Transformer with Controlled Attention for Synchronous Motion Captioning

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

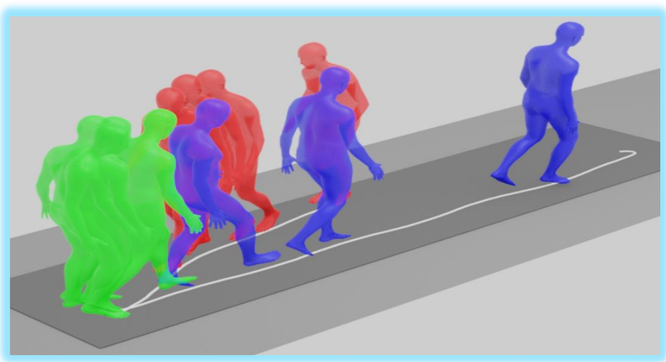
EuroMov DHM-IMT Mines Alès[1], IRIT[2]

QR code

Task and Motivation

- **Synchronous Motion Captioning** aims to generate natural language descriptions in which motion words are temporally synchronized with the corresponding actions in a human motion sequence.
- This task pertains to numerous applications, such as **aligned sign language transcription**, **unsupervised action segmentation** and **temporal phrase grounding**.

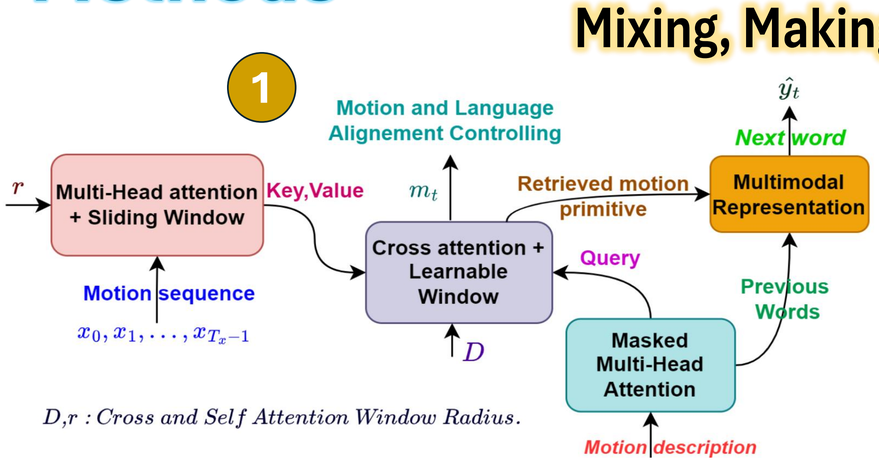
✓ Our approach enable **aligned/synchronized text generation** in time with corresponding actions while performing motion captioning.

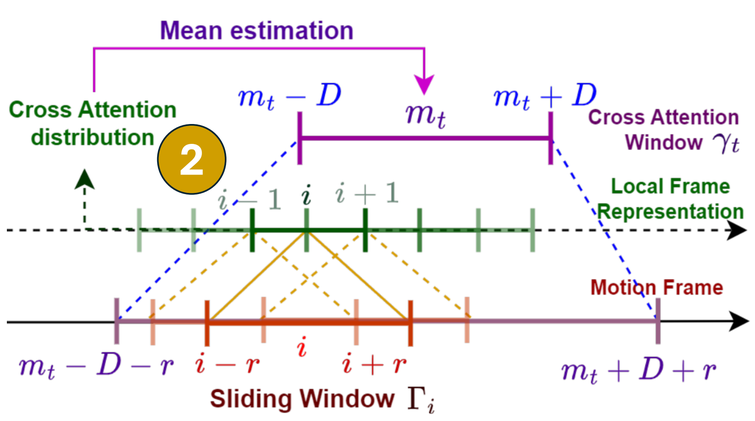


a person **walks forward, turns around,** then **walks back**.

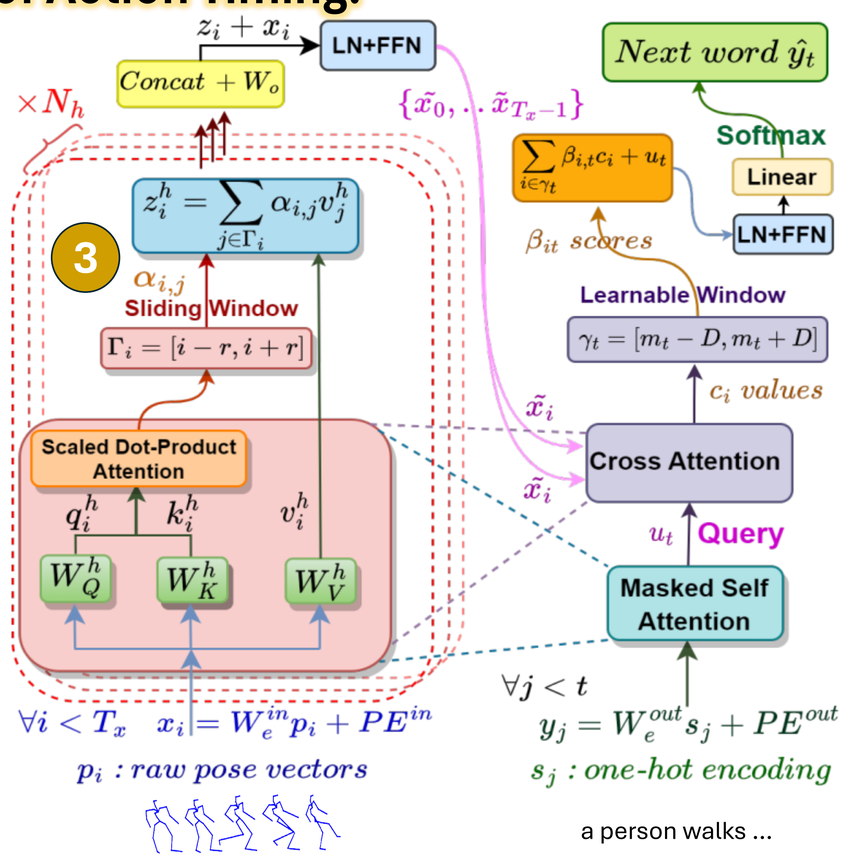
Methods

Controlled Attention and Local Motion Encoding Resolve Source Information Mixing, Making Transformer Attention Informative of Action Timing.





1. A **motion primitive** is retrieved based on a text query via the relevant key, using a weighted sum of the relevant attention scores.
2. The decoder's receptive field during generation spans the primitive motion range, where $\boldsymbol{m_t}$ is the **motion-word alignment** position estimated from the cross-attention distributions.
3. Through a **controlled attention mechanism,** our method enables learning **word-motion temporal alignment** without supervision.



Experimental results

✓ Our model outperforms previous approaches on both **KIT-ML** and **HumanML3D** in text generation performance.

Dataset	Model	BLEU@1	BLEU@4	ROUGE-L	CIDEr	Bertscore
KIT-ML	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	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**

✓ Improved Motion-Word temporal alignment, better synchronicity.

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
MLP+GRU (Radouane et al. 2023)		36.29	53.71	58.33	23.4

Ablations

- Even with masking, **multi-layer Transformers** show **lower synchronization** scores than a **single layer**.
- As the **receptive field grows**, early frame representations **mix distant-frame information**, causing attention to concentrate on regions **uninformative about action timing**.

# Layers	Mask.	BLEU@4 ↑	IoU ↑	IoP ↑	Element of ↑
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

Training loss

Initial position → $Loss_0 = m_0/T_x$

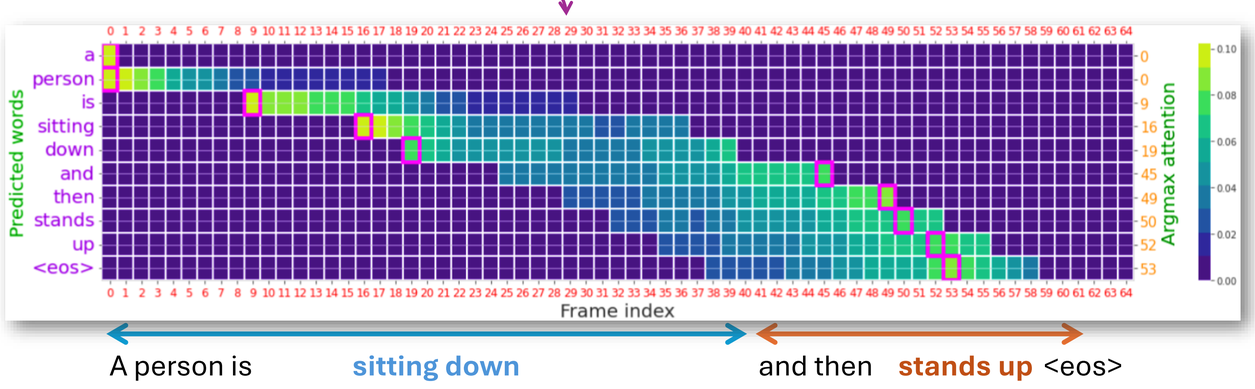
Structuring loss → $Loss_m = \frac{1}{T_x} \sum_{t<T_x-1} \max((m_t + m) - m_{t+1}, 0)^2$

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

Total loss → $Loss = Loss_{lang} + \lambda_0 Loss_0 + \lambda_m Loss_m$

Qualitative analysis

- The model's learned attention enables action-timed **motion word** generation, achieving motion–language synchronization.



- **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.