ClimaCS: A Multi-Genre Dataset of Timed Comments for Music Highlight Detection

Loïs Guerci
IRCAM & Sorbonne Université
1 Place Igor Stravinsky
Paris, 75004
lois.guerci@free.fr

Laure Prétet
Bridge.audio
Paris, France
laure.pretet@bridge.audio

Abstract

Recent advances in deep classifier models achieved a remarkable audience experience not only for describing music but also for generating creative outcomes. Motivated by perspectives in automatic music and video synchronization, we consider the challenge of identifying the most salient moments in a music track. To train a model on this task, an appropriate dataset becomes essential. We introduce ClimaCS, an timed comment dataset designed for training models for highlight detection. We describe our methodology to bring together this set, which features over 2,000 tracks from a wide range of music genres and additional metadata. We expose the main characteristics of the content of the dataset and suggest an initial benchmark experiment. Using a simple CNN-based architecture, we demonstrate that highlights can be predicted using exclusively timed comment for training, without expert manual annotation of the highlights.

1 INTRODUCTION

In audiovisual content, pairing the perfect song with the perfect video is a real challenge, but when done correctly, it opens up exciting possibilities in areas like advertising or music videos. Synchronization is an essential source of revenue for emerging music artists, although it requires editing and self-advertising skills. Using MIR technologies, music-video pairing can be achieved by semantically matching the content (Stewart et al., 2025), but previous studies have shown that the matching of key moments is equally important (Prétet et al., 2021). Thanks to advances in deep learning for classification, it is becoming possible to synchronize both media by aligning their most impactful moments, creating an automatic and dynamic edit. These high points, often referred to as **climaxes** or **highlights**, are crucial to detect, as they represent the most engaging and powerful moments in the media.

In addition to music-video synchronization, music highlights detection has applications in music genre classification (Lee et al., 2014), video-to-music recommendation (Sarakit, 2015), music summarization (Sul et al., 2023), as well as music retrieval and captioning (Stewart et al., 2025).

However, music climaxes are subjective by nature and are not always identifiable directly from harmony or rhythm (Patty, 2009). Data-driven methods such as deep learning are worth exploring for this task, but require large amounts of data, which are difficult to collect manually. To address this, we introduce **ClimaCS**, a CLIMAx estimation Comments Set from social data on music tracks. The dataset consists of pairs comprising released tracks from the SoundCloud streaming platform and related user comments, collected "in the wild", and associated to their temporal localization within the track.

Through this paper and the ClimaCS set, our aim is to provide resources to the research community that works on the identification of emotionally or sonically intense moments. Indeed, we demonstrate that the characteristics of a large number of user comments provide a proxy for highlight detection in a diversity of music genres. The **three main contributions** of our work are the following:

- ClimaCS, the first dataset of music / timed comments pairs on a variety of music genres, and the largest to our knowledge (over 2,000 tracks and 1 million comments). We make it publicly available online [1] by organizing it with track metadata, and the complete list of timed comments associated to each track.
- Analysis of the data and presentation of one possible way of using ClimaCS for music highlight detection,
- Formalization of the music highlight detection task on a reproducible benchmark introduced by Yadati et al. (2018).

We organize our study as follows. In Section 2, we review existing datasets related to ours. In Section 3, we present our proposed ClimaCS set, from our data collection strategy to its content with statistics. In Section 4, we present one highlight detection strategy to illustrate the purpose and usefulness of ClimaCS. Then, in Section 5, we summarize our work and suggest ways to use ClimaCS for future research.

2 RELATED WORK

In this section, we review all existing datasets and tools previously published for highlight detection tasks, as context for our work. This review is divided into two categories: general media and music focused.

2.1 HIGHLIGHT DETECTION AND ASSOCIATED DATASETS IN GENERAL MEDIA

Over the years, numerous datasets have aimed to capture and describe aspects such as energy, peak moments, or remarkable events in multimedia content. Earlier approaches

[1] https://zenodo.org/records/21933494

focused on domain-specific datasets. For instance, Xiong et al. (2004) compiled a private collection of 814 audio clips from sports broadcasts (golf, baseball, soccer), annotated with sound event classes such as applause, cheering, and speech. These were used with Gaussian Mixture Models for highlight extraction, though the dataset remains private and is not music-focused. Later, Yeh et al. (2016) leveraged low-level audio-visual features (e.g., motion, sound energy, shot transitions) to detect highlight scenes in action films, showing the continued relevance of handcrafted features in earlier methodologies.

Several other datasets contributed to the study of highlight detection in video. For example, Sun et al. (2014) put together a dataset of YouTube videos with highlight annotations sourced via Amazon Mechanical Turk. Similarly, Xu and Larson (2014) released a dataset rich in timed comments, though lacking climax annotations. The TVSum dataset (Song et al., 2015), focused on video summarization, includes importance scores derived from crowdsourced annotations, though it lacks audio data. The VTW dataset (Video Titles in the Wild, Zeng et al. (2016)) offers manually labeled highlights from user-generated videos, originally developed for automatic title generation tasks.

In video analysis, the Danmus dataset (Lv et al., 2022) stands out with 7.9 million danmus and 4.8 million video frames across eight different categories. Danmus are real-time user comments that appear on-screen and are synchronized with specific video moments, offering valuable contextual insight. Although relevant, this dataset is not publicly accessible. Building upon this idea, the TSC-MDFFM model (Time-Sync Comment Enhanced Multimodal Deep Feature Fusion Model) (Yin et al., 2025) leverages similar time-aligned comments to precisely capture user engagement. Its evaluation combines self-constructed and public datasets.

Highlight detection from multimodal data is also explored by Mundnich et al. (2021), who fuse audio, visual, and scene dynamics. Using relevance scores designed for summarization annotation, they show that this knowledge can be transferred to the highlight detection task. This demonstrates how diverse the methods can be for collecting highlight annotations. In contrast, the Video Highlight Detection Dataset from Lei et al. (2021) comprises 10,000 YouTube videos annotated directly with saliency scores via Mechanical Turk. More recently, the Mr. HiSum dataset (Sul et al., 2023) has been proposed for large-scale highlight summarization, using YouTube's "Most replayed" statistics to aggregate highlight labels.

2.2 MUSIC HIGHLIGHT DETECTION AND ASSOCIATED DATASETS

In the domain of music highlight detection, several approaches have explored the use of audio features and crowdsourced data. It is worth mentioning that the variations of the audio signal's energy are often considered a first element towards structure analysis. Schlüter and Böck (2014) proposed a CNN-based method for precise musical onset detection using spectrograms as input, while Lee et al. (2014) developed a music emotion classifier that incorporates energy-based highlight detection.

A notable direction involves leveraging timed comments from platforms like SoundCloud: Yadati et al. (2014) curated a dataset of 100 EDM tracks with timed comments to train a SVM to detect *drops*, although the dataset was not released and was genre-specific. This work was later expanded by increasing the dataset size to 500 tracks (Yadati et al., 2018). The event labels were made public, though the comments themselves were not released, and the genre limited again to EDM. With a total of 1950 (640 drops, 760 builds and 550 breaks) events for 500 tracks, the authors expect an average of 4 highlights per track. They conclude that timed comments can be reliable training labels for identifying socially significant musical events.

Other approaches include model-based climax detection without machine learning by Sato et al. (2016). Then, CRAN (Ha et al., 2017) an unsupervised neural architecture combines convolutional, recurrent, and attention layers to extract genre-discriminative music highlights from mel-spectrograms. Wang et al. (2021) later introduced a supervised, end-to-end chorus detector trained on an in-house dataset of 2,480 annotated tracks spanning multiple genres. In this scenario, the highlights are linked to choruses in Pop music, which indicates that several highlights per track are expected. More recently, Tseng et al. (2023) introduced a small dataset for studying music memorability. Finally, it should be noted that timed comment datasets have also found use beyond highlight detection, including applications in audio captioning (Zhang et al., 2022) and emotion recognition (Sarakit, 2015; Yamamoto and Nakamura, 2013).

3 THE CLIMACS DATASET

In this section we present the ClimaCS set, from the data collection process to all relevant details on its content.

3.1 Data collection

Playlist collection. SoundCloud[2] is an online platform where users can upload their music, share it, connect with their favorite artists and critics, and leave comments on the tracks they listen to. Interestingly, the SoundCloud service registers time codes associated with each comment, which allows us to infer to what part of the track the comment refers to. Other platforms such as YouTube provide the possibility of adding time codes to one's comments, but contrary to SoundCloud, it is not systematic, and therefore requires a lot more web scraping to collect enough timed comment. Based on the work of Yadati et al. (2018), we use the SoundCloud API to obtain our tracklist and the associated time-coded comments. We start from the homepage, where editorial playlists can be found, and we target the most listened-to playlists (charts) across the top music genres on the platform (e.g. R&B, Indie, Electronic). We then use the search feature to complete the selection with user-generated playlists from underrepresented genres (e.g. KPop, Bollywood). In total, the dataset includes 48 playlists (mostly based on genres) of 1 to 158 tracks.

Track metadata collection. At this stage, we use browser developer tools on each playlist page to interact with SoundCloud's API via HTTP requests. This allows us to collect the ID of the tracks included in each playlist, and extract their metadata. The output of this process is one `json` file per track, that lists all related social data, namely title, artist name and ID, comment count, creation date, lyrics, download count, duration, genre (specified by the artist at upload) and other user-generated tags.

[2]https://soundcloud.com/

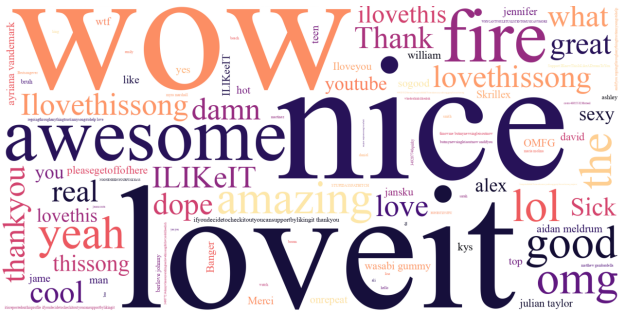

Figure 1: Visual representation of the most used words in the ClimaCS set, after removing common stopwords ("user", "it", "https", ...) and all user names. The size of the word depends on its rank in the frequency ranking, but is not directly proportional to its frequency.

Timed comments collection. We use the API again to collect all listener comments on each track. We obtain one `json` file per track, that lists every single comment along with its time code, content, date of creation, and user details.

Cleaning / curation. We filter out invalid pairs, i.e.:

- tracks which are too long (over 10,000 seconds) or too short (under 30s),
- tracks that have no comments,
- comments which time code happens before 10 seconds (often generic to the track and not specific to a moment of the track), or after the end of track, and we remove the tracks which have no comment left after these operations.

We do not conduct any more curation of the tracks or comments, hence the data can be considered "in the wild". Our final dataset consists in 2,033 pairs of unique tracks and their comments. Fig. 1 illustrates the most frequently used words in the dataset.

3.2 Data analysis

In this subsection, we present several statistics to provide an overview of the dataset, as summarized in Table 1. The median duration of the tracks is 200.4s, which is about 3 minutes and 12 seconds. The median number of comments per track is 24, with some tracks reaching over 25,000 comments. On average, there are 148.6 comments for each minute of audio. To ensure that the number of comments was sufficient, the track selection of the ClimaCS is biased toward high-popularity tracks (no track was played less than 162 times); however, we can still observe a large variation of the number of likes across the tracks, from 7 to over one million.

Fig. 2 shows the distribution of tracks and comments across the different genres represented in ClimaCS. SoundCloud is famous for its hip-hop, rap, and overall contemporary content, as a platform geared towards artists uploading their music themselves. The distribution of genre reflects this, with a predominance of KPop, Electro (French Touch) and Hip-Hop (Phonk), and a smaller proportion of Jazz and Rai tracks. In between, most genres are represented by around 50 samples, a characteristic inherited from the *charts* playlists. Dubstep gathers the highest number of comments (over 100k), followed by R&B, Pop, and Trap. The least represented genre, Rai, still accounts for over a hundred comments.

Table 1: Dataset Statistics Summary.

Metric	Value
Nb. of pairs	2,033
Nb. of comments	1,132,432
Track duration (s)	Avg: 381.5, Min: 46.8, Max: 8,710.9, Med: 200.4
Comments / track	Avg: 557.0, Min: 1, Max: 29,644, Med: 24.0
Comments / min	Avg: 148.6, Min: 0.03, Max: 8,884.9, Med: 8.1
Track notoriety	Likes: [7–1,486,470], Plays: [162–103,445,360]

4 EXPERIMENTS AND RESULTS

In this section, we demonstrate the purpose of the ClimaCS set, by presenting a music highlight detection experiment.

4.1 Proposed definition of the task

Given the subjective nature of music climaxes, detecting them is a complex and ill-defined task. As seen in section 2, previous studies often address a simplified problem by strongly restricting the music genre and definition of a highlight. Similarly, in this benchmark study, we frame our highlight estimation task as binary classification on audio segments. We use fixed-length segments of duration d.

Model input. We use the Constant-Q transform (CQT) as our time-frequency input representation. We compute the CQTs of the entire dataset using of 96 bins (12 bins/octave), and a hop size of 1024 at 44.1 kHz (23ms). We convert them to power amplitude with log-scaled magnitudes. We fix $d = 11.88s$ which corresponds to 512 CQT frames. This duration was selected as a compromise between temporal accuracy of the event detection and limitation of the temporal noise of the time codes (Yadati et al., 2018). We obtain a total of 37,475 valid segments for the entire dataset.

Model targets. Since we use user-generated comments as a proxy for listener reaction, we generate training labels from the time coded comments that are present within each segment. For each segment i, we collect C_i, the list of all comments present within its temporal boundaries. To generate a label y_i for highlight detection, based on the characteristics of C_i, we use rely on the following strategy, which we will refer to as **ComsIntensity**.

We use OpenAI's API and instruct the model `gpt-4o` to return $y_i = 1$ if according to its analysis, C_i represents "a highlight (an engaging, impactful and powerful moment in the music)" and $y_i = 0$ otherwise. To better measure user enthusiasm in segment i, we instruct the LLM to take into consideration the emotional profile of C_i. We are aware that Soundcloud users often write their general opinion on the track, for which the timing of the comment has no validity. Hence, we describe in the LLM's prompt a list of characteristics to pay special attention to: the number of comments, the punctuation (including emojis), upper case characters, strong positive words and reaction keywords (such as "right here", "this moment", "now"), and shortness of comments (that indicates pure reaction, as opposed to longer reflections).

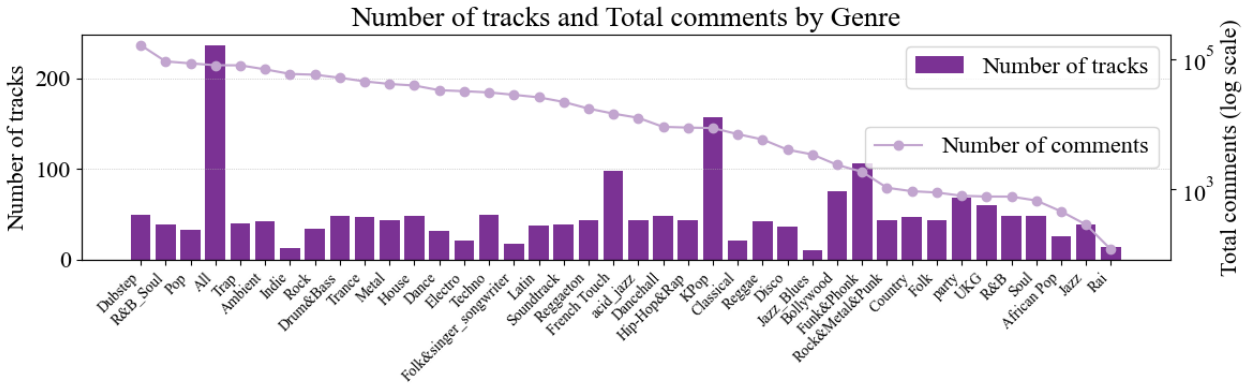

Figure 2: Distribution of the number tracks (bars) and comments (line plot, log scale) across all music genres present in the ClimaCS dataset.

As a result, we obtain 9,001 highlight segments, an average of 4 per track, similar to the ratio found in Yadati et al. (2018). The median number of climaxes per track is 2. We then add an equal amount of randomly selected non-highlight segments to create a balanced set.

We then split it into training (80%) and validation (20%) parts. We illustrate our annotations on an example track in Fig. 3.

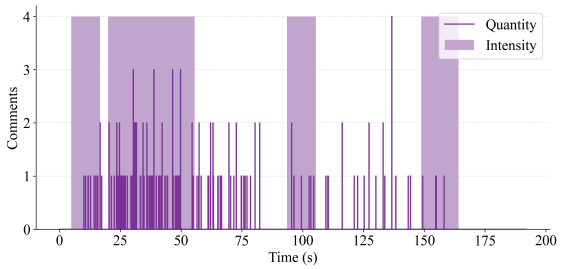

Figure 3: An example of annotations from the training dataset on track *Manu Chao - Me Gustas Tu (JKRS Techno Remix)*. In dark, thin bars, we display the evolution of the number of comments over time. In light, thick bars, we display the segments where the LLM used for **ComsIntensity** indicates a highlight moment. We observe that these moments often coincide with a high number of comments, but not all the time, a hint at the fact that semantics matter, too.

4.2 Benchmark model

To assess the interest of ClimaCS, we train a Fully Convolutional Neural Network with 4 layers (FCN-4) proposed by Choi et al. (2016) for each experiment. We adapt the pooling size to our input ($(96, 512)$, representing the frequency bins and the time frames, respectively) and output shape (1 unit representing the likelihood of a highlight in the segment). We also scale down the filter number by a factor of 2 due to the relative small size of the dataset. Our model, implemented in TensorFlow (Keras), has 6,197,249 trainable parameters and is trained with a Binary Cross-Entropy loss. We use the Adam optimizer with a learning rate of 10^{-7} and early stopping (patience of 50 epochs without improvement).

4.3 Evaluation and results

To evaluate the models, we use the full dataset proposed by Yadati et al. (2018). This dataset consists of 500 tracks from various electronic genres, manually annotated with socially significant event labels such as *Drops*, *Builds*, and *Breaks*. While *Drops* are punctual events, *Builds* and *Breaks* are provided with a start and end date, and can therefore span across multiple segments. With this logic, we divide the full dataset in 9,818 segments, among which 2,873 are labeled as highlights.

We measure the performance of the model using F-score and AUC. We obtain a F-score of 0.450 and an AUC of 0.563. These results are modest, but better than the random baseline ($F_1 = 0.368$, $AUC = 0.500$), confirming that timed comments contain some exploitable information for detecting musical climaxes. Importantly, despite being trained on a more diverse dataset (not just EDM), our models reach performance levels comparable to those reported by Yadati et al. (2018) for their most similar experiments (avg F_1 in the range 0.42–0.47). This highlights the transferability of ClimaCS to other highlight detection contexts, even when annotation sources and training material differ.

In Fig. 4, we present three examples of predictions from the **ComsIntensity** experiment, compared with the annotations provided by Yadati et al. (2018). For each track, we selected one segment that we consider particularly relevant for climax detection. Across these tracks, we observe that the model's predictions often align with salient events in the audio signals. By listening to these excerpts, we can further confirm the relevance of our dataset. On track #72141629, the predicted climax appears to be strongly correlated with the waveform intensity. On track #90268834, around 320 seconds, the break annotation at the end of the motif is reflected by a rise in the prediction curve. Finally, on track #56116804, subjective listening suggests that the build-up begins significantly earlier than the annotations provided by Yadati et al. (2018), and this earlier onset is well captured by our prediction curve. Other listening tests by the authors reveal that the model often flags major changes in the music as highlights. In the context of music-video synchronization (our initial motivation), this is sensible and useful.

4.4 Modeling listeners' anticipation

Another structural analysis direction can be explored with ClimaCS. In this section, we model listeners' anticipation

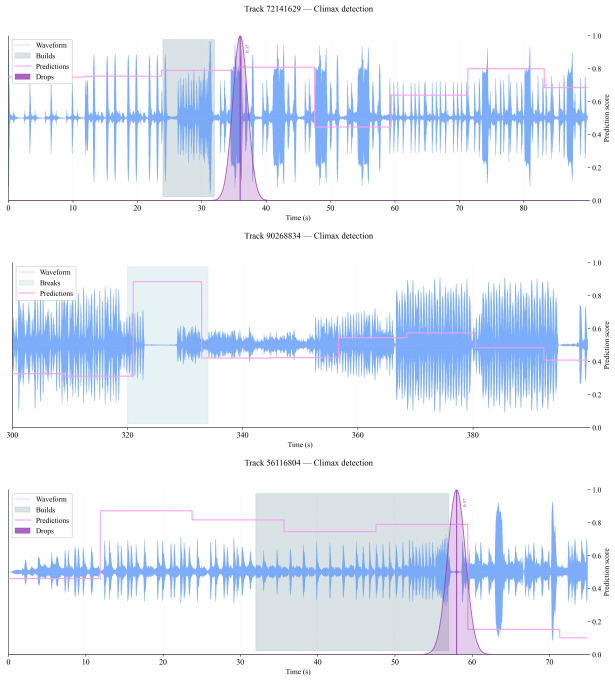

Figure 4: Visualization of the model's predictions across time using strategy **ComsIntensity** on segments of tracks #72141629, #90268834 and #56116804. We display the manually annotated *Drops* as vertical purple lines surrounded by gaussian curve for visibility, and *Builds* and *Breaks* as green step functions. The model's predictions (highlight likelihood) for each segment is represented in pink.

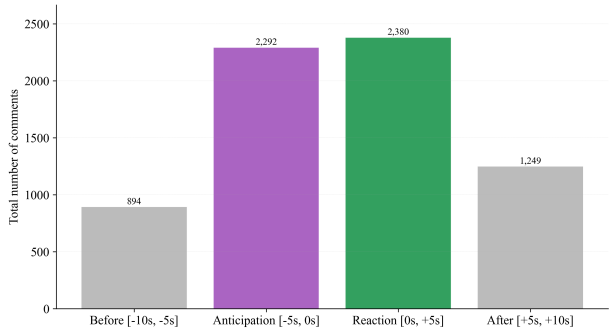

Figure 5: Modeling listeners' anticipation. For each temporal region, the total number of comments is reported. The *Anticipation* region contains more comments than the *Before* and *After* regions.

of climactic moments. Our intuition is that highlight moments in tracks trigger reactions not only at the time they occur, but also before, in anticipation. To investigate this hypothesis, we first select all tracks containing more than 100 comments in order to reduce statistical bias, resulting in a subset of 554 tracks. We then define four temporal regions around the peak of comment activity for each track: *Before* (between -10 and -5 seconds before the climax), *Anticipation* (-5 to 0 seconds before the climax), *Reaction* (0 to 5 seconds after the climax), and *After* (5 to 10 seconds after the climax). For each region, we compute the total number of comments.

As shown in Fig. 5, the number of comments increases as the climax approaches and decreases afterward, supporting the hypothesis that listeners exhibit anticipatory behavior prior to climactic musical events. This experiment provides deeper insight into how musical tension and expectation are formed over time.

5 CONCLUSION AND PERSPECTIVES

Through this paper, we introduced and demonstrated the utility of ClimaCS, the largest dataset of multi-genre music tracks and the corresponding user comments (text and evolution of the activity across time). As shown in our benchmark experiment, a model trained on ClimaCS achieved encouraging highlight detection results on the datasets proposed by Yadati et al. (2018), which underlines their robustness and transferability, despite being trained solely with user comments from SoundCloud.

However, given the underexplored nature of this task, our experiment presents several limitations. First, building a robust training set remains difficult due to the subjective criteria used to identify climaxes from comments. The amount of irrelevant information contained in the comments, as well as the amount of *temporal noise* (information related to a specific part of the music, but which timestamp corresponds to another), can lead to unreliable training labels. Other methods of creating training labels have been explored in preliminary studies. For example, we tried using the density of comments across time or retaining only comments that include climax-related keywords, but without success. Furthermore, the large number of comments prevents exhaustive human verification and analysis, and the relative small size of the dataset can lead to overfitting. Lastly, identifying an optimal test set for evaluating the benchmarks remains challenging, as no other publicly available evaluation data were found to assess the relevance of our dataset. Paradoxically, this lack of existing evaluation data further highlights the motivation for introducing our dataset. Consequently, we encourage our colleagues to use the ClimaCS set in upcoming research, especially on issues related to highlight detection, music-video synchronization, musical structure, and musicology.

We have already shown in section 4.4 that listeners tend to anticipate music highlights. Another avenue would be to investigate whether climax perception is universal or culturally dependent. A salient moment in a track may be interpreted differently depending on the listener's cultural background or nationality (this information can be inferred from the language of the comments). Similarly, climax perception may manifest differently according to the music genre. Finally, the diffusion of musical information through online interactions represents another promising topic. On SoundCloud, users can reply to time-stamped comments, which may contribute to renewed attention toward a track or even to its revival on other platforms (*short/reels* video medias, where highlight moments are preferred during editing). As the timestamp feature on YouTube is also popular, future work can investigate highlights related to the most salient moment on music videos, but will face the risk of estimating video highlights instead of music highlights. Such data could be used to construct propagation graphs and to study how emotional responses are transmitted through comment interactions.

6 ETHICS STATEMENT

As musicians, our work seeks to support the artistic community in their creative process. However, we acknowledge that our approach may raise ethical concerns, particularly regarding the collect of audio tracks. In order to gather the maximum number of comments per track, we focused on highly popular tracks, which are typically not copyright-free, although they are publicly available. Therefore, ClimaCS is made available without the audio files. We do not claim ownership of the songs included in this dataset, as the rights to the original audio content remain with their respective copyright holders. We believe that our work uses this data under fair use for research, to foster collaboration, creativity, and innovation. We are aware that further use of the data may require permission from the artists, who can be contacted via SoundCloud's internal messaging. We are also aware that our work may contribute indirectly to *cultural predictions* technologies, by influencing the way music is recommended and promoted on social media. We claim that our dataset is intended for research purposes and should not be used for commercial or political purposes.

REFERENCES

Choi, K., Fazekas, G., and Sandler, M. (2016). Automatic tagging using deep convolutional neural networks. In *Proceedings. of ISMIR (International Society for Music Information Retrieval Conference)*, Suzhou, China.

Ha, J.-W., Kim, A., Kim, C., Park, J., and Kim, S. (2017). Automatic Music Highlight Extraction using Convolutional Recurrent Attention Networks. *arXiv preprint arXiv:1712.05901*.

Lee, J.-Y., Kim, J.-Y., and Kim, H.-G. (2014). Music Emotion Classification Based on Music Highlight Detection. In *Proceedings of ICISA (International Conference on Information Science and Applications)*, Seoul, South Korea.

Lei, J., Berg, T. L., and Bansal, M. (2021). Detecting Moments and Highlights in Videos via Natural Language Queries. In *Proceedings od NIPS (International Conference on Neural Information Processing Systems)*.

Lv, G., Zhang, K., Wu, L., Chen, E., Xu, T., Liu, Q., and He, W. (2022). Understanding the Users and Videos by Mining a Novel Danmu Dataset. *IEEE Transactions on Big Data*, 8(2).

Mundnich, K., Fenster, A., Khare, A., and Sundaram, S. (2021). Audiovisual Highlight Detection in Videos. In *Proceedings of ICASSP (International Conference on Acoustics, Speech and Signal Processing)*, Toronto, Canada.

Patty, A. T. (2009). Pacing Scenarios: How Harmonic Rhythm and Melodic Pacing Influence Our Experience of Musical Climax. *Music Theory Spectrum*, 31(2).

Prétet, L., Richard, G., and Peeters, G. (2021). Is there a "language of music-video clips"? A qualitative and quantitative study. In *Proceedings of ISMIR (International Conference on Music Information Retrieval)*, Virtual Event.

Sarakit, P. (2015). *A Music Video Recommender System based on Emotion Classification on User Comments*. PhD thesis, Sirindhorn International Institute of Technology.

Sato, H., Hirai, T., Nakano, T., Goto, M., and Morishima, S. (2016). A Soundtrack Generation System to Synchronize the Climax of a Video Clip with Music. In *Proceedings of ICME (International Conference on Multimedia and Expo)*, Seattle, WA, USA.

Schlüter, J. and Böck, S. (2014). Improved musical onset detection with Convolutional Neural Networks. In *Proceedings of ICASSP (International Conference on Acoustics, Speech and Signal Processing)*, Florence, Italy.

Song, Y., Vallmitjana, J., Stent, A., Jaimes, A., Labs, Y., and York, N. (2015). TVSum: Summarizing Web Videos Using Titles. In *Proceedings of CVPR (Conference on Computer Vision and Pattern Recognition)*, Boston, MA, USA.

Stewart, S., KV, G., Lu, L., and Fanelli, A. (2025). Semi-Supervised Contrastive Learning for Controllable Video-to-Music Retrieval. In *Proceedings of ICASSP (International Conference on Acoustics, Speech and Signal Processing)*, Hyderabad, India.

Sul, J., Han, J., and Lee, J. (2023). Mr. HiSum: a large-scale dataset for video highlight detection and summarization. In *Proceedings od NIPS (International Conference on Neural Information Processing Systems)*, Red Hook, NY, USA.

Sun, M., Farhadi, A., and Seitz, S. (2014). Ranking Domain-Specific Highlights by Analyzing Edited Videos. In *Proceedings of ECCV (European Conference on Computer Vision)*, Milan, Italy.

Tseng, L.-Y., Lin, T.-L., Shuai, H.-H., Huang, J.-W., and Chang, W.-W. (2023). A Dataset and Baselines for Measuring and Predicting the Music Piece Memorability. In *Proceedings of ISMIR (International Society for Music Information Retrieval Conference)*, Milan, Italy.

Wang, J.-C., Smith, J. B., Chen, J., Song, X., and Wang, Y. (2021). Supervised Chorus Detection for Popular Music Using Convolutional Neural Network and Multi-Task Learning. In *Proceedings of ICASSP (International Conference on Acoustics, Speech and Signal Processing)*, Toronto, Canada.

Xiong, Z., Radhakrishnan, R., Divakaran, A., and Huang, T. S. (2004). Effective and efficient sports highlights extraction using the minimum description length criterion in selecting GMM structures [audio classification]. In *Proceedings of ICME (International Conference on Multimedia and Expo)*, Taipei, Taiwan.

Xu, P. and Larson, M. (2014). Users Tagging Visual Moments: Timed Tags in Social Video. In *Proceedings the ACM Workshop on Crowdsourcing for Multimedia*, New York, NY, USA.

Yadati, K., Larson, M., Liem, C. C. S., and Hanjalic, A. (2014). Detecting Drops in Electronic Dance Music: Content based approaches to a socially significant music event. In *Proceedings of ISMIR (International Society for Music Information Retrieval Conference)*, Taipei, Taiwan.

Yadati, K., Larson, M., Liem, C. C. S., and Hanjalic, A. (2018). Detecting Socially Significant Music Events Using Temporally Noisy Labels. *IEEE Transactions on Multimedia*, 20.

Yamamoto, T. and Nakamura, S. (2013). Leveraging viewer comments for mood classification of music video clips. In *Proceedings ACM SIGIR (Conference on Research and development in Information Retrieval)*, New York, NY, USA.

Yeh, M.-C., Tsai, Y.-W., and Hsu, H.-C. (2016). A content-based approach for detecting highlights in action movies. *Multimedia Systems*, 22.

Yin, M., Chen, W., Zhu, D., and Jiang, J. (2025). Enhancing video rumor detection through multimodal deep feature fusion with time-sync comments. *Information Processing & Management*, 62(1).

Zeng, K.-H., Chen, T.-H., Niebles, J. C., and Sun, M. (2016). Title Generation for User Generated Videos. In *Proceedings of ECCV (European Conference on Computer Vision)*, Amsterdam, The Netherlands.

Zhang, P., Guo, J., Xu, L., You, M., and Yin, J. (2022). Bridging Music and Text with Crowdsourced Music Comments: A Sequence-to-Sequence Framework for Thematic Music Comments Generation. *arXiv preprint arXiv:2209.01996*.