Language-Guided Music Recommendation for Video via Prompt Analogies

Daniel McKee[1]* Justin Salamon[2] Josef Sivic[2,3] Bryan Russell[2]
[1]University of Illinois at Urbana-Champaign [2]Adobe Research
[3]Czech Institute of Informatics, Robotics and Cybernetics, Czech Technical University
https://www.danielbmckee.com/language-guided-music-for-video

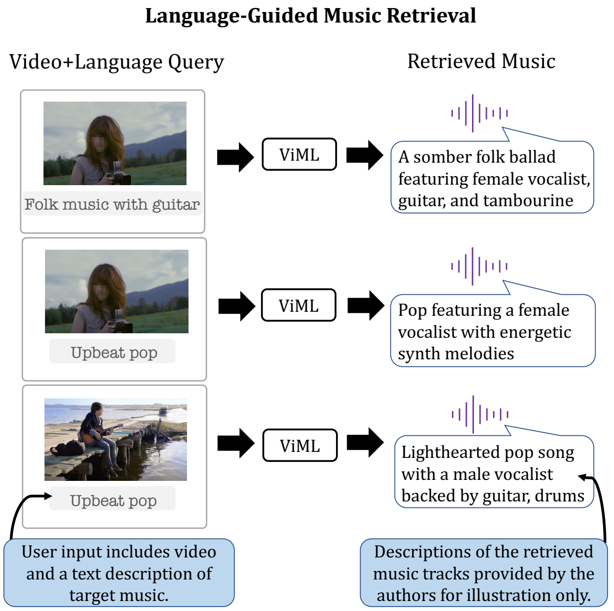
Language-Guided Music Retrieval
Video+Language Query
Retrieved Music
Folk music with guitar
ViML
A somber folk ballad featuring female vocalist, guitar, and tambourine
Upbeat pop
ViML
Pop featuring a female vocalist with energetic synth melodies
Upbeat pop
ViML
Lighthearted pop song with a male vocalist backed by guitar, drums
User input includes video and a text description of target music.
Descriptions of the retrieved music tracks provided by the authors for illustration only.

Figure 1. **Language-guided music retrieval.** Our ViML model takes a video and text prompt as input to retrieve a suitable music track from a database. The model fuses video and language representations to guide retrieval. Note how our approach retrieves audio matching both video and language content. For the same video query (top two rows), we can change the music style to match the language query, and for the same `Upbeat pop` query (bottom two rows) we can change the vocalist to match the video content. **Please view our results and full paper [9] on our website!**

1. Introduction

A key part of the video editing process for creators is choosing a musical soundtrack. Especially given the rise of short-form videos on social media platforms, automated music recommendation systems have become an increasingly common and important part of video editing applications. While these systems can be helpful for finding relevant music, they often provide limited capability for user control over the types of music recommended. In previous work, music is retrieved based solely on the visual content and style from a video [16, 12]. However, music itself can convey critical information about how a video should be perceived[1]. As a result, the lack of user input capability to describe target music for an input video is a key limitation on the utility of current music recommendation methods.

In this work, which is an abbreviated version of our full paper [9], we propose a more flexible music-for-video recommendation approach that allows a user to guide recommendations towards specific musical attributes including mood, genre, or instrumentation, illustrated in Figure 1. To maximize flexibility and user convenience, we propose to take user musical attribute descriptions in the form of *free-form natural language* (*e.g.*, "Folk music with guitar" in Figure 1). There are two key challenges in learning a model for this task. First, while there are datasets which include music+text [5, 11, 6] or music+video [1], there are no available datasets which include music, video, and text together. Second, previous works have explored jointly learning visual, audio, and text embeddings [4, 3, 14, 2], and without careful regularization, a network can overfit and possibly learn to ignore one of the input modalities.

In order to meet the challenges outlined above, our work makes the following contributions:

(1) We propose a text-synthesis approach relying on an analogy-based prompting procedure to generate natural language music descriptions from a large language model given pre-trained music tagger outputs and a small number of human text descriptions as illustrated in Figure 2 (left).

(2) We propose a Transformer-based model, which we call **Vi**deo to **M**usic with **L**anguage (ViML), to fuse text and video inputs and query music samples. For training, we introduce a text dropout regularization which is critical to model performance. Our model design allows for retrieved music audio to agree with both input modalities by matching visual style from the video and musical genre, mood, or instrumentation described in the natural language query.

(3) We release a dataset of 4000 high quality text annotations for clips from a subset of the YT8M-MusicVideo dataset [1] to evaluate language-guided music recommendation. We show that our method can achieve substantial improvements over prior works on music retrieval when incorporating text inputs. Moreover, our model can match or even exceed performance of baseline music-for-video rec-

*Work done as an intern with Adobe Research

[1] https://www.youtube.com/watch?v=iSkJFs7myn0

ommendation models when the text input is ignored.

2. Approach

Our goal is to train a pair of feature encoders f^{vt} and f^m which are capable of predicting the similarity $s(f^{vt}, f^m)$ between an input pair of video and musical text description (v, t) and a music clip m. To train such a model in a supervised manner, we need a dataset of corresponding triplets (v, m, t). While large-scale datasets of videos with paired music are available, these lack the needed natural language descriptions of the paired music tracks. As a result, we investigate a synthesis approach based on a model G which generates text descriptions from available structured data in the form of music tags for each music track. In the following sections, we will first discuss text synthesis approaches then describe an approach to train a model for our task.

2.1. Synthesizing Text Descriptions for Music

Suppose that we are given a set of video and music pairs (v_i, m_i) and that we also have access to structured data $d_i \in \mathcal{T}^D$ describing the music m_i. In our case, this structured data consists of musical tags with confidences predicted by a pretrained tagger with a vocabulary of 41 instrument tags, 20 genre tags, and 28 mood tags. Each music track m_i may be described by a free-form human text description $t_i \in \mathcal{T}^T$. However, it can be prohibitively expensive to obtain high-quality human descriptions on a large scale. Instead, we propose to synthesize text descriptions using a generator $G : \mathcal{T}^D \rightarrow \mathcal{T}^T$ which maps structured data describing an audio track to the space of natural human descriptions. We describe three such synthesis approaches below.

I. Few-shot `prompt2text` approach. We first explore whether the full mapping function G can be encompassed by a single large language model through careful few-shot prompting. This approach relies on a small set of example human-provided descriptions $t_0, ..., t_N$ where $t_i \sim \mathcal{T}^T$. We assume that for each example t_i, we also have a paired structured data output d_i, provided by the automatic music tagger, which describes the same audio track. As shown in Figure 2 (left), the structured data output d_i is converted to text form via a template, and a set of pairs $(d_0, t_0), ..., (d_k, t_k)$ are used to form k input/output components in the prompt. The final segment of the prompt is the structured data d_i corresponding to a new music track. Given d_i, the model will attempt to output a description t_i following the mapping $\mathcal{T}^D \rightarrow \mathcal{T}^T$ suggested by the example inputs. For text generation in this setting, we use BLOOM-176B[2] which is trained on a highly diverse 1.5TB text corpus.

II. Zero-shot `data2text` approach. The second approach we propose is a data-to-text generation process illustrated in Figure 2 (top right). We collect high-confidence tags predicted for each music track grouped into genre, mood, and instrument categories. Next, we insert these into pre-defined templates which are randomly sampled. Finally, to form the templated sentences into more natural free-form descriptions, we make use of pretrained large language models (we follow the Zero-shot D2T approach [7]).

III. `tags` approach. The final setting we use involves a simple concatenation of predicted tags. We take the set of top filtered predicted tags for each music track (typically around 10-15 tags), randomly shuffle, and concatenate into a comma-separated list (bottom right of Figure 2).

2.2. Text Dropout for Music Retrieval Training

Our objective is to retrieve a music track m matching a query video v and natural language query t describing the target music track. This is a challenging task as the model has to fuse together information from both input video and language query to find an appropriate music track. Moreover, the difference in granularity between audio/video and text can significantly hinder training. We design a tri-modal approach, dubbed ViML, for this task and introduce text dropout to address the granularity issue. In a similar manner to the way dropout prevents overfitting by reducing co-adaptation between individual neurons [15], text dropout serves to avoid overfitting to the text inputs and prevent co-adaptations between the video and text encoders.

Model architecture and loss. Our model is trained on a set of (video, music, text) pairings, (v, m, t), corresponding to a music video clip v, labeled with a generated text description t of its music track m, as outlined in Section 2.1. We transform these inputs into sequences of base features $x^v = g^v(v)$ for visual video features, $x^m = g^m(m)$ for music features, and $x^t = g^t(t)$ for text features using pretrained large-scale encoders g^v, g^m, and g^t which are frozen during training. We use CLIP ViT-B/32 [13] to encode for video frames and text, and DeepSim [8] to encode music.

Our model consists of three separate modules corresponding to each modality f^v, f^m, f^t, and a fourth fusion module f^{vt} to combine video and text representations (all modules consist of two-layer Transformers [17]). The modules take respective base features and output embeddings $y^v = f^v(x^v)$, $y^m = f^m(x^m)$, $y^t = f^t(x^t)$. The fusion module outputs a fused embedding for video and text $y^{vt} = f^{vt}(y^v, y^t)$. For training, we use an InfoNCE loss [10] between music and fused video-text embeddings with cosine similarity as our chosen similarity metric.

Text dropout. With probability p, we set the input text embedding x^t to a specific value x^{NULL} defined as the embedding produced by the pretrained g^t model for an empty string. Beyond improving the performance of music retrieval from text and video together, training with text

[2] https://huggingface.co/bigscience/bloom

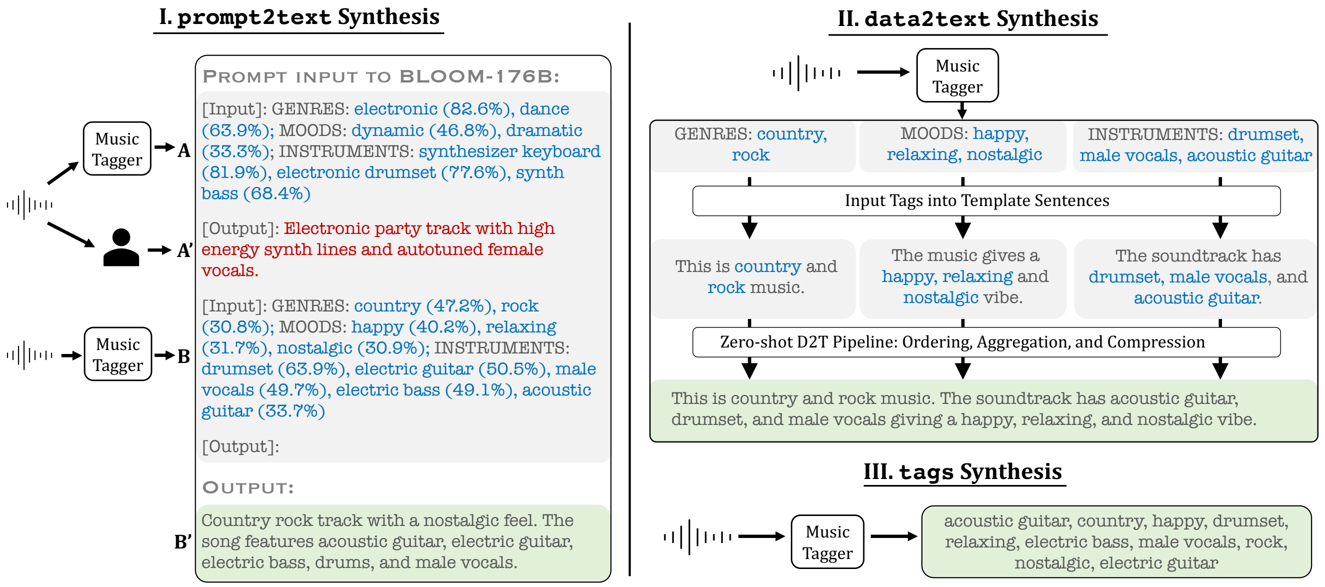
I. prompt2text Synthesis
Music Tagger
A
A'
Music Tagger
B
B'
Prompt input to BLOOM-176B:
[Input]: GENRES: electronic (82.6%), dance (63.9%); MOODS: dynamic (46.8%), dramatic (33.3%); INSTRUMENTS: synthesizer keyboard (81.9%), electronic drumset (77.6%), synth bass (68.4%)
[Output]: Electronic party track with high energy synth lines and autotuned female vocals.
[Input]: GENRES: country (47.2%), rock (30.8%); MOODS: happy (40.2%), relaxing (31.7%), nostalgic (30.9%); INSTRUMENTS: drumset (63.9%), electric guitar (50.5%), male vocals (49.7%), electric bass (49.1%), acoustic guitar (33.7%)
[Output]:
Output:
Country rock track with a nostalgic feel. The song features acoustic guitar, electric guitar, electric bass, drums, and male vocals.
II. data2text Synthesis
Music Tagger
GENRES: country, rock
MOODS: happy, relaxing, nostalgic
INSTRUMENTS: drumset, male vocals, acoustic guitar
Input Tags into Template Sentences
This is country and rock music.
The music gives a happy, relaxing and nostalgic vibe.
The soundtrack has drumset, male vocals, and acoustic guitar.
Zero-shot D2T Pipeline: Ordering, Aggregation, and Compression
This is country and rock music. The soundtrack has acoustic guitar, drumset, and male vocals giving a happy, relaxing, and nostalgic vibe.
III. tags Synthesis
Music Tagger
acoustic guitar, country, happy, drumset, relaxing, electric bass, male vocals, rock, nostalgic, electric guitar

Figure 2. **Overview of three text synthesis approaches explored in our work.** All rely on tag predictions from a pretrained music tagger model. We highlight output text from each method in green, inputs from the tagger in blue font, and inputs from a human annotator in red font. **Left:** We introduce the `prompt2tags` approach for generating natural language descriptions given automatically predicted music tags and a small set of human descriptions. We ask a large language model (BLOOM-176B) to complete an analogy task ($A : A' :: B : B'$) between music tags (A, B) and descriptions (A', B'). **Top right:** The `data2text` pipeline inserts sampled tags into randomly selected template sentences corresponding to each tag category. The Zero-shot D2T model [7] then orders, aggregates, and compresses these templates into a final output description. **Bottom right:** The `tags` approach involves direct concatenation of high confidence tags.

dropout yields a model which also performs well at retrieval from video alone by removing dependence on text inputs.

3. Experiments

In all of our experiments, we train models using the YT8M-MusicVideo dataset which includes around 100k videos with the "music video" tag from the much larger YouTube8M dataset [1]. We synthesize tags and a natural language text describing the music track of each video for the full dataset using the approaches described in Sec. 2.1.

We evaluate music retrieval performance consistently with previous works [12, 16]. However, in our case, a query can be either a video alone or a video and corresponding text annotation together. For each query, we rank music tracks in a pool containing the ground truth track corresponding to an input query. We then compute Recall@K (R@K) for K=1,5,10 and Median Rank (MR), calculating the average of each of these metrics across the full set of test queries.

Method	Train Text	Query	MR ↓	R@1 ↑	R@5 ↑	R@10 ↑
a. Pretét et al. [12]	-	-	234	0.76	3.42	5.90
b. MVPt [16]	-	-	13	6.09	24.91	41.89
c. MVPt+ [16]	-	-	5	27.93	50.64	60.68
d. ViML (ours)	`tags`	-	3	29.43	62.49	75.40
e. ViML (ours)	`tags`	`tags`	**2**	**49.49**	**81.61**	**89.41**
f. Chance			1000	0.05	0.25	0.50

Table 1. **Tag-based music retrieval on full YouTube8M-MusicVideo test set.** We compare ViML against prior methods on video to music retrieval without tag queries (row d.). We also evaluate ViML on video+text to music retrieval using (synthetic) tags at test time (row e.). The text descriptions for both training and evaluation are generated with the `tags` approach.

Tag-Based Retrieval. First, we explore the tag-based retrieval setting. Here the goal is to retrieve a music track given a query video and set of tags from a pre-defined vocabulary, such as "happy", "piano" and "jazz". In these experiments, we train a track-level model and perform retrieval on a track-level in a manner consistent with prior work [12, 16] evaluating on the full YT8M-MusicVideo test set. As shown in Table 1, our baselines include a version of MVPt that we call MVPt+ with significantly improved performance from tuning the temperature parameter in the InfoNCE loss. We evaluate our ViML model in two settings: *without* input texts at test time (an empty text input is used instead) and *with* text inputs at test time. As we do not have track-level human-provided music tag annotations for the full YT8M-MusicVideo test split, we evaluate the model on synthetically generated tags using our `tags` approach. In this setting, our model shows a very substantial performance increase over MVPt+ of 20-30 points in each recall metric. Interestingly, our ViML model evaluated without text at test time not only matches the video-to-music re-

Method	Train Text	MR ↓	R@1 ↑	R@5 ↑	R@10 ↑
a. MVPt+	-	17	12.20	29.43	40.46
b. ViML	`tags`	15	11.95	30.34	42.62
c. ViML	`data2text`	13	13.61	33.94	46.24
d. ViML	`prompt2text`	**12**	**14.09**	**35.04**	**47.88**
Chance		250	0.20	1.00	2.00

Table 2. **Music retrieval with free-form natural language (human annotations) on YT8M-MusicTextClips test set.** Since the MVPt+ model does not take text inputs, it is evaluated on music retrieval from video alone for the same set of 3k video clips.

trieval performance of MVT+ but substantially improves over MVPt+, especially in Recall@5 and Recall@10.

Free-Form Natural Language Retrieval. For the next experiments, we turn to retrieval with video and free-form natural language inputs. For this setting, we introduce a new dataset consisting of a 4,000 sample subset of YT8M-MusicVideo with human-provided text descriptions of the music track accompanying each video. To create these annotations, we sample 10 second audio clips from the middle of each music video, and we ask human annotators to describe the music they hear after listening to the audio clip. This annotated set is meant mainly for evaluation, and we test our models on the 3,000 samples from the test set of YT8M-MusicVideo. We use a similar testing protocol here to the "segment-level" setting reported by Surís et al. [16], but our input video includes only a 30sec clip surrounding the 10sec of audio labeled by a human annotator.

As shown in in Table 2, our baseline is an MVPt+ model trained on 30sec segments. In addition, we report music retrieval using video and free-form human text descriptions as input queries to our ViML model. We show three variants trained on YT8M music videos with text synthesized by each of the approaches described in Sec. 2.1. The model trained with our first `tags` synthesis baseline (b.) improves over retrieval with MVPt+ using only video (a.). Next, the `data2text` approach (c.), which generates more natural phrases while strictly preserving tag semantics, provides a consistent improvement over the ViML `tags` variant (b.). Finally, our `prompt2text` approach (d.) yields the best performance showing that large language models are strong annotators on this task with careful few-shot prompting.

4. Conclusion

In this work, we presented our ViML model which allows language-guided music recommendation for video. The model fuses text and video inputs to find matching music and is trained with a text dropout technique to improve performance. We proposed a music description synthesis approach using a large language model and introduced a new dataset, YouTube8M-MusicTextClips, which includes free-form human descriptions of music in YT8M videos.

References

[1] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016. 1, 3

[2] H. Akbari, L. Yuan, R. Qian, W.-H. Chuang, S.-F. Chang, Y. Cui, and B. Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *NeurIPS*, 2021. 1

[3] J.-B. Alayrac, A. Recasens, R. Schneider, R. Arandjelović, J. Ramapuram, J. De Fauw, L. Smaira, S. Dieleman, and A. Zisserman. Self-supervised multimodal versatile networks. *NeurIPS*, 2020. 1

[4] Y. Aytar, C. Vondrick, and A. Torralba. See, hear, and read: Deep aligned representations. *arXiv preprint arXiv:1706.00932*, 2017. 1

[5] T. Bertin-Mahieux, D. P. Ellis, B. Whitman, and P. Lamere. The million song dataset. In *ISMIR*, 2011. 1

[6] D. Bogdanov, M. Won, P. Tovstogan, A. Porter, and X. Serra. The mtg-jamendo dataset for automatic music tagging. In *Machine Learning for Music Discovery Workshop, ICML*, 2019. 1

[7] Z. Kasner and O. Dušek. Neural pipeline for zero-shot data-to-text generation. In *ACL*, 2022. 2, 3

[8] J. Lee, N. J. Bryan, J. Salamon, Z. Jin, and J. Nam. Disentangled multidimensional metric learning for music similarity. In *ICASSP*, 2020. 2

[9] D. McKee, J. Salamon, J. Sivic, and B. Russell. Language-guided music recommendation for video via prompt analogies. In *CVPR*, 2023. 1

[10] A. v. d. Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2

[11] S. Oramas, O. Nieto, F. Barbieri, and X. Serra. Multi-label music genre classification from audio, text and images using deep features. In *ISMIR*, 2017. 1

[12] L. Prétet, G. Richard, and G. Peeters. Cross-modal music-video recommendation: A study of design choices. In *IJCNN*, 2021. 1, 3

[13] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2

[14] A. Rouditchenko, A. Boggust, D. Harwath, B. Chen, D. Joshi, S. Thomas, K. Audhkhasi, H. Kuehne, R. Panda, R. Feris, et al. Avlnet: Learning audio-visual language representations from instructional videos. *Interspeech*, 2021. 1

[15] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *JMLR*, 2014. 2

[16] D. Surís, C. Vondrick, B. Russell, and J. Salamon. It's time for artistic correspondence in music and video. In *CVPR*, 2022. 1, 3, 4

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *NIPS*, 2017. 2