element of* metrics were proposed by [20].

Motion Frozen in Time. We use static visualizations to illustrate, at a single point in time, the association of motion words with the frames receiving maximum attention. Figure 3 illustrates motion phrases and their sequence of frames at maximum attention. More illustrations are given in Figure 4. However, this static visualization still have their inherent limitation, so we include animations in the code page¹.

4.3. Evaluating word-motion synchronicity

In this section, to quantify the synchronization between a human pose sequence and the corresponding motion-description words we use the metrics *IoP*, *IoU* and *Element*

¹<https://github.com/rd20karim/Synch-Transformer>

Dataset	Model	BLEU@1	BLEU@4	ROUGE-L	CIDEr	Bertscore
KIT-ML	RAEs [24]	30.6	0.10	25.7	8.00	0.40
	Seq2Seq(Att)	34.3	9.30	36.3	37.3	5.30
	SeqGAN [7]	3.12	5.20	32.4	29.5	2.20
	TM2T w/o MT [9]	42.8	14.7	39.9	60.1	18.9
	TM2T [9]	46.7	18.4	44.2	79.5	23.0
	MLP+GRU [20]	56.8	25.4	58.8	125.7	42.1
	[Spat+Adapt](2,3) [19]	58.4	24.7	57.8	106.2	41.3
Ours	58.8	26.5	58.7	132.3	45.8	
HML3D	RAEs [24]	33.3	10.2	37.5	22.1	10.7
	Seq2Seq(Att)	51.8	17.9	46.4	58.4	29.1
	SeqGAN [7]	47.8	13.5	39.2	50.2	23.4
	TM2T w/o MT [7]	59.5	21.2	47.8	68.3	34.9
	TM2T [9]	61.7	22.3	49.2	72.5	37.8
	MLP+GRU [20]	67.0	23.4	53.8	53.7	37.2
	[Adapt](0,3) [19]	67.9	25.5	54.7	64.6	43.2
Ours	69.2	27.1	56.1	70.3	45.5	

Table 2. Text generation performance conditioned on human pose motion sequence. Beyond our motion-language synchronization goal, our approach performs significantly better across different NLP metrics.

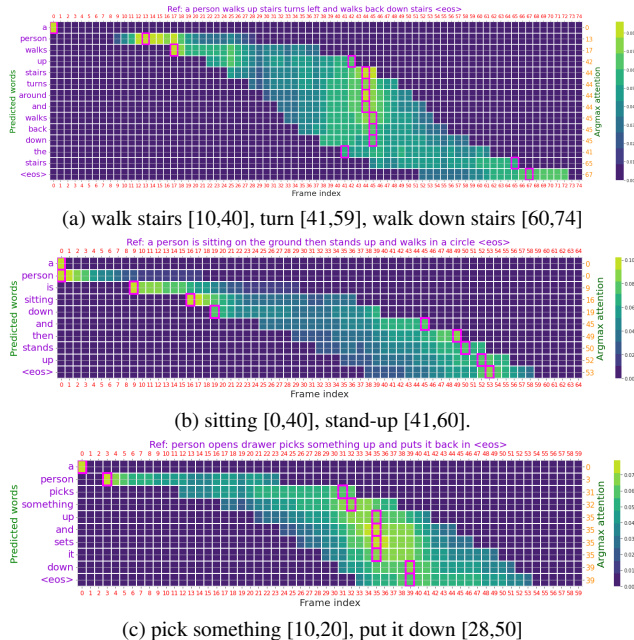


Figure 2. Cross attention map of compositional motions with corresponding frame range of each action ($D=r=10$). Across multiple examples, we observe that the attention distribution of motion words consistently falls within the indicated motion range for each specific action.

of proposed in [20]. However, the subjective nature of captioning process and time labeling make it difficult to consider exclusively metric values, as results, it remains very

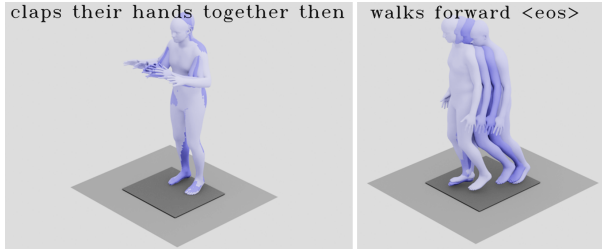
challenging and serves as quantitative complementary measure to visual animations. *These animations of synchronous text generation can be found in our code repository.*

Annotation. First, we annotate a representative subset of the test set from Human-ML3D, which is richer in diverse compositional motions. We select samples from different actions featuring compositional motions, each containing at least two actions, to ensure an effective evaluation of synchronicity.

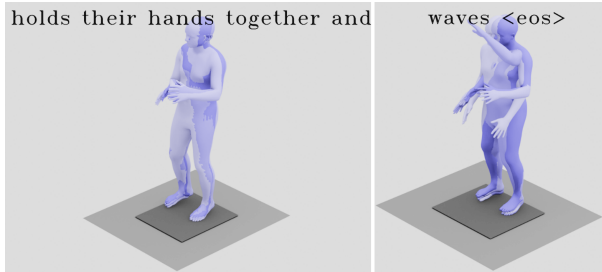
Metrics. We assess the alignment between a primitive human motion and its description based on motion words. We identify the frame time with maximum attention given to a motion and word, and then test whether the frame time falls within the motion action range (utilizing the *Element of* method). Effective synchronization involves outputting each motion word during its corresponding motion execution. In contrast, IoP and IoU metrics primarily gauge the accuracy of localizing the start/end of each action. These metrics were introduced in detail by [20]. Observing the results in Table 3, we can conclude that $D = r = 10$ provides the best tradeoff between the quality of text generation and synchronization.

5. Applications

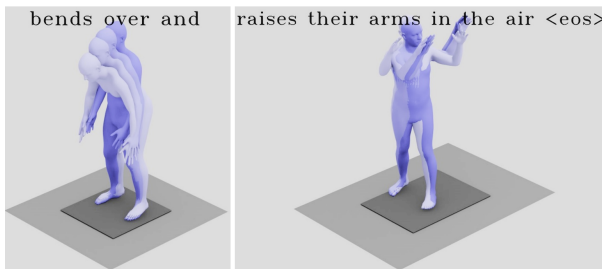
In recent times, substantial advancements have been achieved in the domain of sign language research, focusing on various specific objectives, including *alignment* [3], *temporal localization* [22], and *sign spotting* [14]. In line with these efforts, a related approach in this field, proposed by [22], also uses attention scores to identify and segment



(a) claps then walks.



(b) holds hands then waves.



(c) bends then raises arms.

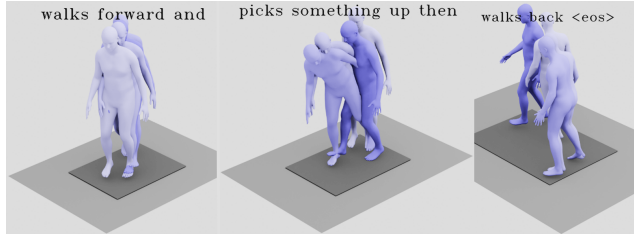
Figure 3. Frozen motion with 4 keyframes of higher attention corresponding to the language segment.

D	r	IoU	IoP	Element of	BLEU@4
20	20	51.35	60.55	71.55	27.1
10	10	46.40	67.96	78.48	26.6
5	10	45.23	62.40	75.62	25.1
∞	∞	39.93	39.96	46.98	26.5

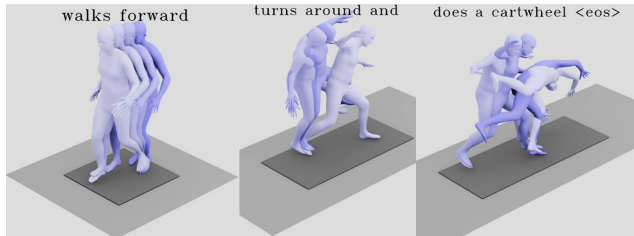
Table 3. Synchronization scores for different D and r values show that these parameters have a more significant effect on action localization (IoP/IoU) and synchronicity (Element of). Our masking approach with ($D = r = 10$) prevents the mixing of information from different actions, enabling a better attention-based localization of action time compared to ($D = r = \infty$), despite having slightly the same BLEU score.

signs in continuous video.

Aligned sign language translation. Building an automated sign language translator with alignment information involves associating sign segments with their corresponding language segments. This task implicitly aims to link a sequence of upper-body pose movements to words, and the proposed approach can be employed to create techniques



(a) a person walks forward picks something up with their right hand and walks back.



(b) a person walks forward turns around and does a cartwheel.

Figure 4. Decomposition of motions and associated descriptions (Animations in the code repository, other visualizations in Supp.C).

for achieving this alignment in an unsupervised manner.

Temporal action localization. For skeleton based action localization within a continuous stream, this task could be formulated as mapping a sequence of poses to a sequence of actions. Utilizing cross attention weights in this scenario enables the unsupervised inference of action start/end times, eliminating the necessity for labeled action time data. When time annotations are accessible, they can guide the supervision of temporal weight distribution, thereby improving the accuracy of action localization and providing more interpretable attention maps.

6. Conclusion

In the future, we may explore more advanced methods for local motion representation, including the incorporation of multiple heads in cross-attention. However, improving synchronous captioning remains challenging, as it requires tracking the interaction between different attention weights sources. We plan to leverage existing attention aggregation methods. Furthermore, it’s worth noting that the presented methodologies hold promise for application in various scenarios beyond our current task, such as alignment for sign language translation and unsupervised action segmentation. We believe that taking steps towards controlling attention weights can lead to more explainable solutions, especially in resolving multiple tasks in unsupervised settings.

References

- [1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197, Online, July 2020. Association for Computational Linguistics. 3
- [2] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. 2020. 3
- [3] Hannah Bull, Triantafyllos Afouras, Gül Varol, Samuel Albanie, Liliane Momeni, and Andrew Zisserman. Aligning subtitles in sign language videos. In *ICCV*, 2021. 7
- [4] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18000–18010, 2023. 2
- [5] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, Oct. 2014. Association for Computational Linguistics. 1
- [6] Anindita Ghosh, Noshaba Cheema, Cennet Oguz, Christian Theobalt, and Philipp Slusallek. Synthesis of compositional animations from textual descriptions. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1376–1386, 2021. 2
- [7] Yusuke Goutsu and Tetsunari Inamura. Linguistic descriptions of human motion with generative adversarial seq2seq learning. In *Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4281–4287. IEEE, 2021. 1, 2, 7
- [8] 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. 1, 2
- [9] 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 Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 580–597, Cham, 2022. Springer Nature Switzerland. 1, 2, 3, 7
- [10] 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. 2
- [11] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 2
- [12] Christian Mandery, Ömer Terlemez, Martin Do, Nikolaus Vahrenkamp, and Tamim Asfour. Unifying representations and large-scale whole-body motion databases for studying human motion. *IEEE Transactions on Robotics*, 32:796–809, 8 2016. 2
- [13] Hosein Mohebbi, Willem Zuidema, Grzegorz Chrupała, and Afra Alishahi. Quantifying context mixing in transformers. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3378–3400, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. 3
- [14] Liliane Momeni, Gül Varol, Samuel Albanie, Triantafyllos Afouras, and Andrew Zisserman. Watch, read and lookup: learning to spot signs from multiple supervisors. In *ACCV*, 2020. 7
- [15] Mathis Petrovich, Michael J. Black, and Gül Varol. TEMOS: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision (ECCV)*, 2022. 1, 2
- [16] Mathis Petrovich, Michael J. Black, and Gül Varol. TMR: Text-to-motion retrieval using contrastive 3D human motion synthesis. In *ICCV*, 2023. 2
- [17] Matthias Plappert, Christian Mandery, and Tamim Asfour. The KIT motion-language dataset. *Big Data*, 4(4):236–252, dec 2016. 1
- [18] 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, 5 2017. 2
- [19] Karim Radouane, Julien Lagarde, Sylvie Ranwez, and Andon Tchechmedjiev. Guided attention for interpretable motion captioning. In *Proceedings of the 35th British Machine Vision Conference*, 2024. 7
- [20] 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, Dec. 2023. 1, 2, 5, 6, 7
- [21] Leonid Schwenke and Martin Atzmueller. Show me what you’re looking for visualizing abstracted transformer attention for enhancing their local interpretability on time series data. *The International FLAIRS Conference Proceedings*, 34, 2021. 3
- [22] Gül Varol, Liliane Momeni, Samuel Albanie, Triantafyllos Afouras, and Andrew Zisserman. Read and attend: Temporal localisation in sign language videos. In *CVPR*, 2021. 7
- [23] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz 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. 1, 3
- [24] Tatsuro Yamada, Hiroyuki Matsunaga, and Tetsuya Ogata. Paired recurrent autoencoders for bidirectional translation between robot actions and linguistic descriptions. *IEEE Robotics and Automation Letters*, 3:3441–3448, 10 2018. 7
- [25] Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Shaoli Huang, Yong Zhang, Hongwei Zhao, Hongtao Lu, and Xi Shen. T2m-gpt: Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

Supplementary

A. Introduction

The motivation of our work stems from our additional goal of synchronizing text with motion using motion captioning as an intermediate task due to its applications such as aligned sign language transcription, unsupervised action segmentation and human motion segmentation. Unsupervised synchronization of text with motion requires a specific focus in architecture design that will be further analyzed in this supplementary material. In addition to animations, this supplementary feature introduces static visualizations as a preliminary to showcase the alignment between attention time for motion words and the corresponding retrieved primitives. Subsequently, we present additional quantitative results, followed by the illustration of qualitative assessments through static visualizations.

B. Ablation analysis

In addition to the results mentioned in the paper, we highlight the following other important analysis.

B.1. Multilayer vs. 1-layer Transformer

To demonstrate the sufficiency of our 1-layer based Transformer design, we compare the results against a multilayer transformer with 3 layers in both the encoder and decoder. The quantitative effect is discussed in Table 4. Qualitative impact is shown in Figure 5. The multilayer setting did not enhance text quality or synchronization performance beyond not being lightweight.

# Layers	Mask.	BLEU@4 \uparrow	IoU \uparrow	IoP \uparrow	Element of \uparrow
1	No	26.5	39.93	39.95	48.98
	Yes	27.1	51.35	60.55	71.55
3	No	25.7	41.88	41.92	39.81
	Yes	25.9	45.16	55.60	49.06

Table 4. **Even with Masking in Multi-layer Transformer**, the synchronization scores remain low compared to a single layer. This occurs because the **receptive field increases** across layers, causing **mixing information** in frame representations at the **final** encoder layer where representations of **early** frames contains information about very distant frames, leading to attention concentration at the beginning (See Figure 5).

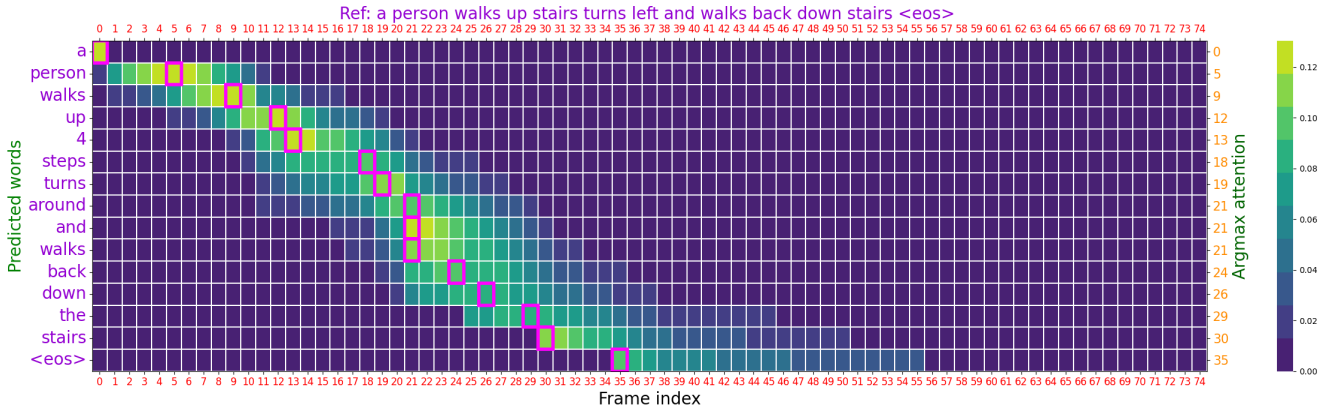


Figure 5. **Multi-Layer Transformer**: Compared to Figure 2a attention distributions, here, are uninformative about action times, attention weights (for 'turns', 'walks back down') are not aligned with action times (same observation for different samples).

B.2. Masking and Attention Controlling Impact

Without masking (See Figure 8) attention weights are uninformative about action times, this highly impacts synchronization scores (See Table 4, case 1-layer) which demonstrates the importance of our masking strategies.

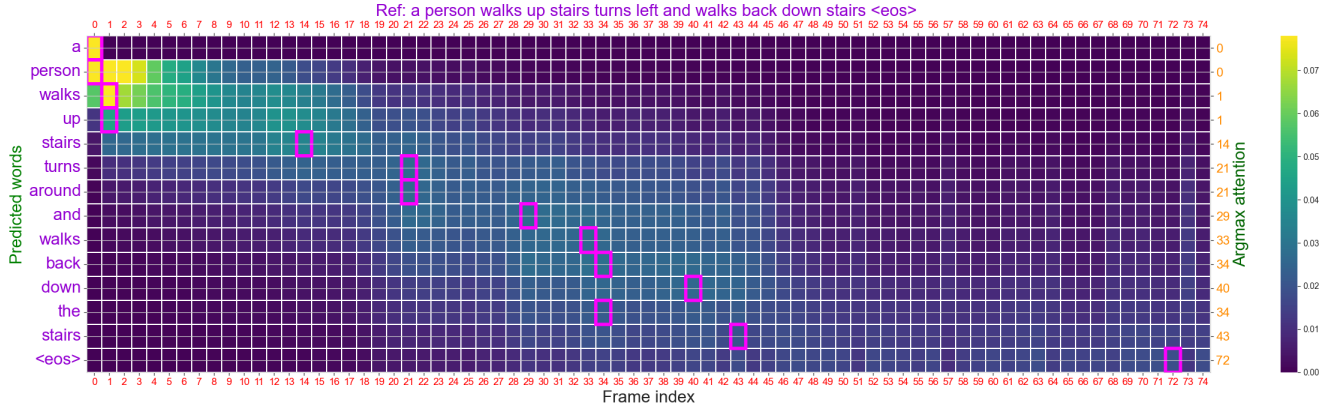


Figure 6. **1-Layer Without Masking:** Compared to Figure 2a 'walks' word attention is maximal around frame 1, while this action starts at 10. 'turns' (frame 21) vs. correct range [41, 59]. 'Walk down stair' highlighted in range [33, 43] vs. [60, 74]. Our masking strategies were crucial in solving these issues.

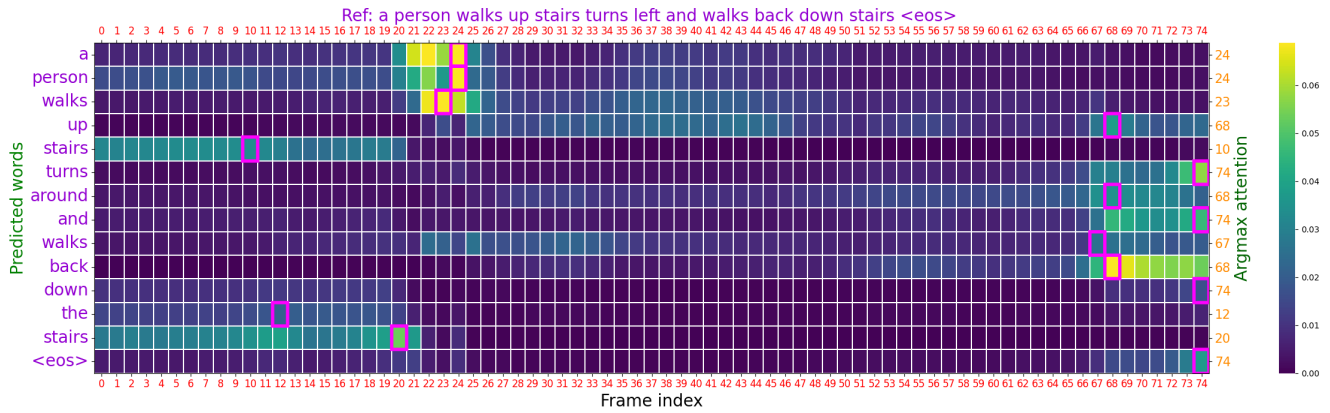


Figure 7. **Without Attention Control and Masking:** attention distributions are disordered and not very informative about action times (the same observation holds for different samples). Our masking and attention control were crucial in solving these issues (Cf.Tab 4-case 1-Layer).

Without Attention Control and Masking attention weights don't carry any information about the action's timing or the order of execution (See Figure 7).

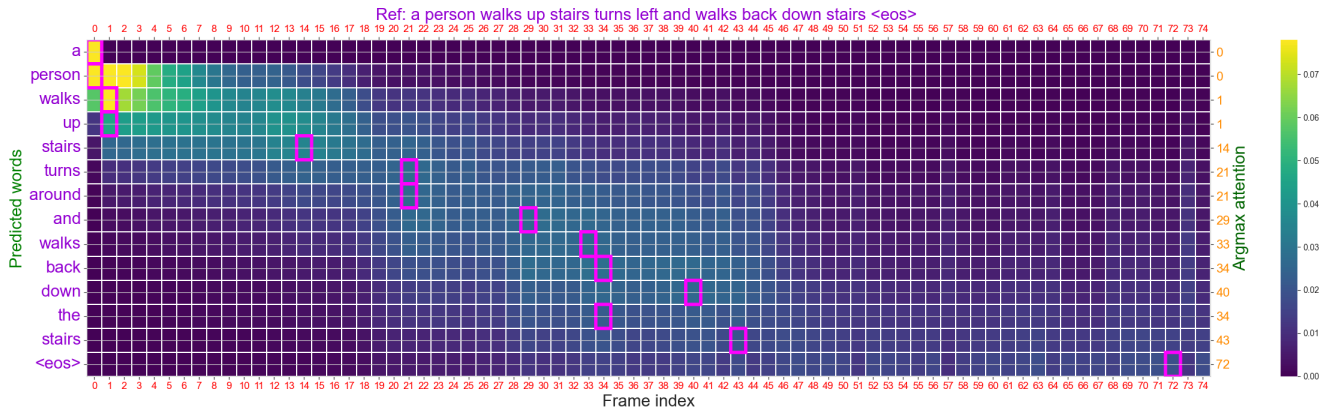


Figure 8. **Controlling Attention without Masking:** *walks* word attention is maximal around frame 1, while this action starts at 10. *turns* (frame 21) vs. correct range [41, 59]. *Walk down stair* highlighted in range [33, 43] vs. [60, 74]. These issues occurs because early frames have access to all distant frames without masking.

C. Visualizations

Our Controlled and Masked Transformer is designed to enable action localization solely through attention. The current synchronization involves *word-events*, but words describing the same event (action) could also be grouped, with attention weights aggregated by averaging across relevant language segments to form *phrase-events* association. In this part, we provide static visualizations with motion frozen in time, given by the key attention frames (motion frames receiving maximum attention) at word- and phrase-level. Then, we visualize some additional cross-attention maps associating human motion sequences and language word descriptions in time.

C.1. Motion Frozen in Time

In this section, we aim to illustrates poses sequence and motion words association based on attention weights. We present static visualizations capturing motion frozen in time accompanied by corresponding descriptive words. Nevertheless, as previously explained, static visualizations inherently possess limitations, making them a complement to animations.

C.1.1 Word level attention

For each word, we visualize *four motion frames receiving maximum attention at inference time*.

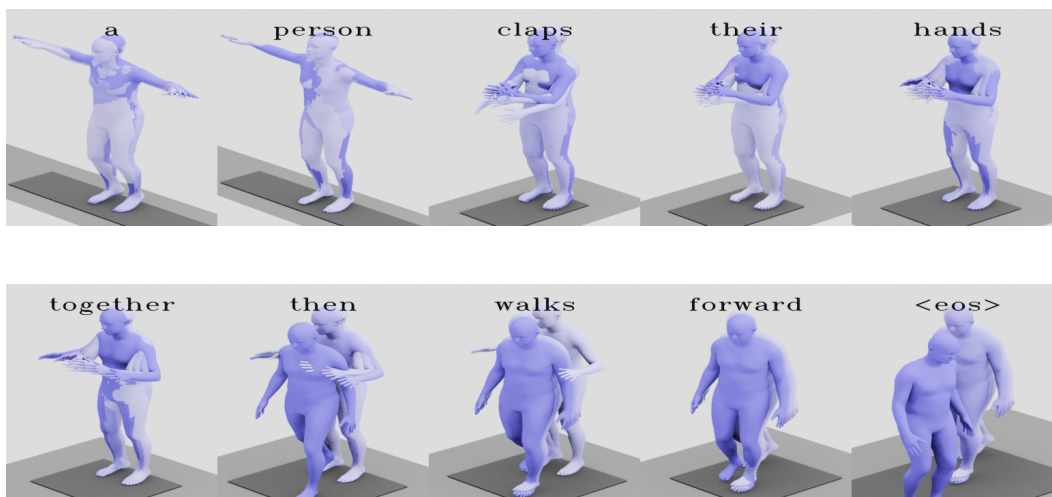


Figure 9. a person **claps** their **hands** then **walks forward**.

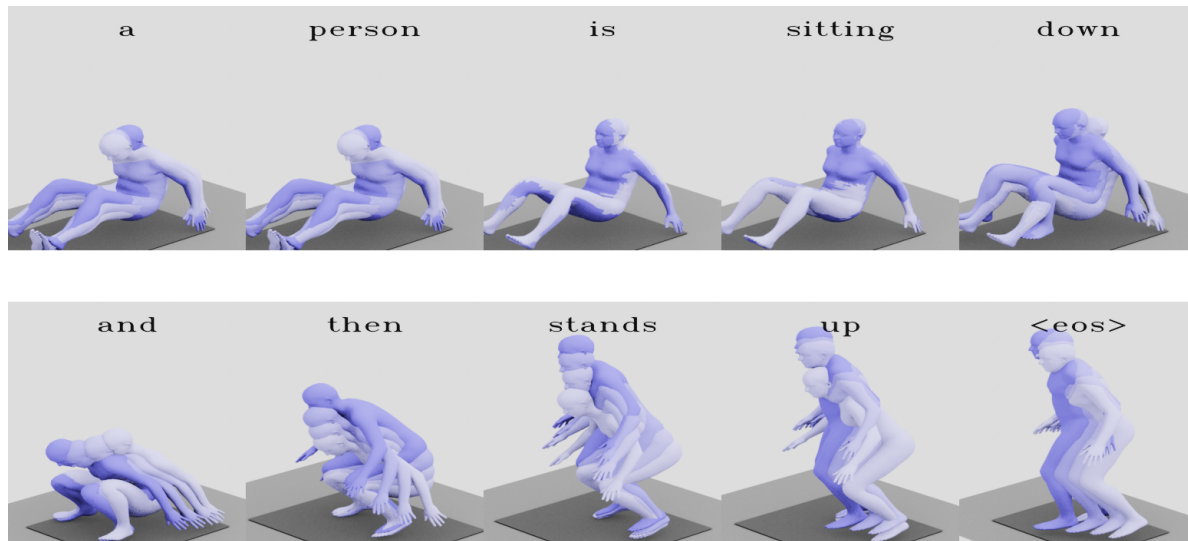


Figure 10. a person is **sitting down** and then **stands up**.

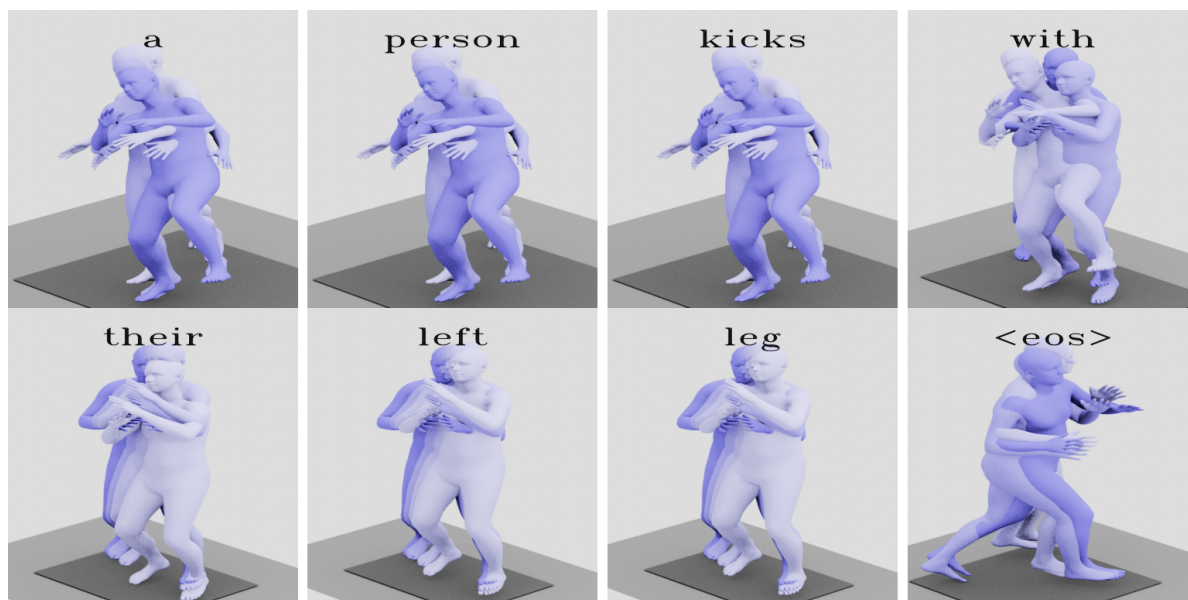


Figure 11. a person **kicks** with their **left leg**.

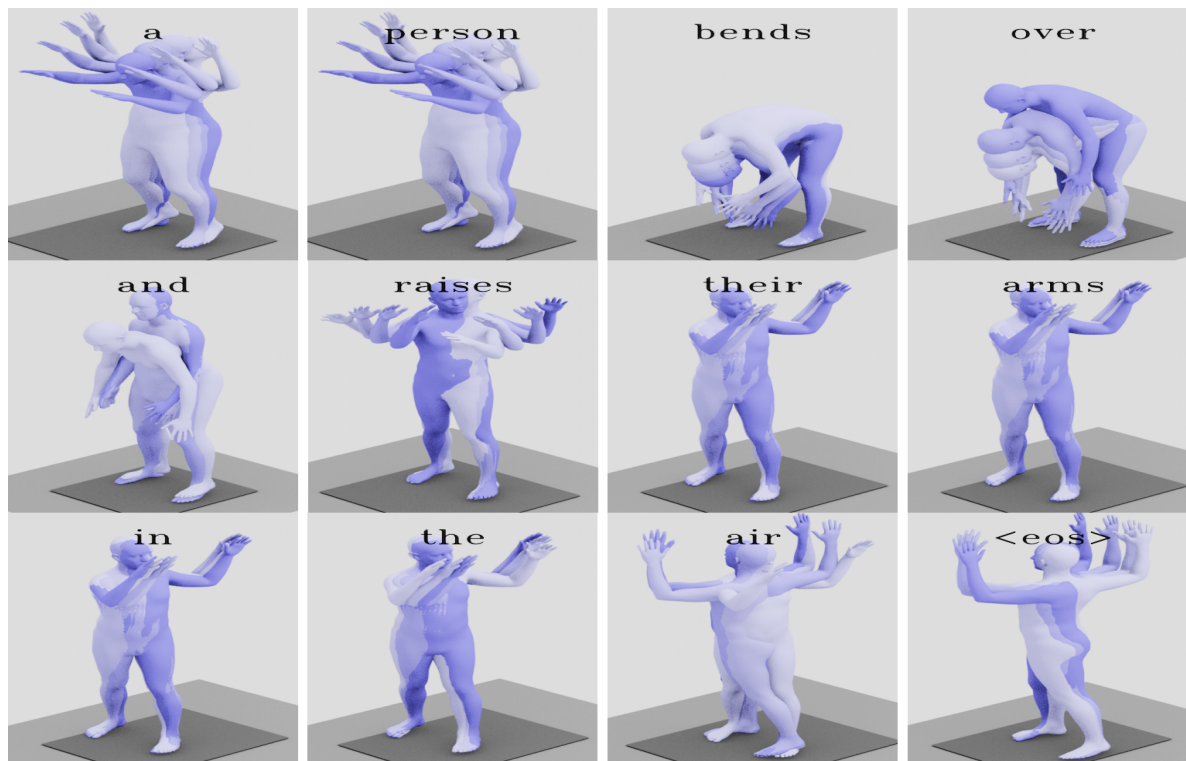


Figure 12. a person **bends** over and **raises** their **arms** in the air.

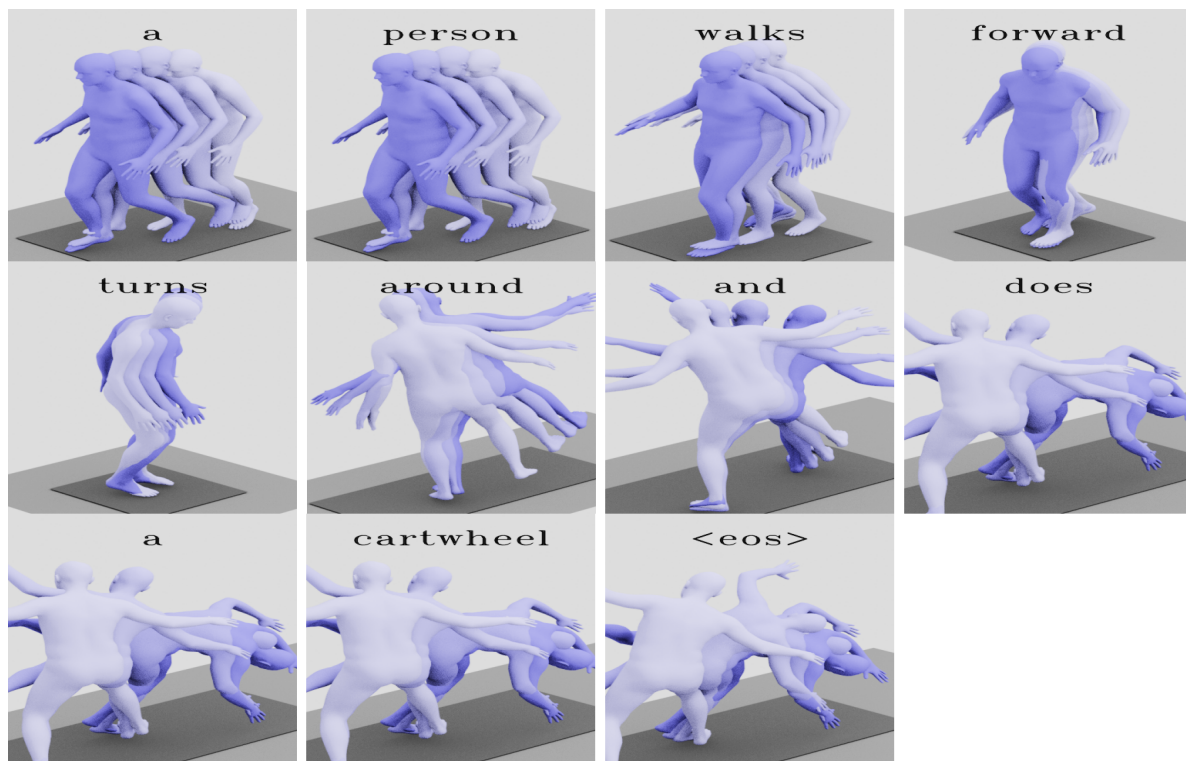


Figure 13. a person **walks forward** **turns around** and does a **cartwheel**.

C.1.2 Phrase level attention

The attention weights are aggregated by averaging across words relative to the primitive motion between motion words, then four frames of higher attention are displayed for each corresponding language segment.

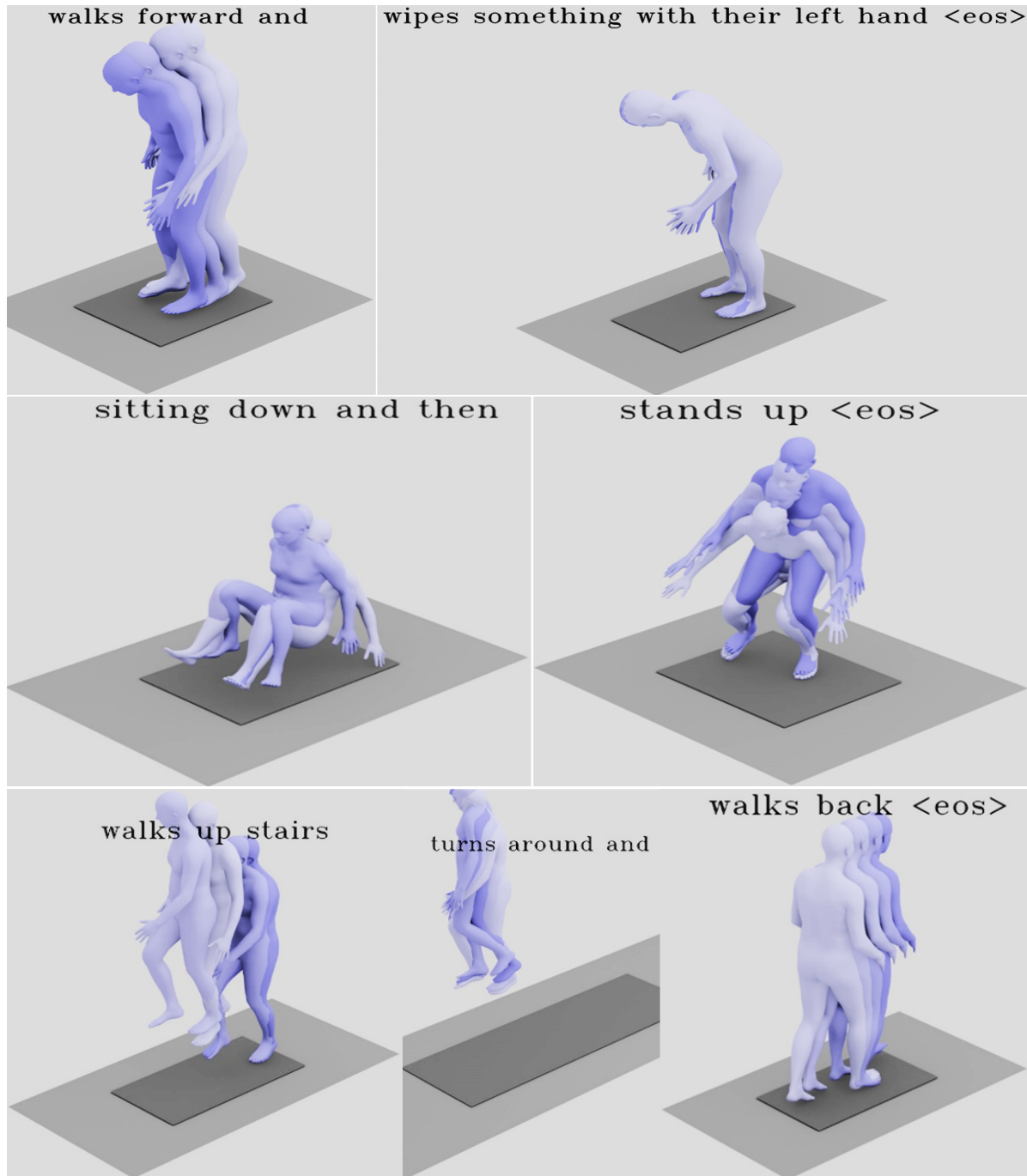
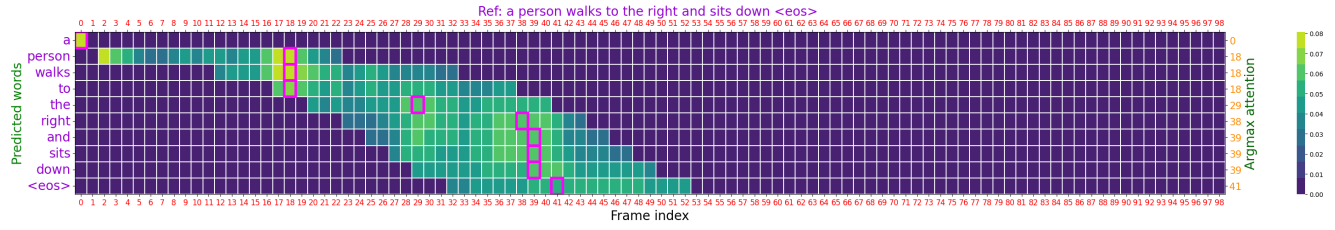


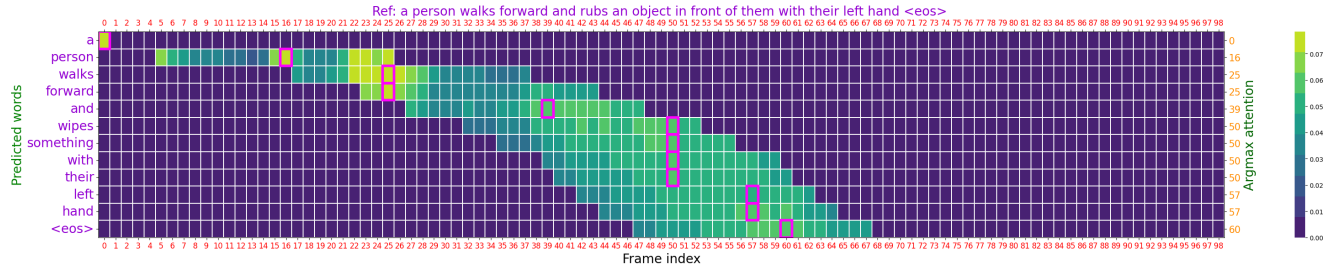
Figure 14. **Phrase-level** attention based association between motion and language segments.

C.2. Cross attention maps

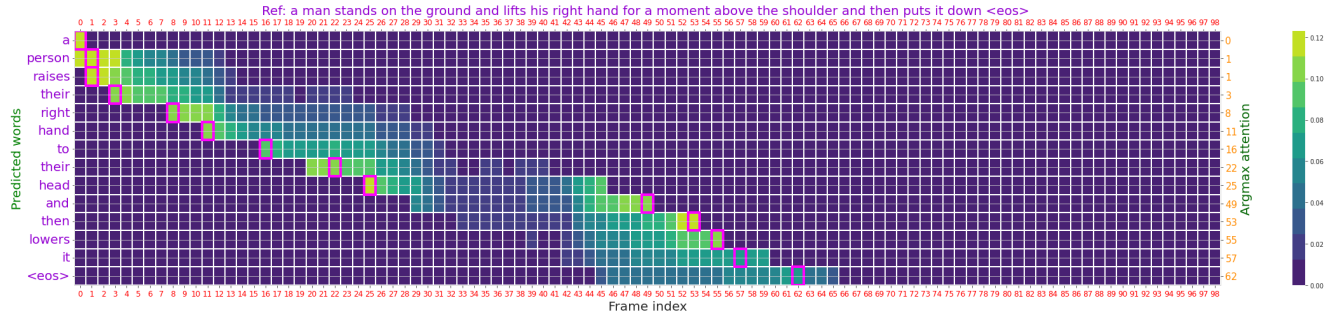
In the following attention maps of different motions (some from the same samples visualized with frozen mesh above and in the paper above for correspondence):



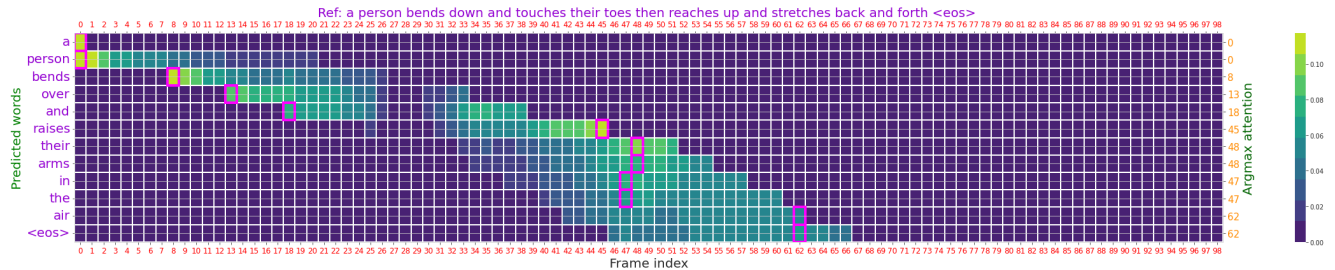
(a) Walks to the right, sits down.



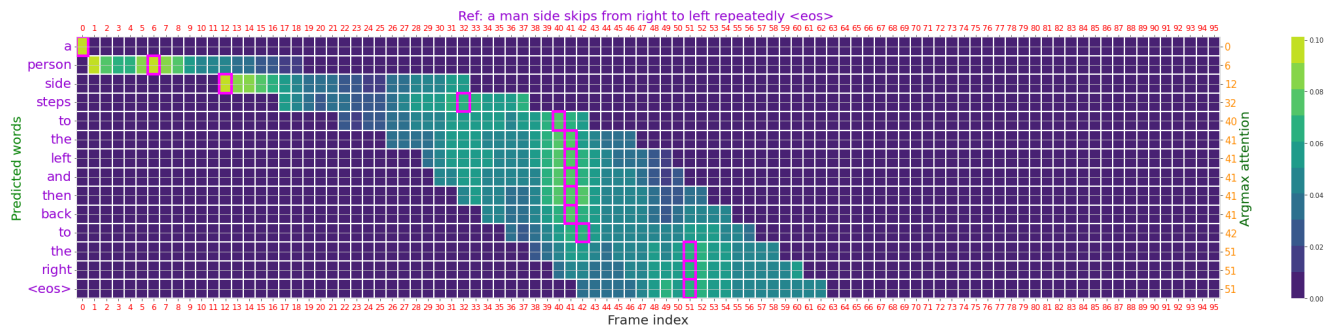
(b) Walks forward, wipes something.



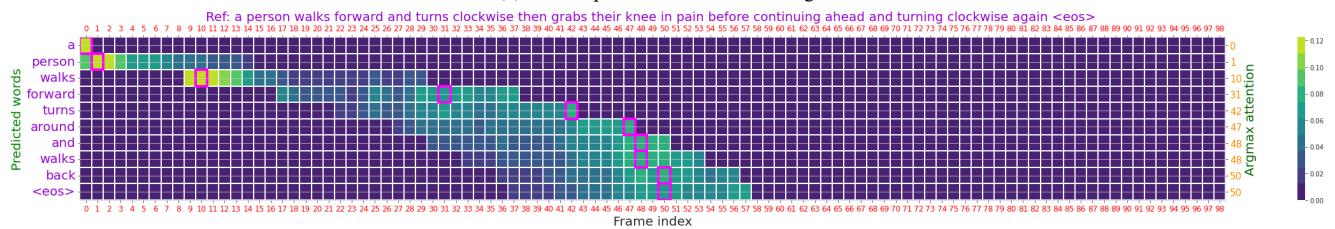
(a) raise right hand to shoulder/head level then put it down (right hand).



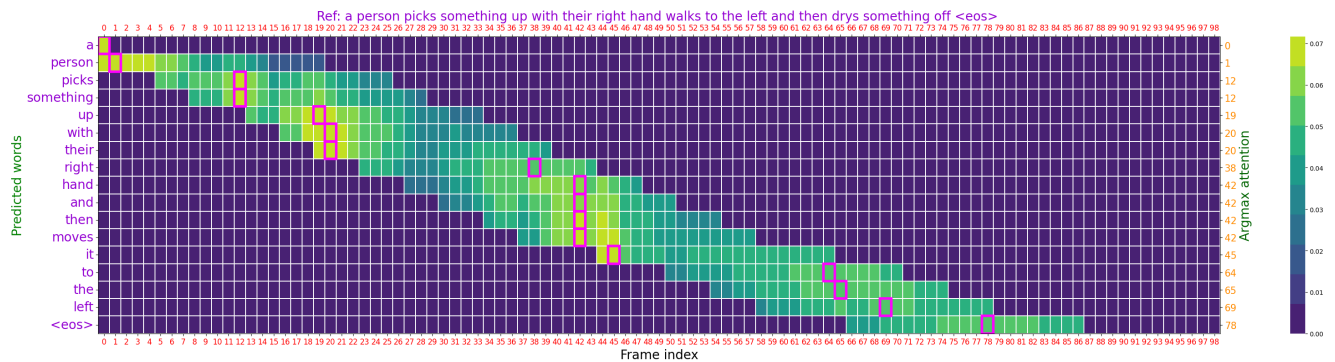
(b) bends over and raises arms in the air.



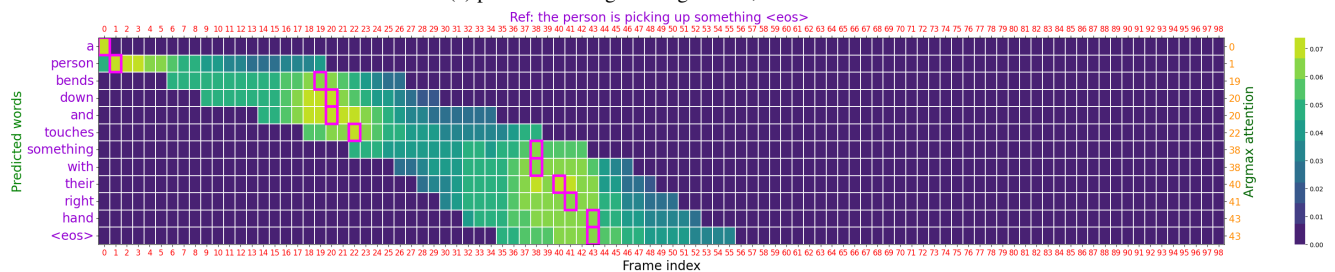
(a) Side step to the left then to the right.



(b) walks forward, turn around then walks back.



(a) picks something with right hand, move it to left.



(b) bends down and touches something.