# Interpreting Song Lyrics with an Audio-Informed Pre-trained Language Model

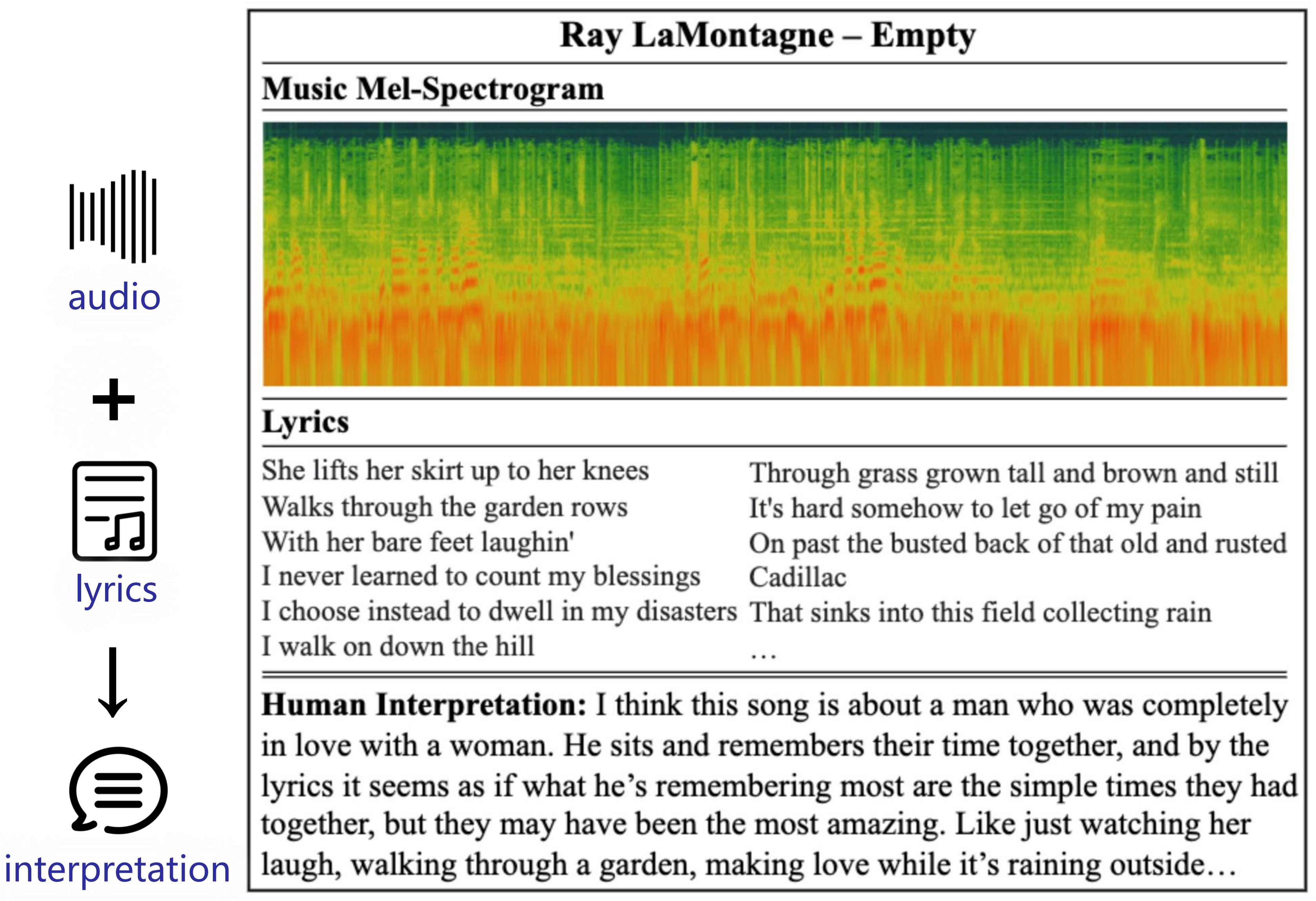
**Yixiao Zhang[1] Junyan Jiang[2,3] Gus Xia[2,3] Simon Dixon[1]**

[1]Centre for Digital Music, Queen Mary University of London
[2]Music X Lab, NYU Shanghai [3]MBZUAI

## Highlights

- We present a novel model, **BART-fusion**, which is the first multimodal generative model for lyric interpretation;
- We find that **audio representation integration** can improve the performance of lyric interpretation models on both the interpretation task and the retrieval task;
- We contribute a large-scale multimodal dataset, **Song Interpretation Dataset**, which contains paired audio, lyrics, and lyric interpretations. It is the first large-scale open-source dataset for lyric interpretation task.

## Task Description



**Ray LaMontagne – Empty**

**Music Mel-Spectrogram**

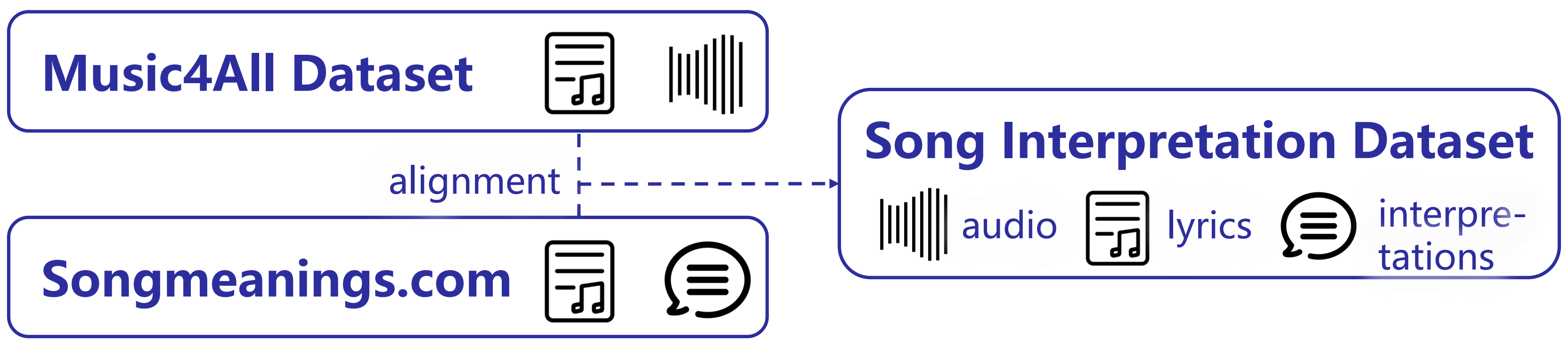
**Lyrics**

She lifts her skirt up to her knees
Walks through the garden rows
With her bare feet laughin'
I never learned to count my blessings
I choose instead to dwell in my disasters
I walk on down the hill
Through grass grown tall and brown and still
It's hard somehow to let go of my pain
On past the busted back of that old and rusted Cadillac
That sinks into this field collecting rain
…

**Human Interpretation:** I think this song is about a man who was completely in love with a woman. He sits and remembers their time together, and by the lyrics it seems as if what he's remembering most are the simple times they had together, but they may have been the most amazing. Like just watching her laugh, walking through a garden, making love while it's raining outside…

## Song Interpretation Dataset

### Data source



Music4All Dataset

Songmeanings.com

alignment

Song Interpretation Dataset

audio lyrics interpre-tations

### Data Size

- 27,834 songs, around 490,000 interpretations;
- Covering various genres: Rock, Pop, Metal, Folk, …

### Data Preprocessing

- We remove overly short interpretations;
- We design a **voting-based filtering mechanism** to improve data quality.

| Dataset Name | Train | Valid. | Test |
|---|---|---|---|
| Raw dataset | 440,000 | 50,000 | 800 |
| Dataset Full | 279,283 | 31,032 | 800 |
| Dataset w/vote $\geq 0$ | 265,360 | 29,484 | 800 |
| Dataset w/vote $> 0$ | 49,736 | 5,526 | 800 |

## Contact

Yixiao Zhang
*yixiao.zhang@qmul.ac.uk*

Paper Code Demo Dataset

## BART-fusion model

### Lyrics encoder

$$\widetilde{H}_t^i = \mathrm{LN}(\mathrm{SA}(H_t^{i-1}W_Q, H_t^{i-1}W_K, H_t^{i-1}W_V)W_a + H_t^{i-1})$$

$$H_t^i = \mathrm{LN}(\mathrm{FFN}(\widetilde{H}_t^i) + \widetilde{H}_t^i),$$

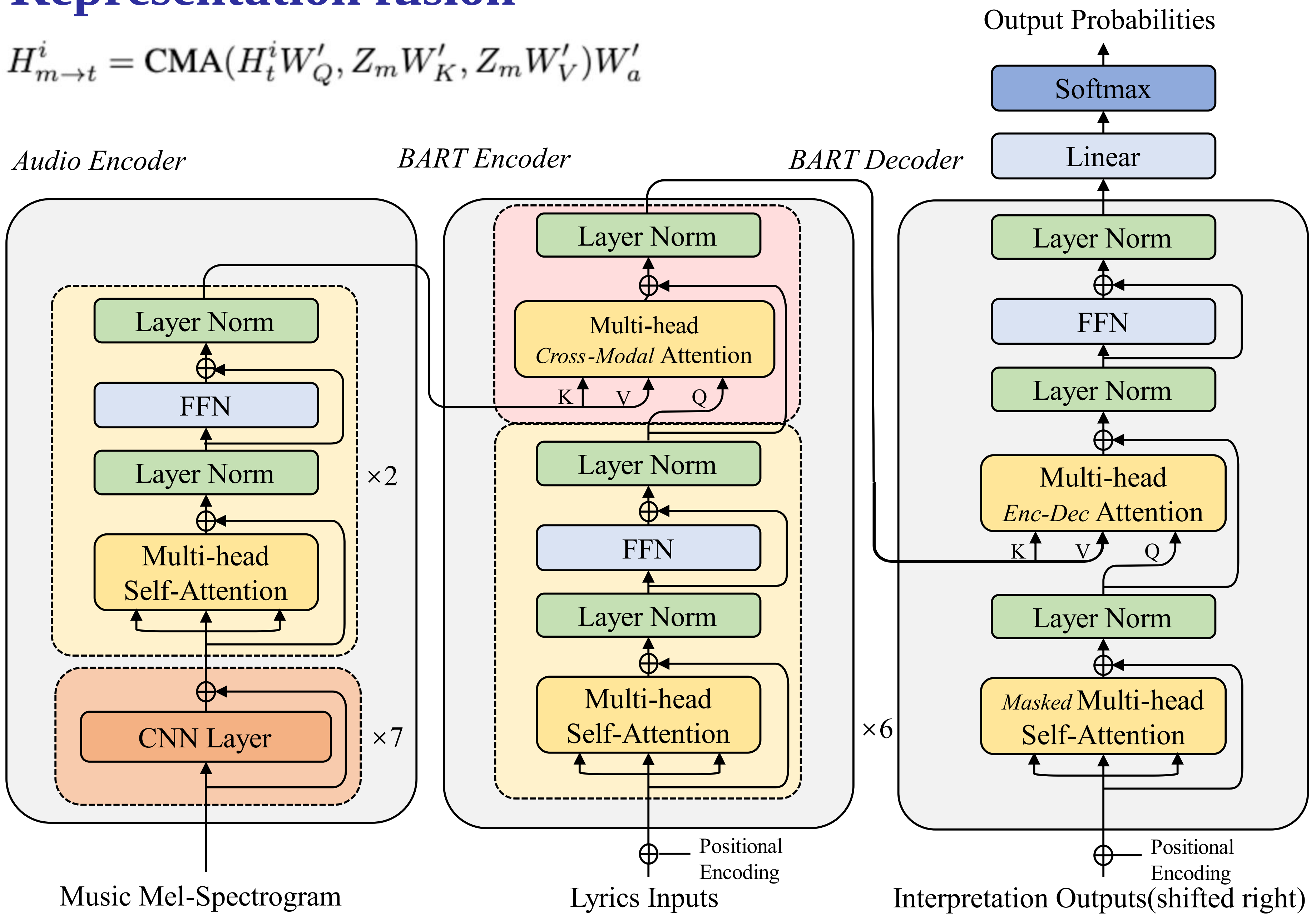
### Audio encoder

$$\widetilde{H}_m^i = \mathrm{BN}(\mathrm{CNN}_2(\mathrm{ReLU}(\mathrm{CNN}_1(H_m^{i-1}))))$$

$$H_m^i = \widetilde{H}_m^i + \mathrm{BN}(\mathrm{CNN}_3(H_m^{i-1})),$$

### Representation fusion

$$H_{m\to t}^i = \mathrm{CMA}(H_t^i W_Q', Z_m W_K', Z_m W_V')W_a'$$



## Results

| Training dataset | Method | Data size | R-1 | R-2 | R-L | METEOR | BERT-Score |
|---|---|---|---|---|---|---|---|
| Dataset w/random | BART | 56,470 | 40.0 | 12.5 | 21.7 | 21.1 | **83.7** |
| Dataset w/random | BART-fusion | 56,470 | **42.1**$^*$ | **13.6**$^*$ | **23.4**$^*$ | **22.0**$^*$ | 83.3 |
| Dataset w/voting $> 0$ | BART | 56,470 | 41.2$_+$ | 13.0$_+$ | 22.8$_+$ | 22.0$_+$ | **83.6** |
| Dataset w/voting $> 0$ | BART-fusion | 56,470 | **44.3**$^*_+$ | **14.6**$^*_+$ | **24.7**$^*_+$ | **22.6**$^*_+$ | 83.3 |
| Dataset Full | BART | 316,478 | 44.1 | 14.0 | 24.5 | 22.5$_+$ | **83.5** |
| Dataset Full | BART-fusion | 316,478 | **46.1**$^*$ | **15.0**$^*$ | **25.1**$^*$ | **23.0**$^*$ | **83.5** |
| Dataset w/voting $\geq 0$ | BART | 300,712 | 44.8$_+$ | 14.9$_+$ | 24.7 | 22.7 | 83.9 |
| Dataset w/voting $\geq 0$ | BART-fusion | 300,712 | **46.7**$^*_+$ | **15.6**$^*_+$ | **25.5**$^*_+$ | **23.4**$^*$ | **84.1** |

## Generation example

*Avril Lavigne -Take Me Away*

*I think this song is about a girl who is in love with a guy but she is afraid to tell him because she doesn't want to ruin the friendship. In the first verse she is talking about how she can't find a way to describe how she feels and how she wishes it would just go away. The second verse is about how the guy is telling her how he feels but she isn't sure if he really feels the same way. The chorus isher saying how she is confused about what to say and how to tell the guy. "I'm unable, come and take me away" means that she is unable to tell her feelings and she is trying to get the guy out of her head. The last verse is her telling the guy that she feels like she is going nowhere in her life, and she just wants him to take her away from all of this. That's just my take on it. I can totally relate to this song. It's a very powerful song.*

## Selected references

[1] K. Watanabe and M. Goto, "Lyrics information processing: Analysis, generation, and applications," in Proceedings of the 1st Workshop on NLP for Music and Audio (NLP4MusA), 2020, pp. 6–12.
[2] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mo-hamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," arXiv preprint arXiv:1910.13461, 2019.
[3] Y. Zhang, Z. Wang, D. Wang, and G. Xia, "BUTTER: A representation learning framework for bi-directional music-sentence retrieval and generation," in Proceedings of the 1st Workshop on NLP for Music and Audio (NLP4MusA), 2020, pp. 54–58.
[4] T. Yu, W. Dai, Z. Liu, and P. Fung, "Vision guided generative pre-trained language models for multimodal abstractive summarization," arXiv preprintarXiv:2109.02401, 2021.
[5] M. Won, S. Chun, and X. Serra, "Toward interpretable music tagging with self-attention," arXiv preprint arXiv:1906.04972, 2019.
[6] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in ACL, vol. 2019. NIH Public Access, 2019, p. 6558.

## Acknowledgement

We want to thank Mark Levy for his great contribution to this work. We also thank Yin-Jyun Luo and Liming Kuang for their enthusiastic help during the writing process of the paper. Yixiao Zhang is a research student at the UKRI Centre for Doctoral Training in Artificial Intelligence and Music, supported jointly by the China Scholarship Council, Queen Mary University of London and Apple Inc.