FUTGA: Towards Fine-grained Music Understanding through Temporally-enhanced Generative Augmentation

Junda Wu[1] Zachary Novack[1] Amit Namburi[1] Jiaheng Dai[1] Hao-Wen Dong[1]
Zhouhang Xie[1] Carol Chen[2] Julian McAuley[1]

[1] Computer Science and Engineering, UC San Diego [2] Computer Science Department, UC Los Angeles

Abstract

We propose FUTGA, a model equipped with fined-grained music understanding capabilities through learning from generative augmentation with temporal compositions. We leverage existing music caption datasets and large language models (LLMs) to synthesize fine-grained music captions with structural descriptions and time boundaries for full-length songs. Augmented by the proposed synthetic dataset, FUTGA is enabled to identify the music's temporal changes at key transition points and their musical functions, as well as generate detailed descriptions for each music segment. We further introduce a full-length music caption dataset generated by FUTGA, as the augmentation of the MusicCaps and the Song Describer datasets. The experiments demonstrate the better quality of the generated captions, which capture the time boundaries of long-form music. Generated temporal-aware music descriptions are illustrated in our demonstration https://namburiamit.github.io/futga.github.io/.

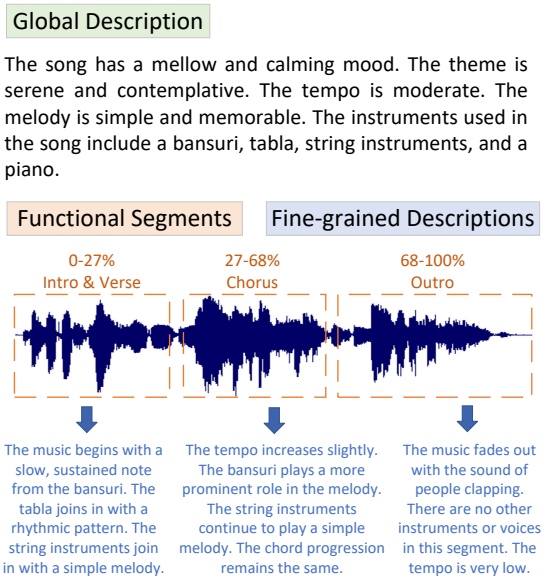

Figure 1: **Overview of FUTGA's capabilities.** Given a long-form audio example, FUTGA is able to provide time-located captions by automatically detecting functional segment boundaries, as well as global captions.

1 Introduction

Natural language music understanding, which extracts music information and generates detailed music captions, is a fundamental task within the MIR community, beneficial for a series of applications including music generation (Copet et al., 2024; Chen et al., 2024; Novack et al., 2024; Melechovsky et al., 2023), editing (Wang et al., 2023; Zhang et al., 2024), question-answering (Deng et al., 2023; Gao et al., 2022), and retrieval (Doh et al., 2023b; Wu et al., 2023; Bhargav et al., 2023). Recent developments in music foundation models (Gardner et al., 2023; Hussain et al., 2023; Tang et al., 2023; Liu et al., 2024) enable free-form music prompts and multitasking. These foundation models are developed based on pre-trained large language models (LLMs) and aligned with the music modality. Although LLM-powered music understanding models can leverage the abundant pre-trained music knowledge in caption generation, the success of modality alignment still requires a large amount of high-quality music caption data.

Restricted by the current music captioning paradigm, available music caption datasets are limited to two major challenges: (1) Conventional music captions focus only on the global description of a (potentially long) music clip, which cannot efficiently capture a piece of music's fine-grained characteristics nor differentiate it from other music within-genre songs. (2) Key structural information, such as time boundaries of functional music segments and time-aware musical changes, is mostly neglected in traditional music understanding and hard to retrieve due to the limitation in the length of music clips.

To address the limitations, we propose **FUTGA**, a generative music understanding model trained with time-aware music caption data and cali-

brated with Music Information Retrieval (MIR) features. We first augment the MusicCaps dataset (Agostinelli et al., 2023) by mixing music clips together into synthetic full-length songs. The corresponding music captions are composed with original short music captions as individual segment descriptions, which are also tagged with temporal segmentation information. To enable more realistic full-length music captioning, we further leverage a text-only LLM for the augmentation of the global music caption, musical changes between segments (*e.g.*, increase of volume, slowing down the tempo, introducing new instruments, etc.), and functional tags of the segments (*e.g.*, intro, verse, chorus, etc.), by paraphrasing and summarizing the template-based captions.

Inspired by existing Large Audio-Language Models (LALMs), we use the open-source SALMONN model (Tang et al., 2023) as the backbone and fine-tune the model with our developed synthetic full-length music caption data. Using our synthetic data augmentation, FUTGA is able to identify key transition points in musical changes and segment full-length songs according to their musical functions. For example, in Figure 1, we illustrate FUTGA's capacities as a novel form of music captioning. Given a song in full length, FUTGA can generate a global caption that summarizes the whole song's characteristics before identifying the music structure with time segments. Following the flow of music structures, FUTGA can further describe each music clip and musical changes between consecutive music clips. In addition, we also discover that the fine-tuned SALMONN model demonstrates a great instruction-following capacity to generate fine-grained music captions conditioned on given time boundaries and MIR features. By injecting the ground-truth information into the instruction prompt, we can accurately guide the model to generate fine-grained music captions corresponding to the time segments. With the final version of FUTGA, we propose automatically annotating the full-length songs in two existing datasets MusicCaps (Agostinelli et al., 2023) and Song Describer (Manco et al., 2023).

2 Temporally-enhanced Generative Augmentation

In this section, we introduce our proposed temporally-enhanced generative augmentation. Due to the limitation of existing music caption datasets, music captioning and understanding models can only generate global music descriptions for short music clips (Manco et al., 2023; Agostinelli et al., 2023; Doh et al., 2023a). To address this limitation, we propose the augmentation of synthetic music and caption composition, which empowers music understanding models with capacities of time-aware music segmentation and fine-grained music description generation.

For each sampled set of music clips C_k, the corresponding music caption set T_k and the clip length information L_k are interleaved and composed by the template,

$$\tilde{X}_k = \left\{ \left(\frac{l_{k,j-1}}{\sum_i l_{k,i}}, \frac{l_{k,j}}{\sum_i l_{k,i}}, t_{k,j} \right) \right\}_{j=1}^{n},$$

in which the specific original time-boundaries L_k are transformed into relative time-boundaries, which are always between $0 - 100\%$ ($l_{k,0} \equiv 0$). We use a relative time-boundary representation approach to minimize training bias towards specific numbers of music lengths in our model. In addition, relative time-boundary representation enables the model's ability to comprehend music of varying lengths, thereby improving the model's generalizability. By generating these relative time boundaries, our generative music understanding model gains a better awareness of the music's overall progression, which further enhances the model's temporal understanding of music.

To further augment rich MIR features in music captions, we further propose to use a text-only large language model (LLM) to augment the template-based caption $\tilde{X}_k$ with natural language descriptions, in which additional information, such as global captions, musical changes, and music structures, can be automatically extracted from the LLM. Since LLMs are pre-trained with abundant domain knowledge including music analysis (Tang et al., 2023; Gardner et al., 2023) and music information retrieval (Tang et al., 2023; Hussain et al., 2023), with enough context provided, such LLMs can accurately extract music information via language-based summarization and reasoning (Doh et al., 2023a). Inspired by LLMs' capacities in language reasoning, we propose to paraphrase and augment additional music information with instructions as follows:

Context: *Music Analysis* $\{\tilde{X}_k\}$. *This is a music analysis of a song. Note that the numbers indicate the time-boundaries of functional segments in this*

MusicCaps: The low quality recording features footsteps, birds chirping, distant traffic and wind sound effects. At the very end, there is a suspenseful, low, widely spread pad chord. It sounds like it is a sound for some kind of video.
Song Describer: A rock song with a French speaking male voice. It has the typical feeling of political/social rock song with a really simple guitar, bass and drum skeleton.
LP-MusicCaps: The low quality recording features a drum solo that consists of a double pedal kick, punchy snare, shimmering cymbals and boomy toms. It sounds energetic and manic, thanks to that kick pattern. The recording is mono and noisy.
FUTGA: The song exhibits a diverse range of musical styles and instrumentation. Overall, the mood is upbeat and energetic, with a theme of youthfulness and optimism. The tempo is moderate to fast, with a catchy melody and a simple chord progression. **0-27%: Intro and Verse** - The song begins with a haunting intro featuring syncopated rhythms and atmospheric chords. The verse introduces the main melody, which is characterized by a catchy hook and a simple chord progression. **27-58%: Chorus and Bridge** - The chorus features a soaring melody and energetic vocals. The bridge introduces a new melody and chord progression, which contrasts with the verse and chorus. **58-79%: Rock Section** - The song transitions into a rock section with a distorted electric guitar and a heavy drum beat. The section has a groovy feel and could be used in the soundtrack of a high school drama TV series. **79-100%: Acoustic Ballad** - The song concludes with an acoustic ballad, sung in a melancholic manner. The melody is simple and the chord progression is repetitive.

Table 1: A comparison example of captions generated or annotated by MusicCaps (Agostinelli et al., 2023), Song Describer (Manco et al., 2023), LP-MusicCaps (Doh et al., 2023a) and FUTGA.

Dataset	# Caption	# Segment	Tokens	Vocab.	# Inst.	# Genre	# Mood
MusicCaps	6k	–	48.9 ± 17.3	6,144	75	267	146
Song Describer	1k	–	21.7 ± 12.4	2,859	39	152	122
LP-MusicCaps	542k	–	45.3 ± 28.0	1,686	65	239	151
FUTGA	7k	4.32	472.419 ± 88.5	3,537	64	187	128

Table 2: Statistics summarization of generated or annotated music captions of baselines and FUTGA.

song.

Paraphrase: *Paraphrase the music analysis to make it sound like a coherent song, instead of a remix. Additionally, remove any mention of sound quality.*

Global Caption: *Start with a general description of the song focusing on subjectivity.*

Musical Change: *Describe the song in detail and explain transitions between parts of the song.*

Music Structure: *Remember to indicate the temporal annotations and music structures when talking about a specific part of the song.*

3 Dataset Creation: FUTGA

Based on the final version of our proposed music captioning model, we automatically generate music captions for whole songs between 2 minutes and 5 minutes in MusicCaps (Agostinelli et al., 2023) and Song Describer (Manco et al., 2023). During inference time, we set the repetition penalty as 1.5 to prevent repetitive descriptions of the same music segments. In addition, we also set the beam search number to 10 to find the statistically best captions. We allow a maximum of 2048 tokens to be generated from FUTGA.

As demonstrated in the comparison example in Table 1, FUTGA provides more fine-grained music understanding descriptions with time boundaries indicating music segments, for which the average segment number and the number of musical changes are reported in Table 2. In addition, we can observe relatively longer global captions with more details, which is also verified by the data statistics in Table 2.

In terms of music caption diversity, we first show that our captions have significantly larger numbers of tokens and vocabulary size, compared to existing music caption datasets. Second, our dataset still maintains good diversity in terms of unique genre, instrument, and music mood vocabularies, which are comparable to human or GPT-3.5 annotations. Thus, the FUTGA dataset can serve to augment existing music captioning models with strong temporal reasoning abilities without harming the model's generalizability, which will be further evaluated in our evaluation section.

	MusicCaps						Song Describer					
Model	B1	B2	B3	M	R	B-S	B1	B2	B3	M	R	B-S
LP-MusicCaps	19.77	6.70	2.17	12.88	13.03	84.51	1.68	0.71	0.27	7.68	2.76	79.62
FUTGA (complete)	9.21	4.18	1.97	**20.85**	11.96	82.62	4.58	1.72	0.61	12.82	6.90	81.21
FUTGA (global)	**26.46**	**10.93**	**4.66**	18.60	**17.40**	**86.48**	**14.23**	**5.04**	**1.75**	**15.04**	**11.67**	**85.42**

Table 3: Comparison results of caption generation for LP-MusicCaps and FUTGA.

4 Experiments and Results

We obtain 5K synthetic training samples by prompting the GEMMA-7B model (Team et al., 2024) with the template-based caption $\tilde{X}_k$ and the designed instructions. Then we adopt LoRA (Hu et al., 2021) instruction finetuing of the SALMONN-7B (Tang et al., 2023) backbone model for 100 epochs and the learning rate of 1e-5, with 2 NVIDIA RTX A6000 GPUs with 48GB each. We use the bfloat16 type for training with the batch size set to 4 and gradient accumulation steps to 8.

We first evaluate the generated data samples' quality by comparing them to existing human annotation datasets, MusicCaps (Agostinelli et al., 2023) and Song Describer (Manco et al., 2023). We follow the previous works (Doh et al., 2023a; Manco et al., 2023) and report the metrics, BLEU (B), METEOR (M), ROUGE (R), and BERT-score (B-S), in Table 3. Since our captions are formally different from original music captions, we report the evaluation metrics for the global and the complete captions in our dataset separately. For a fair comparison, we adopt the zero-shot performance of LP-MusicCaps in (Doh et al., 2023a), since our model is only trained on the synthetic dataset and Harmonixset.

Based on the results in Table 3 on MusicCaps, we observe that the global captions generated from our model consistently show higher quality than the zero-shot results of LP-MusicCaps, which demonstrates that by capturing more details from longer songs, we can obtain more accurate descriptions of the music. In addition, comparing FUTGA and LP-MusicCaps on Song Describer, which is the out-of-domain dataset for both methods, FUTGA shows a significantly larger improvement in the generation results, which demonstrates the model's better capacities in generalizability.

However, the complete music captions generated from FUTGA show relatively inferior performance on MusicCaps, which is mainly due to the different forms of music captions. Since FUTGA focuses on the temporal reasoning of a whole song, the time segment information and musical changes are completely new to both the original MusicCaps captions. Whereas, LP-MusicCaps is directly augmented from MusicCaps, which makes their captions formally more similar. Such observations can motivate future works to explore more fine-grained and complex music caption forms in terms of evaluating the model's generation capacities.

5 Conclusion

In this work, we propose a temporally-enhanced music caption augmentation method through generative large language models. By bootstrapping existing music captions with time boundary tags, MIR features, and musical changes, we fine-tune the pre-trained music understanding model SALMONN-7B, where we observe emerging music segmentation capacities and enable instruction prompting to guide the generation with ground-truth time segments. We use the fine-tuned model to re-annotate the existing MusicCaps and Song Describer datasets with full-length songs. The generated captions are shown to be more fine-grained and beneficial for various downstream tasks.

For future works, since our model is the first to enable end-to-end full-length song captioning with significantly longer context provided (10 times more than conventional music captions), we are motivated to further develop a long-context-based CLAP model, which can enable more complex and longer music retrieval tasks. In addition, with more fine-grained details provided by our captions, we propose to further use such captions for more complex music understanding tasks, including music question-answering and whole-song generation.

References

Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al. 2023. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*.

Samarth Bhargav, Anne Schuth, and Claudia Hauff. 2023. When the music stops: Tip-of-the-tongue retrieval for music. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2506–2510.

Ke Chen, Yusong Wu, Haohe Liu, Marianna Nezhurina, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2024. Musicldm: Enhancing novelty in text-to-music generation using beat-synchronous mixup strategies. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1206–1210. IEEE.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. 2024. Simple and controllable music generation. *Advances in Neural Information Processing Systems*, 36.

Zihao Deng, Yinghao Ma, Yudong Liu, Rongchen Guo, Ge Zhang, Wenhu Chen, Wenhao Huang, and Emmanouil Benetos. 2023. Musilingo: Bridging music and text with pre-trained language models for music captioning and query response. *arXiv preprint arXiv:2309.08730*.

SeungHeon Doh, Keunwoo Choi, Jongpil Lee, and Juhan Nam. 2023a. Lp-musiccaps: Llm-based pseudo music captioning. *arXiv preprint arXiv:2307.16372*.

SeungHeon Doh, Minz Won, Keunwoo Choi, and Juhan Nam. 2023b. Toward universal text-to-music retrieval. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Wenhao Gao, Xiaobing Li, Cong Jin, and Yun Tie. 2022. Music question answering: Cognize and perceive music. In *2022 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pages 1–6. IEEE.

Josh Gardner, Simon Durand, Daniel Stoller, and Rachel M Bittner. 2023. Llark: A multimodal foundation model for music. *arXiv preprint arXiv:2310.07160*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Atin Sakkeer Hussain, Shansong Liu, Chenshuo Sun, and Ying Shan. 2023. M Θ{2} ugen: Multi-modal music understanding and generation with the power of large language models. *arXiv preprint arXiv:2311.11255*.

Shansong Liu, Atin Sakkeer Hussain, Chenshuo Sun, and Ying Shan. 2024. 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)*, pages 286–290. IEEE.

Ilaria Manco, Benno Weck, Seungheon Doh, Minz Won, Yixiao Zhang, Dmitry Bodganov, Yusong Wu, Ke Chen, Philip Tovstogan, Emmanouil Benetos, et al. 2023. The song describer dataset: A corpus of audio captions for music-and-language evaluation. *arXiv preprint arXiv:2311.10057*.

Jan Melechovsky, Zixun Guo, Deepanway Ghosal, Navonil Majumder, Dorien Herremans, and Soujanya Poria. 2023. Mustango: Toward controllable text-to-music generation. *arXiv preprint arXiv:2311.08355*.

Zachary Novack, Julian McAuley, Taylor Berg-Kirkpatrick, and Nicholas J Bryan. 2024. Ditto: Diffusion inference-time t-optimization for music generation. *arXiv preprint arXiv:2401.12179*.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2023. Salmonn: Towards generic hearing abilities for large language models. *arXiv preprint arXiv:2310.13289*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Yuancheng Wang, Zeqian Ju, Xu Tan, Lei He, Zhizheng Wu, Jiang Bian, et al. 2023. Audit: Audio editing by following instructions with latent diffusion models. *Advances in Neural Information Processing Systems*, 36:71340–71357.

Shangda Wu, Dingyao Yu, Xu Tan, and Maosong Sun. 2023. Clamp: Contrastive language-music pre-training for cross-modal symbolic music information retrieval. *arXiv preprint arXiv:2304.11029*.

Yixiao Zhang, Yukara Ikemiya, Gus Xia, Naoki Murata, Marco Martínez, Wei-Hsiang Liao, Yuki Mitsufuji, and Simon Dixon. 2024. Musicmagus: Zero-shot text-to-music editing via diffusion models. *arXiv preprint arXiv:2402.06178*.