MUSICSEM: A DATASET OF MUSIC DESCRIPTIONS ON REDDIT CAPTURING MUSICAL SEMANTICS

Rebecca Salganik[1], **Teng Tu**[2], **Fei-Yueh Chen**[1], **Xiaohao Liu**[2], **Kaifeng Lu**[1], **Ethan Luvisia**[1], **Zhiyao Duan**[1], **Guillaume Salha-Galvan**[3], **Anson Kahng**[1], **Yunshan Ma**[4], **Jian Kang**[1]
[1]University of Rochester [2]National University of Singapore
[3]Kibo Ryoku [4]Singapore Management University
rsalgani@ur.rochester.edu

ABSTRACT

We present MusicSem, a dataset of 32,493 language–audio music descriptions derived from organic discussions on Reddit. What sets MusicSem apart is its focus on capturing a broad spectrum of musical semantics, reflecting how listeners naturally describe music in nuanced, human-centered ways. To structure these expressions, we propose a taxonomy of five semantic categories: descriptive, atmospheric, situational, metadata-related, and contextual. Our motivation for releasing MusicSem stems from the observation that music representation learning models often lack sensitivity to these semantic dimensions, due to the limited expressiveness of existing training datasets. MusicSem addresses this gap by serving as a novel semantics-aware resource for training and evaluating models on tasks such as cross-modal music generation and retrieval.

1. INTRODUCTION

Music representation learning is central to music information retrieval and generation [1, 2]. While prior work has primarily focused on audio-centric models [3–6], recent advances in multimodal learning, particularly in aligning text and audio, have enabled progress in tasks such as cross-modal retrieval [7–9], music-to-text generation [10–12], and text-to-music generation [13–16]. However, recent work has shown that multimodal models often fail to capture the user's expressed intent in text descriptions of music [17, 18]. This interpretation gap suggests that the language-audio datasets used to train these models may not fully reflect the broader and more natural forms of human discourse.

In this paper, we begin by formalizing the notion of musical semantics and introducing a taxonomy that distinguishes five types of music captions. We then confirm that many state-of-the-art generative and retrieval models lack sensitivity to these semantic distinctions, particularly variations in atmosphere, context, situational cues, and metadata-related aspects of user intent. Motivated by this observation, we introduce MusicSem, a semantically rich language-audio dataset derived from organic music discussions on the social media platform Reddit. The dataset comprises 32,493 language-audio music description pairs, with textual annotations that express not only descriptive attributes of the music, but also emotional resonance, contextual and situational usage, and co-listening patterns. MusicSem distinguishes itself by capturing a broader spectrum of musical semantics than prior datasets used for multimodal model training. As demonstrated in an extended version of this work, under review at the time of writing, MusicSem also serves as a novel semantics-aware resource for benchmarking cross-modal retrieval and generation models. The accompanying MusicSem website provides access to the full dataset, detailed documentation, and source code for data construction and experiments at: `https://music-sem-web.vercel.app/`.

2. CAPTURING MUSICAL SEMANTICS

Capturing the nuances that contextualize a listening experience in textual descriptions is crucial for multimodal music understanding. Consider, for example, a text-to-music generation or retrieval model and the following two prompts: *"This song is a ballad. It contains guitar, male vocals, and a piano. It sounds like something I would listen to at church"* versus *"This song is a ballad. It contains guitar, male vocals, and a piano. It sounds like something I would listen to while tripping on acid."* While both descriptions specify identical musical attributes, the situational context drastically shifts our expectations of the corresponding audio.

In Table 1, we organize these contextual elements into five categories, which we term *musical semantics* [19–21]. To quantify a model's sensitivity to variations in semantics, we conduct the following experiment. Given a language-audio pair (t_i, a_i) from a dataset, we construct a counterfactual annotation $\tilde{t}_i^x$ by modifying the text according to a semantic category x, e.g., *"while at church"* versus *"while tripping on acid"*. We sample 50 pairs from MusicCaps [13] and generate counterfactuals for each semantic category present in the captions (data available in our codebase). We consider two metrics to assess sensitivity to

Late-Breaking / Demo Session Extended Abstract, ISMIR 2025 Conference

Table 1. Categorization of different caption elements.

Category	Description	Example
Descriptive	concrete musical attributes	"I like the high pass filter on the vocals in the chorus, really makes harmonies pop."
Contextual	other songs	"Sabrina Carpenter's *Espresso* is just a mix of old Ariana Grande and 2018 Dua Lipa."
Situational	an activity or environment	"I listened to this song on the way to quitting my sh**ty corporate job."
Atmospheric	emotions and expressive adjectives	"This song makes me feel like a manic pixie dream girl in a bougie coffeeshop."
Metadata-related	technical & background information	"This deluxe edition of this song was released in 2013 and it has three bonus hiphop tracks."

Table 2. Semantic sensitivity analysis of generative (top) and retrieval (bottom) models. Best performance is in bold. Superscripts ^{d}, ^{a}, ^{s}, ^{m}, and ^{c} denote descriptive, atmospheric, situational, metadata, and contextual, respectively.

Generative Model	G^d	G^a	G^s	G^m	G^c
AudioLDM2 [16]	0.68	0.37	0.35	0.40	0.34
MusicLM [13]	0.50	0.36	0.42	0.39	0.35
Mustango [22]	0.62	0.27	0.25	0.26	0.32
MusicGen [14]	0.57	0.47	0.39	0.47	0.52
Stable Audio [15]	**0.72**	**0.67**	**0.68**	**0.70**	**0.74**
Retrieval Model ($k = 10$)	R^d	R^a	R^s	R^m	R^c
LARP [23]	**0.98**	0.17	0.06	0.0	**0.56**
CLAP [7]	0.95	0.52	0.35	0.42	0.52
ImageBind [9]	0.84	0.39	0.35	0.38	0.41
CLaMP3 [8]	0.92	**0.58**	**0.49**	**0.62**	0.55

semantic shifts. For text-to-music generation, we define $G^x = \frac{1}{n}\sum_{i=1}^{n}[1 - \text{cosine}(f_i, \tilde{f}_i^x)]$, where n is the number of language-audio pairs, $f_i = \mathcal{M}(t_i)$ and $\tilde{f}_i^x = \mathcal{M}(\tilde{t}_i^x)$ are the outputs of the model $\mathcal{M}$. For text-to-music retrieval, we define $R^x@k = \frac{1}{n}\sum_{i=1}^{n}[1 - \frac{|A_i \cap \tilde{A}_i^x|}{|A_i|}]$, where $A_i = \mathcal{M}(t_i)$ and $\tilde{A}_i^x = \mathcal{M}(\tilde{t}_i^x)$ are the top-k retrieved audio candidates.

Results in Table 2 show that many state-of-the-art models exhibit substantially greater sensitivity to changes in descriptive attributes than to shifts in atmospheric, situational, contextual, or metadata-related information. This observation confirms the relative lack of semantic awareness in their textual conditioning and highlights their limited ability to capture the expectations implied by user intent.

3. THE MUSICSEM DATASET

To address this lack of semantic sensibility, we introduce MusicSem, a novel dataset of language–audio music description pairs extracted from five English-language Reddit threads featuring detailed user discussions across diverse genres: r/electronicmusic, r/popheads, r/progrockmusic, r/musicsuggestions, and r/LetsTalkMusic. The dataset aims to capture more nuanced musical semantics to support the training and evaluation of multimodal models in future work. Its construction involved substantial effort to identify, extract, structure, and validate semantic content from online discourse, combining LLM-assisted extraction with human annotation and verification. A comprehensive description of this process is provided in our extended paper and illustrated in the Demo section of our website.

The released dataset comprises 32,493 entries, each including a Spotify ID and URL for audio retrieval, the source thread, raw text, song and artist names, and semantics structured according to the taxonomy in Table 1. We also constructed an unpublished test set of 480 entries for future leaderboard use on our website. Table 3 shows the proportion of entries containing each of the five semantic categories in MusicSem and two canonical language-audio datasets. MusicSem consistently demonstrates broader coverage across all categories, highlighting its semantic richness. It also exhibits a richer vocabulary, with a higher count of unique words and music genres.

Table 3. Statistics (top) and semantic diversity (bottom) of MusicSem and two other language-audio music datasets.

Statistics	**MusicCaps** [13]	**Song Describer** [24]	**MusicSem (ours)**
# Entries	5,521	1,100	32,493
# Vocab. Words	6,245	2,824	22,738
# Music Genres	267	152	493
Category	**MusicCaps** [13]	**Song Describer** [24]	**MusicSem (ours)**
Descriptive	100%	94%	100%
Contextual	6%	8%	77%
Situational	41%	16%	48%
Atmospheric	57%	33%	64%
Metadata	28%	6%	64%

4. CONCLUSION AND FUTURE WORK

In conclusion, MusicSem reflects how listeners naturally describe music in nuanced and contextualized ways on Reddit. By releasing this novel dataset, we aim to foster the development of semantics-aware multimodal music representation learning models. The website and extended version of this work complement this two-page paper by providing additional comparisons with existing language-audio datasets, detailed documentation of the Reddit thread selection and dataset construction processes, discussion of our proposed musical semantics taxonomy, and reflections on cultural representativeness (e.g., the dataset reflects English-speaking Reddit users, whose discussions may skew toward Western, more opinionated, and niche community perspectives than those of the general population [25, 26]). They also demonstrate how MusicSem can serve as a semantics-aware evaluation resource, reporting comprehensive benchmark analyses of state-of-the-art models on three key tasks: (1) cross-modal music retrieval, (2) text-to-music generation, and (3) music-to-text generation. In future work, we plan to expand the scale and scope of MusicSem by incorporating additional Reddit threads related to music, as well as conversations about lyrics and symbolic representations. We also aim to extend our benchmarking efforts to new tasks such as controllable music generation [27] and text-guided music recommendation [28] using MusicSem.

5. REFERENCES

[1] M. Müller, *Fundamentals of Music Processing: Audio, Analysis, Algorithms, Applications*. Springer, 2015.

[2] M. Schedl, E. Gómez, and J. Urbano, "Music Information Retrieval: Recent Developments and Applications," *Foundations and Trends® in Information Retrieval*, vol. 8, no. 2-3, pp. 127–261, 2014.

[3] Y. Lin, Y. Yang, and H. H. Chen, "Exploiting Online Music Tags for Music Emotion Classification," *ACM Trans. Multim. Comp. Com. App.*, vol. 7, p. 26, 2011.

[4] D. Bogdanov, M. Won, P. Tovstogan *et al.*, "The MTG-Jamendo Dataset for Automatic Music Tagging," in *ICML ML for Music Discovery Workshop*, 2019.

[5] S. Oramas, O. Nieto, F. Barbieri *et al.*, "Multi-Label Music Genre Classification from Audio, Text and Images Using Deep Features," in *ISMIR*, 2017, pp. 23–30.

[6] R. Yuan, Y. Ma, Y. Li *et al.*, "MARBLE: Music Audio Representation Benchmark for Universal Evaluation," in *NeurIPS*, 2023, pp. 39 626–39 647.

[7] Y. Wu, K. Chen, T. Zhang *et al.*, "Large-Scale Contrastive Language-Audio Pretraining with Feature Fusion and Keyword-to-Caption Augmentation," in *ICASSP*, 2023, pp. 1–5.

[8] S. Wu, Z. Guo, R. Yuan *et al.*, "Clamp 3: Universal music information retrieval across unaligned modalities and unseen languages," *arXiv preprint arXiv:2502.10362*, 2025.

[9] R. Girdhar, A. El-Nouby, Z. Liu *et al.*, "ImageBind: One Embedding Space To Bind Them All," in *CVPR*, 2023, pp. 15 180–15 190.

[10] 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, pp. 286–290.

[11] S. Doh, K. Choi, J. Lee *et al.*, "LP-MusicCaps: LLM-Based Pseudo Music Captioning," in *ISMIR*, 2023, pp. 409–416.

[12] J. Wu, Z. Novack, A. Namburi *et al.*, "FUTGA: Towards Fine-grained Music Understanding through Temporally-enhanced Generative Augmentation," in *NLP4MusA*, 2024, pp. 107–111.

[13] A. Agostinelli, T. I. Denk, Z. Borsos *et al.*, "MusicLM: Generating Music From Text," *arXiv preprint arXiv:2301.11325*, 2023.

[14] J. Copet, F. Kreuk, I. Gat *et al.*, "Simple and Controllable Music Generation," in *NeurIPS*, 2023, pp. 47 704–47 720.

[15] Z. Evans, C. Carr, and J. a. Taylor, "Fast Timing-Conditioned Latent Audio Diffusion," in *ICML*, 2024, pp. 12 652–12 665.

[16] H. Liu, Y. Yuan, X. Liu *et al.*, "AudioLDM 2: Learning Holistic Audio Generation With Self-Supervised Pretraining," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 32, pp. 2871–2883, 2024.

[17] Y. Zang and Y. Zhang, "The Interpretation Gap in Text-to-Music Generation Models," *arXiv preprint arXiv:2407.10328*, 2024.

[18] F. Ronchini, L. Comanducci, G. Perego, and F. Antonacci, "PAGURI: a user experience study of creative interaction with text-to-music models," *arXiv preprint arXiv:2407.04333*, 2024.

[19] M. Levy and M. Sandler, "Learning latent semantic models for music from social tags," *Journal of New Music Research*, vol. 37, no. 2, pp. 137–150, 2008.

[20] J. Nam, K. Choi, J. Lee *et al.*, "Deep Learning for Audio-Based Music Classification and Tagging: Teaching Computers to Distinguish Rock from ," *IEEE Signal Process. Mag.*, vol. 36, no. 1, pp. 41–51, 2019.

[21] J. Choi, A. Khlif, and E. Epure, "Prediction of User Listening Contexts for Music Playlists," in *NLP4MusA*, 2020, pp. 23–27.

[22] J. Melechovsky, Z. Guo, D. Ghosal *et al.*, "Mustango: Toward Controllable Text-to-Music Generation," in *NAACL*, 2024, pp. 8286–8309.

[23] R. Salganik, X. Liu, Y. Ma *et al.*, "LARP: Language Audio Relational Pre-training for Cold-Start Playlist Continuation," in *KDD*, 2024, pp. 2524–2535.

[24] I. Manco, B. Weck, S. Doh *et al.*, "The Song Describer Dataset: A Corpus of Audio Captions for Music-and-Language Evaluation," *arXiv preprint arXiv:2311.10057*, 2023.

[25] A. N. Medvedev, R. Lambiotte, and J.-C. Delvenne, "The Anatomy of Reddit: An Overview of Academic Research," *Dynamics on and of Complex Networks*, pp. 183–204, 2017.

[26] E. Epure and R. Hennequin, "A Human Subject Study of Named Entity Recognition in Conversational Music Recommendation Queries," in *EACL*, 2023, pp. 1281–1296.

[27] L. Lin, G. Xia, Y. Zhang *et al.*, "Arrange, Inpaint, and Refine: Steerable Long-term Music Audio Generation and Editing via Content-based Controls," in *IJCAI*, 2024, pp. 7690–7698.

[28] M. Delcluze, A. Khoury, C. Vast *et al.*, "Text2Playlist: Generating Personalized Playlists from Text on Deezer," in *ECIR*, 2025, pp. 164–170.