Classifying Audio Timbre Without Audio Using Text-Only Training

Peter McCabe(✉) and Patrick J. Donnelly

Oregon State University, Corvallis, OR 97330, USA
{mccabepe,donnellp}@oregonstate.edu

Abstract. Text-only training is a promising machine learning paradigm for training multimodal models without requiring data from every modality. However, despite the potential of text-only training, to date few studies have explored its use as an approximation of missing data for supervised learning in data-scarce environments. In this work we define the kinds of problems suited for text-only training and examine techniques to acquire or design text-based training data. We address the modality gap's role in the performance of text-only training, and present a case study on classifying subjective audio timbre descriptions based on three kinds of text-only training data and six augmentation methods on eight audio-timbre datasets. Our work shows that the text-only training paradigm successfully trains audio classifiers without audio and opens the door to future work in examining the effectiveness of text-only training for supervised machine learning problems without available datasets.

Keywords: Timbre · Machine Learning · Text-only Training

1 Introduction

Large multimodal models have been shown to be very effective at a wide variety of tasks in the machine learning community. Lately, an increasing body of research has explored using text as the only data modality during training (text-only training), despite its use in multimodal applications at inference time. This strategy is highly desirable because it circumvents the need for large datasets annotated across multiple modalities. Acquiring text data is often far cheaper and available at much larger scales than other modalities like images, audio, or video. While most research in this area has focused on tasks like captioning [8,36], retrieval [21], and zero-shot classification [17], little work has been done to explore this promising method for direct use in downstream supervised learning. These tasks naturally allow a practitioner to use paired text data, such as captions, metadata, etc., as both labels and input in a self-supervised fashion [8,36], but little work has been done exploring text-only training outside of this narrow context. Instead, we investigate a common frustration encountered by researchers: lacking necessary data. This problem motivates our work: *how effective is Text-only Training at solving supervised machine learning problems*

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2026
P. Machado et al. (Eds.): EvoMUSART 2026, LNCS 16523, pp. 347–366, 2026.
https://doi.org/10.1007/978-3-032-24350-8_23

in which there are no pre-existing supervised data? To explore this question we present a case study on the task of audio timbre classification.

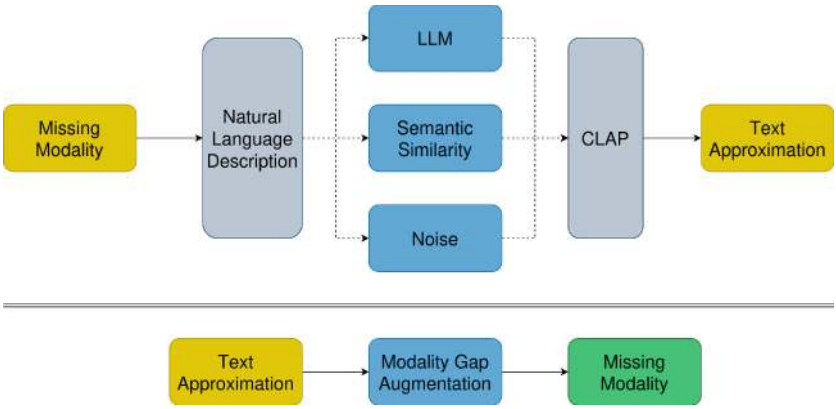


Fig. 1. By describing the desired attributes of the missing modality with words, CLAP [70] can approximate the absent modality. Then, by bridging the gap between its latent text and audio embedding spaces, one increases the fidelity of the approximation.

We must first define text-only training (ToT) and describe the kinds of problems it is well-suited to solve. An overview is provided in Fig. 1. ToT refers to the method of using only text data to train machine learning models by leveraging a pretrained multimodal model capable of comparing data from different modalities. By using the most similar text data to the source material in lieu of the original, one may approximate the same data in the missing modality. In other words, instead of using real audio data, one could use the text data that is most similar to it according to the chosen multimodal model. Therefore, the kinds of supervised learning problems best-suited to ToT have three criteria: (1) there is a lack of labeled data in the original modality, (2) the source modality must have a pretrained multimodal model able to measure semantic similarity, and (3) the desired property or properties should be well-described by language.

First, one can expect the ToT process to be noisy, and therefore would prefer to use cleaner data in the desired source modality. Second, ToT requires a semantic connection between text and desired modalities, usually in the form of a pretrained multimodal model. Third, in order to best approximate it, the desired data's key property (e.g., label, structure, value, etc.) should be easily and fully describable with natural language. In other words, as little information is lost when describing the property with words as possible.

For our case study, we explore classifying subjective labels of non-music audio timbre. This challenge involves the prediction of what a human might typically describe a non-musical sound "sounding like" or the defining characteristic or description of a sound. This task meets the above criteria because there are no large datasets available, there are several multimodal contrastive audio-language

models suitable for our task [14,18,70], and there are a multitude of ways to describe the timbre of sound [4]. We suggest that this task represents a typical use-case for the ToT paradigm, and seek to show the potential efficacy, pitfalls, and practitioner decisions required to make the paradigm succeed. In summary, this case study seeks to answer our research question via the proxy task of non-musical audio timbre classification.

1.1 Background

Together with rhythm, pitch, and volume, timbre is a fundamental aspect of sound that allows humans to differentiate between multiple sounds. In other words, a sound's timbre, or tone color, is usually described as what a sound "sounds" like. While often used to describe the particular characteristics of instruments [22], musical moods and themes [15], or genre [61], every sound has a describable timbre. For example, one might describe the sound of a spoon striking a crystal glass as "bright" or "ringing" [53]. These everyday sounds and subjective descriptors serve as the data and labels for our experiments as they are often passed over by the machine learning community in favor of musical excerpts [3] or musical instruments [19], used as features for other tasks [2,57], or as only parts of long-form labels for captioning [9,35].

In this work we choose the LAION-CLAP implementation [70] of CLAP [13] as our pretrained audio-language model. CLAP is a popular multimodal Audio-Language Model based on the highly influential CLIP [43] model used in computer vision. At a high level, CLAP is trained to encode paired audio and text data as vectors, and predict matching pairs within a given training batch. At inference time, CLAP is used to measure how well an audio and text pair match semantically. Mathematically, CLAP seeks to minimize the symmetric InfoNCE [37] loss. This loss encourages the vector representations of paired or semantically similar audio and text vectors to be near each other in the CLAP embedding space, and to encourage dissimilar pairs to be far apart. A typical use is providing the names of classes, e.g., "A sound of a [class]", and selecting the class label which CLAP predicts to match each sound [14,70].

1.2 Modality Gap

Unfortunately, CLAP and models like it suffer from the effects of the Modality Gap, which refers to the phenomenon observed in multimodal models in which the embeddings from each modality reside in separate subspaces within the overall embedding space [28]. These subspaces are nearly or entirely disjoint.

Despite being trained to restrict cross-modal paired data to be as close as possible and all other data as far as possible, the CLAP embedding space has two distinct clusters of text and audio embeddings. To show CLAP's modality gap, we follow [56] to train a linear classifier to predict the embedding modality on the Clotho [12] dataset. We achieve > 0.998 F_1-score on every trial of a 10-fold cross validation experiment, indicating the text and audio embeddings are almost entirely linearly separable. With respect to ToT, this means that naïvely

using the text embedding instead of the audio embedding for any given pair will not lead to identically trained models. In essence, the modality gap is a specific form of domain adaptation which necessitates methods for mitigating its effect and motivates some of our experiments.

2 Related Work

To better understand the process of ToT and closing the modality gap, we review several pertinent studies. We also summarize relevant research about the perception of timbre and attempts to automatically quantify it.

Text-only training has mostly been explored in the computer vision (CV) community, and has been applied to a variety of CV tasks, including: image captioning [68,73,74], generation [45,79], zero-shot classification [17,71], multimodal retrieval [21], visual question answering [59], category discovery [67], and improving multimodal embedding model alignment [30,72]. In the audio domain, ToT has been used in targeted sound extraction [31,32,51], captioning [25,75], and retrieval [20]. We are inspired by [8] who train an audio captioning model with only text with augmented captions as input instead of audio.

Without addressing the modality gap, ToT faces significant performance degradation. Broadly, these methods tend to be supervised [10,11,43] or unsupervised [17,36,79] mappings that act directly on one modality's embeddings to approximate embeddings in the other, architecture or optimization changes [1,56,63], or changes to the underlying embedding space [5,44,78]. As our work focuses on utilizing pre-trained models, we focus on the first category.

In one of the first works on approximating cross-modal embeddings, Radford *et al.* train a diffusion model to convert between text and image embeddings for image generation [43]. After introducing the modality gap, Liang *et al.* mitigate the gap by vector addition [28]. Nukrai *et al.* compare that with adding Gaussian noise and prefer the latter approach [36], while Tam *et al.* find that it is preferred over using optimal transport or PCA features [60]. Saijo *et al.* compare several gap-closing methods, including Gaussian noise and PCA features, and find that adding noise is the superior method [51]. Similarly, Role *et al.* find both optimal transport and PCA features inferior to a spectral embedding method [50].

The study of human description and perception of timbre is most often found in the musicology [26,48] and psycho-acoustic [4,57,66] literature. Relatively few studies in the machine learning community have focused on the prediction or linguistic description of timbre, choosing to use the timbre of sounds as features rather than labels [2,19,57]. Mannone *et al.* cluster listener associations with timbre and color [33], and Reymore *et al.* differentiate between fine-grained timbral descriptors of noisy instruments [47]. Mapping timbre has been investigated in the context of digital music production [29,54,64] and audio effects [6]. Audealize [55] uses a crowd-sourcing approach to learn connections between words and digital audio effects. Directly predicting timbral descriptors has been performed in K-Pop vocal styles [23], musical instruments [22,49], and general audio [38].

With the rise of LLMs, research into how well LLMs "understand" concepts like timbre has developed [34]. Saitis *et al.* [52] and Siedenburg *et al.* [58] compare

human and LLM timbre ratings of instruments and find similar intra-rater agreements as well as moderate rating correlations. Czedik-Eysenberg *et al.* measure the similarities between human perception of three timbral descriptors and popular audio feature extraction software libraries and five variants of CLAP [7]. Tian *et al.* perform a similar analysis between 18 representations [62]. Finally, Velissaridis *et al.* compare language and audio-language models equipped with timbral descriptors in instrument classification and emotion recognition [65].

3 Datasets

Because we investigate the scenario in which there is no available data whatsoever, arguably the most important decision a practitioner of ToT must make is how to source the text data. Depending on the exact nature of the desired dataset, there are numerous ways to create text-based synthetic data. Wang *et al.* compare methods for closing the modality gap using four datasets with different levels of multimodality [69]. Qi *et al.* scrape text from the web and use filter models to remove unwanted data [42]. A similar approach uses a captioning model and filter to generate synthetic captions for both images [27,78] and audio [24]. As we do in this study, LOVM [71] directly generates text data via LLM by prompting for descriptions based on class labels. Before we discuss the methods we choose, we first introduce our validation datasets.

3.1 Audio Commons

The principal datasets used in our work are the AudioCommons (AC) Timbral Characterisation Tool v0.2 Development Dataset(s) [38,39] developed to create the `Timbral Models` python package [40]. This tool is designed to measure how well seven different timbral descriptors match a given audio recording, as well as to detect the presence of reverb. The dataset is comprised of eight subsets corresponding to each descriptor. For each descriptor a collection of sounds was sourced from FreeSound[1]. Next, for all words[2], human annotators were presented with a set of eight sounds and rated each on a scale from 0–100 on how well that word described the sound relative to the others.

To simplify the creation of new text data (described below), we binarize the continuous-valued annotations into two classes. Additionally, by reducing the difficulty of the downstream task, we can clearly see the effects of ToT on an otherwise simple problem. First, we calculate the mean of all annotations, then set all values less than 50 as the negative class (e.g., *Not Bright*) and those 50 or greater as the positive class (e.g., *Bright*). Finally, we compute the CLAP embeddings for all audio files, and use the pairs of embedding and binary label as our testing data. In Table 1, we provide the specific details on dataset sizes and label distributions of the eight binary datasets.

[1] https://freesound.org/, last accessed 2026/02/01.
[2] The *Reverb* set was created separately by the authors of [40].

Table 1. Sizes and class balance of each timbral descriptor audio dataset. †We use half of the reverb dataset ($n \approx 400$) for training and testing the audio baseline and the other for development and validation.

	Booming	**Bright**	**Deep**	**Hard**	**Reverb†**	**Rough**	**Sharp**	**Warm**
Count	102	210	218	206	220	212	165	184
% Positive	30	62	61	67	53	55	60	42

3.2 Pseudo-labeling Real Captions

Now that we have established our problem and class labels, we can explore different methods for generating labeled text data. The first of the three dataset sourcing methods we explore in this work is labeling real captions with text-based semantic similarity models. The process consists of labeling real audio captions by comparing each caption with a predetermined sentence describing the label. In this work we call these "Anchor Prompts". We write five versions of our anchor prompt to reduce bias and average their embeddings to create a central pseudo-prompt, which we list in Table 2.

Table 2. The prompts we used for semantic similarity labeling with CLAP and SBERT. Each prompt is embedded into the model's embedding space and the mean is calculated and re-normalized for use as the comparison pseudo-prompt.

A {label} sound.
A sound that could be described as: {label}
An audio clip with a {label} quality.
An audio clip that sounds {label}.
A sound/audio clip that has or contains {label}.

As in works like [43], each label's mean anchor prompt embedding is then compared with the caption's embedding vector and the class with the highest similarity is selected. We use two large models that capture semantic relationships between text pairs: LAION-CLAP [70] and the SBERT [46] model all-MiniLM-L12-v2. CLAP is the natural choice for audio-related text-text similarity, since it is explicitly trained to match text and audio. This also represents the natural choice for a practitioner already using CLAP as an encoder to also use it as the pseudo-labeler in their workflow. SBERT juxtaposes CLAP as a text-only semantic similarity model trained on general text data. While similarly designed to correctly identify semantically-similar text pairs, SBERT was not necessarily intended to be used on audio domain-specific tasks, and is selected to compare strategies of using a domain-specific similarity versus a general language similarity model.

We choose WavCaps [35] as our caption dataset because it is both large and widely used in training multimodal models for tasks like audio captioning.

Because much of the data used in WavCaps was also used to train LAION-CLAP, we can be more confident that the concepts that the captions describe are in CLAP's distribution, and the label predictions are likely more accurate than otherwise [70]. It consists of over 400K partially-human annotated caption-audio pairs, and the final distributions of labeled data for each model are shown in Fig. 2. None of the test audio is present in WavCaps.

The advantage to this method is that it uses real text data and is cheaper than sourcing new data altogether. However, this method relies on the performance of the model(s) chosen to label the data. As can be seen in Fig. 2, the bias of the chosen model can dramatically alter the final dataset's distribution. Additionally, because label noise is particularly harmful to machine learning models, one must rely on finding large amounts of real data to overcome the noise present in the labeling process, which may not be available.

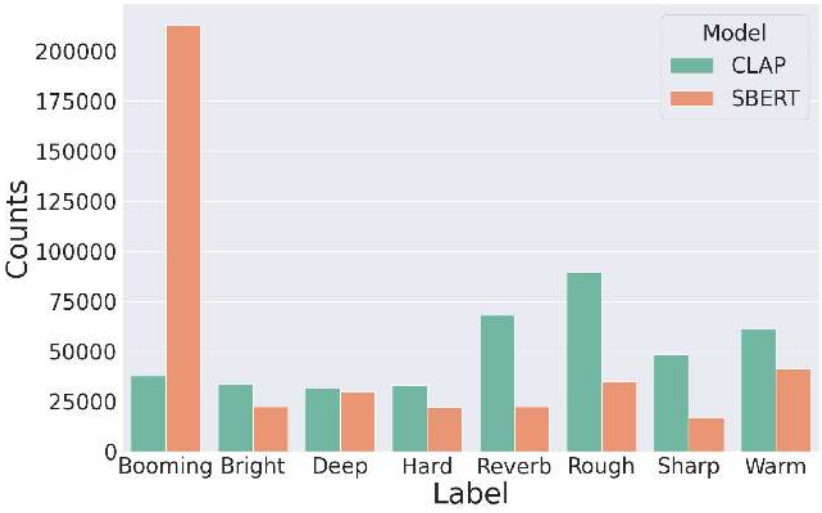


Fig. 2. The class label distributions of WavCaps [35] as determined by text-text similarity between captions and a label prompt. We compare CLAP [70], an audio-language model, and SBERT [46].

3.3 LLM Captions

To contrast with labeling existing captions, another natural choice is to use one of the widely available LLMs to generate pseudo-captions based on a desired class [71]. To test this method, we employ the `Llama-3.1-8B-Instruct` [16] LLM. For each label in the AC dataset we prompt the LLM to generate descriptions of audio events that might be described using that label. For each label we generate

approximately 15k samples after removing duplicates, and assign a class label based on the prompted timbral descriptor.

One might prefer this method over labeling real data because one could theoretically generate an arbitrarily large amount of data. However, generating high-quality text data using LLMs is expensive and ultimately not as reliable as real data. This method suffers from the same possible label noise mentioned above, because the generated text might not actually correspond to the provided label as desired. For example, we empirically found that, despite selecting hyperparameters that promoted varied responses, many of the generated captions described the same scene in slightly different words. This limits the overall distribution of captions and corresponds to an effective dataset size smaller than it seems. And just as we saw above, the exact output distribution depends on the model chosen, but with the added burden of tuning hyperparameters.

3.4 Label-Centered Gaussian Noise

For our final method, we explore training purely on adding noise to our anchor prompts. This way, a practitioner need not bother with the expensive processes of pseudo-labeling or generating captions from scratch. Inspired by [77], who model discrepancies in modality alignment as Gaussian noise, we utilize the anchor prompts and their mean embedding as the centers of six separate Gaussian distributions with a shared variance we measure following the procedure in [8, 36]. In other words, each example is constructed as a mixture of six Gaussians, each with the same variance, but centered around different anchor prompts. We generate approximately 25k samples for each label.

This method is the least labor-intensive of the three we present and is easily scalable to an arbitrary size. However, it assumes that the embedding model has a general "understanding" of the classes in both modalities, and that the anchor prompt(s) act as a good measure of the mean of each class's distribution in the text space. Luckily, we already assume our embedding model has at least some correlation between the desired data and descriptive text labels, otherwise the contrastive-learning objective underpinning the model would have failed. Since we can empirically see the effectiveness of the CLAP model and its variants [18,32], we are able to implicitly verify our assumption.

4 Methodology

In this section, we discuss methods for addressing the modality gap, our experimental design, and the comparative baselines we select.

4.1 Bridging the Modality Gap

Bridging the modality gap is a specific form of domain adaptation, meaning applying machine learning to one's out-of-domain dataset will likely result in poor performance. Therefore, it is critical that a practitioner of ToT address

the modality gap problem. While there are many methods used in practice, we focus on common baselines and the methods with the most theoretical backing. We compare (1) doing nothing, akin to not performing domain adaptation, (2) with the method first explored by Liang *et al.* in their work that introduced the modality gap [28], and (3) the method described by Zhang *et al.* [76,77]. Liang *et al.*'s method (Mean Shifting) involves measuring the mean difference vector between each modality, then using that mean vector to shift samples from one modality towards the other. The mean difference vector is defined as:

$$\Delta_{gap} = \frac{1}{n}\sum_{i=1}^{n} a_i - \frac{1}{n}\sum_{i=1}^{n} t_i \tag{1}$$

where a_i and t_i refer to the i_{th} audio and text embedding respectively. Then, by adding the mean vector to the text embeddings, one can approximate the paired audio vectors: $\hat{a}_i = t_i + \Delta_{gap}$. We store the mean audio embedding of the Clotho dataset [12], and calculate the mean of each text dataset during training.

Our third method is taken from Zhang *et al.* (named C2). They proved under certain empirically verified assumptions that their method, C2, removes the modality gap [77]. C2 involves centering each modality at zero by subtracting the mean embedding vector from each modality during both training and inference. Similar to Mean Shifting, C2 shifts one modality towards the other, but while Mean Shifting pushes one modality towards the empirical mean of the other, C2 shifts both modalities towards zero.

As discussed by Zhang *et al.* the two modalities are nominally aligned at zero, but suffer from "alignment noise" that prevents paired text and audio from being perfectly aligned [77]. To account for alignment noise, the authors introduce "C3" which expands on C2 by injecting Gaussian noise into the training data to mimic additional noisy samples centered near the real data and to train a model robust to noisy data. This also falls well within the pattern of the literature, in which many authors use added Gaussian noise or other random methods for bridging the gap [8,36,51]. For a fair comparison, we investigate adding noise to all three methods listed above: Nothing, Mean Shifting, and C2. We label these as Gaussian Noise, Shift+Noise, and C3 [77].

4.2 Experiments

For each combination of training dataset, data augmentation strategy, and target dataset we perform the following experiment: (1) Randomly sample a balanced training dataset by undersampling the negative classes to match the number of positive examples[3], (2) Augment the data with the augmentation strategy[4]. (3) Calculate the F_1-score on the target audio dataset. We run every experiment 10 times and report the mean and standard deviation of test F_1-scores.

[3] Recall that there are eight total classes but we perform a series of binary classifications.

[4] Note that we add noise only once; we find repeatedly adding noise makes little difference.

Additionally, we perform ablations on training with the full-but-imbalanced training sets and combinations thereof. Training on the large imbalanced datasets accounts for any performance changes that might occur due to the undersampling used in the main experiments, but suffers from significant class imbalances. Then, we train on combinations of the full datasets to examine if one or more strategies compensates for the shortcomings of the other(s).

We select our classification model based on the theoretical guarantees provided by Zhang *et al.* [76]. Under the same conditions as C2, they prove that any regularized linear classifier minimizing a quadratic loss will be identical when trained on either modality. This means that under perfect conditions the model trained on text would exactly match one trained on real audio. We select the RidgeClassifier from `scikit-learn` [41] with default hyperparameters as our model. We utilize CLAP's zero-shot similarity scores between the test audio datasets and the mean embedding of our anchor prompts as a common baseline. For each dataset we calculate the mean of the negative class anchor prompt embeddings and select the larger CLAP scores between the positive and mean negative prompt embedding as the predicted class. Finally, we compare these scores against those produced by simply using the small real audio datasets.

5 Results and Discussion

First, we examine the differences in performance between each dataset creation method. In Fig. 3, we observe that the CLAP labeling of WavCaps and Label-Centered Noise tend to outperform both the LLM-generated captions and the SBERT labeling. Unsurprisingly, the two best methods are based around the CLAP embedding space itself. The Label-Centered Noise dataset assumes that the audio clips that have the described timbral characteristic are embedded in the CLAP space near the anchor prompt, while the WavCaps-CLAP dataset relies on the captions that display semantic similarity to the anchor prompts to be embedded nearby. The former approach relies on between-modality similarity, while the latter relies on within-modality similarity. Viewed through that lens, the better performance on the Label-Centered Noise dataset could be explained by the effects of the modality gap itself, because within-modality similarities are on average higher than cross-modal similarities [28]. Additionally, the WavCaps-CLAP dataset may contain more mislabeled data as captions with similar meanings in the CLAP embedding space are confused with each other.

As shown in Fig. 4, the LLM-generated dataset appears to perform particularly poorly in the audio modality, despite scoring highly during training. It exhibits the greatest deviation between train and test scores, indicating that its examples were easily separable, but failed to emulate real audio data. This could be explained by the kind of language model we chose. By following existing work [71] and using a general language model for dataset creation, we sacrifice domain-specific knowledge that might be available in a specialized LLM. The SBERT labeling exhibits the opposite behavior: underperforming in train scores, but occasionally competing for the second best scores on the test sets. This is somewhat surprising due to the SBERT dataset's heavy bias towards the *Booming*

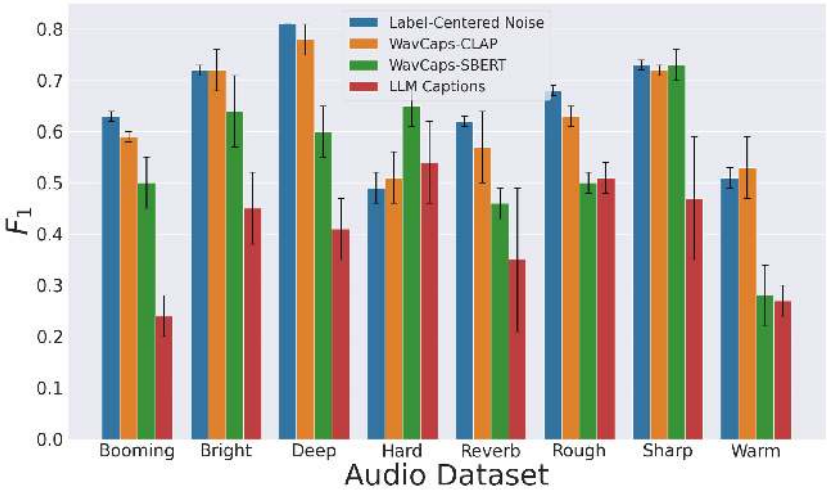


Fig. 3. Mean and std. of test scores averaged across data augmentation methods.

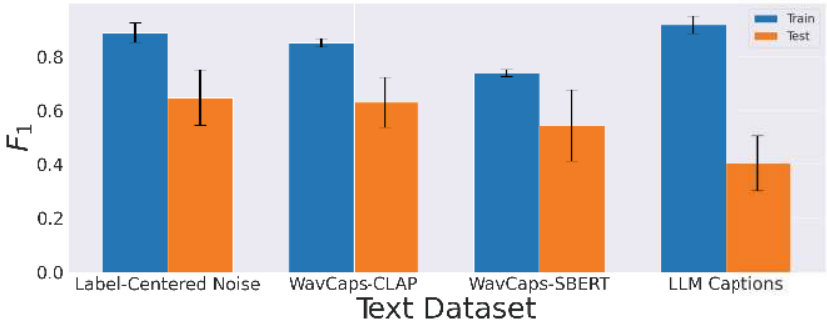


Fig. 4. Mean and std. of train and test scores across all augmentations and targets.

class (see Fig. 2). Since *Booming* is a minority class in the audio modality, one might expect SBERT's labeling to fail to achieve success on real audio.

Next, we analyze the impact of each gap-closing method (see Fig. 5). As stated above, applying no data augmentation method generally harms performance. This is particularly clear in the *Hard* and *Reverb* datasets, in which performance drops significantly when using no data augmentation. C2 and Mean Shift exhibit approximately comparable performance, although Mean Shift does have marginally more variance in its scores. When noise is added, C3 achieves the highest test scores for the majority of audio datasets, but does not show any major improvements. As seen in Table 3, only the change from performing no data augmentation to adding Gaussian noise showed any noteworthy improvement.

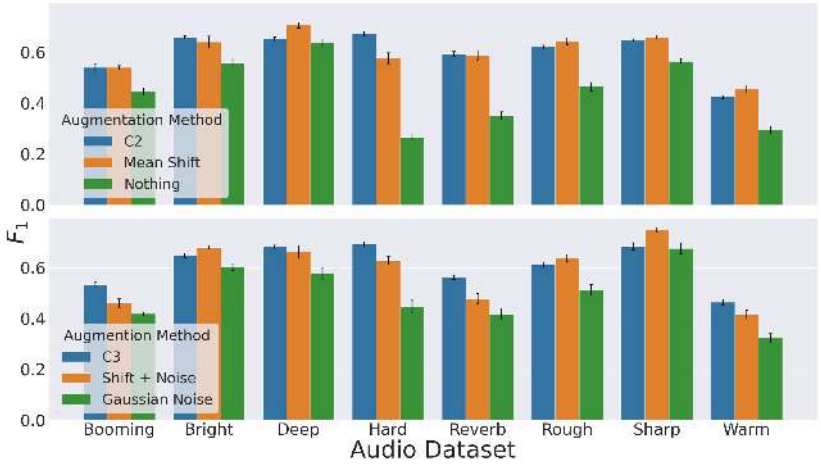


Fig. 5. Mean test scores using methods without (top) and with (bottom) added noise.

Additionally, we explore the effects of the size of the dataset in an ablation study. In Fig. 6, we show additional experiments performed on the full and unbalanced datasets, as well as all combinations of the datasets. Recall that each constructed dataset has approximately equal numbers of each class before binarization, which means each binary dataset has seven times the number of negative examples as positive examples. Additionally, the class distributions of the pseudo-labeled datasets WavCaps-CLAP and WavCaps-SBERT are dependent on the models' outputs and vary widely between them (see Fig. 2).

Table 3. Mean and std. of change between noise-free and added-noise methods.

Method	**C3**	**Shift + Noise**	**Gaussian Noise**
Change	0.009 ± 0.025	-0.012 ± 0.065	0.050 ±0.070

We find that the average F_1-score for all large unbalanced datasets is lower than the average scores of the balanced datasets. Additionally, the variance of the scores across methods and target datasets is much higher. This indicates that naïvely creating large datasets may not be effective for ToT. Combining datasets also harms performance, indicating that dataset size does not overcome class imbalance. This is also likely caused by conflicting labels between datasets. Each method has its own distribution: some examples in the same class from a different datasets may appear very different. Combining datasets may increase the overall proportion of noisy labels.

Finally, Table 4 lists the average and best performing model scores for experiments with data augmentation. For comparison, we show CLAP zero-shot scores

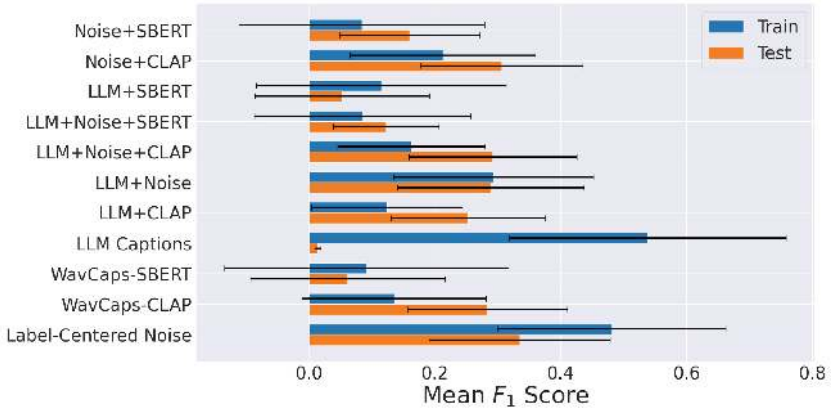


Fig. 6. Ablation study on training with large and unbalanced datasets.

and the mean cross-validation score from training on real audio datasets. While the average model scores are generally less than CLAP's zero-shot performance, we find the best performing model is always at least as good as CLAP, and is better in seven out of eight cases. Notably, for a majority of audio datasets, the best text-only models were within 0.05 F_1-score of the average model trained on real audio. More impressively, ToT meets or exceeds the performance of training on real audio on the *Bright* and *Sharp* datasets, respectively.

Table 4. Test F_1 scores of the mean and best models with data augmentation, CLAP zero-shot, and 10-fold cross validation on real audio. Best text-only results are in bold.

Dataset	Booming	Bright	Deep	Hard	Reverb	Rough	Sharp	Warm
Average Model	0.52	0.66	0.68	0.64	0.56	0.63	0.68	0.44
CLAP Zero-Shot	0.58	0.76	**0.85**	0.25	0.57	0.62	0.75	0.47
Best Model	**0.67**	**0.79**	**0.85**	**0.81**	**0.66**	**0.78**	**0.80**	**0.62**
Audio Datasets	0.72	0.79	0.89	0.85	0.82	0.81	0.78	0.77

Overall, we demonstrate that dataset creation techniques which leverage the embedding space required by text-only training tend to lead to better downstream performance. Addressing the modality gap is necessary for ToT to succeed, but the exact method is not as important. Adding noise can help but is less effective than other methods at bridging the modality gap. Despite label noise, training on small balanced datasets proved to be more successful than training on large and unbalanced data. Altogether, we find that text-only training has the potential to outperform CLAP and training on the audio itself.

6 Future Work and Conclusion

While we demonstrated that text-only training can be effective under the right circumstances, there are still many avenues of future research to explore. In this work we limited ourselves to binary classification to simplify and align the downstream task with text-only training. However, timbre is inherently multi-label and continuous-valued, and future work in timbre classification should explore these more difficult versions of the problem. As our study serves as an introduction to the problem and not as a definitive source for the state of the art, we do not perform hyperparameter tuning on models and noise, nor claim to have found the best-performing paradigm for text-only training. Likewise, we only explored common augmentation techniques and classification models with theoretical backing, but in practice, text-only training is most often used in conjunction with deep learning architectures [8,32].

We define three criteria for problems that best suit text-only training, but provide no empirical justification. As future work, we intend to explore the effect of how much mutual information is shared between the missing modality and the language used to describe it. In this work, we did not compare using different LLMs to generate training data, or different embedding models. As mentioned above, each LLM could have a different biased output distribution with respect to desired properties, and each embedding model has its own biases that might affect performance on downstream tasks. Yet, even when using the same LLM, there are more variations of dataset creation methods to be explored. For simplicity we restricted our generated datasets to have labels that matched our validation sets, but in the real world many different descriptors might describe a sound. For that reason we recommend future work to investigate using multiple labels as prompts, out-of-domain labels as negative examples, and other techniques to better mimic real data. While we explored three methods for creating text data, many more approaches are possible. For example, we did not exhaustively explore methods for adding noise. As future work, we intend to consider Gaussian noise with non-constant covariance matrices, as there is no guarantee of non-correlated alignment noise.

In this study we introduce a text-only supervised learning paradigm and discuss the challenges of dataset creation and addressing the modality gap. We present a case study on audio timbre classification and compare three training data generation techniques and six augmentation strategies on eight timbre datasets. We show that leveraging the embedding model's inherent knowledge is superior to using dedicated text-text similarity methods or generating text using an LLM, and that addressing the modality gap is a necessary step in the process. Finally, we show that text-only training is a viable machine learning paradigm by meeting or exceeding the performance of models trained on real audio.

References

1. Al-Jaff, M.: Messing with the Gap: On the Modality Gap Phenomenon in Multimodal Contrastive Representation Learning. Ph.D. thesis, Uppsala University (2023)
2. Aucouturier, J., Pachet, F.: Improving timbre similarity: how high's the sky? J. Negat. Res. Speech Audio Sci. (JNRSAS) (2004)
3. Bogdanov, D., Minz, W., Tovstogan, P., Porter, A.: Mtg-jamendo dataset (2019). https://doi.org/10.5281/ZENODO.3826813
4. Carron, M., Rotureau, T., Dubois, F., Misdariis, N., Susini, P.: Speaking about sounds: a tool for communication on sound features. J. Des. Res. **15**(2), 85 (2017). https://doi.org/10.1504/JDR.2017.086749
5. Chen, H., He, L., Liu, Y., Yang, L.: Visual neural decoding via improved visual-EEG semantic consistency (2024). https://doi.org/10.48550/arXiv.2408.06788. Accessed 26 Jun 2025
6. Chu, A., O'Reilly, P., Barnett, J., Pardo, B.: Text2FX: harnessing clap embeddings for text-guided audio effects. In: ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5 (2025). https://doi.org/10.1109/ICASSP49660.2025.10890334, iSSN: 2379-190X
7. Czedik-Eysenberg, I., Reuter, C.: Evaluating timbre-related audio descriptors across different libraries and multimodal embeddings. In: Proceedings of DAS|DAGA 2025 (2025). https://doi.org/10.71568/dasdaga2025.657
8. Deshmukh, S., Elizalde, B., Emmanouilidou, D., Raj, B., Singh, R., Wang, H.: Training audio captioning models without audio. In: ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Republic of Korea, pp. 371–375. IEEE (2024). https://doi.org/10.1109/ICASSP48485.2024.10448115
9. Doh, S., Choi, K., Lee, J., Nam, J.: LP-MusicCaps: LLM-based pseudo music captioning. In: Proceedings of the 24th International Society for Music Information Retrieval Conference, pp. 409–416 (2023)
10. Dong, H.W., et al.: CLIPSonic: text-to-audio synthesis with unlabeled videos and pretrained language-vision models. In: 2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA, pp. 1–5. IEEE (2023). https://doi.org/10.1109/WASPAA58266.2023.10248160
11. D'Orazio, A., Briglia, M.R., Crisostomi, D., Loi, D., Rodolà, E., Masi, I.: Implicit inversion turns CLIP into a decoder (2025). https://doi.org/10.48550/arXiv.2505.23161. Accessed 26 Jun 2025
12. Drossos, K., Lipping, S., Virtanen, T.: Clotho: an audio captioning dataset. In: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, pp. 736–740. IEEE (2020). https://doi.org/10.1109/ICASSP40776.2020.9052990
13. Elizalde, B., Deshmukh, S., Ismail, M.A., Wang, H.: CLAP learning audio concepts from natural language supervision. In: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5 (2023). https://doi.org/10.1109/ICASSP49357.2023.10095889, ISSN: 2379-190X
14. Elizalde, B., Deshmukh, S., Wang, H.: Natural language supervision for general-purpose audio representations. In: ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Republic of Korea, pp. 336–340. IEEE (2024). https://doi.org/10.1109/ICASSP48485.2024.10448504

15. Ferrer, R., Eerola, T.: Semantic structures of timbre emerging from social and acoustic descriptions of music. EURASIP J. Audio Speech Music Process. **2011**(1), 11 (2011). https://doi.org/10.1186/1687-4722-2011-11
16. Grattafiori, A., et al.: The Llama 3 herd of models (2024). https://doi.org/10.48550/arXiv.2407.21783. Accessed 25 Oct 2025
17. Gu, S., Clark, C., Kembhavi, A.: I can't believe there's no images!: learning visual tasks using only language supervision. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 2672–2683 (2023). https://doi.org/10.1109/ICCV51070.2023.00252, ISSN: 2380-7504
18. Guinot, J., Riou, A., Quinton, E., Fazekas, G.: SLAP: Siamese language-audio pretraining without negative samples for music understanding (2025). https://doi.org/10.48550/arXiv.2506.17815, ISMIR 2025, Accessed 26 Jun 2025
19. Herrera-Boyer, P., Peeters, G., Dubnov, S.: Automatic classification of musical instrument sounds. J. New Music Res. **32**(1), 3–21 (2003). https://doi.org/10.1076/jnmr.32.1.3.16798
20. Hu, R., et al.: Vela: scalable embeddings with voice large language models for multimodal retrieval. In: Interspeech 2025, pp. 2640–2644. ISCA (2025). https://doi.org/10.21437/Interspeech.2025-159
21. Jang, Y.K., Lim, S.N.: Towards cross-modal backward-compatible representation learning for vision-language models (2024). https://doi.org/10.48550/arXiv.2405.14715. Accessed 26 Jun 2025
22. Jiang, W., Liu, J., Li, Z., Zhu, J., Zhang, X., Wang, S.: Analysis and modeling of timbre perception features of Chinese musical instruments. In: 2019 IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS), Beijing, China, pp. 191–195. IEEE (2019). https://doi.org/10.1109/ICIS46139.2019.8940168
23. Kim, K.L., Lee, J., Kum, S., Park, C.L., Nam, J.: Semantic tagging of singing voices in popular music recordings. IEEE/ACM Trans. Audio Speech Lang. Process. **28**, 1656–1668 (2020). https://doi.org/10.1109/TASLP.2020.2993893
24. Kong, Z., et al.: Improving text-to-audio models with synthetic captions (2024). https://doi.org/10.48550/arXiv.2406.15487. Accessed 29 Oct 2024
25. Kouzelis, T., Katsouros, V.: Weakly-supervised automated audio captioning via text only training. In: Proceedings of the 8th Detection and Classification of Acoustic Scenes and Events 2023 Workshop (DCASE2023), Tampere, Finland, pp. 81–85 (2023)
26. Lavengood, M.: A New Approach to the Analysis of Timbre. Ph.D. thesis, CUNY Academic Works (2017)
27. Li, J., Li, D., Xiong, C., Hoi, S.: BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., Sabato, S. (eds.) Proceedings of the 39th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 162, pp. 12888–12900. PMLR (2022). https://doi.org/10.48550/arXiv.2201.12086. Accessed 26 Jun 2025
28. Liang, W., Zhang, Y., Kwon, Y., Yeung, S., Zou, J.: Mind the gap: understanding the modality gap in multi-modal contrastive representation learning. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. NIPS '22, Red Hook, NY, USA, pp. 17612–17625. Curran Associates Inc. (2022)
29. Limberg, C., Schulz, F., Zhang, Z., Weinzierl, S.: Pitch-conditioned instrument sound synthesis from an interactive timbre latent space. In: Gabrielli, L., Cecchi,

S. (eds.) Proceedings of the 28-th International Conference on Digital Audio Effects (DAFx25), Italy, pp. 433–440. Ancona (2025)
30. Liu, Z., Liu, J., Ma, F.: Improving cross-modal alignment with synthetic pairs for text-only image captioning. In: Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence. AAAI'24/IAAI'24/EAAI'24, vol. 38, pp. 3864–3872. AAAI Press (2024). https://doi.org/10.1609/aaai.v38i4.28178
31. Ma, H., et al.: Language-queried target sound extraction without parallel training data. In: ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, pp. 1–5. IEEE (2025). https://doi.org/10.1109/ICASSP49660.2025.10887543
32. Ma, H., Peng, Z., Li, X., Shao, M., Wu, X., Liu, J.: CLAPSep: leveraging contrastive pre-trained model for multi-modal query-conditioned target sound extraction. IEEE/ACM Trans. Audio Speech Lang. Process. **32**, 4945–4960 (2024). https://doi.org/10.1109/taslp.2024.3497586, publisher: Institute of Electrical and Electronics Engineers (IEEE)
33. Mannone, M., Distefano, V., Santini, G.: Classes of colors and timbres: a clustering approach. Electron. J. Appl. Stat. Anal. **15**(3), 588–605 (2022). https://doi.org/10.1285/i20705948v15n3p588, number: 3
34. Marjieh, R., Sucholutsky, I., Van Rijn, P., Jacoby, N., Griffiths, T.L.: Large language models predict human sensory judgments across six modalities. Sci. Rep. **14**(1), 21445 (2024). https://doi.org/10.1038/s41598-024-72071-1
35. Mei, X., et al.: WavCaps: a ChatGPT-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. IEEE/ACM Trans. Audio Speech Lang. Process. **32**, 3339–3354 (2024). https://doi.org/10.1109/taslp.2024.3419446, publisher: Institute of Electrical and Electronics Engineers (IEEE)
36. Nukrai, D., Mokady, R., Globerson, A.: Text-only training for image captioning using noise-injected CLIP. In: Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, pp. 4055–4063. Association for Computational Linguistics (2022). https://doi.org/10.18653/v1/2022.findings-emnlp.299
37. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding (2019). https://doi.org/10.48550/arXiv.1807.03748. Accessed 26 Jun 2025
38. Pearce, A., Brookes, T., Mason, R.: Modelling timbral hardness. Appl. Sci. **9**(3), 466 (2019). https://doi.org/10.3390/app9030466
39. Pearce, A., Brookes, T., Mason, R.: Timbral Characterisation Tool v0.2 Development Dataset (2019). https://doi.org/10.5281/zenodo.2545496. Accessed 02 Aug 2024
40. Pearce, A., Safavi, S., Tim, B., Mason, Russell Wang, W., Plumbley, M.: D5.8: release of timbral characterisation tools for semantically annotating non-musical content (2019). https://audiocommons.github.io/materials/. Accessed 1 Feb 2026
41. Pedregosa, F., et al.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
42. Qi, D., Su, L., Song, J., Cui, E., Bharti, T., Sacheti, A.: ImageBERT: cross-modal pre-training with large-scale weak-supervised image-text data (2020). https://doi.org/10.48550/arXiv.2001.07966. Accessed 22 Jul 2024
43. Radford, A., et al.: Learning transferable visual models from natural language supervision (2021). https://doi.org/10.48550/arXiv.2103.00020. Accessed 30 Sept 2025

44. Ramasinghe, S., Shevchenko, V., Avraham, G., Thalaiyasingam, A.: Accept the modality gap: an exploration in the hyperbolic space. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, pp. 27253–27262. IEEE (2024). https://doi.org/10.1109/CVPR52733.2024.02574
45. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip Latents (2022). https://doi.org/10.48550/arXiv.2204.06125. Accessed 13 Jul 2025
46. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, pp. 3980–3990. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/D19-1410
47. Reymore, L., Beauvais-Lacasse, E., Smith, B.K., McAdams, S.: Modeling noise-related timbre semantic categories of orchestral instrument sounds with audio features, pitch register, and instrument family. Front. Psychol. **13**, 796422 (2022). https://doi.org/10.3389/fpsyg.2022.796422
48. Reymore, L., Huron, D.: Using auditory imagery tasks to map the cognitive linguistic dimensions of musical instrument timbre qualia. Psychomusicology: Music, Mind, and Brain **30**(3), 124–144 (2020). https://doi.org/10.1037/pmu0000263, place: US Publisher: Educational Publishing Foundation
49. Reymore, L., Noble, J., Saitis, C., Traube, C., Wallmark, Z.: Timbre semantic associations vary both between and within instrumentsan empirical study incorporating register and pitch height. Music. Percept. **40**(3), 253–274 (023). https://doi.org/10.1525/mp.2023.40.3.253, publisher: University of California Press
50. Role, F., Meyer, S., Amblard, V.: Fill the Gap: quantifying and reducing the modality gap in image-text representation learning (2025). https://doi.org/10.48550/ARXIV.2505.03703. Accessed 12 Apr 2025
51. Saijo, K., Ebbers, J., Germain, F.G., Khurana, S., Wiehern, G., Roux, J.L.: Leveraging audio-only data for text-queried target sound extraction. In: ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Hyderabad, India, pp. 1–5. IEEE (2025). https://doi.org/10.1109/icassp49660.2025.10888769
52. Saitis, C., Siedenburg, K.: When ChatGPT talks timbre. In: Proceedings of the Third International Conference on Timbre, pp. 73–77. The School of Music Studies Aristotle University of Thessaloniki, Aristotle University of Thessaloniki (2023)
53. Saitis, C., Weinzierl, S.: The semantics of timbre. In: Siedenburg, K., Saitis, C., McAdams, S., Popper, A.N., Fay, R.R. (eds.) Timbre: Acoustics, Perception, and Cognition, vol. 69, pp. 119–149. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-14832-4_5
54. Salas, J.: Generating music from literature using topic extraction and sentiment analysis. IEEE Potentials **37**(1), 15–18 (2018). https://doi.org/10.1109/MPOT.2016.2550015
55. Seetharaman, P., Pardo, B.: Audealize: crowdsourced audio production tools. J. Audio Eng. Soc. **64**(9), 683–695 (2016). https://doi.org/10.17743/jaes.2016.0037
56. Shi, P., Welle, M.C., Björkman, M., Kragic, D.: Towards understanding the modality gap in CLIP. In: ICLR 2023 Workshop on Multimodal Representation Learning: Perks and Pitfalls (2023). https://openreview.net/forum?id=8W3KGzw7fNI, last accessed 2025/10/27

57. Siedenburg, K., Fujinaga, I., McAdams, S.: A comparison of approaches to timbre descriptors in music information retrieval and music psychology. J. New Music Res. **45**(1), 27–41 (2016). https://doi.org/10.1080/09298215.2015.1132737
58. Siedenburg, K., Saitis, C.: The language of sounds unheard: exploring musical timbre semantics of large language models (2023). https://doi.org/10.48550/arXiv.2304.07830. Accessed 20 Jul 2025
59. Song, H., Dong, L., Zhang, W., Liu, T., Wei, F.: CLIP models are few-shot learners: empirical studies on VQA and visual entailment. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Dublin, Ireland. Association for Computational Linguistics (2022). https://doi.org/10.18653/v1/2022.acl-long.421
60. Tam, D., Raffel, C., Bansal, M.: Simple weakly-supervised image captioning via CLIP's multimodal embeddings. In: The AAAI-23 Workshop on Creative AI Across Modalities (2023). https://openreview.net/forum?id=UMxeP-FuwyC. Accessed 21 Jul 2025
61. Thompson, A.E.: Playing Tag: An Analysis of Vocabulary Patterns and Relationships Within a Popular Music Folksonomy. Master's thesis, University of North Carolina at Chapel Hill (2008). https://doi.org/10.17615/08PE-W163
62. Tian, H., Lattner, S., Saitis, C.: Assessing the alignment of audio representations with timbre similarity ratings (2025). https://doi.org/10.48550/arXiv.2507.07764, ISMIR 2025. Accessed 24 Sept 2025
63. Udandarao, V.: Understanding and Fixing the Modality Gap in Vision-Language Models. Master's thesis, University of Cambridge (2022)
64. Vaillant, G.L., Molle, Y.: Contrastive timbre representations for musical instrument and synthesizer retrieval (2025). https://doi.org/10.48550/arXiv.2509.13285. Accessed 24 Sept 2025
65. Velissaridis, G., Athwal, R., Musharaf, M., Fazekas, G., Saitis, C.: Evaluating foundation models on timbre-related cognitive tasks. In: 1st Workshop on Large Language Models for Music & Audio (LLM4MA) (2025). https://openreview.net/forum?id=tXyh8CY9kZ. Accessed 27 Oct 2025
66. Wallmark, Z.: Semantic crosstalk in timbre perception. Music Sci. **2**, 205920431984661 (2019). https://doi.org/10.1177/2059204319846617
67. Wang, E., Peng, Z., Xie, Z., Yang, F., Liu, X., Cheng, M.M.: GET: unlocking the multi-modal potential of CLIP for generalized category discovery. In: Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), pp. 20296–20306. arXiv (2025). https://doi.org/10.48550/arXiv.2403.09974, CVPR 2025. Accessed 26 Jun 2025
68. Wang, J., Zhang, Y., Yan, M., Zhang, J., Sang, J.: Zero-shot image captioning by anchor-augmented vision-language space alignment (2022). https://doi.org/10.48550/arXiv.2211.07275. Accessed 26 Jun 2025
69. Wang, Y., Luo, H., Xu, J., Sun, Y., Wang, F.: Text data-centric image captioning with interactive prompts (2024). https://doi.org/10.48550/arXiv.2403.19193. Accessed 26 Jun 2025
70. Wu, Y., Chen, K., Zhang, T., Hui, Y., Berg-Kirkpatrick, T., Dubnov, S.: Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In: ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, pp. 1–5. IEEE (2023). https://doi.org/10.1109/ICASSP49357.2023.10095969
71. Yi, C., He, Y.H., Zhan, D.C., Ye, H.J.: Bridge the modality and capability gaps in vision-language model selection. In: Globerson, A., et al. (eds.) Advances in Neural

Information Processing Systems. vol. 37, pp. 34429–34452. Curran Associates, Inc. (2024)
72. Yu, Y., et al.: Multimodal knowledge alignment with reinforcement learning. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2023). https://doi.org/10.48550/arXiv.2205.12630. Accessed 30 Jun 2025
73. Zeng, D., Shen, Y., Lin, M., Yi, Z., Ouyang, J.: Zero-shot image captioning with multi-type entity representations. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, no. 21, pp. 22308–22316 (2025). https://doi.org/10.1609/aaai.v39i21.34386, number: 21
74. Zeng, Z., Xie, Y., Zhang, H., Chen, C., Chen, B., Wang, Z.: MeaCap: memory-augmented zero-shot image captioning. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 14100–14110 (2024). https://doi.org/10.1109/CVPR52733.2024.01337, ISSN: 2575-7075
75. Zhang, Y., Xu, X., Du, R., Liu, H., Dong, Y., Tan, Z.H., Wang, W., Ma, Z.: Zero-shot audio captioning using soft and hard prompts. IEEE Trans. Audio Speech Lang. Process. **33**, 2045–2058 (2025). https://doi.org/10.1109/TASLPRO.2025.3567770
76. Zhang, Y., HaoChen, J.Z., Huang, S.C., Wang, K.C., Zou, J., Yeung, S.: Diagnosing and rectifying vision models using language. In: The Eleventh International Conference on Learning Representations (ICLR) (2023). https://openreview.net/forum?id=D-zfUK7BR6c. Accessed 30 Jun 2025
77. Zhang, Y., Sui, E., Yeung, S.: Connect, collapse, corrupt: learning cross-modal tasks with UNI-modal data. In: The Twelfth International Conference on Learning Representations (ICLR) (2023). https://openreview.net/forum?id=ttXg3SKAg5. Accessed 09 Jul 2025
78. Zhao, P., Luan, R., Zhang, W., Wu, P., He, S.: Guiding cross-modal representations with MLLM priors via preference alignment (2025). https://doi.org/10.48550/arXiv.2506.06970. Accessed 26 Jun 2025
79. Zhou, Y., et al.: Towards language-free training for text-to-image generation. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, pp. 17886–17896. IEEE (2022). https://doi.org/10.1109/cvpr52688.2022.01738