ALCAP: Alignment-Augmented Music Captioner

Zihao He[1], **Weituo Hao**[2], **Wei-Tsung Lu**[2], **Changyou Chen**[3]
Kristina Lerman[1], **Xuchen Song**[2]
[1]USC Information Sciences Institute; [2]TikTok Inc.
[3]Department of Computer Science and Engineering, University of Buffalo
`zihaoh@usc.edu`, `changyou@buffalo.edu`, `lerman@isi.edu`
`{weituohao, weitsung.lu, xuchen.song@bytedance.com}@bytedance.com`

Abstract

Music captioning has gained significant attention in the wake of the rising prominence of streaming media platforms. Traditional approaches often prioritize either the audio or lyrics aspect of the music, inadvertently ignoring the intricate interplay between the two. However, a comprehensive understanding of music necessitates the integration of both these elements. In this study, we delve into this overlooked realm by introducing a method to systematically learn multimodal alignment between audio and lyrics through contrastive learning. This not only recognizes and emphasizes the synergy between audio and lyrics but also paves the way for models to achieve deeper cross-modal coherence, thereby producing high-quality captions. We provide both theoretical and empirical results demonstrating the advantage of the proposed method, which achieves new state-of-the-art on two music captioning datasets. Our code is publicly available at `https://github.com/zihaohe123/ALCAP`.

1 Introduction

Learning to interpret music based on audio and lyrics has become an increasingly attractive research area for researchers in the filed music understanding and natural language processing (Manco et al., 2021; Zhang et al., 2022b). The insights gained from this research into multimodal representation learning enable a wide range of applications such as streaming media discovery (Salha-Galvan et al., 2021) and music recommendation with detailed and human-like descriptions (Andjelkovic et al., 2019), making the dynamics of search and recommendation engines more explainable. However, captioning music is a challenging task. The duality of music, with its often nebulous and repetitious lyrics intertwined with intricate audio compositions harboring multiple layers of information, introduces a sophisticated web of complexities for models to navigate and understand.

Previous works on music captioning have primarily focused on refining singular facets of the encoder-decoder paradigm, from the enhancement of the music encoder to the incorporation of sophisticated attention mechanisms and beam search strategies. However, little effort has been directed towards leveraging the correspondence between audio and lyrics, which could potentially provide useful information for generating high-quality captions. Zhang et al. (2022b) leverage the multimodal information from both lyrics and music through a cross-modal attention module, but the two modalities are not aligned before fusion. In reality, audio and lyrics are loosely aligned, as it is common for composers and lyricists to work separately in the music industry, resulting in different lyrics fitting the same melody. Additionally, the same words with different song patterns and styles can express diametrically opposite emotions. Therefore, the loose alignment between music and lyrics make them imperfect sources of data for existing multimodal learning methods that are not equipped with alignment mechanisms (Nichols et al., 2009; Zhang et al., 2022a). In this regard, accurate and comprehensive music interpretation should leverage the subtle connections between music and lyrics.

In addressing the complexities of music understanding, we propose to align audio and lyrics pairs with contrastive learning before modality fusion and caption generation. Intuitively, paired audio and lyrics should be brought close together in the latent space, while non-paired ones should be pulled apart. By adding a contrastive loss, the multimodal input pairs are forced to be more aligned, which in turn guides the model to achieve stronger cross-modal consistency for a more meaningful fused latent space, thus generating better music captions. To this end, we propose **Al**ignment Augmented

*Work done when Zihao He interned at TikTok Inc.

Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 16501–16512
December 6-10, 2023 ©2023 Association for Computational Linguistics

Music **Cap**tioner (ALCAP), which is an extension of BART-fusion (Zhang et al., 2022b) for music captioning with a contrastive learning based alignment augmentation module. We provide a theoretical explanation of why the proposed alignment module results in improved generalization from an information bottleneck perspective. Extensive experiments on the Song Interpretation Dataset (Zhang et al., 2022b) and the NetEase Cloud Music Review Dataset demonstrate ALCAP's superiority with marked improvements in key metrics (ROUGE-L and METEOR) over previous benchmarks. We also observe performance gain of ALCAP in cross-modal text-music retrieval, which is a common application in industry, providing an indirect perspective to evaluate the caption quality. Lastly, we explore the effect of contrastive loss weights on the model performance via grid search and conclude our ablation study by showing that our proposed multimodal alignment module leads to more concentrated attention on language tokens through visualization analysis.

Our contributions are summarized as follows:

- To the best of our knowledge, we are the first to propose an alignment augmentation module through cross-modal contrastive learning between music and lyrics for music captioning. By learning the interactions between the two modalities in an unsupervised manner, the model is guided to learn better cross-modal attention weights for meaningful fused latent space, leading to high-quality music interpretation generations.
- We provide a theoretical justification for the improved generalization of the proposed multimodal alignment module from an information theory perspective.
- Extensive experiments on two music captioning datasets demonstrate the effectiveness of our proposed alignment augmentation module, and we set the new state-of-the-art on the Song Interpretation Dataset.

2 Related Work

2.1 Multimodal Alignment

Multimodal representation learning has been increasingly important as modern intelligent applications require a comprehensive understanding of vision, language and speech (Yin et al., 2022). To learn meaningful latent spaces, unsupervised alignment between different modality inputs has been proven effective as an additional layer of structural information about the data. In the work of pretraining for speech synthesis (Bai et al., 2022), aligning the acoustic and phoneme inputs makes the model more capable of learning cross-modal attention weights, thereby improving the quality of acoustic signal reconstruction. ALBEF (Li et al., 2021) proposes to align vision and language before the modality fusion, purifying the multimodal input pairs, thus resulting in a more grounded vision and language representation. This approach can be interpreted as maximizing mutual information among different views of the same vision and language pair. μ-VLA (Zhou et al., 2022) introduces image-text level and region-phrase level alignment in vision and language pretraining so as to make the most of unpaired data. Goyal et al. (2022) propose a retrieval process operating on past experiences to provide the agent with contextual relevant information, improving sample efficiency and representation learning of the policy function. It proves the effectiveness of retrieval-augmented module in continuous decision making process which also applies to the sequence of words generation (Ren et al., 2017; Guo et al., 2018; Yu et al., 2022; Humphreys et al., 2022).

Challenges in music-language alignment. While the majority of work in this field focuses on the alignment of vision and language (Radford et al., 2021), the complexities of aligning audio, especially music, with text present a set of challenges fundamentally different from those of image-text alignment, as illustrated by the following reasons. 1) Richness of audio signals: Music, as an auditory medium, encompasses a rich variety of signals ranging from melodies, rhythms, timbres, to harmonic structures. These multifaceted signals, when combined, deliver a sonic experience that often possesses layers of meaning, emotional depth, and narrative nuances. 2) Ambiguity and subjectivity of lyrics: Lyrics, while being textual, are laden with poetic devices, metaphors, and often abstract representations. They can be open-ended, prompting multiple interpretations even without the musical context. When paired with music, lyrics can either complement the musical message or introduce added layers of ambiguity. 3) Synchronization challenges: Music and lyrics evolve synchronously over time. The alignment is not just about mapping the overall theme of a song to its lyrics; it is also about understanding how specific musical passages correspond to specific lyrical segments, reflecting shifts in emotion, intensity, or narrative. These challenges make the alignment of music and lyrics a unique and intricate problem. Hence, while the underlying principle of using contrastive learning might resemble existing models applied in image captioning, the intricacies of our domain necessitate tailored approaches. The application of such techniques to music captioning is relatively nascent, and our work aims to pave the way for more explorations in this direction.

Comparison between ALCAP and CLIP. While both models employ contrastive learning for multimodal alignment, their objectives and applications are distinct. CLIP (Radford et al., 2021) is designed for image-text understanding and generation, leveraging a large dataset of images with their corresponding textual captions. In contrast, ALCAP focuses specifically on music captioning by aligning audio and lyrics pairs using cross-modal contrastive learning before modality fusion. This novel alignment augmentation module in ALCAP is tailored to address the unique challenges posed by the ambiguous and repetitive nature of lyrics, as well as the complexity of audio signals in music.

2.2 Multimodal Music Captioning

Music captioning is a challenging task as it requires the model to not only comprehensively understand both music and corresponding lyrics but also to avoid overfitting on limited music-lyrics pairs due to copyright restrictions. MusCaps (Manco et al., 2021) firstly addresses the music captioning task from an audio captioning perspective, using a multimodal input encoder-decoder architecture based on LSTM (Hochreiter and Schmidhuber, 1997). While MusCaps achieves a performance boost in caption generation, its predictive word sequence is limited to 20 tokens, which narrows down the approach's applicability, or at least not suitable for our long and human-like language composition scenario. One of the most relevant works to ours is BART-fusion (Zhang et al., 2022b) which is built on top of BART (Lewis et al., 2020), adding a music encoder and modality fusion module. However, BART-fusion fails to fully mine the relationship between the music and lyrics input data. Inspired by works from retrieval augmented representation learning, we propose to improve the generalization ability of BART-fusion by introducing music and lyrics alignment before modality fusion.

3 Methodology

In this section, we introduce the architecture of ALCAP, which is based on BART-fusion (Zhang et al., 2022b). We first state the problem definition, then go through each module of the architecture. The overall framework of ALCAP is shown in Figure 1.

3.1 Problem Definition

Given a song represented as a music-lyrics pair $\boldsymbol{x}_i$, with a music track $\boldsymbol{m}_i$ and its corresponding lyrics $\boldsymbol{t}_i$, we aim to generate the caption (or interpretation) $\hat{\boldsymbol{y}}_i$ of the song, consisting of a sequence of word tokens. In a typical setting of captioning, the attention-based encoder-decoder architecture is adopted to learn the mapping function from multimodal input to text output $f_{\boldsymbol{\theta}} : \{\boldsymbol{m}_i, \boldsymbol{t}_i\} \rightarrow \hat{\boldsymbol{y}}_i$. The model parameters $\boldsymbol{\theta}$ are optimized to generate the caption that is most consistent with the human annotated caption $\boldsymbol{y}_i$.

3.2 Multimodal Encoding

Music Encoder To obtain the representation of the music track, we use a pre-trained music encoder that includes a convolutional front-end and Transformer encoder layers (Won et al., 2019). The model was originally trained to classify music audio into 50 tags under a multi-class setting using the Million Song Dataset (Bertin-Mahieux et al., 2011). These tags cover various musical characteristics, such as the genre (e.g., Jazz and Blues), mode, and the presence of specific instruments (e.g., piano or guitar). To perform the classification, the mel-spectrogram of a music track $\boldsymbol{m}_i$ is first passed through a series of CNN layers for local feature aggregation in the time and frequency axis. The intermediate features are then fed into two Transformer encoder layers to model the information along the time axis, taking into account that elements of music can appear at different moments within a music clip. In Won et al. (2019), the output embedding series from the Transformer layer is further pooled to perform the classification task. However, in this paper, the embedding series $\boldsymbol{h}_i^m \in \mathbb{R}^{l_m \times d_m}$ is used directly, where l_m is the length of the music sequence and d_m is the hidden dimension.

Lyrics Encoder The representation of lyrics $\boldsymbol{t}_i$ is obtained following standard BART encoder (Lewis

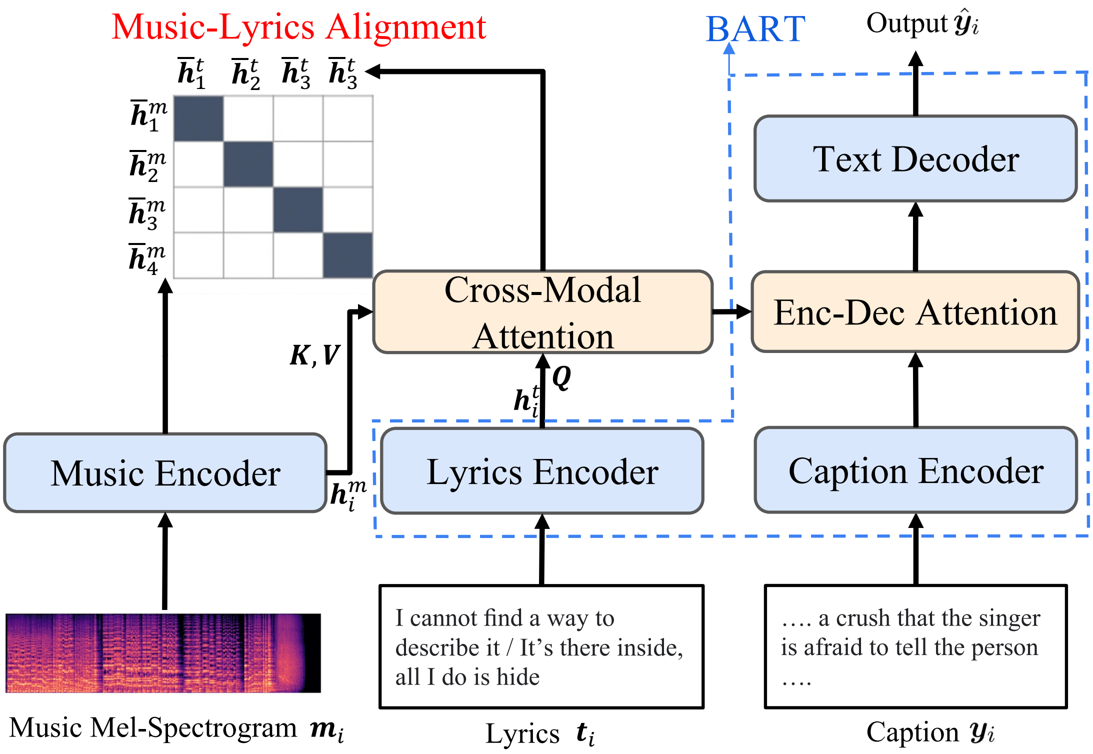

Figure 1: An overview of ALCAP. The encoded representations of music and lyrics are first aligned using contrastive learning, then the aligned representations are fused using cross-attention, and further decoded through the text decoder. The architecture is based on BART (Lewis et al., 2020).

et al., 2020), and denoted as $h_i^t \in \mathbb{R}^{l_t \times d_t}$, where l_t is the length of the lyrics sequence and d_t is the hidden dimension. The encoder consists of six multi-head self-attention layers.

3.3 Multimodal Representation Alignment

Music and lyrics are not inherently connected, as different lyrics can fit the same melody, and the same lyrics can convey different emotions when paired with dynamic, rhythmic music. To fully capture the interactions between music and lyrics, we propose using contrastive learning before modality fusion to explicitly align the two modalities. This is expected to result in improved performance due to increased interactions between the two modalities, as has been previously shown to be effective in the vision and language domain (Li et al., 2021).

To be specific, given a batch of input music-lyrics pairs $\{(\boldsymbol{m}_1, \boldsymbol{t}_1), (\boldsymbol{m}_2, \boldsymbol{t}_2), ..., (\boldsymbol{m}_n, \boldsymbol{t}_n)\}$, we first obtain the music representations $\{\boldsymbol{h}_1^m, \boldsymbol{h}_2^m, ..., \boldsymbol{h}_n^m\}$ by the music encoder, and lyrics representations $\{\boldsymbol{h}_1^m, \boldsymbol{h}_2^m, ..., \boldsymbol{h}_n^m\}$ by the lyrics encoder respectively. As both music and lyrics are sequences, we denote $\bar{\boldsymbol{h}}$ as the mean aggregation of $\boldsymbol{h}$ along the sequence length dimension. Through a linear transform on $\bar{\boldsymbol{h}}$, we obtain the latent code $\boldsymbol{z}$ and use the InfoNCE loss (Oord et al., 2018) as the contrastive learning objective in latent space, as

$$\mathcal{L}_{contrast} = -\sum_{i=1}^{n} \log \frac{\sigma(\boldsymbol{z}_i^m \cdot \boldsymbol{z}_i^t / \tau)}{\sum_k \sigma(\boldsymbol{z}_i^m \cdot \boldsymbol{z}_k^t / \tau)}, \quad (1)$$

where $\boldsymbol{z}_i^m$ and $\boldsymbol{z}_i^t$ are the latent code of music and lyrics respectively, and $\sigma(\cdot)$ is the exponential function. For simplicity, we ignore the symmetric version by switching $\boldsymbol{z}_i^m$ and $\boldsymbol{z}_i^t$ in Equation 1, which is also applicable for the purpose of modality alignment. Note that InfoNCE can be interpreted as an estimator of a lower bound of mutual information (Belghazi et al., 2018; Oord et al., 2018; Cheng et al., 2020). We will incorporate this to prove the effectiveness of out proposed alignment module both theoretically and empirically, which is supposed to be non-trivial. We will revisit this in § 4 and § 6.

3.4 Multimodal Fusion and Decoding

Before decoding, the aligned representations of music tracks $\boldsymbol{h}_i^m$ and lyrics $\boldsymbol{h}_i^t$ are further fused by a cross-attention module, where the lyrics representations are linearly projected as queries, and the music representations are projected as keys and values. The process can be described as

$$\boldsymbol{h}_i^f = \mathcal{T}(\mathbf{Q}, \mathbf{K}, \mathbf{V}),$$
$$\mathbf{Q} = \mathbf{W}^Q \boldsymbol{h}_i^t, \mathbf{K} = \mathbf{W}^K \boldsymbol{h}_i^m, \mathbf{V} = \mathbf{W}_\ell^V \boldsymbol{h}_i^m, \quad (2)$$

where $\boldsymbol{h}_i^f$ is the final fused representation, $\mathbf{W}^Q \in \mathbb{R}^{d_t \times d_k}$, $\{\mathbf{W}_\ell^K, \mathbf{W}_\ell^V\} \in \mathbb{R}^{d_m \times d_k}$ are linear transform parameters, respectively; d_k is the projection dimension.

The fused representation contains semantics from both the music track and the lyrics, as the alignment by contrastive learning ensures sufficient interactions between them. While the multimodal encoder fused the text and music as a whole, the decoding process follows a teacher-forcing fashion to predict each caption words, *i.e.*, the ground-truth word token of the ith sample $\boldsymbol{y}_{i,t}$ are provided at every step t during training. We use the BART decoder (Lewis et al., 2020) to generate the caption autoregressively and maximize the factorized conditional likelihood. The caption loss is defined as

$$\mathcal{L}_{\text{cap}} = -\frac{1}{n} \sum_{i=1}^{n} \sum_{t=1}^{T} \log P(\boldsymbol{y}_{i,t} | \boldsymbol{y}_{i,<t}, \boldsymbol{h}_i^f), \quad (3)$$

where $\boldsymbol{y}_{i,<t}$ is the ground-truth word token before step t and P indicates the probability of the token at step t conditioning on previous tokens and fused multimodal representation.

3.5 Overall Learning Objective

To this end, we define the final loss to be the weighted sum of the caption loss and the contrastive learning loss as follows:

$$\mathcal{L} = \mathcal{L}_{\text{cap}} + \alpha * \mathcal{L}_{\text{contrast}}, \quad (4)$$

where α is the weight of the contrastive learning loss, balancing the contribution of captioning and multimodal alignment.

4 An Information Theoretical Perspective

In this section, we explain the performance improvement of our alignment module based on contrastive learning from a mutual information perspective.

Given an input pair $\boldsymbol{x}_i := \{\boldsymbol{m}_i, \boldsymbol{t}_i\}$, information bottleneck (IB) (Alemi et al., 2016) encourages the model to find minimal but sufficient information about the input $\boldsymbol{x}_i$ with respect to the target caption words $\boldsymbol{y}_i$. In other words, the objective of the training process in IB can be formulated as

$$\max_{p_\theta(\boldsymbol{z}|\boldsymbol{x})} \mathrm{I}(\boldsymbol{y}; \boldsymbol{z}) - \beta \mathrm{I}(\boldsymbol{x}; \boldsymbol{z}), \quad (5)$$

where $\mathrm{I}(\boldsymbol{y}; \boldsymbol{z})$ is the mutual information between the output and the latent code, $\mathrm{I}(\boldsymbol{x}; \boldsymbol{z})$ is the mutual information between the input and the latent code, and $p_\theta(\boldsymbol{z}|\boldsymbol{x})$ is the conditional distribution of latent code parameterized by the encoder θ. To optimize the IB, an upper bound on $\mathrm{I}(\boldsymbol{x}; \boldsymbol{z})$ is typically taken for generalization ability of a model (Tishby et al., 2000; Alemi et al., 2016). From the information perspective, we show the following lower bound on the mutual information of $(\boldsymbol{x}, \boldsymbol{z})$ in our setting.

Proposition 4.1. *The mutual information of $(\boldsymbol{x}, \boldsymbol{z})$ in our setting is upper bounded by*

$$I(\boldsymbol{x}; \boldsymbol{z}) \leq \mathcal{R}(\boldsymbol{z}) - I(m; t),$$

where $\mathcal{R}(\boldsymbol{z}) \triangleq \mathbb{E}_{p((\boldsymbol{m},\boldsymbol{t})|\boldsymbol{z})} \left[\log \frac{\mathbb{E}_{p(\boldsymbol{z})}[p((\boldsymbol{m},\boldsymbol{t})|\boldsymbol{z})]}{p(\boldsymbol{m})p(\boldsymbol{t})} \right]$ *depends only on $\boldsymbol{z}$ and is independent of $\boldsymbol{x}$.*

In light of the fact that contrastive learning tends to maximize mutual information between $(\boldsymbol{m}, \boldsymbol{t})$ pairs, the above lower bound suggests that it can be considered as an approximate implementation of information bottleneck. Furthermore, if the music-lyrics pairs used in contrastive learning are not well aligned, one can actually prove that the learning will fail.

Proposition 4.2. *If the music-lyrics pairing in the learning process is random such that the music and lyrics are sampled independently, then the mutual information between the input $\boldsymbol{x}$ and the representation $\boldsymbol{z}$ will be zero, and thus the encoder cannot learn anything useful.*

The proof is provided in Appendix A. To sum up, based on the InfoNCE loss (Gutmann and Hyvärinen, 2010), the proposed alignment module can be interpreted as maximizing the mutual information lower bound between the music $\boldsymbol{m}$ and corresponding text $\boldsymbol{t}$, which translates to minimizing the mutual information between the input $\boldsymbol{x}$ and the latent code $\boldsymbol{z}$, and consequently improving the generalization ability of the model.

5 Data

In this paper, we experiment on two datasets – the Song Interpretation Dataset (Zhang et al., 2022b) and the NetEase Cloud Music Review Dataset.

5.1 Song Interpretation Dataset

The Song Interpretation (SI) Dataset dataset (Zhang et al., 2022b) contains audio excerpts from 27,834 songs from Music4All Dataset (Santana et al., 2020) and 490,000 user interpretations of the songs. Each song is in 30 seconds and recorded at 44.1 kHz. Based on user votes of the interpretations, Zhang et al. (2022b) create three variants of the dataset, as 1) SI Full: the full dataset after some preprocessing; 2) SI w/voting ≥ 0: the subset with only interpretations that received non-negative votes; 3) SI w/voting > 0: the subset with only interpretations that received positive votes. The sizes of the training splits of the three datasets are 279,283, 265,360 and 49,736 respectively. All three datasets share the same test split consisting of 800 instances.

5.2 NetEase Cloud Music Review Dataset

In addition to the Song Interpretation Dataset where the interpretations were mostly written by people who grew up under the influence of European and American culture, we curate another dataset - the **N**etEase **C**loud **M**usic (NCM) Review Dataset, where the reviews were written by people from China. NCM is a free music streaming service that is immensely popular in China. One of its most prominent features is that users can create their own playlists, write reviews and share the playlists with other users.

We collect user-created playlists from NCM and keep those consisting of only English songs. Because our model generates captions at an individual song level, for each playlist, we keep one song from it that has the highest popularity, i.e., the song that has been collected to most playlists[1]. As a result, from each playlist, we have an instance of the song-review pair. For each song, we keep the middle 30 seconds excerpt and sample it at 22.05kHz. Since BART (Lewis et al., 2020) is pretrained in English, we translate the Chinese reviews into English using Google Translate.

The NCM Review Dataset contains 22,210 playlists (songs) and their reviews. An example is shown in Figure 2. We randomly split the dataset into train/val/test splits, with sizes of 15,547, 3,331, and 3,332.

Title: I Feel Lucky
Artist: Mary Chapin Carpenter
Lyrics: Well I woke up this morning stumbled out of my rack; I opened up the paper to the page in the back; It only took a minute for my finger to find; My daily dose of destiny, under my sign; My eyes just about popped out of my head; It said "the stars are stacked against you girl, get back in bed"; I feel lucky, I feel lucky, yeah; No Professor Doom gonna stand in my way….
Review: As soon as you listen to the style of the song, you will know that it is the familiar style of the American West and the South, the taste of country rock. How could such a delicacy be missing from the music feast? Let's enjoy it together.

Figure 2: An example in NetEase Cloud Music (NCM) Review Dataset.

6 Experiments

6.1 Experimental Setup

We resample each song at 16kHz and take a 15s excerpt. The maximum caption length is 512. The model is implemented in PyTorch (Paszke et al., 2019). We use the BART implementation *facebook/bart-base* from Huggingface (Wolf et al., 2019). We use a batch size of 26 and a learning rate of $5e-5$. The weight of contrastive learning α loss is set to 0.02. For better computation efficiency we freeze the parameters in the music encoder and precompute the music representations. We train the model for 20 epochs and report the results on the test split using the checkpoint with the best evaluation performance. All hyperparameter tuning is based on grid search. All models are trained on a Tesla A100 GPU with 40GB memory. The training time for SI-Full, SI w/voting ≥ 0, SI w/voting > 0, and NCM Review are 28h, 28h, 5h, and 3h respectively.

We use ROUGE-1,2,L (ROUGE, 2004) and METEOR (Banerjee and Lavie, 2005) as evaluation metrics. ROUGE measures the overlap of n-grams between the referenced text and the generated text. On top of ROUGE, METEOR complementarily measures the semantic similarity between the two pieces of text by taking into account synonyms through WordNet. For both metrics, we use the implementation with default parameters from Huggingface Datasets library. We use three random seeds and report the average performance on the test set.

[1]Admittedly this is not the best way to create the song-review pairs given that the reviews were written at the entire playlist level. Nevertheless, the main goal of this paper is NOT to introduce this curated dataset, but to demonstrate the effectiveness of ALCAP in generating better song captions on different datasets.

Dataset	Method	R-1	R-2	R-L	Meteor
SI Full	BART	44.1	14.0	24.5	22.5
	BART-fusion	46.1	15.0	25.1	23.0
	ALCAP	**48.2**±0.3	**15.7**±0.1	**26.4**±0.2	**27.8**±0.2
SI w/voting ≥ 0	BART	44.8	14.9	24.7	22.7
	BART-fusion	46.7	**15.6**	25.5	23.4
	ALCAP	**47.7**±0.3	**15.6**±0.1	**26.0**±0.1	**27.7**±0.1
SI w/voting > 0	BART	41.2	13.0	22.8	22.0
	BART-fusion	44.3	14.6	24.7	22.6
	ALCAP	**49.8**±0.1	**16.0**±0.1	**27.1**±0.1	**27.7**±0.1
NCM Review	BART-fusion	18.2±0.3	1.9±0.1	13.6±0.1	10.9±0.3
	ALCAP	**20.6**±0.2	**2.6**±0.1	**15.3**±0.3	**11.8**±0.2

Table 1: Results of music captioning. The best results are highlighted in bold.

6.2 Baselines

BART is a model that utilizes only unimodal textual information from lyrics. The BART-fusion model, on the other hand, fuses representations from music and lyrics, but the two representations are not aligned prior to modality fusion. The results of these two baselines are reported in Zhang et al. (2022b). We do not compare with Manco et al. (2021), which focuses on short-length music descriptions with a maximum of 22 tokens.

6.3 Experiments I: Music Captioning

The results are presented in Table 1. We have found that ALCAP outperforms both BART-fusion and BART on all four datasets, in terms of all four metrics, thereby setting a new state-of-the-art. Specifically, the improvement on METEOR is more pronounced than on ROUGE metrics, which demonstrates that ALCAP is capable of capturing the semantics of the song for music captioning, not just memorizing the syntax. Furthermore, the results on the NCM Review for both models are overall worse than those on SI datasets. We believe this is due to the weaker correspondence between the music tracks and reviews in the NCM Review, as the reviews were originally created at the playlist level. Despite this, ALCAP is still able to capture such weak correspondence and achieve a significant improvement over the baseline.

6.4 Experiments II: Text-Music Retrieval

One of the most practical applications of music captioning is text-music retrieval, where given a piece of music description, the goal is to retrieve the most relevant music according to the text. In light of this, in this analysis, we test the retrieval capability of ALCAP and the baseline model. The setting of cross-modal retrieval in this experiment is different from previous works such as Yu et al. (2022), where the retrieval is performed on the two modalities that are directly aligned through contrastive learning.

As proposed in Zhang et al. (2022b), we randomly select one sentence from the human-generated interpretation or review in the test split, and use it as a query. The queries are used to retrieve the corresponding songs through their generated captions by our models. Specifically, we compute the representations of the queries and generated captions using Sentence-BERT (Reimers and Gurevych, 2019). Thus, for each query, we obtain a ranked list of retrieved songs through the cosine similarities between the query representation and generated caption representations. We use precision@k and recall@k as the evaluation metrics. The results are shown in Table 2.

We observe that ALCAP outperforms BART-fusion on most datasets and metrics, indicating the superiority of cross-modal alignment between music tracks and lyrics that makes the generated captions more semantically aligned with human-written texts. This is apart from several cases where ALCAP ties with BART-fusion. Compared to SI datasets, the relatively low performance on NCM Review of both models is due to 1) the weak correspondence between the song and the review as we mentioned in previous sections, and 2) the retrieval pool (test split) is much larger – 3,332 for NCM Review vs. 800 for SI. Nevertheless, ALCAP

Dataset	Method	p@5	p@10	p@20	p@30	r@5	r@10	r@20	r@30
SI Full	BART-fusion	3.2%	1.9%	1.2%	0.9%	16.0%	19.0%	24.0%	27.0%
	ALCAP	**3.6%**	**2.1%**	1.2%	**1.0%**	**18.0%**	**21.0%**	24.0%	**31.0%**
SI w/voting >= 0	BART-fusion	2.2%	2.0%	1.0%	0.7%	11.0%	17.0%	20.0%	23.0%
	ALCAP	**4.2%**	**2.6%**	**1.5%**	**1.1%**	**21.0%**	**26.0%**	**30.0%**	**32.0%**
SI w/voting >0	BART-fusion	2.2%	1.2%	0.9%	0.7%	11.0%	12.0%	18.0%	20.0%
	ALCAP	**3.0%**	**1.6%**	**1.0%**	**0.8%**	**15.0%**	**16.0%**	**20.0%**	**23.0%**
NCM Review	BART-fusion	0.2%	0.1%	0.1%	0.1%	1.0%	1.0%	2.0%	2.0%
	ALCAP	0.2%	**0.2%**	0.1%	0.1%	1.0%	**2.0%**	**3.0%**	**4.0%**

Table 2: Results of text-music retrieval. The best results are highlighted in bold.

still outperforms the baseline in such a challenging scenario.

6.5 Case Study I: Visualizing the Attention Weights

To better understand the mechanism within the cross-attention module, we plot the attention weights of BART-fusion and ALCAP on five input examples from the training set in Figure 3. Both models are trained on the SI w/voting > 0 dataset.

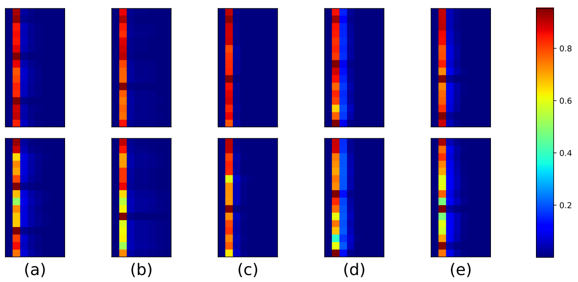

Figure 3: Illustration of the cross-modal weights for five samples (a)$\sim$(e). The first row shows the cross-modal attention weights output by BART-fusion and the second row shows the weights by ALCAP. The y-axis and x-axis in each sub-graph indicates the text tokens and music segments respectively.

The attention weights from ALCAP appear to be more focused on specific text tokens, in contrast to BART-Fusion, which has a more evenly distributed attention across all tokens. This phenomenon suggests that ALCAP, equipped with the cross-modal alignment module, is more effective at learning the interactions between the music audio and text domains.

6.6 Case Study II: Examples of Generated Caption

In this case study we show a representative example of generated captions from ALCAP and BART-fusion on *Child In Time* by Deep Purple, as in Figure 4. The song is from the test split of SI, and both models are trained on SI w/voting > 0.

From the lyrics and the reference interpretation, we can infer that the song is about war, which is captured by ALCAP. The generated caption contains "shot" and "sniper", which indicates that the model has correctly understood the theme of the song. However, BART-fusion fails to interpret the song correctly, instead interpreting it as a love song. We propose that this is due to the song's 70s Rock music style being too typical, and the lack of cross-modal alignment in BART-fusion. This allows the unimodal information from the sound track to dominate and confuse the model. As 70s Rock encompasses a wide range of topics, including love, it becomes harder to identify the correct topic of war. However, the alignment module in ALCAP manages to capture the semantics of the song and provide a more accurate interpretation.

6.7 Ablation Study: Effect of Contrastive Learning Weight

To further investigate the effect of multimodal alignment through contrastive learning, we show the performances of using different weights of contrastive learning α on SI w/voting > 0 on music captioning (Figure 5) and text-music retrieval (Figure 6).

We observe that in both figures, the scores peak at $\alpha = 2e-2$, and decrease with higher weights or lower weights. When the weight is below $2e-2$, the model fails to learn sufficient alignment between the two modalities; on the other hand, when the weight is greater than $2e-2$, the model suffers because the overly large weight of contrastive learning loss negatively affects the optimizing of caption

Lyrics: Sweet child in time; You'll see the line; The line that's drawn between; Good and bad; See the blind man; Shooting at the world; Bullets flying; Oh taking toll; If you've been bad; Oh Lord I bet you have; And you've not been hit; Oh by flying lead; You'd better close your eyes; Oh bow your head; Wait for the ricochet; I wanna hear you sing; Sweet child in time; You'll see the line; The line that's drawn between; Good and bad; See the blind man; Shooting at the world ....
Reference: I don't think this song is about the Vietnamese war, but rather war in general. The blind man is not just a person, or a country - it's mankind. Almost every nation has been in a conflict or war. And yes, some countries are more powerful than others, but in the end - what has war resolved? Every country has sinned, some more than others. Nobody is innocent. Every action has a reaction - you shoot bullets blindly for unjustified reasons, and you are bound to have it ricochet back to you. The lyrics alone make this song great, but what really sends it to one of the all-time best works of genius art is everything else in the song - from the captivating organ intro, to Blackmore's unreal guitar skills, to Gillain's extraordinary vocals, and the passion that brought it all together. Probably the best song ever.
BART-fusion: I think this song is about a guy who is in the midst of a breakup with his girlfriend. He is trying to figure out what to do with his life, but he ends up killing himself. He tries to convince himself that he can do it, and he can't do it. He wants to be able to show her that he still loves her, and is willing to do anything to help her, even if it means killing herself. He can't do it by himself, because he knows she will never be happy, and will never see him again. He will never give up on her, no matter how hard he tries. He knows that she will always love him, and it will be a long time before he does, but she is not willing to give up, and she is going to be with him forever.
ALCAP: I think this song is about a man who was shot in the head by a sniper. He was shot by the sniper. The sniper was shot at the end of the night, and he was killed. He is still alive, but he is in a coma, and the sniper is still there. He has been shot. He's dying. It's a sad song, but it's a beautiful song. I think it's about a guy who is dying. I'm not sure what this song means, but I think this is a good song.

Figure 4: An example of generated captions from ALCAP and BART-fusion on *Child In Time* by Deep Purple.

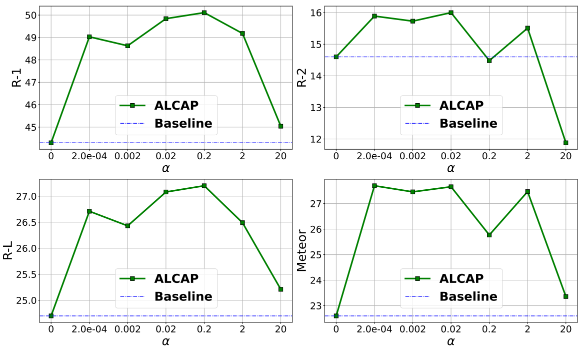

Figure 5: Results of music captioning using different weights of contrastive learning α on SI w/voting.

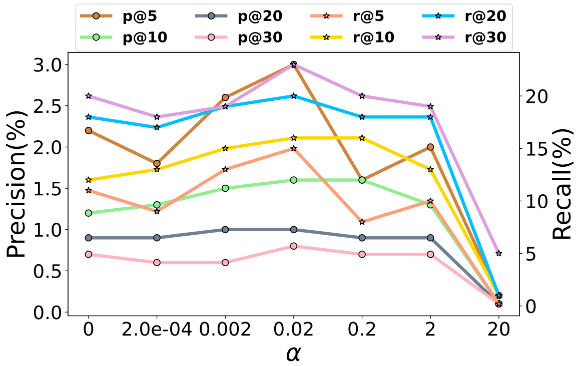

Figure 6: Results of text-music retrieval using different weights of contrastive learning α on SI w/voting.

loss, which is the most prominent at $\alpha = 20$.

7 Conclusions

In this paper, we introduce the **Al**ignment augmented music **Cap**tioner (ALCAP), a pioneering model designed to enhance music captioning by incorporating an alignment augmentation module using cross-modal contrastive learning. Our model's distinctiveness, particularly in the under-researched music domain, stems from its ability to successfully bridge the gap between music and its linguistic interpretation. We provide a theoretical analysis of the improved generalization of our model from an information bottleneck perspective. Experiments on two music captioning datasets demonstrate the effectiveness of ALCAP, and we achieve the new state-of-the-art on both of them.

Our next steps will focus on collecting extensive music song data from the Web and pretrain the music-lyrics alignment module, after which the model is further finetuned on small-scale music captioning. This model would be tailored to cater to large-scale song interpretation generation on various genres and styles of music. In addition, ground-truth interpretations, as drawn from user-generated content, inherently bring the risk of biases. We will be committed to developing strategies to effectively mitigate biases from these interpretations, ensuring that our model's outputs are as neutral and accurate as possible.

Limitations

Due to computational limitations, the parameters of the music encoder in ALCAP were fixed, and the music representations were precomputed, following Zhang et al. (2022b). This approach may result in a decrease in performance compared to a model where the music encoder is fully fine-tuned for the music captioning task. Additionally, the Song Interpretation dataset, being the only publicly available music captioning dataset, is limited in scope, making it challenging to pretrain a large music captioning model that is suitable for various genres and styles of music. Furthermore, user-generated song interpretations and reviews may contain biases or even hate speech, which could be perpetuated during training of the model.

Acknowledgements

Changyou Chen is partially supported by NSF AI Institute-2229873, NSF RI-2223292, an Amazon research award, and an Adobe gift fund. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation, the Institute of Education Sciences, or the U.S. Department of Education.

References

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. 2016. Deep variational information bottleneck. *arXiv preprint arXiv:1612.00410*.

Ivana Andjelkovic, Denis Parra, and John O'Donovan. 2019. Moodplay: interactive music recommendation based on artists' mood similarity. *International Journal of Human-Computer Studies*, 121:142–159.

He Bai, Renjie Zheng, Junkun Chen, Mingbo Ma, Xintong Li, and Liang Huang. 2022. A 3t: Alignment-aware acoustic and text pretraining for speech synthesis and editing. In *International Conference on Machine Learning*, pages 1399–1411. PMLR.

Satanjeev Banerjee and Alon Lavie. 2005. 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*, pages 65–72.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. 2018. Mutual information neural estimation. In *International conference on machine learning*, pages 531–540. PMLR.

Thierry Bertin-Mahieux, Daniel PW Ellis, Brian Whitman, and Paul Lamere. 2011. The million song dataset.

Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin. 2020. Club: A contrastive log-ratio upper bound of mutual information. In *International conference on machine learning*, pages 1779–1788. PMLR.

Anirudh Goyal, Abram Friesen, Andrea Banino, Theophane Weber, Nan Rosemary Ke, Adria Puigdomenech Badia, Arthur Guez, Mehdi Mirza, Peter C Humphreys, Ksenia Konyushova, et al. 2022. Retrieval-augmented reinforcement learning. In *International Conference on Machine Learning*, pages 7740–7765. PMLR.

Tszhang Guo, Shiyu Chang, Mo Yu, and Kun Bai. 2018. Improving reinforcement learning based image captioning with natural language prior. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 751–756.

Michael Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pages 297–304.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Peter C Humphreys, Arthur Guez, Olivier Tieleman, Laurent Sifre, Théophane Weber, and Timothy Lillicrap. 2022. Large-scale retrieval for reinforcement learning. *arXiv preprint arXiv:2206.05314*.

D. Koller and N. Friedman. 2009. *Probabilistic Graphical Models: Principles and Techniques*. Adaptive computation and machine learning. MIT Press.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. 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*, pages 7871–7880.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705.

Ilaria Manco, Emmanouil Benetos, Elio Quinton, and György Fazekas. 2021. Muscaps: Generating captions for music audio. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Eric Nichols, Dan Morris, Sumit Basu, and Christopher Raphael. 2009. Relationships between lyrics and melody in popular music. In *ISMIR 2009-Proceedings of the 11th International Society for Music Information Retrieval Conference*, pages 471–476.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks.

In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Zhou Ren, Xiaoyu Wang, Ning Zhang, Xutao Lv, and Li-Jia Li. 2017. Deep reinforcement learning-based image captioning with embedding reward. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 290–298.

Lin CY ROUGE. 2004. A package for automatic evaluation of summaries. In *Proceedings of Workshop on Text Summarization of ACL, Spain*.

Guillaume Salha-Galvan, Romain Hennequin, Benjamin Chapus, Viet-Anh Tran, and Michalis Vazirgiannis. 2021. Cold start similar artists ranking with gravity-inspired graph autoencoders. In *Fifteenth ACM Conference on Recommender Systems*, pages 443–452.

Igor André Pegoraro Santana, Fabio Pinhelli, Juliano Donini, Leonardo Catharin, Rafael Biazus Mangolin, Valéria Delisandra Feltrim, Marcos Aurélio Domingues, et al. 2020. Music4all: A new music database and its applications. In *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*, pages 399–404. IEEE.

Naftali Tishby, Fernando C Pereira, and William Bialek. 2000. The information bottleneck method. *arXiv preprint physics/0004057*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Minz Won, Sanghyuk Chun, and Xavier Serra. 2019. Toward interpretable music tagging with self-attention. *arXiv preprint arXiv:1906.04972*.

Yufeng Yin, Jiashu Xu, Tianxin Zu, and Mohammad Soleymani. 2022. X-norm: Exchanging normalization parameters for bimodal fusion. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, pages 605–614.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*.

Chen Zhang, Luchin Chang, Songruoyao Wu, Xu Tan, Tao Qin, Tie-Yan Liu, and Kejun Zhang. 2022a. Relyme: Improving lyric-to-melody generation by incorporating lyric-melody relationships. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1047–1056.

Yixiao Zhang, Junyan Jiang, Gus Xia, and Simon Dixon. 2022b. Interpreting song lyrics with an audio-informed pre-trained language model. In *Ismir 2022 Hybrid Conference*.

Mingyang Zhou, Licheng Yu, Amanpreet Singh, Mengjiao Wang, Zhou Yu, and Ning Zhang. 2022. Unsupervised vision-and-language pre-training via retrieval-based multi-granular alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16485–16494.

A Proofs of Proposition 4.1 and 4.2

$$
\begin{aligned}
\mathrm{I}(\boldsymbol{x};\boldsymbol{z}) =& \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{z})}\left[\log \frac{p(\boldsymbol{x},\boldsymbol{z})}{p(\boldsymbol{x})p(\boldsymbol{z})}\right] \\
=& \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{z})}\left[\log \frac{p(\boldsymbol{x}|\boldsymbol{z})}{p(\boldsymbol{x})}\right] \\
=& \mathbb{E}_{p(\boldsymbol{m},\boldsymbol{t},\boldsymbol{z})}\left[\log \frac{p((\boldsymbol{m},\boldsymbol{t})|\boldsymbol{z})}{p(\boldsymbol{m},\boldsymbol{t})}\right] \\
=& \mathbb{E}_{p(\boldsymbol{m},\boldsymbol{t},\boldsymbol{z})}\left[\log \frac{p((\boldsymbol{m},\boldsymbol{t})|\boldsymbol{z})}{p(\boldsymbol{m})p(\boldsymbol{t})}\right] \\
&- I(m;t) \\
=& \mathbb{E}_{p((\boldsymbol{m},\boldsymbol{t})|\boldsymbol{z})p(\boldsymbol{z})}\left[\log \frac{p((\boldsymbol{m},\boldsymbol{t})|\boldsymbol{z})}{p(\boldsymbol{m})p(\boldsymbol{t})}\right] \\
&- I(m;t) \\
\leq& \mathbb{E}_{p((\boldsymbol{m},\boldsymbol{t})|\boldsymbol{z})}\left[\log \frac{\mathbb{E}_{p(\boldsymbol{z})}[p((\boldsymbol{m},\boldsymbol{t})|\boldsymbol{z})]}{p(\boldsymbol{m})p(\boldsymbol{t})}\right] \\
&- I(m;t) \\
=& \mathbb{E}_{p((\boldsymbol{m},\boldsymbol{t})|\boldsymbol{z})}\left[\log \frac{p(\boldsymbol{m},\boldsymbol{t})}{p(\boldsymbol{m})p(\boldsymbol{t})}\right] \\
&- I(m;t)\ ,
\end{aligned}
$$

where the inequality follows by Jensen inequality. This completes the proof of Proposition 4.1.

Based on the above derivation, if $(\boldsymbol{m},\boldsymbol{t})$ pairs are sampled randomly, in the probabilistic graphical model language (Koller and Friedman, 2009), this corresponds to a V-structure between $(\boldsymbol{m},\boldsymbol{t})$ and $\boldsymbol{z}$. And a V-structure indicates the marginal independency between $\boldsymbol{m}$ and $\boldsymbol{t}$ (Koller and Friedman, 2009). Thus, we have

$$
\begin{aligned}
\mathrm{I}(\boldsymbol{x};\boldsymbol{z}) \leq& \mathbb{E}_{p((\boldsymbol{m},\boldsymbol{t})|\boldsymbol{z})}\left[\log \frac{p(\boldsymbol{m},\boldsymbol{t})}{p(\boldsymbol{m})p(\boldsymbol{t})}\right] - I(m;t) \\
=& \mathbb{E}_{p((\boldsymbol{m},\boldsymbol{t})|\boldsymbol{z})}\left[\log \frac{p(\boldsymbol{m})p(\boldsymbol{t})}{p(\boldsymbol{m})p(\boldsymbol{t})}\right] - I(m;t) \\
=& -I(\boldsymbol{m};\boldsymbol{t})
\end{aligned}
$$

Since we know that both $I(\boldsymbol{x},\boldsymbol{z})$ and $I(\boldsymbol{m};\boldsymbol{t})$ must be non-negative, we have

$$I(\boldsymbol{x};\boldsymbol{z}) = I(\boldsymbol{m};\boldsymbol{t}) = 0\ .$$

Consequently, this leads to the independency of $\boldsymbol{x}$ and $\boldsymbol{z}$, *i.e,*, $\boldsymbol{z}$ contains zero information of $\boldsymbol{z}$. This completes the proof of Proposition 4.2.