It's Time for Artistic Correspondence in Music and Video

Supplementary Material - Appendix

We divide the Appendix in three sections. First, in Appx. A we describe the datasets in more detail. Second, in Appx. B we provide more implementation details about the model. And finally, in Appx. C we describe the experiments setup in more detail, and provide all the results that were not provided in the main paper.

A. Datasets Details

A.1. YT8M-MusicVideos

In this section we give details about the order to study the distribution of music genres in our YT8M-MusicVideos dataset (in the music part), and of gender, race, and age (in its visual part).

We compute genre information using musicnn [50]. We predict a single category from the MSD dataset [12] for every music track, and show its distribution in Figure 9.

We study the gender, race, and age information using the FairFace dataset [36]. First, we collect a random subset of 2079 samples in our dataset, and we extract faces from the raw frames (we sample one frame per second in the selected videos) using a pre-trained face detection model. Second, we train a classification model on the FairFace dataset. Note that the FairFace dataset is uniform across races and genres, which means the trained model is not biased. However, we note that the test-time accuracy of our model is not perfect (72% accuracy for race, 93% for gender), which makes this analysis just orientative. Finally, we classify the previously extracted faces using this model, and represent each video with the most common category across all the faces extracted from it. If a video does not contain faces, we do not use it for the analysis. We show results in Figures 10 and 11. In those figures, the frequencies are normalized for every gender.

The distribution of races, genders and ages in the random subset of 2079 videos of our dataset is the following:

- Gender: {Male: 1368, Female: 711}
- Race: {Indian: 21, White: 1098, Middle Eastern: 283, Black: 414, East Asian: 192, Latino_Hispanic: 57, Southeast Asian: 14}
- Age: { 0-2: 2, 3-9: 48, 10-19: 67, 20-29: 1644, 30-39: 287, 40-49: 17, 50-59: 10, 60-69: 4}

A.2. MovieClips

In this section, we describe the procedure we used to create the MovieClips dataset. First, we downloaded all the videos from the YouTube channel MovieClips [45], which

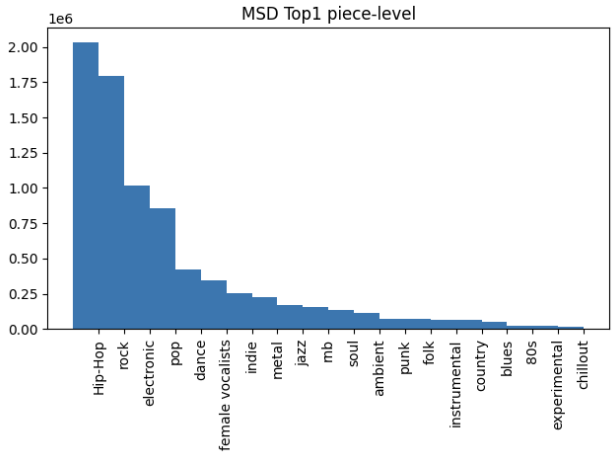

Figure 9. Distribution of the top-20 genres in the YT8M-MusicVideos dataset.

contains 37k videos. From these videos, we selected the parts that contain music consecutively for at least 20s. We did so by training a PANN model [37] on AudioSet [30] and used it to detect regions with music in the data. Specifically, we consider a segment of audio a "music" one if any of the music-related classes in AudioSet is predicted with at least a probability of 0.3 (the probabilities for different classes are predicted independently). We consider music-related classes all the ones under the "Music" category in the AudioSet ontology. In order to filter out non-music audio, we additionally ignore the segments that (even though they may contain music) contain any other audio category with a probability of 0.2 or more. These predictions are done at a very small temporal resolution, resulting in small gaps of silence between music segments. In order to not treat these segments as separate, but as a single one, we finally process the binary music predictions using a morphologic closing of 3 seconds.

B. Implementation Details

In videos whose total duration is longer than $K \cdot t$, clips are sampled from a subset of the video, of duration $K \cdot t$. During training, this subset is sampled randomly, and at test time the subset is found deterministically by centering it in the middle of the video.

The Transformer architecture follows the same design as the Transformer Encoder in the original paper [63]. We use hidden dimensionality $d_h = 256$, two layers, and two heads. We adapt the dimensionality d_{in} of the input features

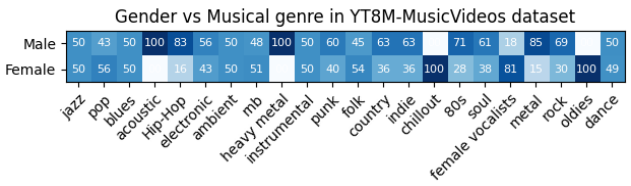

Figure 10. Distribution of genders per music genre, normalized for every genre, in the YT8M-MusicVideos dataset. Note that this is *not* the same as Figure 4 in the main paper. Here we show dataset statistics, and in the main paper we show our model's retrieval results. Most of the very skewed distributions are skewed because those categories do not contain a lot of examples in the dataset (see Figure 9). However, we note the very skewed (while representative) distribution of the hip-hop, female vocalists, and metal genres.

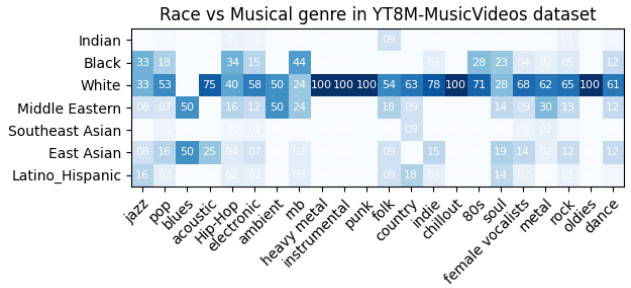

Figure 11. Distribution of races per music genre, normalized for every genre, in the YT8M-MusicVideos dataset. Note that this is *not* the same as Figure 4 in the main paper. Here we show dataset statistics, and in the main paper we show our model's retrieval results.

to the hidden size d_h using a linear projection layer.

We train the model using backpropagation, and following the optimization parameters in [61]. Specifically, we train all the models using AdamW optimizer and a cosine learning rate annealing strategy with an initial value of 1e-3. We set the total batch size to 2048. We implement our models using PyTorch [49].

C. Other Experiments

C.1. Experiment Setup Details

In order to obtain audios of guitars (used in Figure 8 of the main paper), we use the MedleyDB dataset [13], which has separate music tracks for every instrument, for a series of 196 songs (including both version 1 and 2 of the dataset).

C.2. Human Experiment Details

Each human was shown 20 example video pairs for each task (video-to-music, and music-to-video), where the pair consisted of the result provided by our model and the result

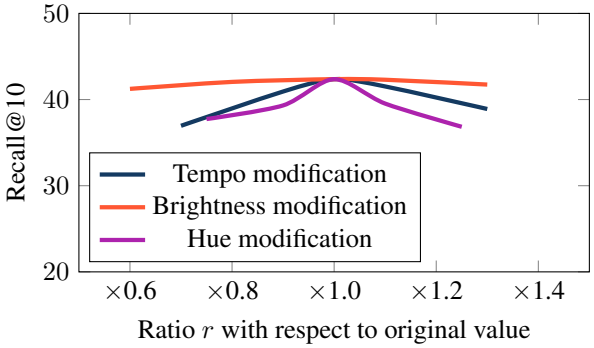

Figure 12. **Manipulation of tempo and color brightness**. Both attributes influence, but are not critical to, model performance.

provided by the baseline. The two clips in the pair contained either the same video and different music (for the video-to-music task), or two different videos with the same music (for the music-to-video task). The task consisted of a binary choice between the two clips, according to the best fit between music and video in the clip. Every human was shown different examples from the test set.

We collected a total of 296 human evaluations for the video-to-music task, with 220 of them (74.3%) preferring our model's result over the baseline. Similarly, we collected 372 human responses for the music-to-video task, with 257 (69.1%) preferring our model's output. Using the binomial test for statistical significance, we obtain in both cases a p-value well below 0.01, validating the claim that our model is preferable to humans.

C.3. Quantitative Analysis Extended

In this section, we extend the analysis in Section 5.1 in the main paper ("Quantitative Analysis").

First, we show the curves of accuracy versus change rate r in color brightness, color hue, and music tempo in Figure 12. In the case of hue, which is not an intensity parameter, a change of r is a change in the hue following the equation $h_{\text{new}} = (h_{\text{original}} + 360(r-1)) \mod 360$, where h represents hue, which is an angle with values in $[0, 360)$. This hue change formulation allows us to represent all three modifications in a common scale.

We also show the matrices relating age, visual scene, and visual objects (instruments) to music genre in Figures 13 to 15. For the visual scene analysis on Places, from the 201 we only show the classes for which the sum of all the genres provides at least a sum of 20%. To obtain the balanced dataset of musical instruments, we download the images in the Open Images Dataset [39] that contain bounding box annotations of musical instruments and crop the instruments in those images. Then, we randomly sample 50 images from every instrument class in order to have a balanced dataset, and proceed as in the other attribute studies.

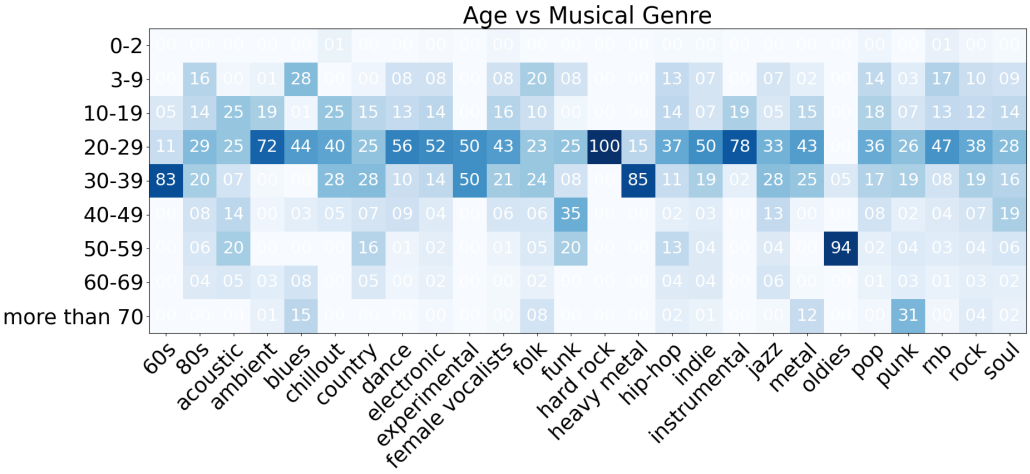

Figure 13. Age vs genre. This figure is equivalent to Figure 4 in the main paper, but showing age targets, instead of gender or race.

We also implemented a same-genre baseline, where we assume the genre label for each video and music track is known (for the video, we use the genre of its associated GT music). From the full set of target candidates, we take just the ones with the same genre as the query, and randomly select one as the prediction. The average R@10 values are 10.00% for segment-level (compared to our 42.37%), and 7.27% for track-level (compared to our 42.27%). This result shows that, while genre is an important attribute, our model captures a variety of other relevant audiovisual cues (*on top of* capturing genre information). Note that the model does not have access to the ground truth (GT) genre. Recognizing the genre is a hard task by itself, and only well defined for music (not for video).

C.4. Qualitative comparison to baselines

We perform some qualitative comparisons between the GT audios and videos, the suggested baseline, and our model results, on the examples shown in Figure 1 of the main paper. In the first video-to-music example, the ground truth genre is "rock". Compared to the ground truth audio, the music retrieved by our method has a very similar tempo, an equal reliance on pronounced beats, and a very similar language (Spanish vs. Portuguese). In contrast, the same-genre baseline (a random retrieval of a rock track) returns a music track which, while having similar instrumentation to the ground truth (electric guitar and drums), is otherwise very different in terms of mood (sadder, less active), tempo (much slower), and language (Slovenian), and is a poor overall fit for the query video. We generally found that the same-genre baseline, unlike our model, is incapable of capturing important attributes not directly related to genre.

C.5. t-SNE visualizations

We computed t-SNE visualizations [62], which show videos and music tracks are generally clustered by genre, although not exclusively. See Fig. 16.

C.6. Addressing Potential Learning Shortcut

Transformers take in a sequence of clips, instead of a single average of all clips. This makes them capable of obtaining information about the length of the sequence. Therefore, a potential learning shortcut Transformers may exploit is the use of sequence length as a signal to match music to a video (music associated to a specific video will have the same length than the video). In order to show that our Transformer model is not relying solely on this shortcut and it is mostly exploiting other cues, we run a test that controls for the sequence length. Namely, we constrain the candidate retrieval set to be the same length as the query by setting the similarity scores between different-length sequences to zero. This would be equivalent to detecting the shortcut and exploiting it with perfect accuracy.

We perform this test for the YT8M-MusicVideos dataset, at the segment level (equivalently to Table 1 in the main paper). The numbers are in Table 4. Overall, the numbers imply that 1) our model is not relying on this shortcut, and 2) even when controlling for the bias (and providing the MLP baseline with a tool that it cannot have by construction), the Transformer model is still significantly better than the MLP baseline, showing that modeling of temporal context in a non-trivial way is highly beneficial for the performance of the model.

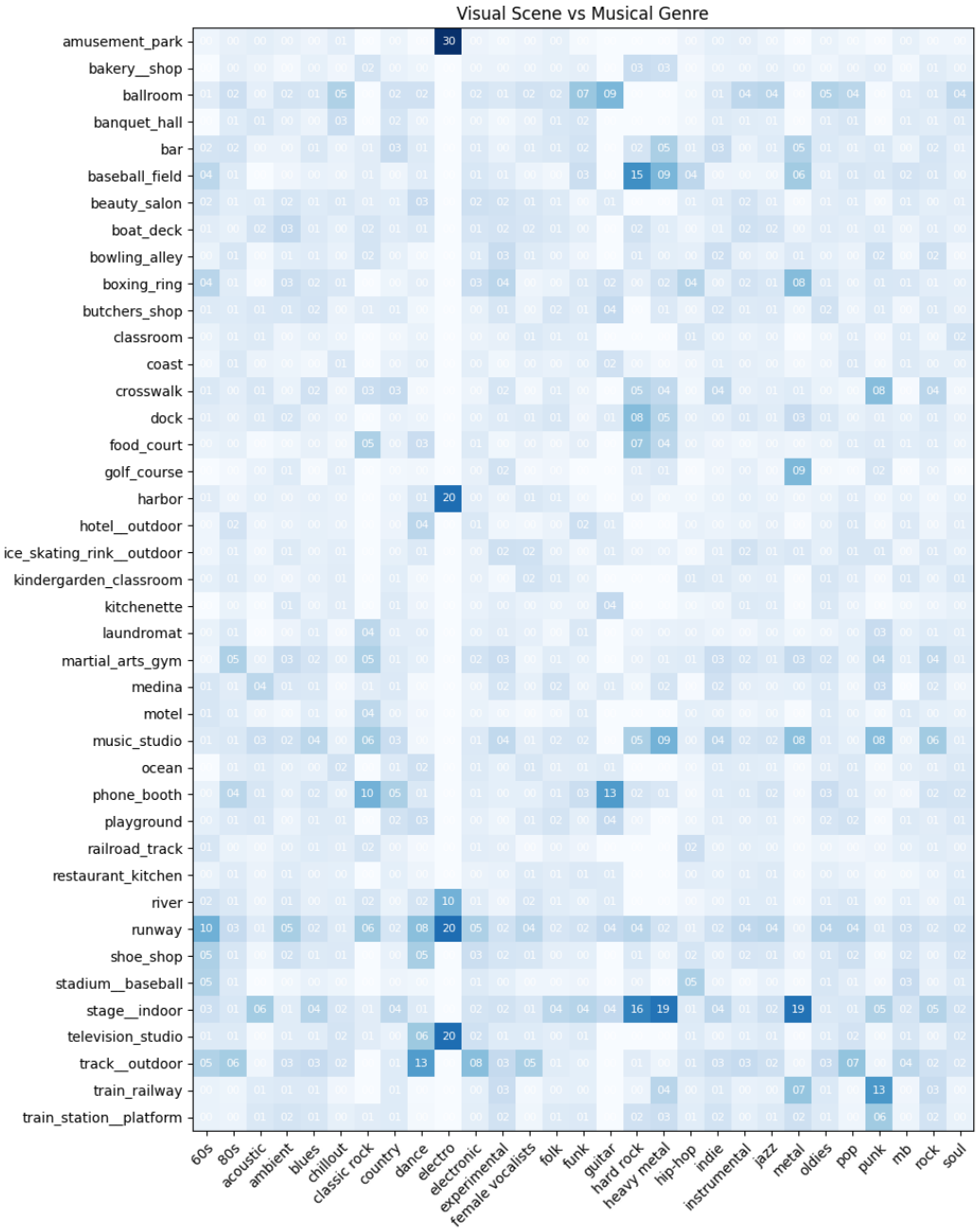

Figure 14. Scene vs genre. This figure is equivalent to Figure 4 in the main paper, but showing visual scene targets, instead of gender or race. While hard to associate specific genres to specific scenes, we note that there is a correlation, and the model is picking up on some signal regarding visual scene.

Table 4. Track-level retrieval results for MusicVid-YT8M when exploiting the potential shortcut.

	Median Rank ↓		Recall ↑						
			V→M			M→V			Average
	V→M	M→V	R@1	R@5	R@10	R@1	R@5	R@10	R@10
Baseline + CLIP and DeepSim features	5	5	20.31	50.96	67.43	23.49	55.52	70.56	71.64
MVPt (ours)	**1**	**1**	48.07	85.44	90.13	48.08	85.74	90.39	**90.26**
Chance	48	48	3.01	9.86	17.40	3.01	9.86	17.40	17.40

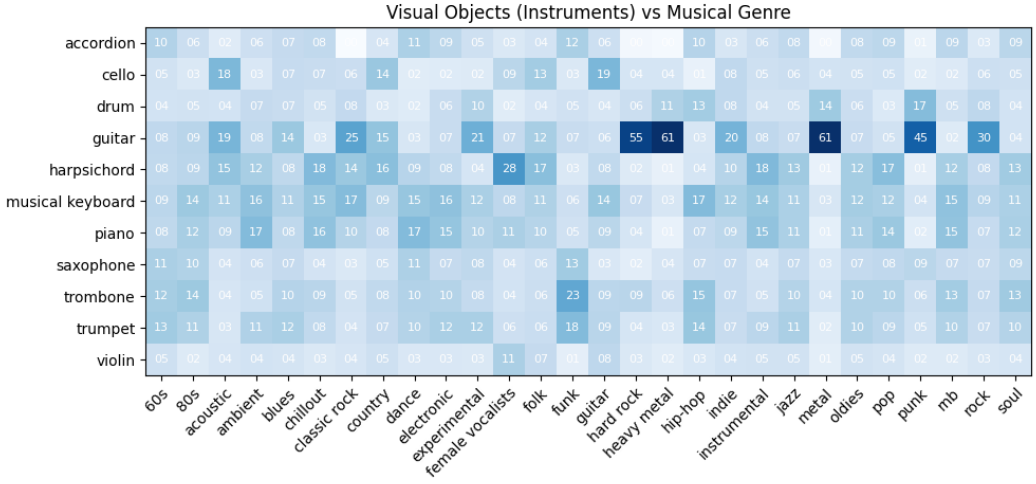

Figure 15. Instrument vs genre. This figure is equivalent to Figure 4 in the main paper, but showing instrument targets, instead of gender or race. The most retrieved instrument is the guitar, suggesting that the model has a good representation of guitars, and therefore we can use guitars for our conditioning experiments.

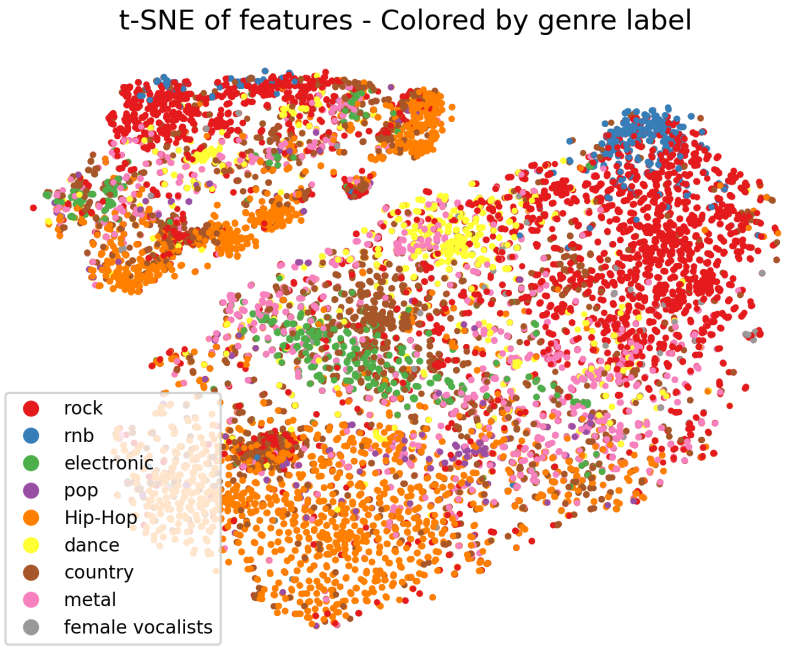

Figure 16. We project the music features into a 2-dimensional space using t-SNE [62], and color the data points by genre label. Music tracks are generally clustered by genre.