Language-Guided Music Recommendation for Video via Prompt Analogies
Supplementary Materials

A. Video Examples

To see our qualitative examples, please watch the demo video on our webpage: https://www.danielbmckee.com/language-guided-music-for-video.

B. Fusion Module Architecture Study

We evaluate different architectures for the fusion module that is responsible for combining encoded visual and text inputs. In Table 1, we present five architecture variants. First, we benchmark fusion by direct addition of the visual and text Transformer encoder outputs (a.) which removes learned parameters from the fusion module entirely. Next, we evaluate three different learned fusion module architectures which involve passing the concatenated visual and text features as input to: (b.) a single linear layer (c.) a two-layer MLP, (d.) a 1-layer Transformer network, and (e.) a 2-layer Transformer network. We find that the size of the fusion module does not significantly change performance. We use the 2-layer Transformer fusion architecture in our main results given the slightly higher performance in recall metrics, but similar performance can be achieved with the other fusion architectures including the "addition" fusion module which does not include learned parameters.

C. Training without Video

We also experimented with models trained only on music and text but found these models to significantly underperform other baselines at music retrieval on the YT8M-MusicTextClips test set. This is not surprising as the input video contains a great deal more information than the short human text descriptions in the dataset. Performance of our music+text model trained on `prompt2text` data and evaluated on human texts in the YT8M-MusicTextClips test set was Recall@1/5/10=2.52/9.27/15.52 and MR=56 (compare to Table 2 results from the main paper). We report the results of our track-level music+text model trained with tag inputs as MT in Table 2 (a.) (compare to Table 1 results from the main paper). This model also performed substantially below MVPt+ or ViML.

D. Ensembled Models

In addition to the baselines reported in Table 1 of the main text, we also investigated forming a stronger baseline by combining MVPt+ and the music+text model from Sec. C into an ensemble.

More specifically, for a music track m and a corresponding video, text pair (v, t), we compute the total similarity score as a weighted sum $(1-\alpha) \cdot s(y^v, y^m) + \alpha \cdot s(z^t, z^m)$ where y^v, y^m are the video and music embeddings generated by MVPt+, z^t, z^m are the text and music embeddings generated by our music+text model, and α is a coefficient which we tuned.

As shown in Table 2 (d.), we found this music+text model and MVPt+ ensemble to reach strong performance, exceeding Recall@1 performance of ViML and achieving similar Recall@5/10 ViML performance. However, we found that such ensembling could be used to improve the performance of ViML as well. In particular, computing scores for music retrieval as a weighted sum of similarity scores from ViML and MVPt+ led to substantial improvements over ViML performance as shown in Table 2 (e.). An ensemble of ViML and our music+text model led to the highest performance in Table 2 (f.).

E. Music Matching the Pace of Videos

In our qualitative results, we did not observe many examples where the music beats per minute (BPM) does not match the video pace. We hypothesize that a given music genre lives in a limited tempo range. Therefore, being able to match effectively the music genre may return a well-matching tempo for free. Note that we do not have fine-grained tempo alignment, *e.g.*, depicted dance motions may not be perfectly in sync with the music. One possible future direction could be to refine the alignment between the music and depicted action in the video.

F. Text Synthesis Examples Outputs

In Figures 1 and 2, we present generated outputs from our text synthesis approaches along with real human annotations for randomly selected examples from the

Method	# Parameters	Median Rank ↓	R@1 ↑	R@5 ↑	R@10 ↑
a. Addition	0	12	13.73	34.12	46.52
b. Linear	131K	13	13.19	33.45	45.68
c. MLP	1.6M	13	13.20	32.69	44.94
d. Transformer (1 layer)	1.4M	12	13.98	34.95	46.85
e. Transformer (2 layer)	2.8M	12	**14.09**	**35.04**	**47.88**

Table 1. **Study of fusion layer architecture.** All models are trained on the synthesized `prompt2text` data, and we report results on the YT8M-MusicTextClips 3k test set.

Method	Train Text	Query Text Input	Median Rank ↓	R@1 ↑	R@5 ↑	R@10 ↑
a. MT	`tags`	`tags`	15	11.51	30.36	42.72
b. MVPt+ [37]	-	-	5	27.93	50.64	60.68
c. ViML (ours)	`tags`	`tags`	2	49.49	81.61	89.41
d. MT & MVPt+ Ens.	`tags`	`tags`	1	55.95	81.73	88.82
e. ViML & MVPt+ Ens.	`tags`	`tags`	1	59.86	85.14	91.43
f. ViML & MT Ens.	`tags`	`tags`	1	**63.05**	**91.32**	**96.59**
g. Chance			1000	0.05	0.25	0.50

Table 2. **Tag-based music retrieval on full YouTube8M-MusicVideo test set for music+text model and model ensembles.** For convenient comparison, we also report the MVPt+ and ViML results from Table 2 in the main text. We denote the music+text model described in Sec. C as MT in the table.

YouTube8M-MusicVideo dataset. The text synthesis approaches show different tradeoffs between tag accuracy and diversity of form/language.

The `prompt2text` setting is the most free-form text synthesis approach but will sometimes generate outputs which are not true to the original tag predictions for a track. In general, the `prompt2text` descriptions tend to be shorter and often omit information in the input tags. The language model used in the `prompt2text` approach can also sometimes hallucinate information which is completely wrong (e.g. "female vocal and a piano section" in Fig. 1 ex. 3). However, the diversity of vocabulary and structure in the outputs produced by `prompt2text` makes this approach most similar to real human annotations.

1) https://youtu.be/xcZIwXABBKA?t=134
TAGS: rock, lead vocals, electric bass, frantic, powerful, vocals, electric guitar, angry, slow, male vocals, acoustic drumset
DATA2TEXT: The soundtrack is rock with a mood of powerful, slow, and frantic. Acoustic drums, male vocals, vocals, lead vocals, and electric guitar are playing in this song.
PROMPT2TEXT: Rock song with electric guitars, electric bass, drums, and male vocals. The track has a powerful and angry vibe.
HUMAN: Alternative hard rock song with grungy male vocals, heavy drumming, guitar riffs and bass conveying high energy.

2) https://youtu.be/A_jOjhJnHmA?t=100
TAGS: dance, synthesizer keyboard, hip-hop, male vocals, sad, electronic, electronic drumset, vocals, synth bass
DATA2TEXT: This music is hip-hop, electronic, and dance with vocals, synthesizer keyboard, electronic drumset, and male vocals.
PROMPT2TEXT: Hip-hop instrumental with a male vocal. Great for rap music.
HUMAN: Rebellious hip-hop track with electronic music and impactful male vocals.

3) https://youtu.be/ZUyZP9ASO1c?t=82
TAGS: pop, hip-hop, synth bass, happy, dance, vocals, electronic, sad, male vocals, synthesizer keyboard, lead vocals, electronic drumset, dreamy
DATA2TEXT: This song is electronic and hip-hop, with synthesizer keyboard, electronic drumset, male vocals, synth bass, vocals, and lead vocals. The feeling is happy, dreamy, and sad.
PROMPT2TEXT: Hip-hop track with a female vocal and a piano section. Great for a sad scene.
HUMAN: A jivey male led track with electronic music and groovy trap beats making the song a dance retreat.

4) https://youtu.be/Z1UhPHPjE10?t=185
TAGS: vocals, synthesizer keyboard, male vocals, synth bass, happy, lead vocals, sad, hip-hop, angry, electronic drumset
DATA2TEXT: This song sounds sad, but is a song with electronic drumset, male vocals, synthesizer keyboard, and synth bass.
PROMPT2TEXT: Hip-hop track with a slow beat, male vocals, and a synth bass. The track has a sad and melancholic feel.
HUMAN: Foreign heavy rap with low bass effects and trap synth line. Groovy jamming with friends feeling.

5) https://youtu.be/kXPxUMv-S9I?t=148
TAGS: electric guitar, vocals, slow, rock, dance, lead vocals, electric bass, blues, frantic, electronic, acoustic drumset, male vocals, quirky, angry, powerful
DATA2TEXT: This is blues and electronic music with the sound of vocals, male vocals, acoustic drumset, lead vocals, and electric bass giving an angry, powerful, and frantic feel.
PROMPT2TEXT: A rock song with a bluesy feel, with electric guitar and drums.
HUMAN: Grunge rock rebel kind of music like Greenday. A male vocal with distorted guitars, bass and drums.

6) https://youtu.be/enHWyaXrcfI?t=134
TAGS: hip-hop, electronic drumset, electronic, sad, synthesizer keyboard, male vocals, lead vocals, dance, synth bass, vocals
DATA2TEXT: The music sounds like hip-hop, electronic, and dance, and sad, including synthesizer keyboard, electronic drumset, male vocals, synth bass, and lead vocals.
PROMPT2TEXT: Hip-hop track with male vocal, electronic drumset, synthesizer keyboard and bass.
HUMAN: Trap song with male rapper, very piercing hi-hat beats and bass line that sounds indulgent and addictive.

7) https://youtu.be/kE2wuQT4J14?t=117
TAGS: electric bass, blues, dynamic, acoustic drumset, happy, vocals, male vocals, frantic, electric guitar, rock
DATA2TEXT: This is blues and rock music featuring acoustic drums, electric bass, vocals, male vocals, and electric guitar with a dynamic, happy, and frantic feeling.
PROMPT2TEXT: Blues rock track with a male vocalist. The song is happy and has a strong electric guitar and electric bass.
HUMAN: Soulful rock track with male vocals backed by synth layers, over driven electric guitars and drums. The song has a relaxing note.

8) https://youtu.be/2dFMqtk1ieM?t=118
TAGS: synthesizer keyboard, hip-hop, vocals, pop, electronic drumset, happy, dreamy, dance, electronic, male vocals, lead vocals
DATA2TEXT: This song is electronic, dance, and hip-hop, sounding dreamy. The instrumentation includes male vocals and an electronic drumset.
PROMPT2TEXT: A dreamy pop track with a strong bass line and a catchy melody. Best for commercials, advertisements, and videos.
HUMAN: Mid tempo low sounding hip hop track with rap male vocal and drum groove.

9) https://youtu.be/PKJ7p_8tj1c?t=115
TAGS: synth bass, synthesizer keyboard, electronic, electronic drumset, vocals, happy, female vocals, slow, powerful, frantic, male vocals, dance, lead vocals
DATA2TEXT: The music gives a slow, happy, and powerful feel with synthesizer keyboards, electronic drumset, vocals, male vocals, and female vocals playing.
PROMPT2TEXT: A dance track with a strong beat and a lot of synth sound effects. Ideal for a party.
HUMAN: Electropop track with strong bass and impactful electronic beats. Great for disco.

Figure 1. **Synthesized text examples using our approach**. We randomly select examples from the YouTube8M-MusicTextClips test set. We show the output from our `tags`, `data2text`, and `prompt2text` approaches for each example video. We also show real `human` annotations collected for each example. The text synthesis approaches show tradeoffs between preserving tag accuracy and increasing diversity of vocabulary and phrase structure. The `prompt2text` approach generates outputs with the highest diversity and most closely resembling human annotations. Additional examples shown in Figure 2.

https://youtu.be/QJIw6v5uuYc?t=97
TAGS: vocals, synth bass, electronic, female vocals, frantic, lead vocals, pop, synthesizer keyboard, dance, happy, electronic drumset, male vocals

10) DATA2TEXT: This sounds like dance and electronic music with a happy sound. The instrumentation includes male vocals, lead vocals, electronic drumset, vocals, female vocals, synthesizer keyboard, and synth bass.
PROMPT2TEXT: EDM Pop track with a synth pad layer, heavy bass drops and beats. Sounds like a party track.
HUMAN: Melodious pop song with rhythm guitars, drums and bass. The track adds a happy vibe to the setting.

https://youtu.be/w3RzK9PEyHs?t=74
TAGS: electronic, synth bass, vocals, hip-hop, male vocals, lead vocals, dance, happy, synthesizer keyboard, electronic drumset

11) DATA2TEXT: This song is happy. Electronic drumset, lead vocals, synth bass, vocals, male vocals, and synthesizer keyboard are playing in this song.
PROMPT2TEXT: Hip-hop music with electronic drums, bass, and synthesizer.
HUMAN: Impactful freestyle male rap with backing prominent synth lines, guitars and drums.

https://youtu.be/Jkb-MeVp4cY?t=129
TAGS: electric guitar, dreamy, dance, synth bass, hip-hop, lead vocals, electronic, pop, synthesizer keyboard, electronic drumset, vocals, happy

12) DATA2TEXT: This is some electronic, dance, and pop music with a happy sound. The soundtrack has synth bass, vocals, synthesizer keyboard, electric guitar, electronic drumset, and lead vocals.
PROMPT2TEXT: A very upbeat track with a lot of energy. Great for a party or a nightclub. Has a lot of synth bass and a lot of synths.
HUMAN: A classical coordinated track with acoustic guitar, wind chimes, soft rock drum beats, and a male artist in a foreign language. Very nostalgic and elegant piece.

https://youtu.be/Jvjgq9LpDZA?t=124
TAGS: electronic drumset, sad, lead vocals, hip-hop, happy, synthesizer keyboard, vocals, synth bass, male vocals

13) DATA2TEXT: This sounds like hip-hop music and sad, including male vocals, synthesizer keyboard, vocals, lead vocals, and electronic drumset.
PROMPT2TEXT: Hip-hop song with male vocals, synth, drum, bass and an impactful drop.
HUMAN: Dark rap song with male vocals and electronically produced layers.

https://youtu.be/7SzppacIY1M?t=94
TAGS: dance, hip-hop, synthesizer keyboard, electronic drumset, sad, vocals, electronic, synth bass, lead vocals

14) DATA2TEXT: This is some hip-hop and electronic music featuring an electronic drumset and synth bass. The feeling is sad.
PROMPT2TEXT: Hip-hop track with male lead, bass, drums and synths. The song is a stress buster.
HUMAN: Hip-hop track with with syncopated beats and synth sound effects that sounds like a Jay-Z song.

https://youtu.be/y2GXHr7P3D0?t=63
TAGS: synth bass, electronic, novelty, angry, synthesizer keyboard, dynamic, electronic drumset, dance

15) DATA2TEXT: This is a song with synth bass and electronic drumset. The mood is dynamic, dance, novelty, and electronic.
PROMPT2TEXT: A dance track with a robotic female vocal and a synth bass line. The track is perfect for a club.
HUMAN: Death-metal track with chaotic music that sounds like guns being fired.

https://youtu.be/CzVcyff_gc4?t=147
TAGS: vocals, dance, synth bass, electronic drumset, male vocals, electronic, happy, pop, synthesizer keyboard, lead vocals
DATA2TEXT: The genre of this music is electronic, sounds happy, electronic drumset, synthesizer keyboard, lead vocals, synth bass, and male

16) vocals playing in this song.
PROMPT2TEXT: A happy and upbeat dance track with a catchy melody and a strong beat. The vocals are sung in an unfamiliar language.
HUMAN: Lounge chill hip hop music with rap male vocals, drum and bass.

https://youtu.be/u6KD0-jBfx4?t=117
TAGS: dance, female vocals, lead vocals, happy, synth bass, electronic drumset, vocals, pop, synthesizer keyboard, electronic, dynamic
DATA2TEXT: The music is dynamic and happy with a pop, electronic, and dance feel, featuring synthesizer keyboard, lead vocals, synth bass,

17) and female vocals.
PROMPT2TEXT: A dance track with a strong beat, a catchy melody and a lot of energy.
HUMAN: Very passionate love ballad track with sensual female vocals, keys, guitar, drum and bass.

https://youtu.be/qwCWz4BFNuk?t=107
TAGS: relaxing, male vocals, lead vocals, dreamy, nostalgic, acoustic guitar, sad, vocals, piano, acoustic

18) DATA2TEXT: This music sounds dreamy and nostalgic, including acoustic guitar, lead vocals, male vocals, and piano.
PROMPT2TEXT: Acoustic ballad with female vocals and piano.
HUMAN: A sorrowful pop song with a melancholic melody. The passionate female vocals add a feeling of separation and longingness.

Figure 2. **Synthesized text examples using our approach**. Continued from Figure 1.