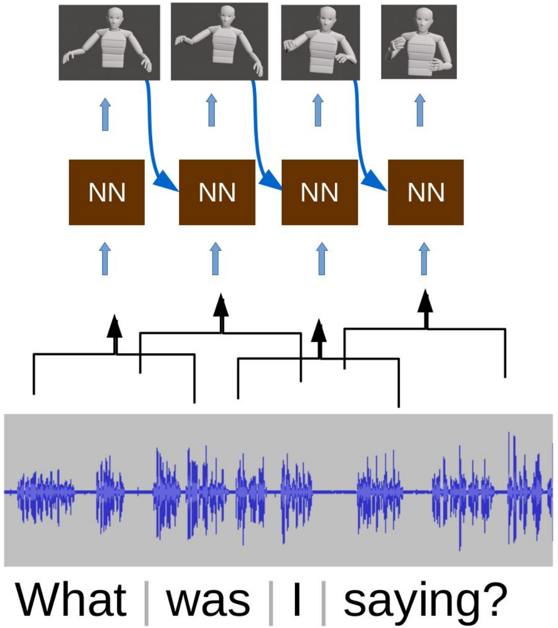
Gesticulator: A framework for semantically-aware speech-driven gesture generation

Taras Kucherenko, Patrik Jonell, Sanne van Waveren, Gustav Eje Henter, Simon Alexanderson, Iolanda Leite, and Hedvig Kjellström

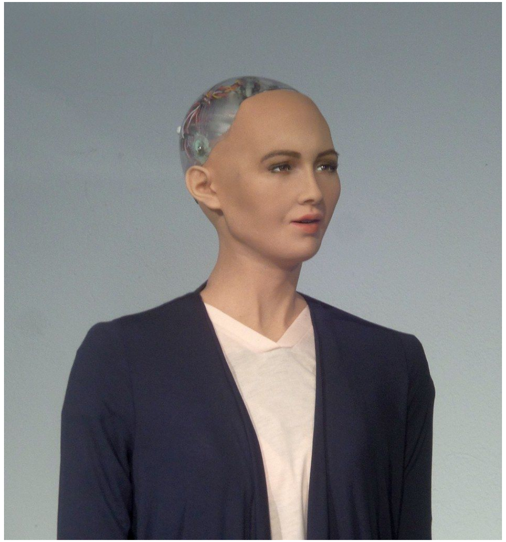
KTH Royal Institute of Technology,
Stockholm, Sweden



ICMI 2020
UTRECHT
25-29 October
Utrecht, NL

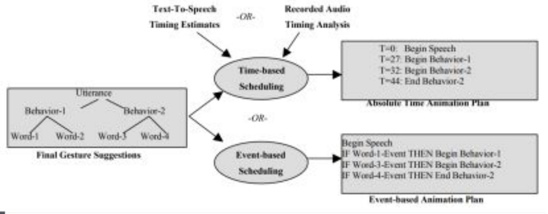
ICMI 2020

Importance of Gestures

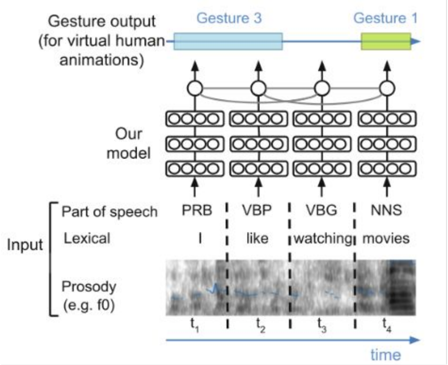


≠

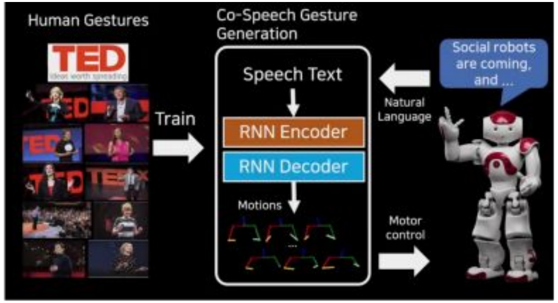
Previous work



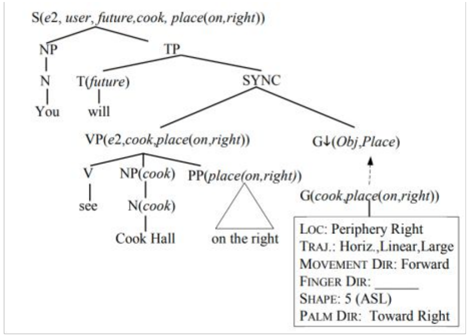
Cassell et al. "BEAT: the Behavior Expression Animation Toolkit" In SIGGRAPH, 2001.



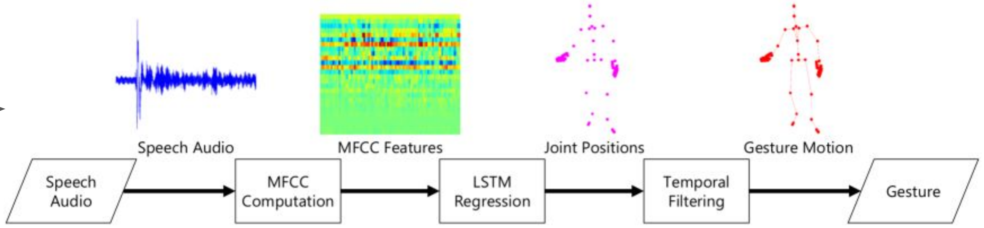
Chung-Cheng Chiu, Louis-Philippe Morency, and Stacy Marsella.
Predicting co-verbal gestures: a deep and temporal modeling approach.
International Conference on Intelligent Virtual Agents. 2015.



Yoon et al. "Robots Learn Social Skills: End-to-End Learning of Co-Speech Gesture Generation for Humanoid Robots." In ICRA. 2019

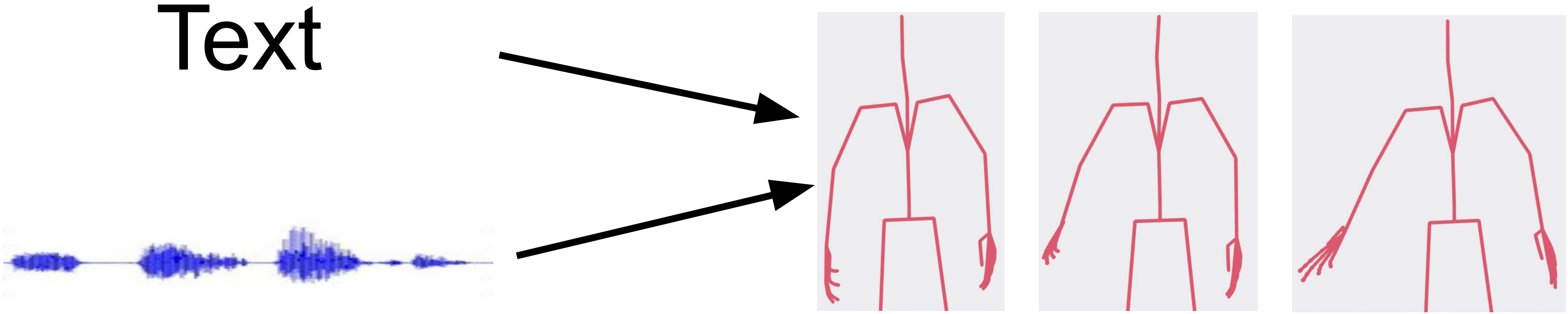


Stefan Kopp, Paul Tepper, and Justine Cassell. 2004.
Towards integrated microplanning of language and iconic gesture for multimodal output.
In Proceedings of the 6th international conference on Multimodal interfaces (ICMI '04).

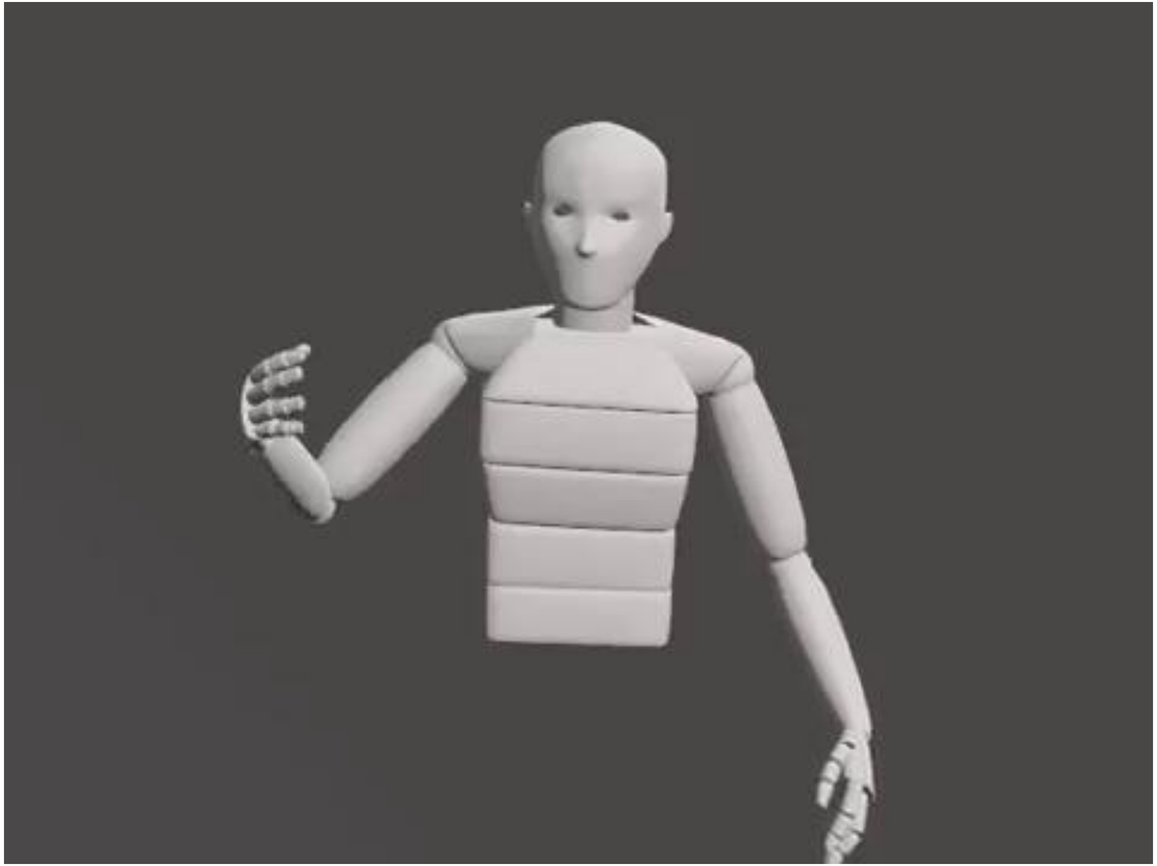


Dai Hasegawa, Naoshi Kaneko, Shinichi Shirakawa, Hiroshi Sakuta, and Kazuhiko Sumi
"Evaluation of Speech-to-Gesture Generation Using Bi-Directional LSTM Network."
International Conference on Intelligent Virtual Agents. 2018.

Multi-modal Gesture Generation



Text

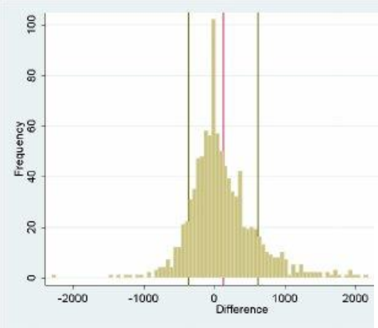


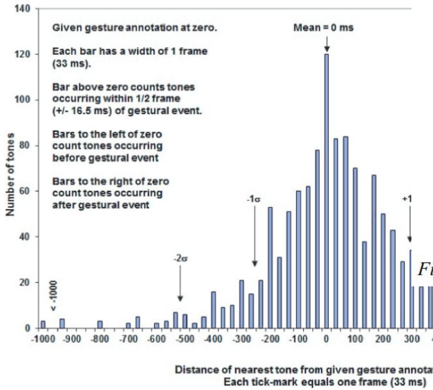
Example of generated gestures

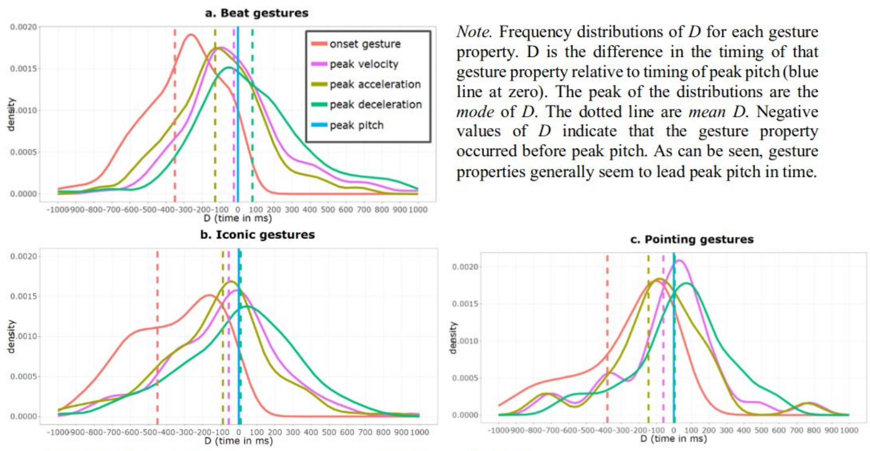
Gesture-Speech Alignment

Table 4 Mean percent (*SD*) of gestures accompanying fluent speech by timing relationship for each language group

	English		French	
				Bilingual
				86.1 (9.0)
				55.9 (28.1)







Note. Frequency distributions of *D* for each gesture property. D is the difference in the timing of that gesture property relative to timing of peak pitch (blue line at zero). The peak of the distributions are the *mode* of *D*. The dotted line are *mean D*. Negative values of *D* indicate that the gesture property occurred before peak pitch. As can be seen, gesture properties generally seem to lead peak pitch in time.

Figure 2. Distribution of D's: Gesture properties relative to peak pitch.

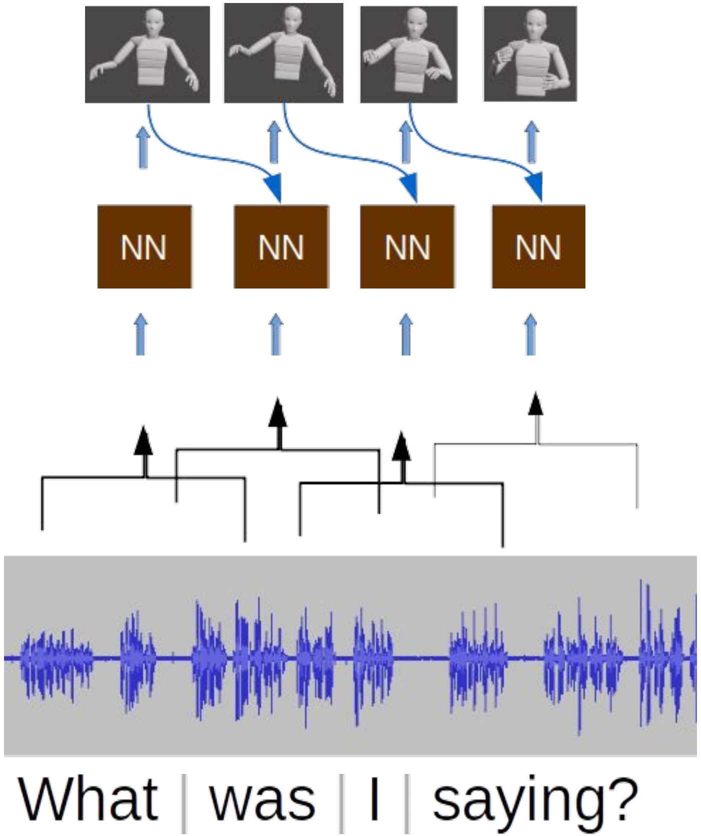
d Paula
align
ual and
ng 70.1

Bergmann, Kirsten, Volkan Aksu,
Kopp. "The relation of speech anc
Temporal synchrony follows sema
synchrony." *Proceedings of the 2r*
Workshop on Gesture and Speec
Interaction (GeSpIn 2011). 2011.

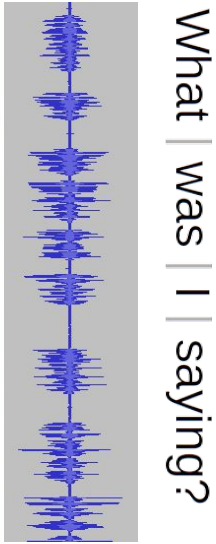
Loehr, Daniel P. "Temporal, structural, and pragmatic synchrony between intonation and gesture." *Laboratory Phonology* 3.1 (2012): 71-89.

Pouw, Wim, and James A. Dixon. "Quantifying gesture-speech synchrony." *the 6th Gesture and Speech in Interaction Conference*. Universitaetsbibliothek Paderborn, 2019.

Model Overview

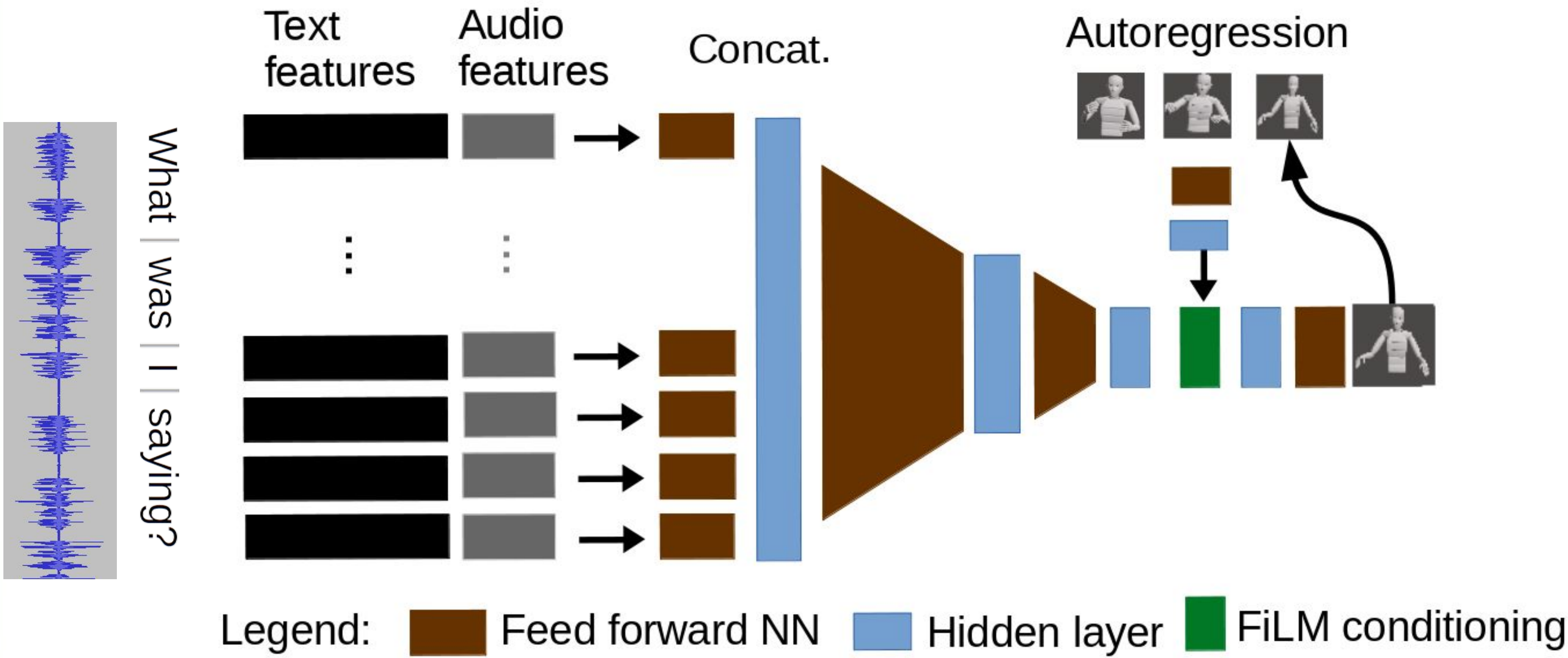


Gesticulator Framework



Legend: 🟫 Feed forward NN 🟦 Hidden layer 🟩 FiLM conditioning

Gesticulator Framework



Text features

Audio features

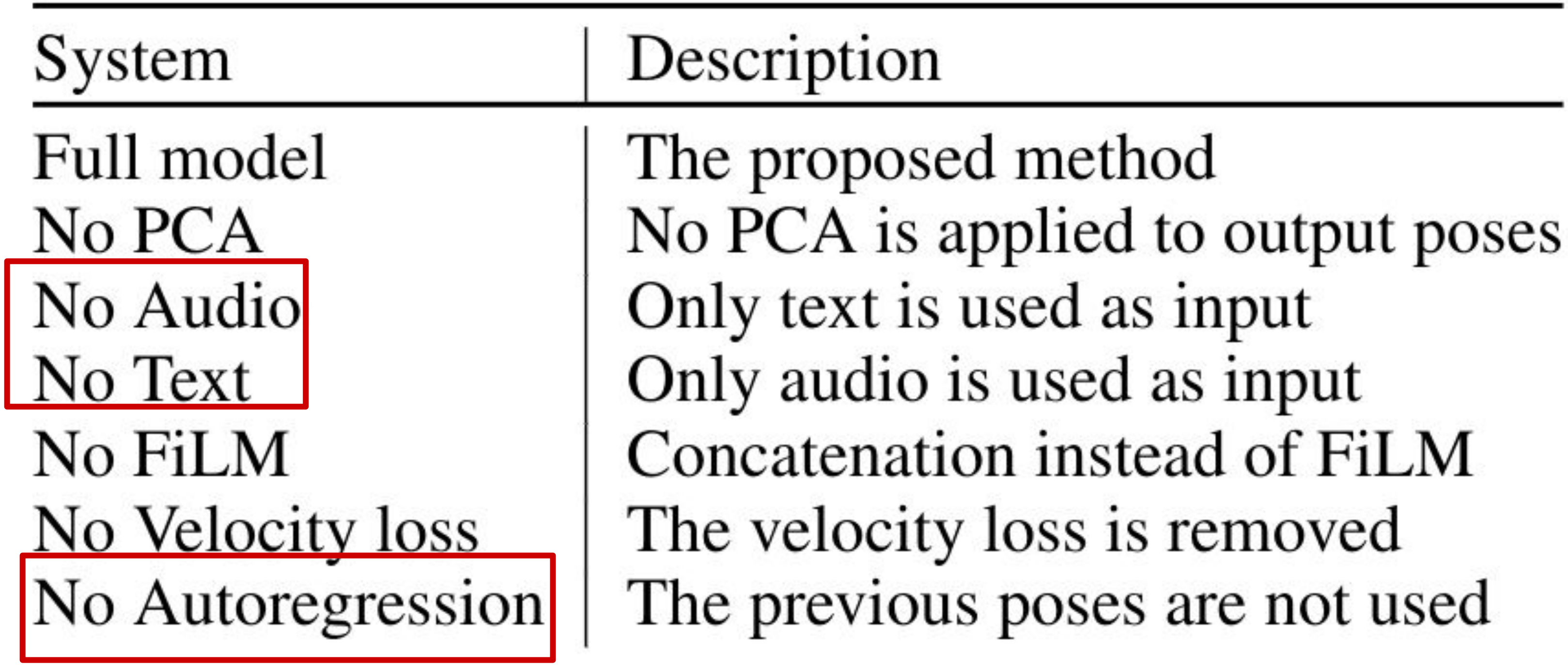
Concat.

Autoregression

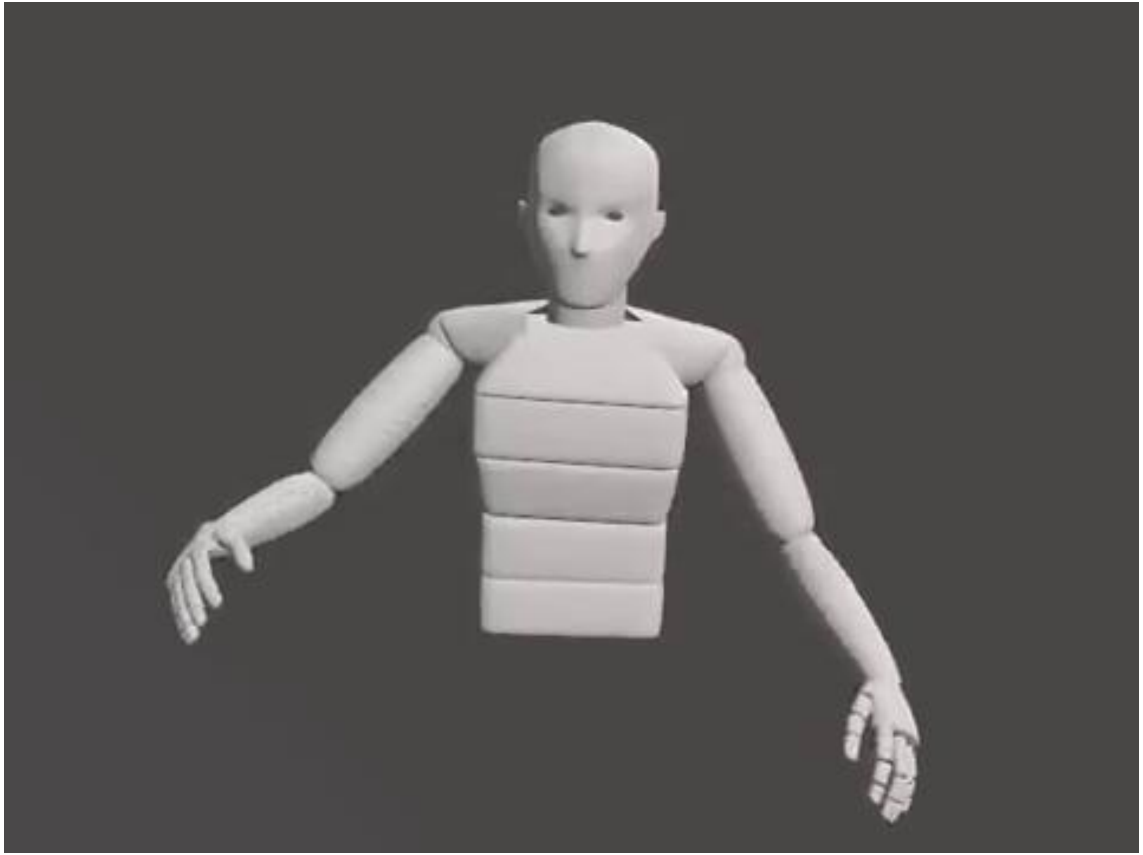
What | was | I | saying?

Legend: 🟫 Feed forward NN 🟦 Hidden layer 🟩 FiLM conditioning

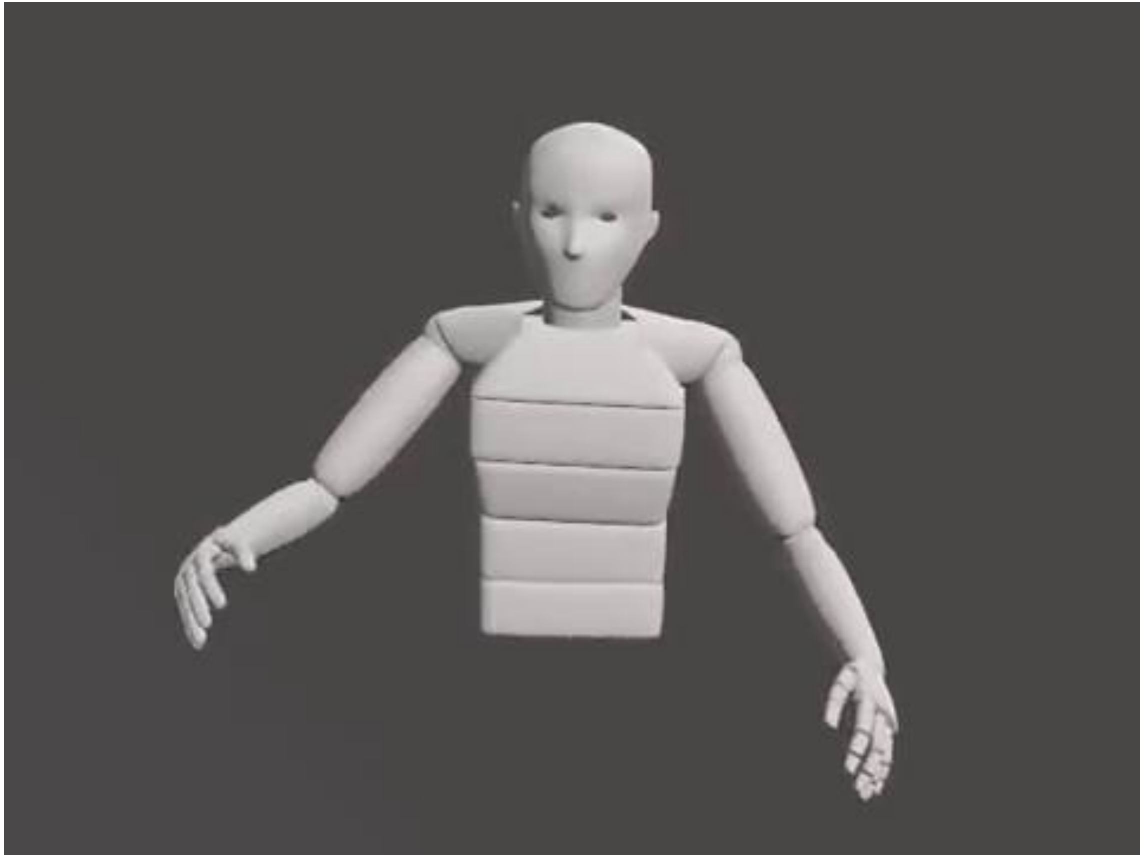
Ablations



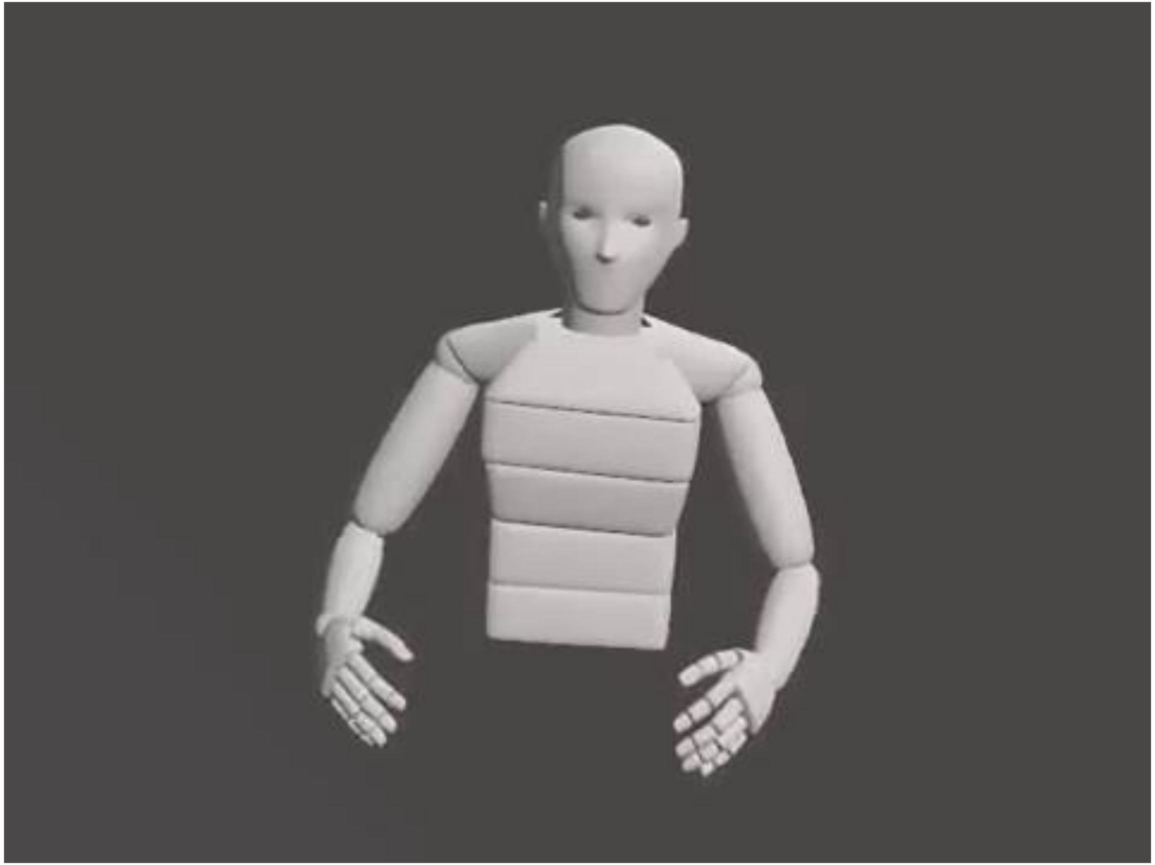
System	Description
Full model	The proposed method
No PCA	No PCA is applied to output poses
No Audio	Only text is used as input
No Text	Only audio is used as input
No FiLM	Concatenation instead of FiLM
No Velocity loss	The velocity loss is removed
No Autoregression	The previous poses are not used



Full Model

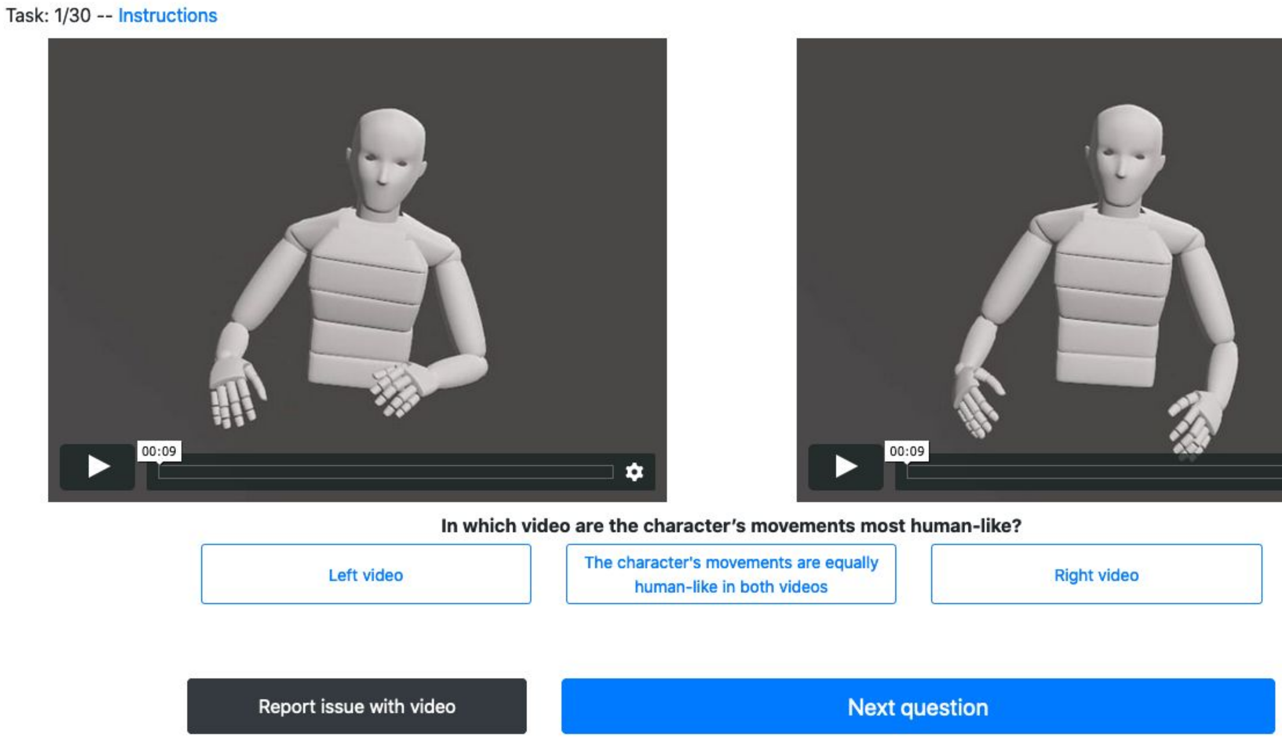


No Autoregression



No Text

User Study Setup



Task: 1/30 -- Instructions

In which video are the character's movements most human-like?

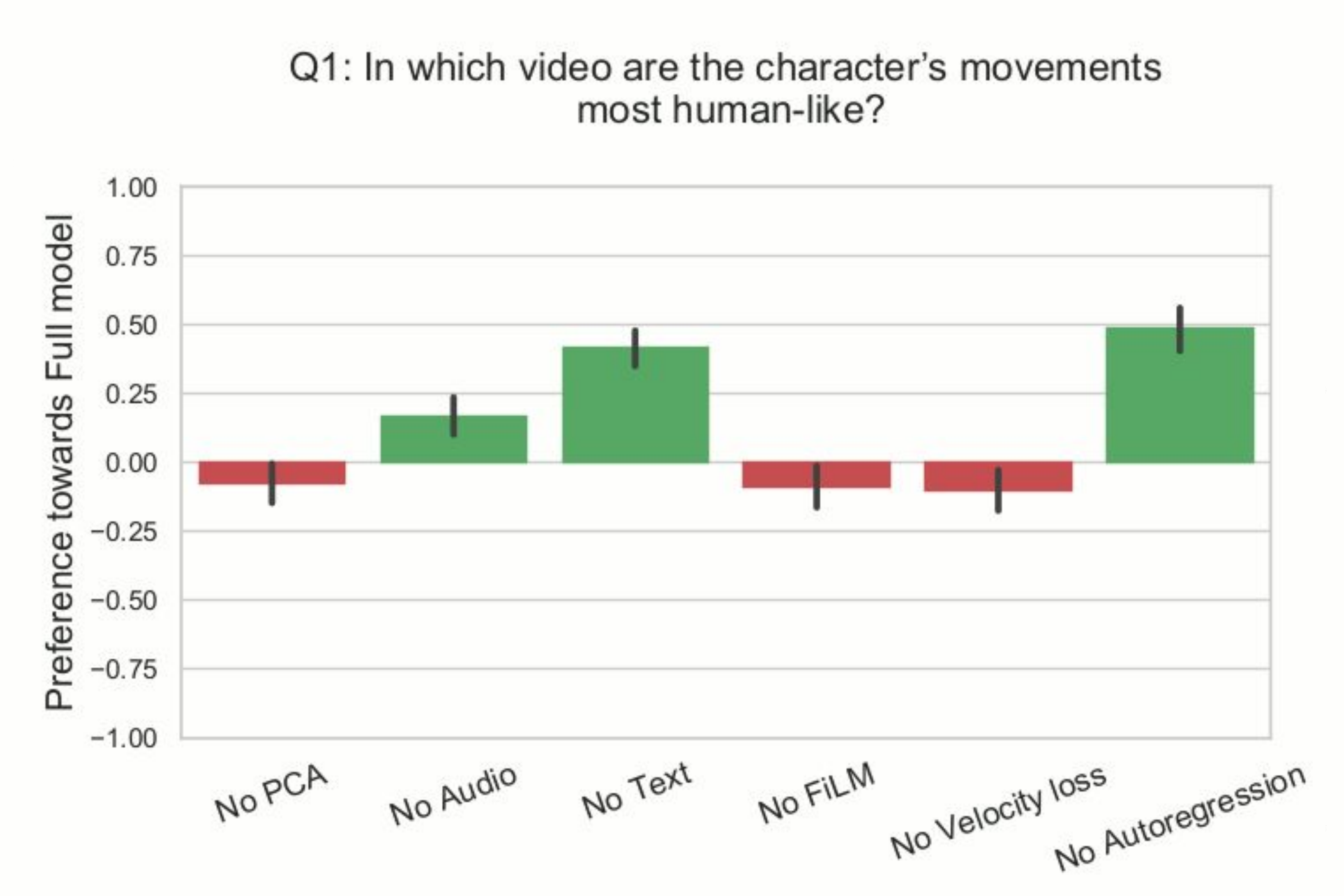
User Study

- 6 ablations compared against the full model

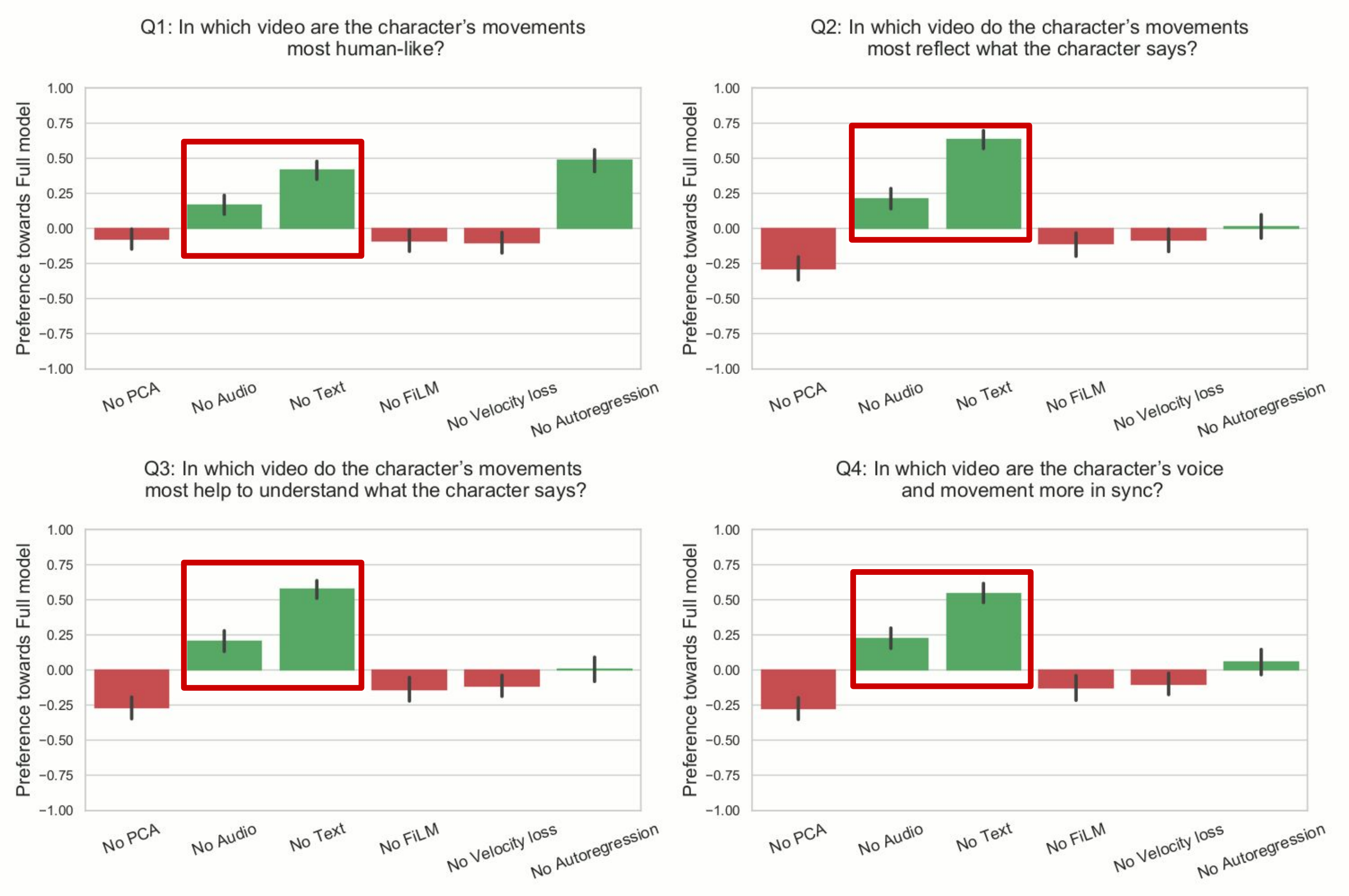
- 123 participants

- 4 questions:
  - *In which video are the character's movements most human-like?*
  - *In which video do the character's movements most reflect what the character says?*
  - *In which video do the character's movements most help to understand what the character says?*
  - *In which video are the character's voice and movement more in sync?*

User Study Results



User Study Results



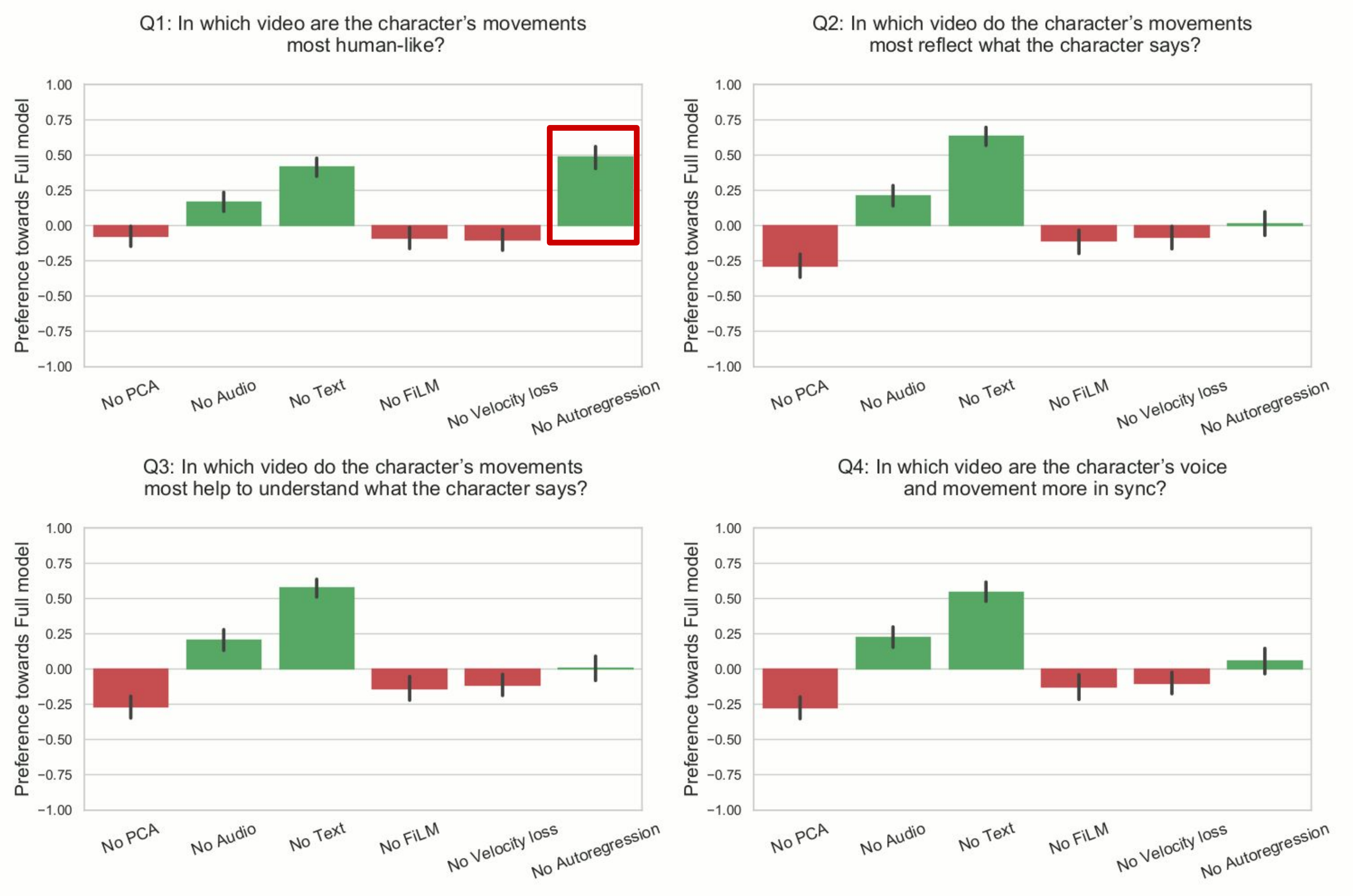
Q1: In which video are the character's movements most human-like?

Q2: In which video do the character's movements most reflect what the character says?

Q3: In which video do the character's movements most help to understand what the character says?

Q4: In which video are the character's voice and movement more in sync?

User Study Results



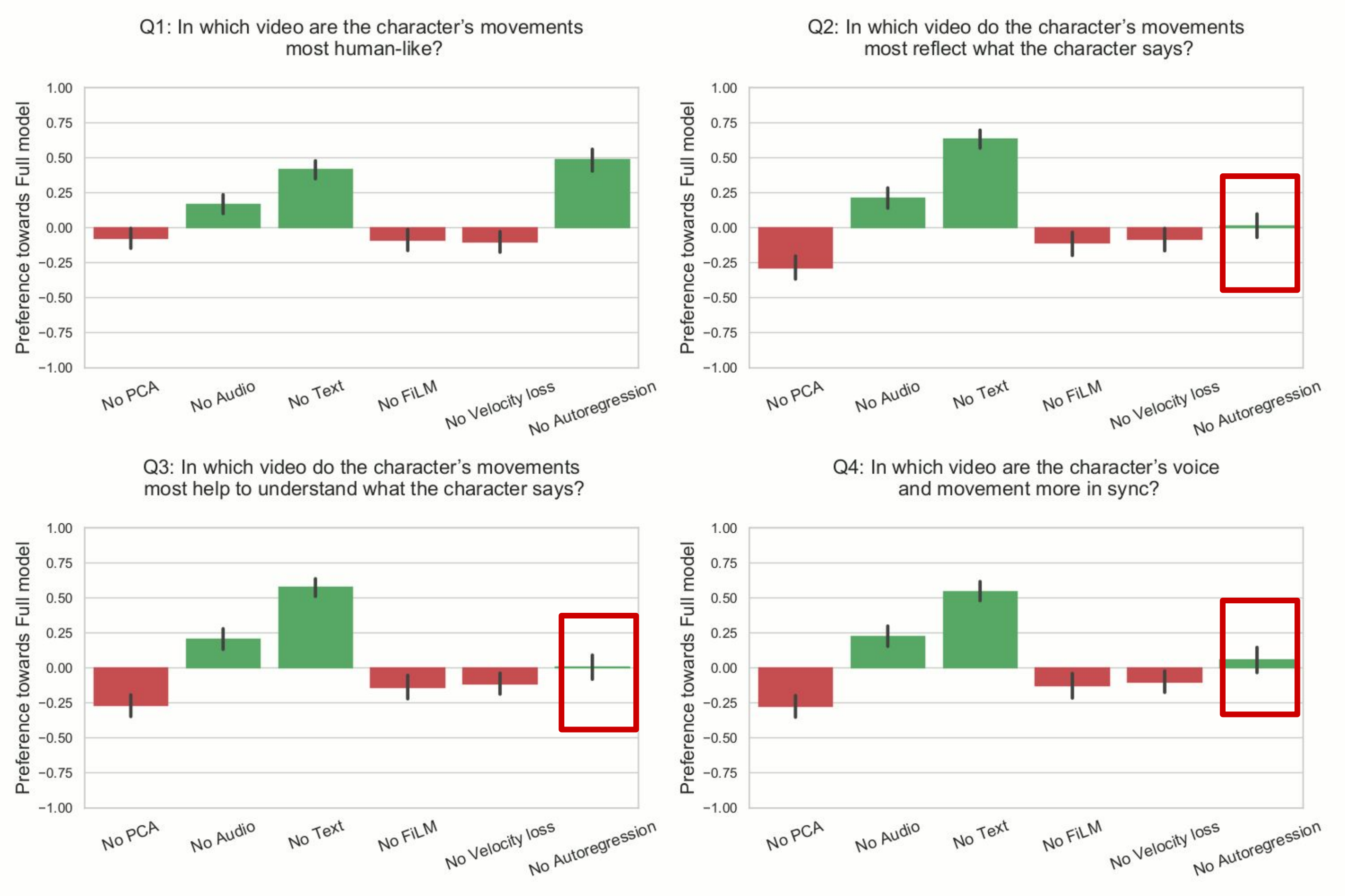
Q1: In which video are the character's movements most human-like?

Q2: In which video do the character's movements most reflect what the character says?

Q3: In which video do the character's movements most help to understand what the character says?

Q4: In which video are the character's voice and movement more in sync?

User Study Results



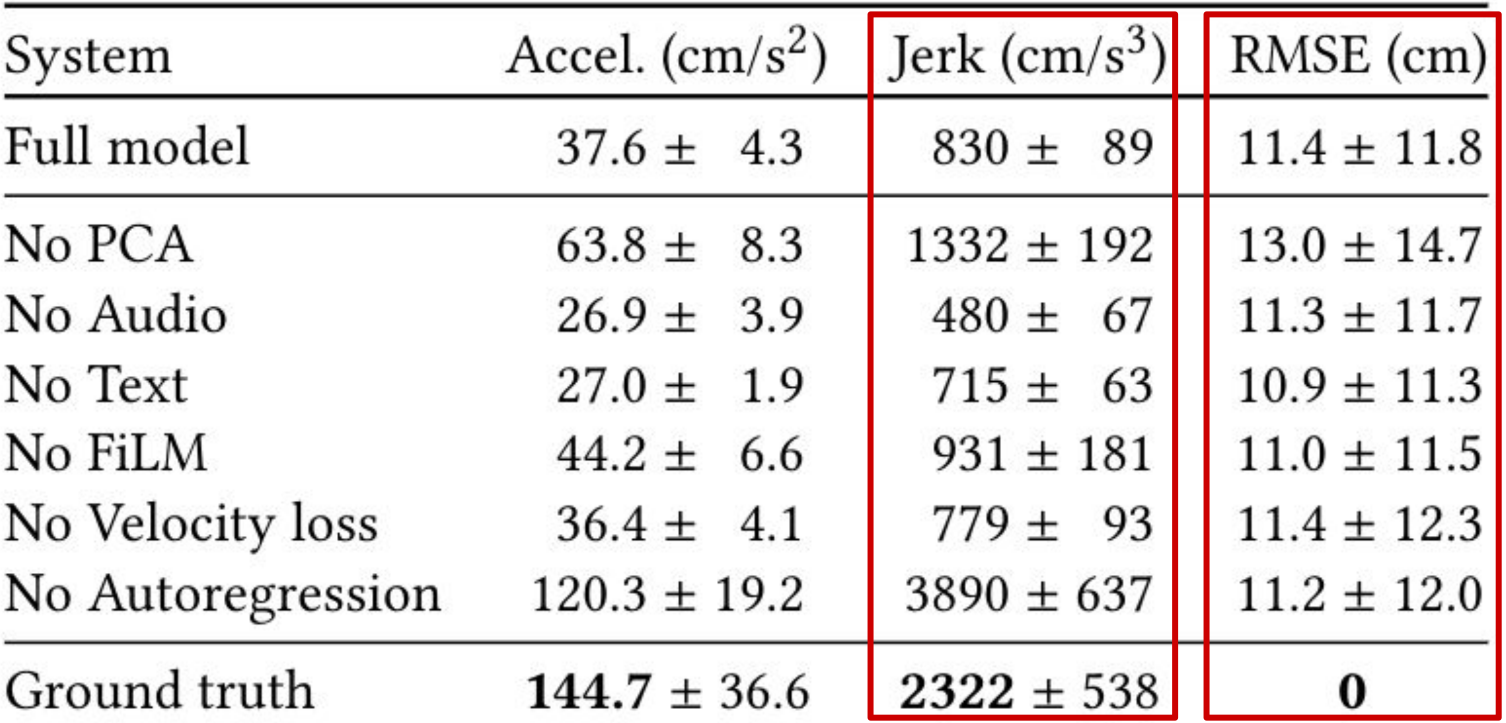
Q1: In which video are the character's movements most human-like?

Q2: In which video do the character's movements most reflect what the character says?

Q3: In which video do the character's movements most help to understand what the character says?

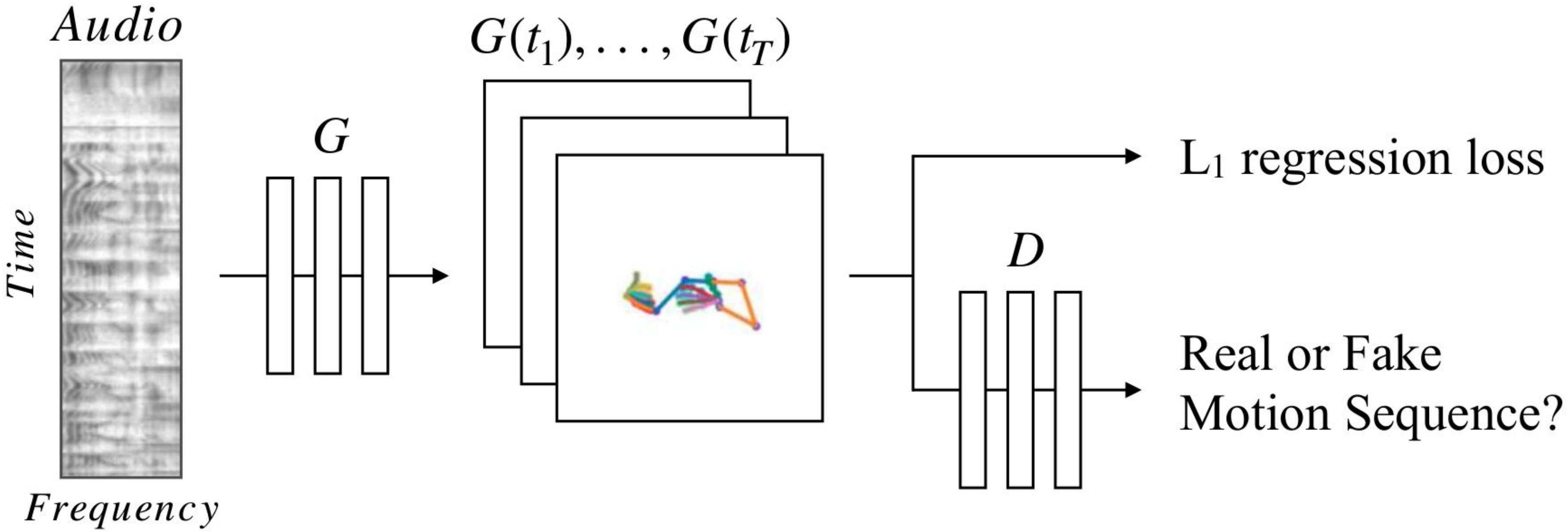
Q4: In which video are the character's voice and movement more in sync?

Numerical Results

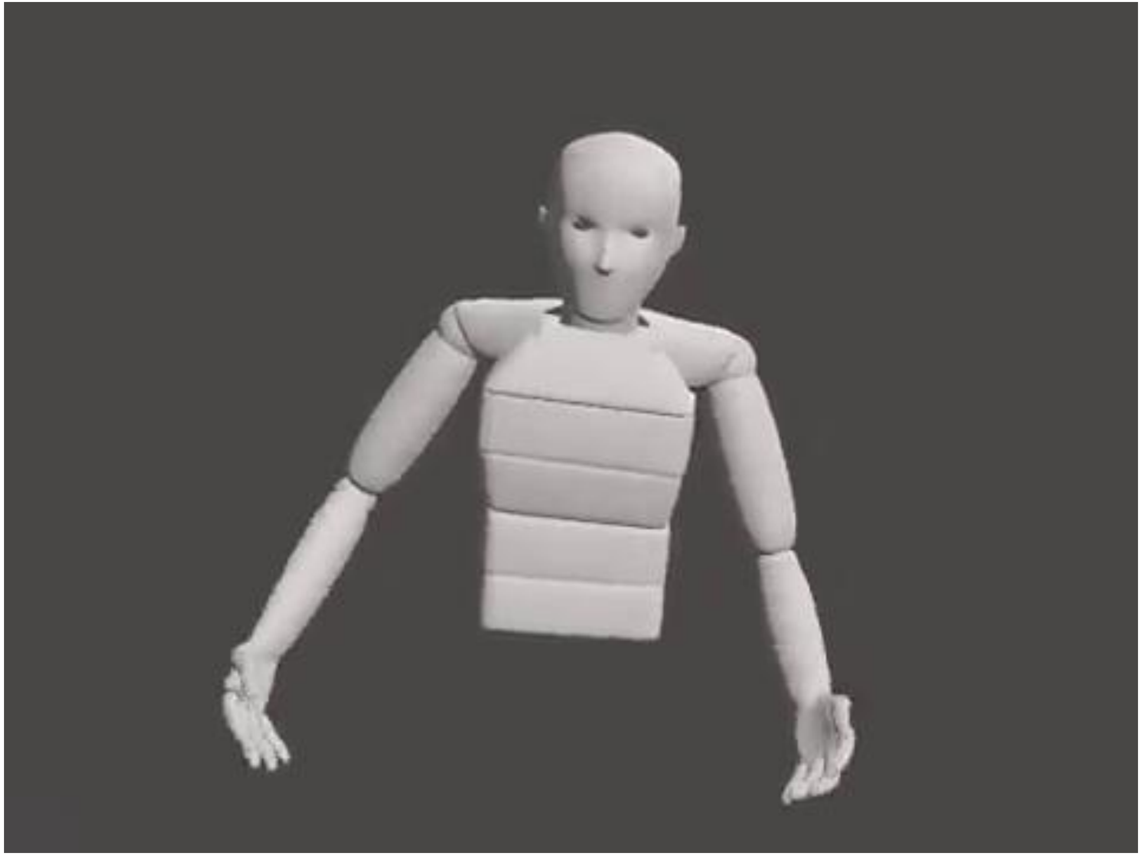


System	Accel. (cm/s^2)	Jerk (cm/s^3)	RMSE (cm)
Full model	37.6 ± 4.3	830 ± 89	11.4 ± 11.8
No PCA	63.8 ± 8.3	1332 ± 192	13.0 ± 14.7
No Audio	26.9 ± 3.9	480 ± 67	11.3 ± 11.7
No Text	27.0 ± 1.9	715 ± 63	10.9 ± 11.3
No FiLM	44.2 ± 6.6	931 ± 181	11.0 ± 11.5
No Velocity loss	36.4 ± 4.1	779 ± 93	11.4 ± 12.3
No Autoregression	120.3 ± 19.2	3890 ± 637	11.2 ± 12.0
Ground truth	**144.7** ± 36.6	**2322** ± 538	**0**

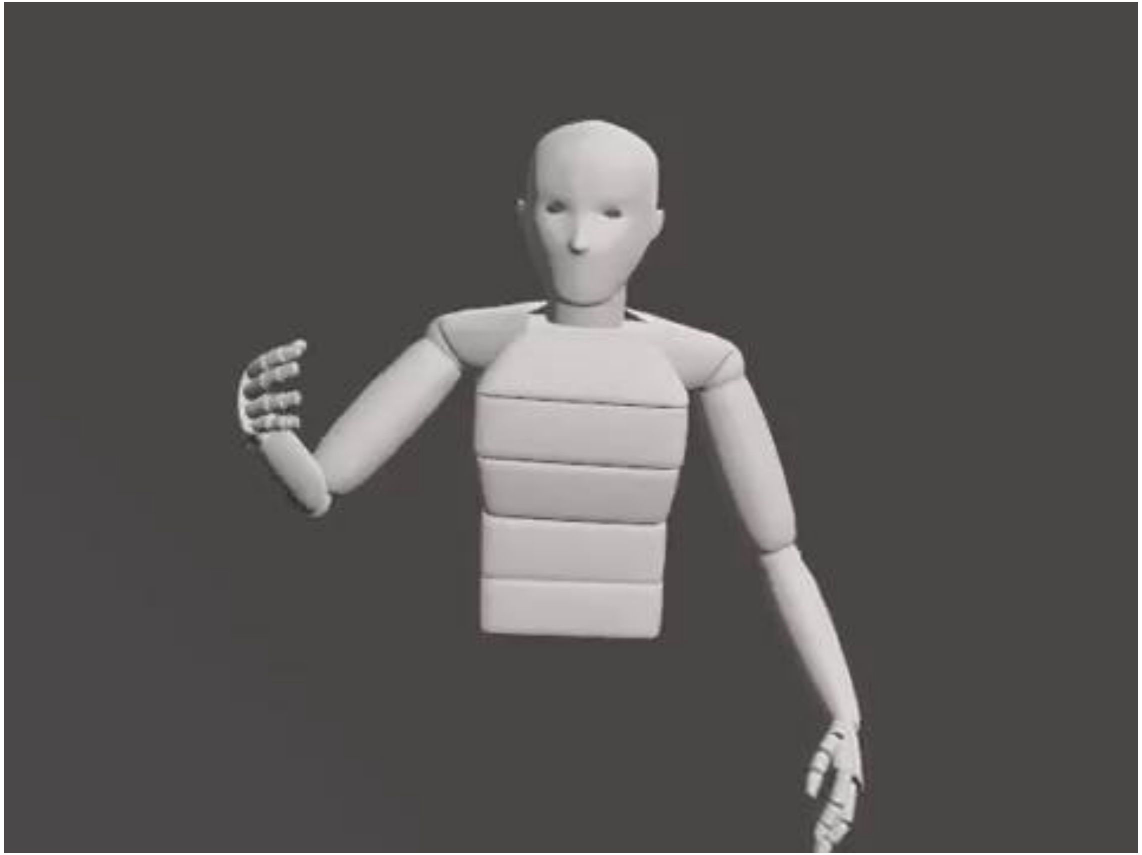
Baseline model



Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, Jitendra Malik
"Learning Individual Styles of Conversational Gesture". CVPR. 2019



CNN-GAN

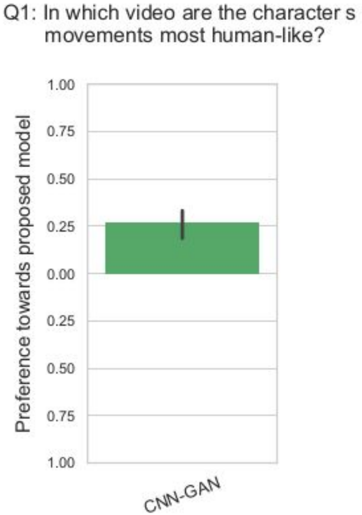


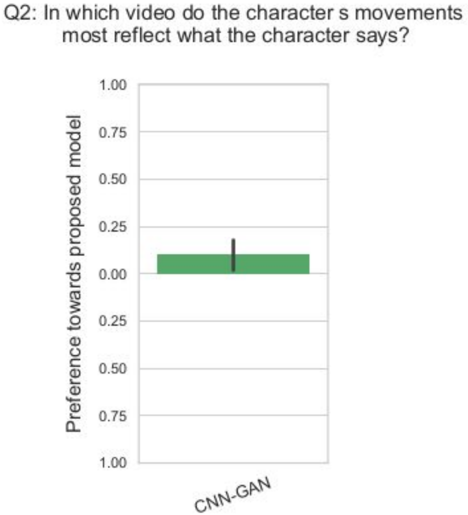
Proposed model

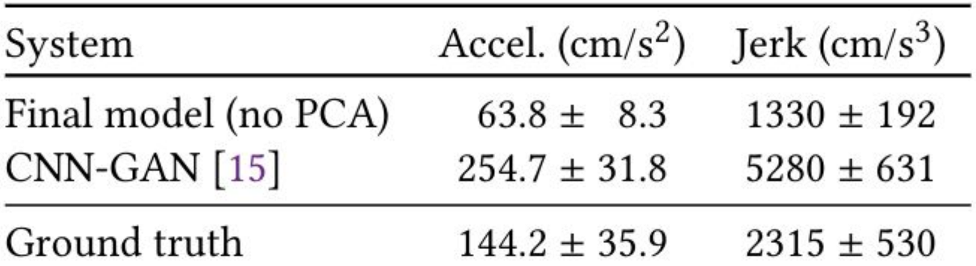
Baselining User Study

- “No PCA” model compared to CNN-GAN [15] baseline
- 27 participants
- 2 questions:
  - *In which video are the character’s movements most human-like?*
  - *In which video do the character’s movements most reflect what the character says?*

Baselining


Q1: In which video are the character s movements most human-like?


Q2: In which video do the character s movements most reflect what the character says?



System	Accel. (cm/s^2)	Jerk (cm/s^3)
Final model (no PCA)	63.8 ± 8.3	1330 ± 192
CNN-GAN [15]	254.7 ± 31.8	5280 ± 631
Ground truth	144.2 ± 35.9	2315 ± 530

Contributors

Taras Kucherenko

Patrik Jonell

Sanne van Waveren

Gustav Eje Henter

Simon Alexanderson

Iolanda Leite

Hedvig Kjellström

Gesticulator: A framework for semantically-aware speech-driven gesture generation

Taras Kucherenko, Patrik Jonell, Sanne van Waveren, Gustav Eje Henter, Simon Alexanderson, Iolanda Leite, and Hedvig Kjellström

KTH Royal Institute of Technology,
Stockholm, Sweden

ICMI 2020
UTRECHT
25-29 October
Utrecht, NL

ICMI 2020

Best Paper Award

https://svito-zar.github.io/gesticulator