# Transformer with Controlled Attention for Synchronous Motion Captioning

**Karim Radouane**[1,2], **Sylvie Ranwez**[1], **Julien Lagarde**[3], **Andon Tchechmedjiev**[1]

[1] EuroMov Digital Health in Motion, University of Montpellier, IMT Mines Ales, Ales, France
[2] University of Toulouse, IRIT UMR 5505, Toulouse, France
[3] University of Pau and the Adour Region, Pau, France

**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 and unsupervised action segmentation and temporal grounding. Our method introduces mechanisms to control self- and cross-attention distributions of the Transformer, allowing interpretability and 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 of the output of synchronous text generation in the code repository.

**Code** — 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 (Vaswani et al. 2017). 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 (Guo et al. 2022b; Plappert, Mandery, and Asfour 2016).



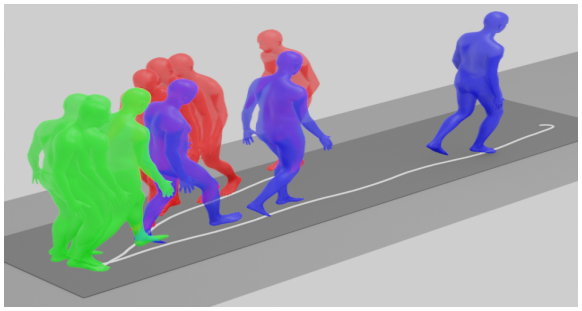


Figure 1: The goal of our approach is to enable aligned/synchronized text generation in time with corresponding actions while performing motion captioning, as example: *"a person walk forward turns around then walks back"*. Through a controlled attention mechanism our method allows seemlessly to infer time alignment without requiring time annotations for training.

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 (Guo et al. 2022a). Although such systems have been a significant focus of research (Plappert, Mandery, and Asfour 2016; Petrovich, Black, and Varol 2022; Guo et al. 2022b), there has also been an interest in motion-to-language generation (Goutsu and Inamura 2021; Guo et al. 2022b; Radouane et al. 2024), that has picked up steam with recent papers addressing synchronous motion to language generation (Radouane et al. 2023). The first motion captioning architecture (Radouane et al. 2023) aiming to synchronize the generation of descriptions with human actions was based on a very simple pre-Transformer model (RNN) and introduced extensions to the canonical attention mechanism. Their experiments were

mainly conducted on the original version of KIT-ML (Plappert, Mandery, and Asfour 2016) before augmentation (Guo et al. 2022a). 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 (Vaswani et al. 2017).

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 (Plappert, Mandery, and Asfour 2017; Guo et al. 2022b), it is evident that the field of motion generation has witnessed significant advancements, with extensive research efforts dedicated to this task (Guo et al. 2022a; Zhang et al. 2023; Ghosh et al. 2021; Petrovich, Black, and Varol 2022; Chen et al. 2023). In contrast, progress in language generation from motion has been comparatively less substantial (Goutsu and Inamura 2021). 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) (Mandery et al. 2016). 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 (Guo et al. 2022a), 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 Transformer-based architectures in recent years (Guo et al. 2022a,b), and MotionGPT (Jiang et al. 2024), which involves multi-task learning but achieves low performance ($12.47\%$) on HumanML3D and reports no results on the 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 (Krishna et al. 2017). But the alignment is performed at the phrase level instead of the word level, and, thus, it doesn't involve progressive word generation.

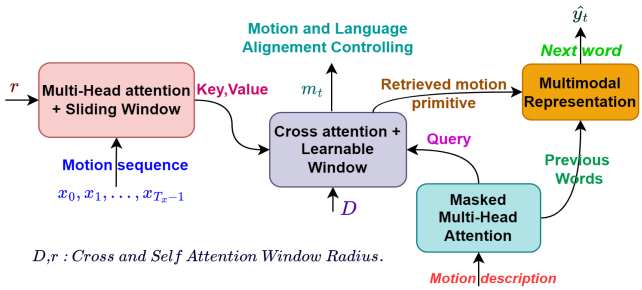
**Motion primitives and description.** Synchronizing motion and language involve implicitly to localize motion primitives and their part of description in the complete sentence. This process intersect with moment retrieval that was presented as use case of *text-to-motion retrieval* (`TMR`) task introduced by (Petrovich, Black, and Varol 2023). `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 (Radouane et al. 2023), involve automatic unsupervised alignment, enabling a simultaneous *progressive text generation* and *motion segmentation*.

## 3 Methods

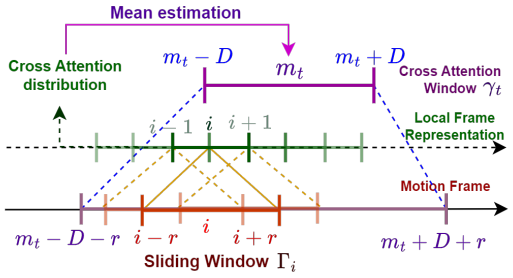
We aim for a motion to language system generating text synchronously while being fed a movement sequence. Like in the approach by (Radouane et al. 2023) who used a modified NMT architecture to enable synchronicity, we propose an evolution of the Transformer architecture to achieve the same objective. In this section, we describe our contributions by going over the main components of our approach. Figure 2 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

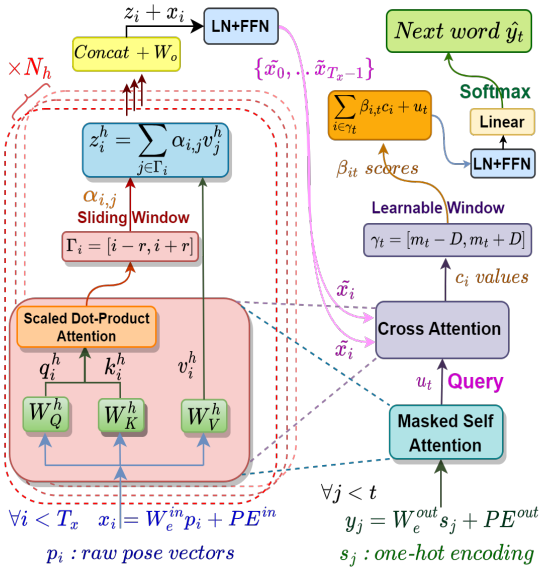
In the context of Neural Machine Translation (NMT), the Transformer employs a Multi-Head Attention Mechanism to learn contextualized token representations. Within each encoder layer, an input token's representation is formed as an aggregated representations of input tokens with different contributions (attention weights). This process results in context mixing (Mohebbi et al. 2023). Several studies have explored information mixing in the Transformer and its influence on predictions (Schwenke and Atzmueller 2021), 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




(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 $\Gamma_i$ is incorporated in the encoder side and learnable mask $\gamma_t$ 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 2: Overview of general proposed framework with relevant details.

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 (Abnar and Zuidema 2020) 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 2. We use only one layer in the Encoder/Decoder for the reasons elicited above.

## 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 2a, 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 (Beltagy, Peters, and Cohan 2020), 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. 2c. In the following, we discuss the window definition for both cases in detail.

**Self Attention Window $\Gamma_i$.** Self attention in its original form, as proposed by (Vaswani et al. 2017) 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 $\gamma_t$.** 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. 2b.

### 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 $h \in \{1, ..., H\}$ in the self-attention block, $x_i$ is transformed into a query $q_i^h$, a key $k_i^h$, and a value $v_i^h$.

$$x_i = (p_i W_e + b_e) + PE \tag{1}$$

$$q_i^h = x_i W_Q^h + b_Q^h \quad k_i^h = x_i W_K^h + b_K^h \quad v_i^h = x_i W_V^h + b_V^h \tag{2}$$

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

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

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

$$\alpha_{i,j}^h = \frac{\exp\left({q_i^h}^T \cdot k_i^h / \sqrt{d}\right)}{\sum_{j \in \Gamma_i} \exp\left({q_i^h}^T \cdot k_j^h / \sqrt{d}\right)} \tag{4}$$

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

$$z_i = \textsc{Concat}(z_i^1, ..., z_i^{N_h}) W_O \tag{5}$$

**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).

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

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

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

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 $\Gamma_i$.

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

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

**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).

$$r_t = \sum_{i \in \gamma_t} \beta_{i,t} c_i \tag{10}$$

**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. 2c).

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

### 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.

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

**Constraint on Alignment Position $\mathbf{m_t}$.** In order to obtain synchronous generation, inspired by (Radouane et al. 2023), 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*:

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

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.

$$Loss = Loss_{lang} + \lambda_0 Loss_0 + \lambda_m Loss_m \tag{13}$$

Where $(\lambda_0, \lambda_m)$ 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.

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

**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 | B-1 | B-4 | CIDEr | ROUGE | 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** |
| | $\infty$ | $\infty$ | 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 |
| | $\infty$ | $\infty$ | 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. (B:BLEU)

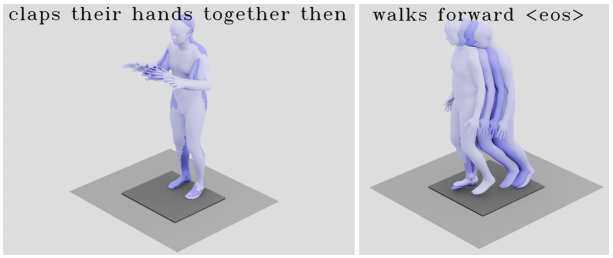
## 4.1 Ablation and Evaluation Study

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.

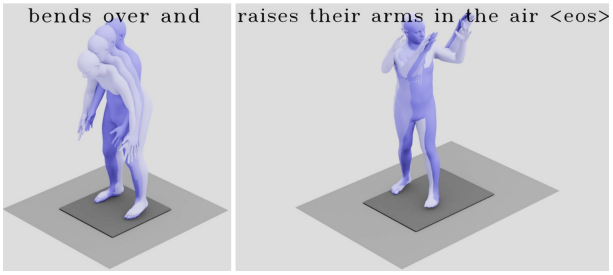
Consequently in our context the ablation study concern only two aspects : i) Effect of motion and language alignment controlling (structuring losses) and ii) Effect of masking approach: learnable and sliding window.

**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. table 1) but not in terms of synchronization between motion and language (cf.table 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 qual-




(a) claps then walks.




(b) bends then raises arms.

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

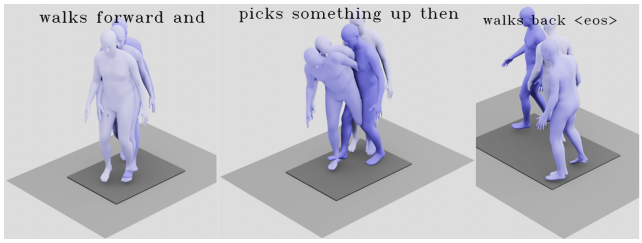
ity generation. On KIT-ML, our model significantly outperforms the TM2T model which is also Transformer-based model but with 3 layers in the Encoder and Decoder. In contrast, our model employs only one layer with fewer parameters and does not utilize beam searching, while achieving synchronous captioning. Compared to MLP+GRU and the model proposed in (Radouane et al. 2024) that uses spatial-temporal encoder and guided attention, our approach achieves superior results on both datasets.

## 4.2 Qualitative analysis

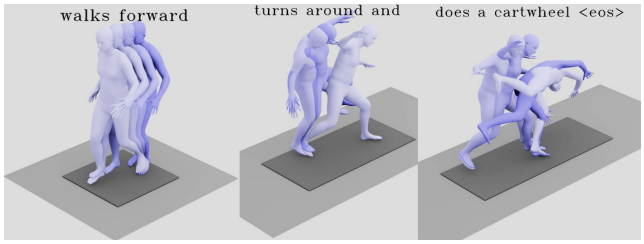
In this part we discuss qualitative results at the level of attention maps and human motion sequences frozen in time.

**Cross Attention Maps.** Example of compositional motion is shown in Figure 5 with corresponding motion ranges. The violet rectangles represents the position of maximum attention. Each word is generated at it's corresponding position (Animations in Code). In fig. 5, considering the motion words, the spread of attention for the phrase *sitting down* is in the range [16, 39], as compared to the manuel observation [0, 40]. When the subject stands up around frame 49, the predicted attention for the word *stands* peaks at frame 50. Similar analyses could be conducted on other samples. 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 (Radouane et al. 2023).

**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 in Figure 4.




(a) a person walks forward and picks something up then walks back.




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

Figure 4: Decomposition of motions and associated descriptions.

## 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 of* proposed in (Radouane et al. 2023). 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.*

**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. 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

Recent work in sign language research targets *alignment* (Bull et al. 2021), *temporal localization* (Varol et al. 2021),

| Dataset | Model | BLEU@1 | BLEU@4 | ROUGE-L | CIDEr | Bertscore |
|---|---|---|---|---|---|---|
| **KIT-ML** | RAEs (Yamada, Matsunaga, and Ogata 2018) | 30.6 | 0.10 | 25.7 | 8.00 | 0.40 |
| | Seq2Seq(Att) | 34.3 | 9.30 | 36.3 | 37.3 | 5.30 |
| | SeqGAN (Goutsu and Inamura 2021) | 3.12 | 5.20 | 32.4 | 29.5 | 2.20 |
| | TM2T w/o MT (Guo et al. 2022b) | 42.8 | 14.7 | 39.9 | 60.1 | 18.9 |
| | TM2T (Guo et al. 2022b) | 46.7 | 18.4 | 44.2 | 79.5 | 23.0 |
| | MLP+GRU (Radouane et al. 2023) | 56.8 | 25.4 | **58.8** | 125.7 | 42.1 |
| | Spat+Adapt (Radouane et al. 2024) | 58.4 | 24.7 | 57.8 | 106.2 | 41.3 |
| | **Ours** | **58.8** | **26.5** | 58.7 | **132.3** | **45.8** |
| **HML3D** | RAEs (Yamada, Matsunaga, and Ogata 2018) | 33.3 | 10.2 | 37.5 | 22.1 | 10.7 |
| | Seq2Seq(Att) | 51.8 | 17.9 | 46.4 | 58.4 | 29.1 |
| | SeqGAN (Goutsu and Inamura 2021) | 47.8 | 13.5 | 39.2 | 50.2 | 23.4 |
| | TM2T w/o MT (Guo et al. 2022b) | 59.5 | 21.2 | 47.8 | 68.3 | 34.9 |
| | TM2T (Guo et al. 2022b) | 61.7 | 22.3 | 49.2 | **72.5** | 37.8 |
| | MLP+GRU (Radouane et al. 2023) | 67.0 | 23.4 | 53.8 | 53.7 | 37.2 |
| | Adapt (Radouane et al. 2024) | 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.

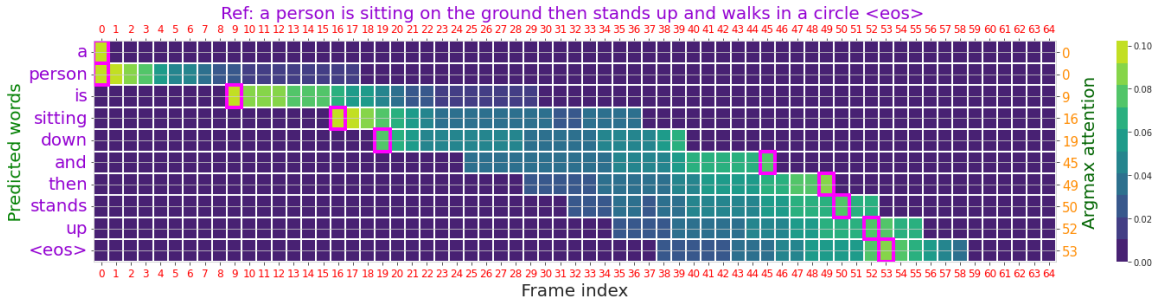



Figure 5: Cross-attention map of compositional motion with frame ranges for each action (D = r = 10): Sitting [0–40], Stand-up [41–60]. Attention for motion words aligns with their corresponding ranges.

| 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 |
| $\infty$ | $\infty$ | 39.93 | 39.96 | 46.98 | 26.5 |
| MLP+GRU (Radouane et al. 2023) | | 36.29 | 53.71 | 58.33 | 23.4 |

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. In comparison, the final line shows the synchronization performance of the MLP+GRU model.

and *sign spotting* (Momeni et al. 2020). Our approach can be leveraged in different contexts, such as: **Aligned sign language translation.** Links signs and language segments by mapping upper-body poses to words, enabling unsupervised alignment. **Temporal grounding and action localization.** Cross-attention maps pose streams to actions, infers unlabeled temporal boundaries, leverages attention weights for precision and interpretability, and supports phrase-motion grounding, an emerging domain for human poses (Fujiwara, Tanaka, and Yu 2024; Wang, Kang, and Mu 2024).

## 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


Abnar, S.; and Zuidema, W. 2020. Quantifying Attention Flow in Transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4190–4197. Online: Association for Computational Linguistics.

Beltagy, I.; Peters, M. E.; and Cohan, A. 2020. Longformer: The Long-Document Transformer.

Bull, H.; Afouras, T.; Varol, G.; Albanie, S.; Momeni, L.; and Zisserman, A. 2021. Aligning Subtitles in Sign Language Videos. In *ICCV*.

Chen, X.; Jiang, B.; Liu, W.; Huang, Z.; Fu, B.; Chen, T.; and Yu, G. 2023. Executing your Commands via Motion Diffusion in Latent Space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18000–18010.

Fujiwara, K.; Tanaka, M.; and Yu, Q. 2024. Chronologically Accurate Retrieval for Temporal Grounding of Motion-Language Models. In *ECCV*.

Ghosh, A.; Cheema, N.; Oguz, C.; Theobalt, C.; and Slusallek, P. 2021. Synthesis of Compositional Animations from Textual Descriptions. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 1376–1386.

Goutsu, Y.; and Inamura, T. 2021. Linguistic Descriptions of Human Motion with Generative Adversarial Seq2Seq Learning. In *Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA)*, 4281–4287. IEEE. ISBN 978-1-7281-9077-8.

Guo, C.; Zou, S.; Zuo, X.; Wang, S.; Ji, W.; Li, X.; and Cheng, L. 2022a. Generating Diverse and Natural 3D Human Motions From Text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5152–5161.

Guo, C.; Zuo, X.; Wang, S.; and Cheng, L. 2022b. TM2T: Stochastic and Tokenized Modeling for the Reciprocal Generation of 3D Human Motions and Texts. In *ECCV*.

Jiang, B.; Chen, X.; Liu, W.; Yu, J.; Yu, G.; and Chen, T. 2024. MotionGPT: Human Motion as a Foreign Language. *Advances in Neural Information Processing Systems*, 36.

Krishna, R.; Hata, K.; Ren, F.; Fei-Fei, L.; and Carlos Niebles, J. 2017. Dense-Captioning Events in Videos. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Mandery, C.; Ömer Terlemez; Do, M.; Vahrenkamp, N.; and Asfour, T. 2016. Unifying Representations and Large-Scale Whole-Body Motion Databases for Studying Human Motion. *IEEE Transactions on Robotics*, 32: 796–809.

Mohebbi, H.; Zuidema, W.; Chrupała, G.; and Alishahi, A. 2023. Quantifying Context Mixing in Transformers. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 3378–3400. Dubrovnik, Croatia: Association for Computational Linguistics.

Momeni, L.; Varol, G.; Albanie, S.; Afouras, T.; and Zisserman, A. 2020. Watch, read and lookup: learning to spot signs from multiple supervisors. In *ACCV*.

Petrovich, M.; Black, M. J.; and Varol, G. 2022. TEMOS: Generating diverse human motions from textual descriptions. In *European Conference on Computer Vision (ECCV)*.

Petrovich, M.; Black, M. J.; and Varol, G. 2023. TMR: Text-to-Motion Retrieval Using Contrastive 3D Human Motion Synthesis. In *ICCV*.

Plappert, M.; Mandery, C.; and Asfour, T. 2016. The KIT Motion-Language Dataset. *Big Data*, 4(4): 236–252.

Plappert, M.; Mandery, C.; and Asfour, T. 2017. Learning a bidirectional mapping between human whole-body motion and natural language using deep recurrent neural networks. *Robotics and Autonomous Systems*, 109: 13–26.

Radouane, K.; Lagarde, J.; Ranwez, S.; and Tchechmedjiev, A. 2024. Guided Attention for Interpretable Motion Captioning. In *Proceedings of the 35th British Machine Vision Conference*.

Radouane, K.; Tchechmedjiev, A.; Lagarde, J.; and Ranwez, S. 2023. Motion2language, unsupervised learning of synchronized semantic motion segmentation. *Neural Computing and Applications*, 36(8): 4401–4420.

Schwenke, L.; and Atzmueller, M. 2021. 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.

Varol, G.; Momeni, L.; Albanie, S.; Afouras, T.; and Zisserman, A. 2021. Read and Attend: Temporal Localisation in Sign Language Videos. In *CVPR*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*.

Wang, X.; Kang, Z.; and Mu, Y. 2024. Text-controlled Motion Mamba: Text-Instructed Temporal Grounding of Human Motion. *arxiv:2404.11375*.

Yamada, T.; Matsunaga, H.; and Ogata, T. 2018. Paired Recurrent Autoencoders for Bidirectional Translation Between Robot Actions and Linguistic Descriptions. *IEEE Robotics and Automation Letters*, 3: 3441–3448.

Zhang, J.; Zhang, Y.; Cun, X.; Huang, S.; Zhang, Y.; Zhao, H.; Lu, H.; and Shen, X. 2023. 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)*.