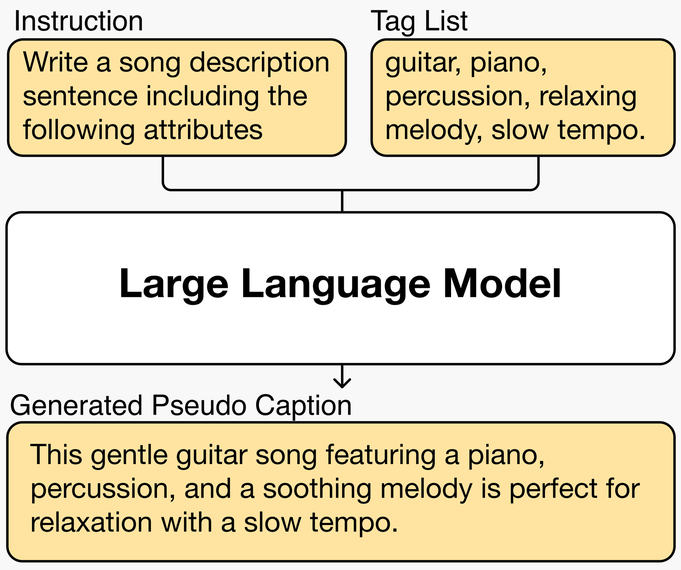
LP-MusicCaps: LLM-Based Pseudo Music Captioning

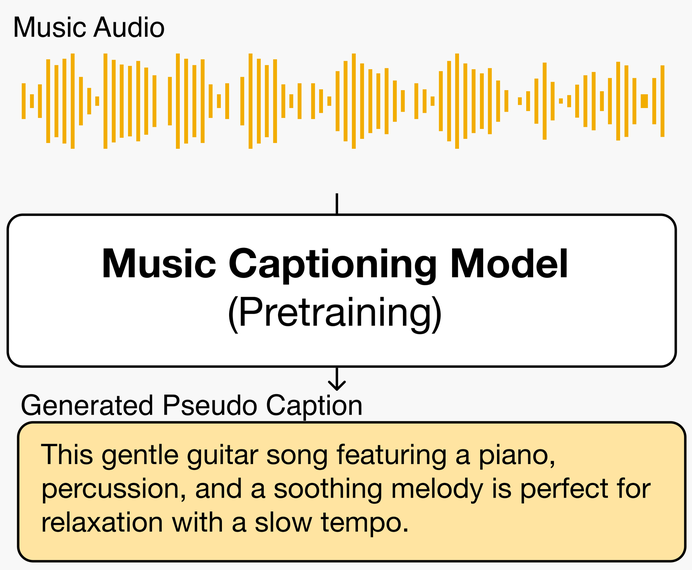
SeungHeon Doh[1], Keunwoo Choi[2], Jongpil Lee[3], Juhan Nam[1]

Graduate School of Culture Technology, KAIST, South Korea[1]
Gaudio Lab, Inc., South Korea[2]
Neutune, South Korea[3]

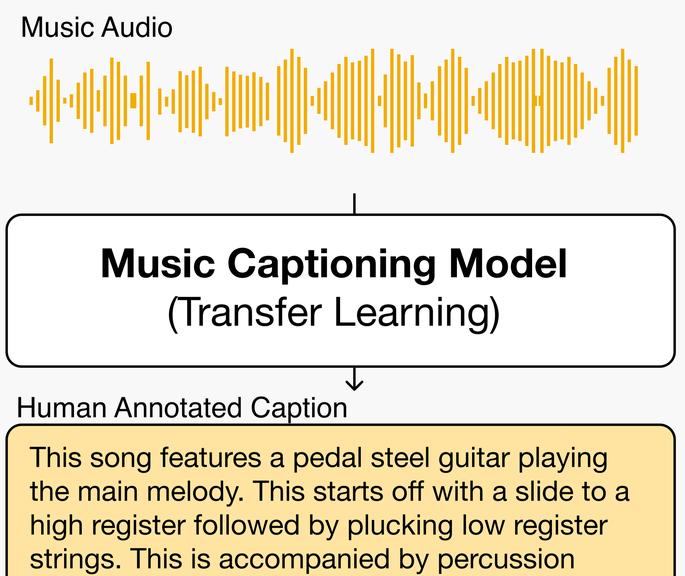
Generate Pseudo Caption Using Instruction & Large Language Model



Pre-train Music Captioning Model Using Pseudo-Caption Dataset



Transfer Learning Using Small Human-Annotated Caption Dataset



Part 1. Tag-to-Caption

LLM-based caption generation shows higher performance than other methods.

LLM can generate various captions depending on the type of instruction and they have different advantage as follows:

- Writing: N-Gram overlapping
- Summary: Semantic similarity
- Paraphrase: Diversity
- Attribute prediction: Diversity

LLM pseudo captions are comparable to human (experts) captions .

Methods	LM	Params	Supervised Metrics B1↑	B2↑	B3↑	B4↑	M↑	R-L↑	BERT-S↑	Diversity Metrics Vocab↑	$Novel_v$ ↑	Length Avg.Token
Baseline												
Tag Concat [2, 13]	-	-	20.25	13.57	8.64	5.42	23.24	19.52	86.24	3506	46.92	20.6±11.2
Template [14]	-	-	25.41	16.15	10.00	6.15	25.57	21.36	87.92	3507	46.93	25.6±11.2
K2C Aug. [22]	T5	220M	6.07	3.01	1.58	0.85	14.23	17.92	86.33	3760	**67.66**	14.7±5.1
Proposed Instruction												
Writing	GPT3.5	175B+	**36.84**	**19.85**	**11.37**	**6.74**	31.44	25.36	89.26	5521	56.17	44.4±17.3
Summary	GPT3.5	175B+	26.12	14.58	8.80	5.52	27.58	**25.83**	**89.88**	4198	49.52	28.6±10.7
Paraphrase	GPT3.5	175B+	36.51	18.73	10.33	5.87	30.36	23.40	88.71	6165	59.95	47.9±18.7
Attribute Prediction	GPT3.5	175B+	35.26	18.16	9.69	5.41	**34.09**	23.19	88.56	**6995**	63.16	66.2±21.6

Table 2. Performance of existing pseudo caption generation methods and the proposed method. LM stand for the language model. Avg.Token stand for the average number of token per caption.

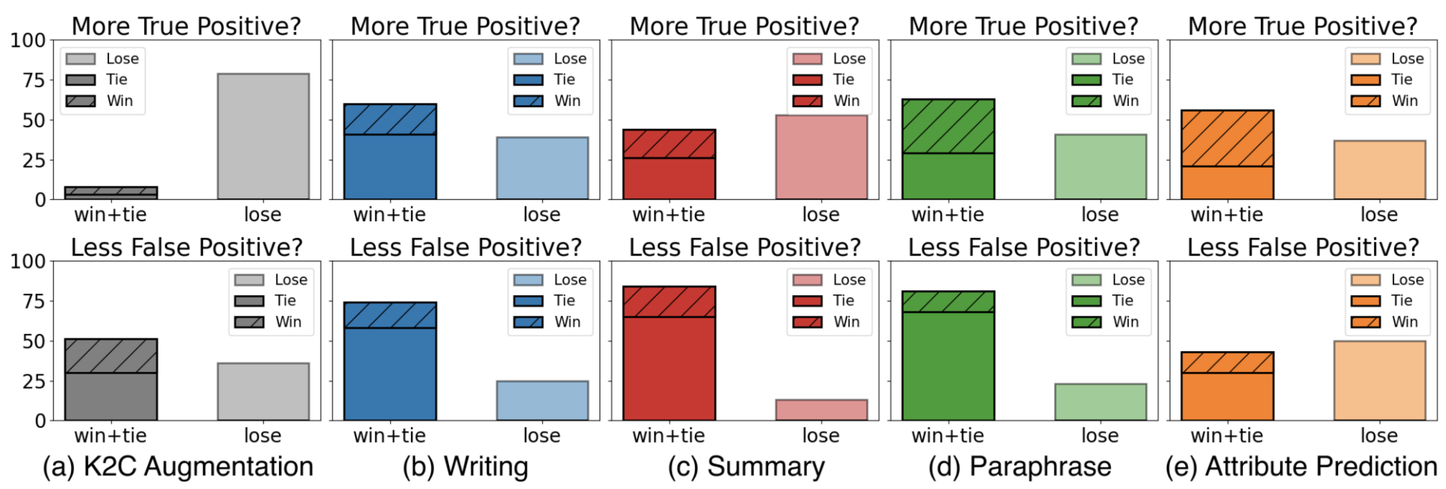


Figure 2. A-vs-B test results. Each method is compared to ground truth in terms of having more true positives and fewer false positives. The proposed methods (b, c, d, e) show comparable **win+tie** performance to ground truth.

Part 2. Audio-to-Caption

Pseudo captions are effective in music captioning in both zero-shot and transfer-learning settings

We released 0.5M audio and 2.2M captions paired with MSD, MTT, and MusicCaps audio clips.

Model	Supervised Metrics B1↑	B2↑	B3↑	B4↑	M↑	R-L↑	BERT-S↑	Diversity Metrics Vocab↑	$Novel_v$↑	$Novel_c$↑	Length Avg.Token
Baseline											
Supervised Model	28.51	13.76	7.59	4.79	20.62	19.22	87.05	2240	0.54	69.00	46.7±16.5
Zeroshot Captioning											
Tag Concat [2, 13]	4.33	0.84	0.26	0.00	3.10	2.01	79.30	802	46.38	100.00	23.8±12.1
Template [14]	7.22	1.58	0.46	0.00	5.28	6.81	81.69	787	45.24	100.00	25.8±12.4
K2C-Aug [22]	7.67	2.10	0.49	0.10	7.94	11.37	82.99	**2718**	**81.97**	100.00	19.9±7.6
LP-MusicCaps (**Ours**)	**19.77**	**6.70**	**2.17**	**0.79**	**12.88**	**13.03**	**84.51**	1686	47.21	100.00	45.3±28.0
Tansfer Learning											
Tag Concat [2, 13]	28.65	14.68	8.68	5.82	21.88	21.31	87.67	1637	3.30	96.07	41.8±14.3
Template [14]	28.41	14.49	8.59	5.78	21.88	21.25	87.72	1545	**3.62**	**96.77**	41.1±13.2
K2C-Aug [22]	**29.50**	**14.99**	8.70	5.73	21.97	20.92	87.50	**2259**	1.42	84.95	44.1±15.0
LP-MusicCaps (**Ours**)	29.09	14.87	**8.93**	**6.05**	**22.39**	**21.49**	**87.78**	1695	1.47	96.06	42.5±14.3

Table 4. Music captioning results on the MusicCaps eval-set. Avg.Token stands for the average number of token per caption.