Guided Attention for Interpretable Motion Captioning

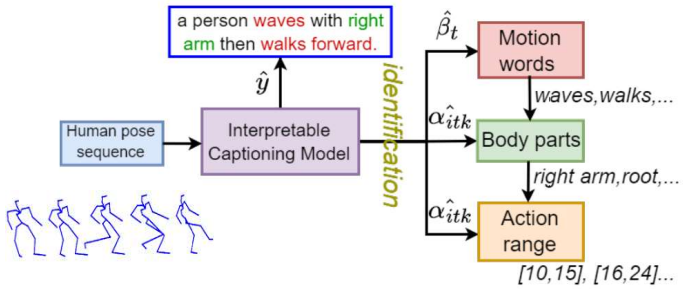
Karim Radouane[1,2], Julien Lagarde[1], Sylvie Ranwez[1], Andon Tchechmedjiev[1]
EuroMov DHM-IMT Mines Alès[1], University Sorbonne Paris Nord[2]

QR code

Task and Goal

- **Motion Captioning** task aims to generate a natural language description from a sequence of human poses.
- Our goal is to design an architecture with a **transparent reasoning process**, aligned with human-like attention perception and analysis, thereby improving performance through interpretability.
- We **qualitatively** demonstrate the interpretability of our approach and leverage it for unsupervised action localization, motion word extraction and fine-grained event description.

Model Inference



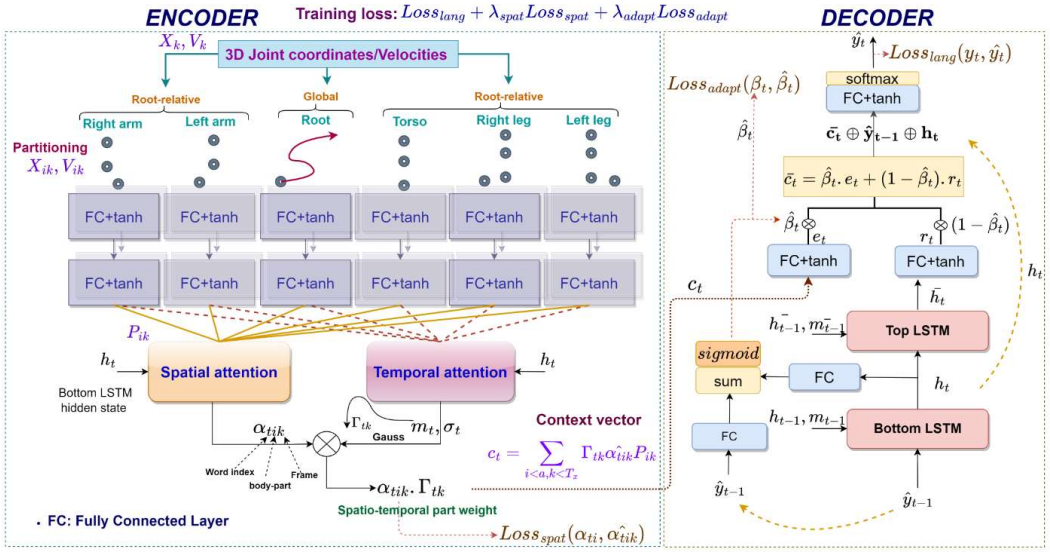
Method

Body-part encoding with **separate local and global motion information** for improved interpretability and efficient feature extraction.

Frame wise encoding for enhanced attention-based action localization.

Human inspired spatial-temporal attention mechanism with **action-aware guidance** to emphasize involved body parts and motion words for fine-grained motion description.

Adaptive gating of motion information over language for text generation.



Experimental results

Evaluation with a beam size of 2, while * indicates a greedy search. Our model outperforms previous approaches on the HumanML3D dataset, whereas MLP+GRU performs better on the smaller KIT-ML dataset.

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

Ablations

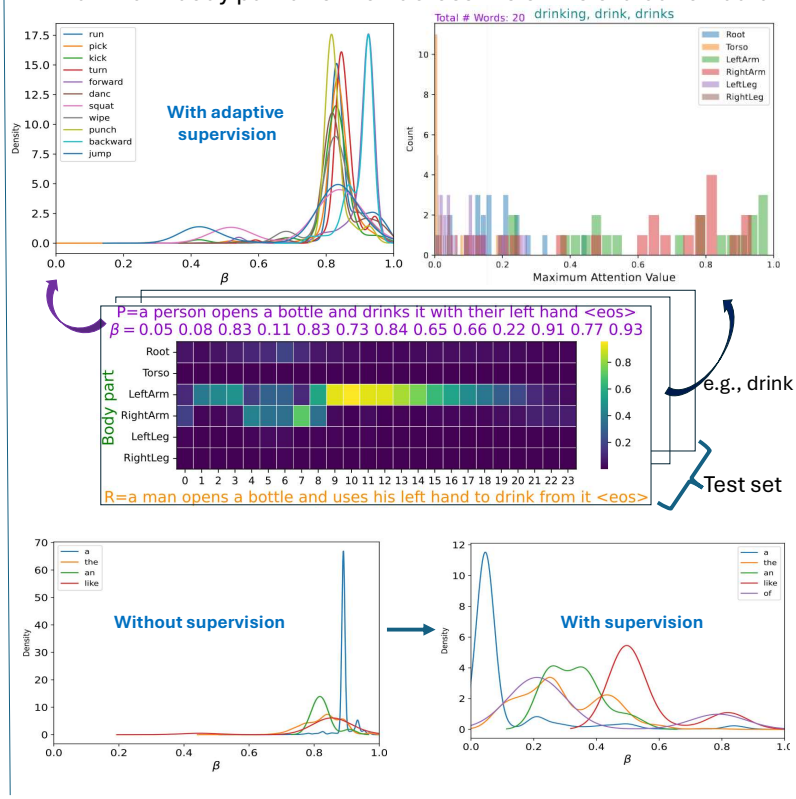
- Spatial and adaptive attention enhances performance on the KIT-ML dataset.
- Interestingly for larger HML3D dataset, which has richer annotations with body part information, the model design enables the learning of more effective interpretable spatial attention with only adaptive guidance.

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

Interpretability analysis

We analyze the correspondence level between learned attentions and human perception.

- An **attention map** for each **motion word** shows relevant body parts and keyframes for action localization.
- We **quantify** the level of interpretability using density distribution of adaptive gate and the histogram of temporal maximum body part attention across the entire evaluation data.



References

[13] Motion2language, unsupervised learning of synchronized semantic motion segmentation.

[5] TM2T: Stochastic and Tokenized Modeling for the Reciprocal Generation of 3D Human Motions and Texts.