Data-Efficient Music Captioning via Contrastive and Semantic Alignment

Leekyung Kim* and Jonghun Park*
* Seoul National University, Korea
E-mail: klk97@snu.ac.kr, jonghun@snu.ac.kr

***Abstract*—We propose a music captioning framework designed for low-resource scenarios by leveraging an encoder-decoder architecture that incorporates the audio encoder of the pre-trained audio encoder of Contrastive Language-Audio Pre-training (CLAP) and two auxiliary training objectives. First, we introduce a contrastive modality alignment loss to align audio and text embeddings in the CLAP latent space. Second, we propose a semantic alignment loss that encourages the decoder to generate captions that are semantically close to the ground-truth captions, measured in the CLAP text embedding space. Experiments on the MusicCaps dataset demonstrate that the proposed model and training objectives improve caption generation performance and enable the proposed model to achieve competitive results compared to a baseline model trained on a large-scale dataset, despite using significantly less training data. These results highlight the effectiveness of cross-modal representation alignment in improving music captioning under data scarcity.**

I. Introduction

Music captioning is a multimodal task that generates natural language descriptions for input music audio. Unlike music tagging, which assigns pre-defined labels such as genre, mood, or instrumentation, music captioning expresses musical concepts in the form of complete sentences[1], enabling richer semantic interpretation and comprehensive understanding of musical content. It has potential applications in music retrieval[2] and music recommendation[3]

However, generating captions for music is a highly challenging task. A music captioning model is expected to simultaneously recognize concrete musical attribute such as tempo and instrumentation as well as more abstract and subjective elements like mood and atmosphere. Moreover, understanding the complex, multidimensional relationships between these elements is critical for producing coherent and meaningful descriptions. Unlike traditional classification tasks, it requires not only accurate perception of audio content but also the ability to express it in well-formed, descriptive natural language, similar to how a human listener might describe a piece of music.

Deep understanding of musical structure and linguistic expression is required to train music captioning model effectively, which makes music captioning task a uniquely multimodal challenge. Training robust and generalizable music captioning model requires large-scale and high-quality datasets that reflect the diversity of real-world music and the detailed explanation. However, there remains a severe lack of high-quality paired audio-text datasets for music captioning.

Generating music captions requires careful listening, musical expertise, and semantic interpretation, making human annotation costly and time-consuming[4]. Most existing datasets for music captioning are relatively small in scale and often suffer from limited descriptive diversity or inconsistent annotation quality. As of now, MusicCaps[5] is the only publicly available human-annotated dataset specifically designed for track-level music captioning. It contains approximately 5,500 music-caption pairs annotated by professional musicians. While MusicCaps offers high-quality and semantically rich descriptions, its limited size constrains its usefulness for model training[6].

To address data scarcity, recent studies have explored several approaches to generate more captions. For example, LP-MusicCaps[6] proposed a large-scale pseudo-captioning dataset generated using large language models (LLMs), leveraging existing music tagging datasets as input. This approach resulted in approximately 2.2 million captions for 0.5 million audio clips, significantly increasing dataset size. However, since these captions are machine-generated, they do not ensure the quality or semantic level of human annotations, limiting their effectiveness for high-quality caption generation.

In this work, we propose several methods to improve music captioning performance in data scarcity scenarios. We validate the effectiveness of the proposed methodologies through experiments on MusicCaps, a high-quality but small-scale human-annotated dataset. While our methods are validated under data scarcity, they are also readily applicable to larger-scale training settings such as when additional human-annotated data becomes available or large volumes of high-quality synthetic captions are generated through alternative approaches.

Specifically, we adopt the pre-trained audio encoder of Contrastive Language-Audio Pre-training (CLAP)[7] to provide robust audio representations and explore an effective decoder design suitable for musical caption generation. To bridge the modality gap and enhance semantic alignment between audio and text, we introduce a contrastive modality alignment loss using audio and text encoders of CLAP. It encourages the audio encoder to produce representations that are semantically closer to the corresponding caption in the CLAP latent space.

Furthermore, to ensure that the decoder generates semantically coherent captions, we propose a semantic alignment loss. Unlike standard maximum likelihood training, which predicts the next token solely based on token probabilities, this loss encourages the decoder to generate sequences whose embeddings are close those of the ground-truth caption at each decoding step. This facilitates semantically meaningful generation throughout the sequence, improving both fluency

979-8-3315-7206-8/25/$31.00

and relevance.

In summary, our contributions are as follows:

- We propose a music captioning framework that integrates the pre-trained CLAP audio encoder with a Transformer decoder under limited supervision.
- We introduce a contrastive modality alignment loss that aligns audio and text embeddings in the CLAP latent space, enabling the model to better capture cross-modal semantic relationships.
- We propose a semantic alignment loss to guide the decoder toward generating semantically faithful captions throughout the decoding process.
- We demonstrate the effectiveness of the proposed methods on MusicCaps, showing improved caption quality despite limited training data.

II. Background

A. Music Captioning

Recent studies on music captioning have explored various neural network architectures. A common approach employs encoder-decoder frameworks[8], as music captioning can be regarded as a cross-modal translation task from audio to natural language. In this framework, an audio encoder extracts features from the input audio, and a decoder generates captions auto-regressively, conditioned on the encoded audio representations.

MusCaps[1] was the first to address the music captioning from an audio captioning perspective, using a multimodal encoder-decoder architecture based on long short-term memory (LSTM) network[9]. Audio encoders are often pre-trained on audio understanding tasks such as BEATs[10] or CLAP[7], while text decoders are typically initialized from pre-trained LLMs such as BART[11] or GPT-2[12].

Due to the limited availability of large-scale, high-quality paired datasets, existing music captioning models often suffer from poor generalization and limited semantic expressiveness. To address this challenge, several studies proposed methods to alleviate the data scarcity problem. These include generating captions from music tagging datasets via tag concatenation[13], [14], manually defined text templates[15], or LLMs[6]. JamendoMaxCaps[16] generates captions from free-licensed instrumental tracks by a music captioning model[17], enhanced with imputed metadata.

In this paper, we propose methods to improve music captioning performance under data-scarce conditions by enhancing the alignment between audio and text. Specifically, we leverage a pre-trained audio encoder and introduce two training objectives that improve both cross-modal and semantic alignment.

B. CLAP

CLAP[7], [18] is a large-scale audio-language model designed to learn a shared embedding space between audio and natural language. It consists of an audio encoder and a text encoder, each responsible for processing audio and text inputs, respectively. Inspired by Contrastive Language-Image Pre-Training (CLIP) for vision-language tasks, CLAP employs contrastive learning to align audio with their corresponding textual descriptions.

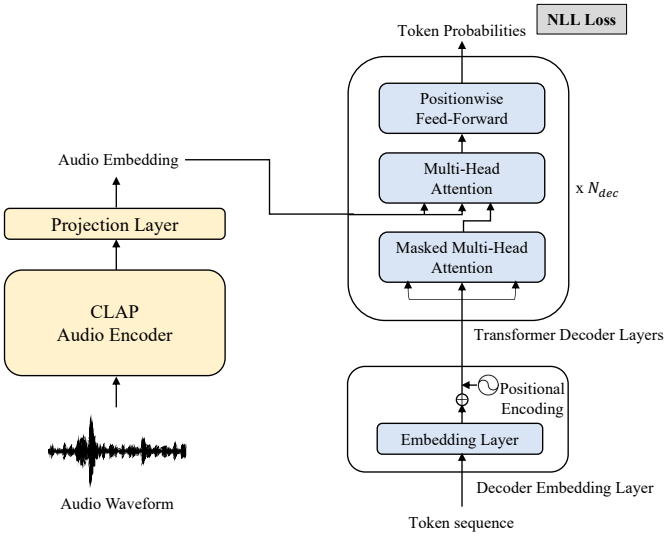

Fig. 1. An overview of our captioning model based on an encoder-decoder architecture.

The model is pre-trained on a large and diverse collection of audio-text pairs, enabling it to capture semantic relationships across modalities without the need for explicit supervision. CLAP supports a wide range of audio domains, including environmental sounds, speech, and music. In particular, the music-specialized version of CLAP has been widely adopted as an audio encoder in tasks such as music tagging and music-text retrieval, demonstrating its effectiveness as a general-purpose audio representation model for music understanding.

In this work, we utilize the audio encoder of CLAP as a fixed feature extractor and leverage its cross-modal alignment capabilities to enhance caption generation for music audio.

III. Proposed Method

A. Problem Specification

Music captioning is the task of generating a caption $\hat{y}$ for a given music track m. Training a music captioning model corresponds to optimizing model parameters θ of a mapping function from audio input to text output, defined as $f_\theta : m \rightarrow \hat{y}$, with the goal of generating captions that are semantically similar to the ground-truth caption y. A music caption consists of a sequence of word tokens y_i, where i denotes the index of each token in the caption. These tokens are predicted auto-regressively by the decoder. This generation process can be formulated as:

$$\hat{y}_i = \arg\max_y P(y_i \mid y_{<i}, m; \theta) \tag{1}$$

B. Model Structure

An overview of our captioning model is depicted in Fig. 1. We adopt the multimodal latent space of CLAP[7] to align audio and text representations for music captioning.

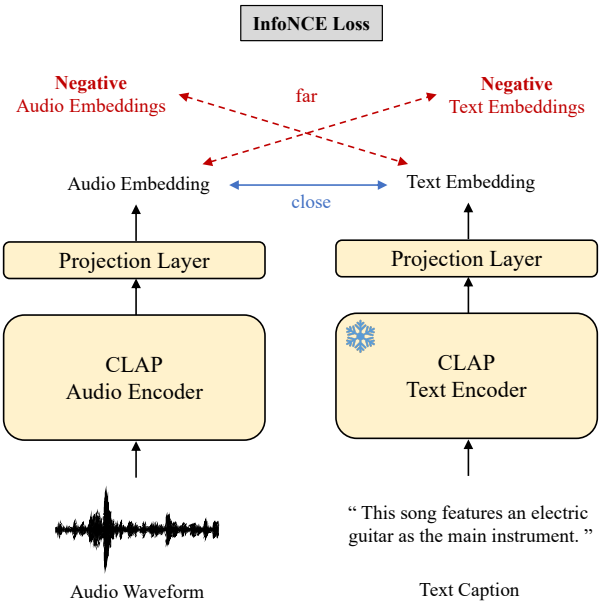

Fig. 2. Contrastive modality alignment based on InfoNCE loss.

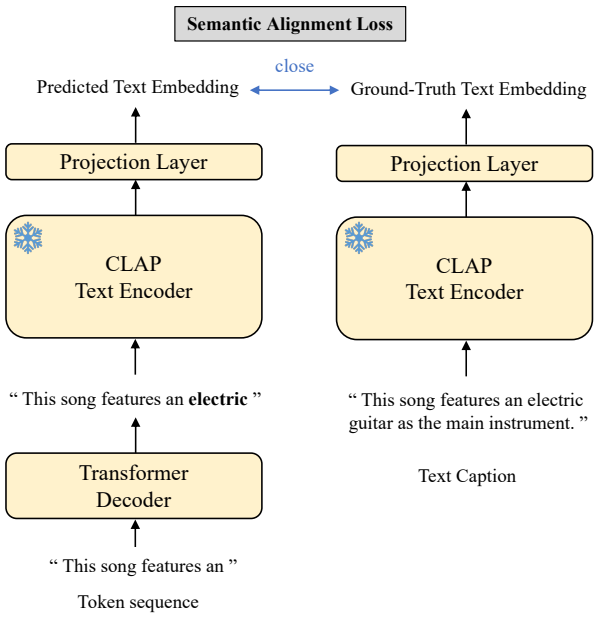

Fig. 3. Semantic alignment loss between the predicted token sequence and the ground-truth caption.

For an audio-text pair (a, t), CLAP jointly trains an audio encoder $f_a(\cdot)$ and a text encoder $f_t(\cdot)$ to align their respective embedding $E_A = f_a(a)$ and $E_T = f_t(t)$.

In our framework, we use $f_a(\cdot)$ as the encoder in an encoder-decoder architecture. Specifically, we adopt the HTSAT[19]-based audio encoder following the HTSAT-RoBERTa configuration, which achieved the best performance in the CLAP paper[7]. $f_a(\cdot)$ consists of Swin Transformer[20] blocks followed by projection layers. Among CLAP variants, we select the music-specific version trained on music datasets, AudioSet[21], and LAION-Audio-630k, as it demonstrated the best performance in musical genre classification.

For the text decoder, we employ vanilla Transformer decoder[22] blocks. In preliminary experiments, we compared the vanilla Transformer decoder with the BART decoder and found that the vanilla Transformer consistently outperformed BART in music captioning metrics under their respective best configurations. We attribute this performance gap in part to tokenizer mismatches between CLAP and BART. Specifically, RoBERTa tokenizer and Bart tokenizer are utilized for CLAP and BART, respectivley.

C. Contrastive Modality Alignment Loss

As music captioning is a multimodal task, the audio encoder is required not only to extract meaningful features from the audio but also to produce embeddings that are useful for generating text captions. In the CLAP latent space, the embedding distance between a positive audio-text pair (a, t) is minimized, while distance between negative pair is maximized.

To ensure that the embeddings from the audio encoder are semantically aligned with the text captions, they should be positioned close to the corresponding embeddings produced by the text encoder within the shared latent space. To this end, we utilize the frozen text encoder from the music-specific version of CLAP, identical to the variant used for the audio encoder, in order to encourage consistent audio-text modality alignment.

Specifically, for each paired sample (a, t), we enforce that the audio embedding of a is closest to the text embedding of t among all text embeddings in the batch and the text embedding of t should be closest to the audio embedding of a among all audio embeddings, as illustrated in Fig. 2. This bidirectional alignment objective is implemented using InfoNCE loss[23] similar to the contrastive training approach in [24].

Let $\cos(a, b)$ denote the cosine similarity between vectors a and b. The contrastive modality alignment loss $\mathcal{L}_{\mathrm{C}}$ is defined as the sum of the audio-to-text alignment loss $\mathcal{L}_{\mathrm{C\_a}}$ and the text-to-audio alignment loss $\mathcal{L}_{\mathrm{C\_t}}$. Following the InfoNCE formulation, the loss is computed as:

$$
\begin{aligned}
\cos(a, b) &= \frac{a \cdot b}{\|a\|\|b\|}, \\
\mathcal{L}_{\mathrm{C\_t}} &= \mathbb{E}_{\mathcal{B} \in \mathcal{D}_{\text{train}}} \left[\sum_{i=1}^{|\mathcal{B}|} -\log \frac{\cos\left(E_A^i, E_T^i\right)}{\sum_{j=1}^{|\mathcal{B}|} \cos\left(E_A^j, E_T^i\right)} \right], \\
\mathcal{L}_{\mathrm{C\_a}} &= \mathbb{E}_{\mathcal{B} \in \mathcal{D}_{\text{train}}} \left[\sum_{i=1}^{|\mathcal{B}|} -\log \frac{\cos\left(E_A^i, E_T^i\right)}{\sum_{j=1}^{|\mathcal{B}|} \cos\left(E_A^i, E_T^j\right)} \right], \\
\mathcal{L}_{\mathrm{C}} &= \mathcal{L}_{\mathrm{C\_a}} + \mathcal{L}_{\mathrm{C\_t}},
\end{aligned}
\tag{2}
$$

where $\mathcal{B}$ denotes a mini-batch sampled from the training set $\mathcal{D}_{\text{train}}$, and i and j index the samples within the batch.

D. Semantic Alignment Loss

In music captioning, the objective is to generate captions that are semantically similar to the ground-truth descriptions

rather than to produce exact matches. Therefore, it is important for the text decoder to consider semantic alignment when predicting the next token. Specifically, the token sequence extended with the newly predicted token should be closer to the ground-truth caption in the CLAP latent space than the previous token sequence, as shown in Fig. 3. To measure semantic closeness, we utilize the same frozen text encoder as used in the contrastive modality alignment loss.

The semantic alignment loss $\mathcal{L}_{\mathrm{S}}$ is defined as follows:

$$\mathcal{L}_{\mathrm{S}} = \mathbb{E}_{\mathcal{B}\in\mathcal{D}_{\text{train}}}\left[\sum_{t=1}^{|\hat{y}|}\left(1-\cos\left(f_t(\hat{y}_{1:t}), f_t(y)\right)\right)\right], \quad (3)$$

where $f_t(\cdot)$ denotes the text encoder from CLAP, $\hat{y}_{1:t}$ is the partially generated token sequence up to step t, and y is the full ground-truth caption.

We conducted preliminary experiments comparing multiple similarity measures, including cosine similarity $(1-\cos)$, mean squared error (MSE), and mean absolute error (MAE), for $\mathcal{L}_{\mathrm{S}}$. Among these, the cosine-based loss consistently yielded superior performance across captioning evaluation metrics. Therefore, we adopt the $1-\cos$ formulation as the final definition of the semantic alignment loss in our methods.

E. Overall Learning Objective

The primary loss function of our model is the negative log-likelihood of the target text tokens, defined as:

$$\mathcal{L}_{\mathrm{NLL}} = \mathbb{E}_{(m,y)\in\mathcal{D}_{\text{train}}}\left[\sum_{i=1}^{|y|} -\log P(y_i \mid y_{<i}, m)\right], \quad (4)$$

where $\mathcal{D}_{\text{train}}$ denotes the training dataset, and y_i is the target word token.

To incorporate both cross-modal and semantic alignment, we define the final training objective $\mathcal{L}_{\text{final}}$ as a weighted sum of the main NLL loss, the contrastive modality alignment loss, and the semantic alignment loss:

$$\mathcal{L}_{\text{final}} = \mathcal{L}_{\mathrm{NLL}} + \alpha\cdot\mathcal{L}_{\mathrm{C}} + \beta\cdot\mathcal{L}_{\mathrm{S}}, \quad (5)$$

where α and β are hyperparameters that control the relative importance of each auxiliary loss term.

IV. Experiments

A. Datasets

We use MusicCaps[5] for training, validation, and testing. It consists of 5,521 music clips, each paired with a caption written by expert musicians. Each caption typically comprises multiple sentences that describe the audio content in detail. For example:

> "The low quality recording features a ballad song that contains sustained strings, mellow piano melody and soft female vocal singing over it. It sounds sad and soulful, like something you would hear at Sunday services."

TABLE I
Hyperparameters of the Transformer decoder in our model

Hyperparameter	Value
Number of layers	8
Dimension of embedding	1024
Number of self-attention heads	8
Dimension of query, key, value	512
Dropout probability	0.1
Maximum length of output sequence	300

All audio clips are 10-second excerpts from the AudioSet dataset[21], with 2,858 examples in the evaluation split and 2,663 in the training split. In our experiments, we follow the official train and evaluation splits provided in the dataset.

B. Input Pre-processing

All audio inputs are pre-processed following the procedure used in CLAP[7] to ensure compatibility with the audio encoder of CLAP. Each audio file is first converted to mono-channel FLAC format with a sampling rate of 48,000 Hz. The duration of each input is fixed to 10 seconds: if the audio clip exceeds 10 seconds, it is truncated to the first 10 seconds; if it is shorter, zero-padding is applied at the end.

Mel-spectrograms are computed using a short-time Fourier transform (STFT) with a window size of 1,024, a hop size of 480, and 64 mel frequency bins. The resulting mel-spectrogram serves as the input to the audio encoder.

The audio encoder in our model follows the HTSAT-RoBERTa configuration from CLAP[7], where audio features are processed by Swin Transformer blocks[20] followed by a projection multilayer perceptron. The output dimensionality of the Swin Transformer blocks is 768, and that of the projection layer is 512.

C. Training Details

We use the loss functions described in SectionIII-E, with the weighting hyperparameters for the auxiliary losses set to $\alpha = 0.5$ and $\beta = 1$. The model hyperparameters used in our experiments are summarized in Table I. As the audio encoder follows the CLAP configuration such as using RoBERTa tokenizer, only the decoder architecture is included.

The model is trained using the Adam optimizer with a learning rate of 1×10^{-4}. Early stopping is applied if the validation loss does not decrease for 10 consecutive epochs. Pitch augmentation is applied by randomly shifting the pitch of each audio sample by a semitone uniformly sampled from the range $[-6, 6]$. The batch size is set to 32.

The decoder is trained using the teacher forcing method, where at each time step t, the decoder receives the ground-truth token sequence $y_{<t}$ as input. During inference, the decoder instead uses the previously generated token sequence as an input. Decoding process begins with a start-of-sequence token $\langle\text{SOS}\rangle$ and terminates when an end-of-sequence token $\langle\text{EOS}\rangle$ is generated or when a pre-defined maximum length of 300 tokens is reached.

TABLE II
MUSIC CAPTIONING PERFORMANCE OF THE MODELS FOR ABLATION STUDY AND THE COMPARISON MODEL TRAINED ON LARGE-SCALE DATASETS

Models	B1↑	B2↑	B3↑	B4↑	M↑	R-L↑	BS↑	Vocab↑	Novel_v ↑	Novel_c ↑
M1	0.284	0.157	0.095	0.065	0.246	0.215	0.873	282	0.09	**1.0**
M2	0.299	0.165	0.101	0.069	0.245	0.223	0.874	322	0.04	**1.0**
M3	0.311	0.168	0.104	0.072	0.237	0.218	0.876	396	0.05	**1.0**
M4	**0.313**	**0.174**	**0.106**	**0.073**	**0.249**	**0.225**	**0.877**	375	0.05	**1.0**
Comparison	0.285	0.138	0.076	0.048	0.206	0.192	0.871	**2240**	**0.54**	0.69

D. Evaluation Metrics

To evaluate the quality of generated captions, we use a combination of n-gram-based, embedding-based, and diversity-oriented metrics, following [6].

For n-gram-based evaluation, we report BLEU-1 (B1) to BLEU-4 (B4)[25], METEOR (M)[26], and ROUGE-L (R-L)[27]. BLEU and METEOR measure n-gram overlap between the generated and reference captions, while ROUGE-L captures the longest common subsequence. METEOR further considers semantic similarity by incorporating synonym matching using WordNet[28], making it more robust to lexical variation.

To assess semantic similarity beyond n-gram-based metrics, we use BERTScore (BS) [29], which computes token-level similarity based on pre-trained BERT embeddings of the generated and reference captions. BERTScore is more robust to paraphrasing, synonym usage, and variations in word order.

We additionally measure caption diversity using three metrics: (1) Vocab, the number of unique words in all the generated captions; (2) Novel_v, the percentage of vocabulary tokens in the generated captions that do not appear in the training set; (3) Novel_c, the percentage of generated captions that were not seen during training. These metrics assess the ability of the model to generate novel and diverse captions rather than just reproducing training data.

V. RESULTS

Table II presents the performance of various models evaluated using standard music captioning metrics. For all metrics, higher values indicate better performance. Models M1 to M4 are designed for the ablation study. M1 serves as the base model, employing our proposed encoder-decoder architecture consisting of the CLAP audio encoder and a Transformer decoder. M2 freezes the Transformer blocks in the CLAP audio encoder while keeping the rest of the architecture identical to M1. Both M1 and M2 are trained solely with the NLL loss. M3 extends M2 by incorporating the contrastive modality alignment loss, while M4 further adds the semantic alignment loss on top of M3. Performance consistently improves from M1 to M4, demonstrating the effectiveness of each proposed training objective.

As a baseline for comparison, we also evaluated a standard encoder-decoder model trained on the large-scale LP-MusicCaps from [6]. Although the baseline and our models, M1 to M4, are trained on datasets of different scales, they are evaluated on the same test set in MusicCaps to ensure a fair comparison. Despite being trained on significantly less data, M4 achieves competitive results in terms of semantic metrics, highlighting the benefit of integrating the CLAP audio encoder and the two auxiliary loss functions. In particular, M4 achieves the highest METEOR score among all models, which can be attributed to the introduction of the semantic alignment loss.

However, M4 shows relatively lower diversity performance, which we attribute to the limited amount of the training data. This result suggests a trade-off between caption quality and diversity: while M1 to M4 focus on generating semantically faithful captions through alignment with ground-truth descriptions, they tend to produce less varied expressions compared to the baseline model. The relatively small vocabulary size and low token-level novelty observed in the proposed model indicate that, under data-scare conditions, the model is exposed to a limited vocabulary set compared to models trained on larger-scale data. Consequently, it tends to favor semantically aligned expressions rather than exploring diverse word choices.

This behavior reflects a common tendency in generation tasks where semantic constraints can limit linguistic variability. Although this trade-off results in high semantic performance, it can be suboptimal in applications that demand expressive or stylistically diverse outputs. We expect that training on larger datasets would mitigate this issue, enabling the model to achieve both high semantic accuracy and greater diversity. It could be addressed by combining our alignment-based objectives with data augmentation or diverse decoding strategies such as top-k sampling or diverse beam search.

VI. CONCLUSION

This paper presents a music captioning framework that integrates the pre-trained audio encoder of CLAP and a Transformer decoder, trained with two proposed objectives: a contrastive modality alignment loss and a semantic alignment loss. The proposed approach improves caption quality under limited supervision and demonstrates competitive performance compared to the model trained on large-scale datasets. These findings suggest that aligning audio and text representations in the multimodal CLAP latent space is effective for music captioning, and that further performance improvements can be achieved by training a on larger datasets in future work.

REFERENCES

[1] I. Manco, E. Benetos, E. Quinton, and G. Fazekas, "Muscaps: Generating captions for music audio," in *2021 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2021, pp. 1–8.

[2] I. Manco, E. Benetos, E. Quinton, and G. Fazekas, "Contrastive audio-language learning for music," in *Ismir 2022 Hybrid Conference*, 2022.

[3] I. Andjelkovic, D. Parra, and J. O'Donovan, "Moodplay: Interactive music recommendation based on artists' mood similarity," *International Journal of Human-Computer Studies*, vol. 121, pp. 142–159, 2019.

[4] S. Liu, A. S. Hussain, C. Sun, and Y. Shan, "Music understanding llama: Advancing text-to-music generation with question answering and captioning," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, pp. 286–290.

[5] A. Agostinelli, T. I. Denk, Z. Borsos, *et al.*, "Musiclm: Generating music from text," *arXiv preprint arXiv:2301.11325*, 2023.

[6] S. Doh, K. Choi, J. Lee, and J. Nam, "Lp-musiccaps: Llm-based pseudo music captioning," in *Ismir 2023 Hybrid Conference*, 2023.

[7] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[8] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *Advances in neural information processing systems*, vol. 27, 2014.

[9] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[10] S. Chen, Y. Wu, C. Wang, *et al.*, "Beats: Audio pre-training with acoustic tokenizers," in *Proceedings of the 40th International Conference on Machine Learning*, 2023, pp. 5178–5193.

[11] M. Lewis, Y. Liu, N. Goyal, *et al.*, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, 2020, p. 7871.

[12] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *et al.*, "Language models are unsupervised multitask learners,"

[13] T. Cai, M. I. Mandel, and D. He, "Music autotagging as captioning," in *Proceedings of the 1st Workshop on NLP for Music and Audio (NLP4MusA)*, 2020, pp. 67–72.

[14] S. Doh, M. Won, K. Choi, and J. Nam, "Toward universal text-to-music retrieval," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[15] T. Chen, Y. Xie, S. Zhang, S. Huang, H. Zhou, and J. Li, "Learning music sequence representation from text supervision," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 4583–4587.

[16] A. Roy, R. Liu, T. Lu, and D. Herremans, "Jamendomaxcaps: A large scale music-caption dataset with imputed metadata," *arXiv preprint arXiv:2502.07461*, 2025.

[17] Y. Chu, J. Xu, Q. Yang, *et al.*, "Qwen2-audio technical report," *arXiv preprint arXiv:2407.10759*, 2024.

[18] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, "Clap learning audio concepts from natural language supervision," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.

[19] K. Chen, X. Du, B. Zhu, Z. Ma, T. Berg-Kirkpatrick, and S. Dubnov, "Hts-at: A hierarchical token-semantic audio transformer for sound classification and detection," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 646–650.

[20] Z. Liu, Y. Lin, Y. Cao, *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[21] J. F. Gemmeke, D. P. Ellis, D. Freedman, *et al.*, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 2017, pp. 776–780.

[22] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[23] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[24] S.-L. Wu, X. Chang, G. Wichern, *et al.*, "Improving audio captioning models with fine-grained audio features, text embedding supervision, and llm mix-up augmentation," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2024, pp. 316–320.

[25] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: A method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[26] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, pp. 65–72.

[27] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[28] C. Fellbaum, *WordNet: An electronic lexical database*. MIT press, 1998.

[29] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," in *International Conference on Learning Representations*.