HARNESSING THE POWER OF DISTRIBUTIONS: PROBABILISTIC REPRESENTATION LEARNING ON HYPERSPHERE FOR MULTIMODAL MUSIC INFORMATION RETRIEVAL

Takayuki Nakatsuka **Masahiro Hamasaki** **Masataka Goto**
National Institute of Advanced Industrial Science and Technology (AIST), Japan
{takayuki.nakatsuka, masahiro.hamasaki, m.goto}@aist.go.jp

ABSTRACT

Probabilistic representation learning provides intricate and diverse representations of music content by characterizing the latent features of each content item as a probability distribution within a certain space. However, typical Music Information Retrieval (MIR) methods based on representation learning utilize a feature vector of each content item, thereby missing some details of their distributional properties. In this study, we propose a probabilistic representation learning method for multimodal MIR based on contrastive learning and optimal transport. Our method trains encoders that map each content item to a hypersphere so that the probability distributions of a positive pair of content items become close to each other, while those of an irrelevant pair are far apart. To achieve such training, we design novel loss functions that utilize both probabilistic contrastive learning and spherical sliced-Wasserstein distances. We demonstrate our method's effectiveness on benchmark datasets as well as its suitability for multimodal MIR through both a quantitative evaluation and a qualitative analysis.

1. INTRODUCTION

Multimodal representation learning of music content, such as music audio and a video [1] and music audio and text [2], has been an important topic of research, given its wide applications to Music Information Retrieval (MIR) tasks. Previous studies have typically used a deterministic approach to train encoders, where the trained encoders are utilized to map each content item to a latent space as a single vector. However, representing an arbitrary content item as a vector has various drawbacks. For example, one-to-many and many-to-many relationships need to be handled in multimodal content, such as those between an album cover image and a set of songs in that album, and between different songs that have the same title and their title text. It is difficult to represent such complex relationships in vectors. To address this challenge, *probabilistic representation learning*, in which each content item is represented as a probability distribution in a latent space, has been studied [3–5].

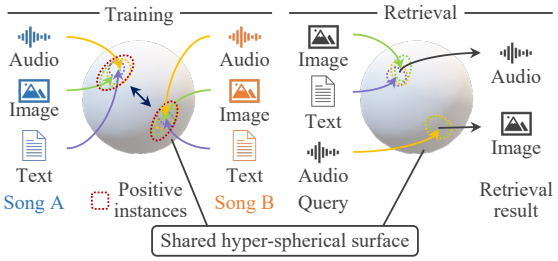

Figure 1. Probabilistic representation learning on hypersphere. (**Left**) Encoders are trained so that the probability distributions of the positive instances (music audio, an image, and text for the same song) are close to each other on the shared hyper-spherical surface, while those of irrelevant instances (different songs, artists, etc.) are far apart. (**Right**) The trained encoders are helpful for multimodal MIR. Given a single-modal query or a multimodal query such as a query that combines an image and text, our method can retrieve content items that match the query by calculating the distance between their probability distributions.

Probabilistic representation learning (Figure 1) is a promising approach that can provide intricate and diverse representations by characterizing each content item as a probability distribution. This approach requires training encoders that estimate the optimal distribution for each content item. The key here is how to design an appropriate loss function for that training. In the literature, three approaches have been proposed, and in this paper, we propose a fourth approach. The first approach uses the probability product kernel [6], which calculates the expected value between distributions. This is used in probabilistic word embedding [7], face recognition [8], and image classification [9]. The second approach uses Hedged Instance Embeddings (HIB) [10]. It formulates a contrastive loss of the match probability, which calculates the distance between a pair of vectors randomly sampled from each distribution. This is used in cross-modal retrieval of text and images [3, 4], as well as in self-supervised video representation learning [11]. The third approach is to replace variables in an existing loss function (e.g., triplet loss) with probability distributions. For example, a loss function designed for deterministic methods can be calculated by using samples obtained from a Gaussian distribution [5, 12, 13] or a von Mises-Fisher distribution [14–16] via a reparameterization trick [17, 18].

© T. Nakatsuka, M. Hamasaki, and M. Goto. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** T. Nakatsuka, M. Hamasaki, and M. Goto, "Harnessing the Power of Distributions: Probabilistic Representation Learning on Hypersphere for Multimodal Music Information Retrieval", in *Proc. of the 25th Int. Society for Music Information Retrieval Conf.,* San Francisco, United States, 2024.

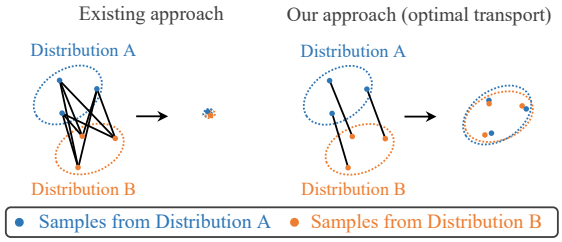

Figure 2. Advantage of optimal transport. (**Left**) To match two positive instance pairs of distributions, A and B, existing approaches ([4, 5, 16], etc.) simply calculate distances between randomly sampled pairs and cannot precisely match distributional shapes, possibly resulting in an undesirable single point when distances of positive sample pairs are minimized for probabilistic representation learning. (**Right**) Optimal transport can select optimal sample pairs to appropriately match their distributional shapes, thereby harnessing the power of rich probability distributions.

These approaches have been applied to text-to-image (or vice versa) cross-modal retrieval [3, 4], and more recently, to multimodal image retrieval [5]. Chun et al. [4] proposed Probabilistic Cross-Modal Embedding (PCME), which is a pioneering work on probabilistic representation learning for cross-modal retrieval. Li et al. [3] proposed Average Semantic Precision (ASP), which can calculate the semantic correlation scores of a dataset, and differentiable ASP approximation, which utilizes ASP as a loss function. Neculai et al. [5] proposed Multimodal Probabilistic Composer (MPC), which can use a multimodal query combining image and text for image retrieval. However, these approaches calculate distances based on sample-wise similarity, with an arbitrary sample pair randomly selected from each distribution (left side of Figure 2). This manner often results in most sample pairs being non-optimal, and as a result, the details of the distributional properties are lost. This disadvantage leads to a decrease in performance.

In light of this background, we propose two novel loss functions, one based on contrastive learning and the other on optimal transport, to be used together for multimodal MIR on a hypersphere. Contrastive learning is an effective tool to jointly map each content item of multiple modalities to a shared latent space [19, 20]. In the context of probabilistic representation learning, utilizing the angular distance between distributions has proven more effective than using their Euclidean distance [15]. Furthermore, the von Mises-Fisher (vMF) distribution (i.e., the distribution on a hypersphere) exhibits a better performance than the Gaussian distribution since vMF-based methods simplify the variance estimation by using a single scalar κ, thereby avoiding the dimension-wise estimation in Gaussian-based methods [14]. Given these insights, we propose a contrastive loss function on a hypersphere for multiple modalities based on probabilistic contrastive learning [16]. In addition, optimal transport [21] offers a robust and effective tool for calculating distances between probability distributions. It allows encoders to bring the probability distributions of a positive pair closer together, thus ensuring a more accurate representation learning (right side of Figure 2). This unique attribute of optimal transport can benefit multimodal MIR tasks. Hence, we propose a loss function based on a Spherical Sliced-Wasserstein (SSW) [22] p-distance, contemplating the compatibility between contrastive learning and optimal transport on a hypersphere.

By using the proposed loss functions, we can train encoders that map each content item to a hypersphere, as shown in Figure 1. During training, we assume that pairwise combinations of music audio of a song, a cover image for the song, and text generated from the song's metadata are positive, and that those for irrelevant ones (different songs, music genres, or artists, etc.) are negative (left side of Figure 1). Once the encoders are trained, we can utilize them to obtain the probabilistic representation of each content item for multimodal MIR (right side of Figure 1). The main advantage of probabilistic representation lies in its ability to seamlessly integrate multiple content items in a latent space as a *multimodal query*, which is a great benefit in retrieval tasks. We conduct both a quantitative evaluation and a qualitative analysis on the public YT8M-MusicVideo dataset and a private dataset to demonstrate the effectiveness of our proposed method.

2. PRELIMINARY

2.1 Problem Specification

We use a mel spectrogram of music audio as the input of an audio encoder, an RGB image as the input of an image encoder, and a tokenized text as the input of a text encoder. We follow previous studies [19, 20] regarding the setup of the input representations.

Let $\mathbf{A} = \{\mathbf{a}_n \in \mathbb{R}^{D^{\mathbf{a}}}\}_{n=1}^N$, $\mathbf{I} = \{\mathbf{i}_n \in \mathbb{R}^{D^{\mathbf{i}}}\}_{n=1}^N$, and $\mathbf{T} = \{\mathbf{t}_n \in \mathbb{R}^{D^{\mathbf{t}}}\}_{n=1}^N$ be a set of spectrograms, a set of images corresponding to $\mathbf{A}$, and a set of tokenized texts corresponding to $\mathbf{A}$, respectively, where $D^{\mathbf{a}}$ is the number of dimensions of each spectrogram, $D^{\mathbf{i}}$ is the number of dimensions of each image, $D^{\mathbf{t}}$ is the number of dimensions of each tokenized text, and N is the number of songs.

Next, let $\mathbf{Z}^{\mathbf{A}} = \{\mathbf{z}_n^{\mathbf{a}} \in \mathbb{R}^d\}_{n=1}^N$, $\mathbf{Z}^{\mathbf{I}} = \{\mathbf{z}_n^{\mathbf{i}} \in \mathbb{R}^d\}_{n=1}^N$, and $\mathbf{Z}^{\mathbf{T}} = \{\mathbf{z}_n^{\mathbf{t}} \in \mathbb{R}^d\}_{n=1}^N$ be sets of the latent variables of spectrograms, images, and tokenized texts, respectively, where d is the number of dimensions of each latent variable.

We train the audio encoder $f_{\mathbf{A}}$ that maps $\mathbf{A}$ to $\mathbf{Z}^{\mathbf{A}}$, the image encoder $f_{\mathbf{I}}$ that maps $\mathbf{I}$ to $\mathbf{Z}^{\mathbf{I}}$, and the text encoder $f_{\mathbf{T}}$ that maps $\mathbf{T}$ to $\mathbf{Z}^{\mathbf{T}}$ so that probability distributions $p(\mathbf{z}_n^{\mathbf{a}}|\mathbf{a}_n)$, $p(\mathbf{z}_n^{\mathbf{i}}|\mathbf{i}_n)$, and $p(\mathbf{z}_n^{\mathbf{t}}|\mathbf{t}_n)$ are close to each other on a shared hyper-spherical surface $S_{\text{shared}}^{d-1} = \{||\mathbf{z}_n^{\cdot}|| = 1\}$.

2.2 Probabilistic Contrastive Learning

Contrastive learning is an established deep learning technique widely utilized in recent research [23]. In particular, methods like N-pairs loss [24], InfoNCE loss [25], and MoCo [26], which calculate the loss based on N-pairs of instances (i.e., one positive pair and $N-1$ negative (or irrelevant) pairs), serve as powerful tools for multimodal representation learning [1, 2, 19, 20]. However, these contrastive loss functions are designed for deterministic methods and cannot be directly applied to probabilistic approaches.

Recently, Kirchhof et al. [16] introduced MCInfoNCE, an adaptation of InfoNCE for probabilistic contrastive learning that uses Monte-Carlo samples from each distribution. The MCInfoNCE loss $\mathcal{L}_{MC}$ is defined as follows:

$$\mathcal{L}_{MC} = -\frac{1}{m}\sum_{j=1}^{m}\sum_{l=1}^{L}\log\frac{e^{\mathrm{sim}(\mathbf{z}_j^l,\mathbf{z}_+^l)/\tau}}{e^{\mathrm{sim}(\mathbf{z}_j^l,\mathbf{z}_+^l)/\tau}+\sum_{\mathbf{z}_-}e^{\mathrm{sim}(\mathbf{z}_j^l,\mathbf{z}_-^l)/\tau}}, \tag{1}$$

where m is a mini-batch size, L is the number of samples, τ is a hyperparameter called temperature scaling, which controls the scale of the loss function, $\mathbf{z}_n \sim p(\mathbf{z}_n)$ is an anchor, $\mathbf{z}_+ \sim p(\mathbf{z}_+|\mathbf{z}_n)$ and $\mathbf{z}_- \sim p(\mathbf{z}_-|\mathbf{z}_n)$ respectively indicate positive and negative samples of the anchor, and $\mathrm{sim}(\cdot,\cdot)$ is a function that calculates the similarity (such as cosine similarity) between two distributions. Since MCInfoNCE is originally designed as the single-modal loss, we are the first to modify it for our multimodal loss in Section 3.1.

2.3 Optimal Transport

Optimal transport has been gaining popularity for a variety of computer vision tasks [27–29], but calculating the optimal transport distance between distributions is known to be computationally intensive [30]. This problem can be solved when the distributions are on a particular manifold [22, 30].

We therefore delve into a recent powerful innovation, the Spherical Sliced-Wasserstein (SSW) [22] p-distance, which is specialized on a hypersphere and is highly efficient and useful, but has not yet been used for representation learning.

2.3.1 Definition of Spherical Sliced-Wasserstein (SSW)

The SSW p-distance for $p \geq 1$ is defined between two probability measures $\mu, \nu \in \mathcal{P}_{p,ac}(S^{d-1})$, the set of absolutely continuous probability measures on a hypersphere S^{d-1} with a finite p-th moment, as follows:

$$SSW_p(\mu,\nu) = \int_{\mathbb{V}_{d,2}} W_p\left(\mu\circ P^{U^{-1}}, \nu\circ P^{U^{-1}}\right)\,\mathrm{d}\sigma, \tag{2}$$

where $\mathbb{V}_{d,2} = \{U \in \mathbb{R}^{d\times 2},\ U^\top U = I_2\}$ is the Stiefel manifold [31], σ is the uniform distribution over $\mathbb{V}_{d,2}$, P^U is the function that projects a point $\mathbf{z} \in S^{d-1}$ onto a great circle S^1 generated by U (for $a.e.\ \mathbf{z} \in S^{d-1}$, P^U can be written in a practical form of $P^U(\mathbf{z}) = \frac{U^\top \mathbf{z}}{\|U^\top \mathbf{z}\|_2}$ [22]), and W_p is the optimal transport distance on S^1 [32, 33]. To avoid any effects stemming from the choice of U, Bonet et al. [22] calculated the SSW distance several times for a set of random U, and we also calculate it in the same way.

2.3.2 Optimal Transport Distance on Great Circle

We focus on the simplest $p = 1$ in Equation (2) to calculate $W_p|_{p=1}$ between two probability measures $\mu', \nu' \in \mathcal{P}(S^1)$ that are after being projected from a hypersphere S^{d-1} onto one of the generated great circles S^1. The W_1 is defined as

$$W_1(\mu',\nu') = \int_0^1 |F_{\mu'}(t) - F_{\nu'}(t) - \mathrm{LevMed}(F_{\mu'} - F_{\nu'})|\,\mathrm{d}t, \tag{3}$$

where $F_{\mu'}, F_{\nu'}$ are the cumulative distribution function of μ', ν', respectively, and $\mathrm{LevMed}(\cdot)$ is the level median [34], defined as follows:

$$\mathrm{LevMed}(f) = \min\left\{\arg\min_{\alpha\in\mathbb{R}}\int_0^1 |f(t)-\alpha|\mathrm{d}t\right\}, \tag{4}$$

where α is a shift parameter. The SW_1, which is utilized in our proposed loss functions (Section 3), can thus be calculated by using Equations (2)–(4). Surprisingly, we can approximate the integral in Equation (3) simply by sorting the samples on S^1 in order to calculate $F_{\mu'}, F_{\nu'}$, and $\mathrm{LevMed}(\cdot)$. To illustrate this intuitively, the optimal sample pairing on the right of Figure 2 is dramatically expedited by this sorting on the *one-dimensional* great circle without examining many pairings. We present the algorithm and pseudocode of SW_1 in our supplementary materials [1].

3. PROPOSED METHOD FOR MULTIMODAL MIR

We design two novel loss functions for probabilistic representation learning: a *multimodal probabilistic contrastive loss function* for multiple modalities (Section 3.1) and an *SSW-based loss function* (Section 3.2) based on optimal transport. To train the encoders as shown in Figure 1, we assign them different roles. The former loss is designed for distancing irrelevant instance pairs of probability distributions on $S^{d-1}_{\mathrm{shared}}$, resulting in closer positive instance pairs. The latter loss focuses on placing positive instance pairs close to each other by matching their distributional shapes, and does not deal with irrelevant pairs at all. Their integration is therefore important. The trained encoders can be applied to multimodal MIR (Section 3.3).

The standard approach for probabilistic representation learning assumes that the latent variables of each content item have a probability distribution of a given form, such as a Gaussian distribution [5, 12, 13] or a von Mises-Fisher (vMF) distribution [14–16]. We use the vMF distribution as the probability distribution on $S^{d-1}_{\mathrm{shared}}$ as follows:

$$p(\mathbf{z}_n^{\mathbf{a}}|\mathbf{a}_n) = \mathrm{vMF}(\mathbf{z}_n^{\mathbf{a}}; \mu(\mathbf{a}_n), \kappa(\mathbf{a}_n)), \tag{5}$$
$$p(\mathbf{z}_n^{\mathbf{i}}|\mathbf{i}_n) = \mathrm{vMF}(\mathbf{z}_n^{\mathbf{i}}; \mu(\mathbf{i}_n), \kappa(\mathbf{i}_n)), \tag{6}$$
$$p(\mathbf{z}_n^{\mathbf{t}}|\mathbf{t}_n) = \mathrm{vMF}(\mathbf{z}_n^{\mathbf{t}}; \mu(\mathbf{t}_n), \kappa(\mathbf{t}_n)), \tag{7}$$

where the variables are as defined in Section 2.1. Using the proposed loss functions, we train three encoders so that they can estimate the appropriate parameters, the mean direction $\mu(\cdot)$ and the concentration $\kappa(\cdot)$, of each vMF distribution.

During training, we utilize L samples taken from each vMF distribution via a rejection-sampling reparameterization trick [18] in practice. Our proposed loss functions in Sections 3.1 and 3.2 use the following notations:

$$\zeta_n \sim \mathrm{vMF}(\mathbf{z}_n^*; \mu(*_n), \kappa(*_n)), \tag{8}$$
$$\eta_n \sim \mathrm{vMF}(\mathbf{z}_n^\star; \mu(\star_n), \kappa(\star_n)), \tag{9}$$

where ζ_n and η_n ($*, \star \in \{\mathbf{a}, \mathbf{i}, \mathbf{t}\}, * \neq \star$) are L samples from the vMF distribution of respective content items.

3.1 Multimodal Probabilistic Contrastive Loss Function for Probabilistic Contrastive Learning

Contrastive learning is an effective approach to jointly train encoders for the representation learning of multiple modal-

[1] https://github.com/T39Nakatsuka/ISMIR2024

ities [1, 2, 19, 20]. By modifying Equation (1), we design our own multimodal loss function $\mathcal{L}_C$ for all pairwise combinations of multiple modalities (we name this *multimodal probabilistic contrastive loss*) as follows:

$$\mathcal{L}_C = -\frac{1}{m} \sum_{<\zeta,\eta>} \sum_{j=1}^{m} \log \frac{e^{\mathrm{sim}(\zeta_j,\eta_+)/\tau}}{\sum_{k=1}^{m} e^{\mathrm{sim}(\zeta_j,\eta_k)/\tau}}, \tag{10}$$

where m is a mini-batch size, τ is a temperature scaling, $+$ indicates a positive sample of an anchor, and $\mathrm{sim}(\cdot,\cdot)$ is a function that calculates the similarity between two distributions by leveraging the L samples as follows:

$$\begin{aligned}\mathrm{sim}(\zeta_j,\eta_k) &\simeq \mathrm{sim}\left(\left\{\mathbf{z}_j^{*,l}\right\}_{l=1}^{L}, \left\{\mathbf{z}_k^{\star,l}\right\}_{l=1}^{L}\right) \\ &= \frac{1}{L}\sum_{l=1}^{L} \frac{\mathbf{z}_j^{*,l\top}\mathbf{z}_k^{\star,l}}{\|\mathbf{z}_j^{*,l}\|\|\mathbf{z}_k^{\star,l}\|}.\end{aligned} \tag{11}$$

This loss $\mathcal{L}_C$ can thus distance the centroids of the distributions of irrelevant instance pairs for the contrastive learning.

3.2 SSW-based Loss Function for Optimal Transport

We formulate our SSW-based loss function $\mathcal{L}_S$ using the SSW_1 distance (Equations (2)–(4)) as follows:

$$\mathcal{L}_S = \frac{1}{m} \sum_{<\zeta,\eta>} \sum_{j=1}^{m} SSW_1(\zeta_j,\eta_j). \tag{12}$$

Intuitively, both the L samples from ζ_j and the L samples from η_j on $S^{d-1}_{\mathrm{shared}}$ are projected onto S^1, sorted (paired), and used to calculate the cumulative distribution functions, resulting in the optimal transport distance between those positive instance pairs. This loss $\mathcal{L}_S$ can thus make the distributions of positive instance pairs closer.

To leverage the advantages of both $\mathcal{L}_C$ and $\mathcal{L}_S$, our method uses a loss function that integrates them as follows:

$$\mathcal{L} = \mathcal{L}_C + \lambda_S \mathcal{L}_S, \tag{13}$$

where λ_S is a weight.

3.3 Probabilistic Multimodal MIR

Once the encoders have been trained, we can leverage them to map each content item as a probability distribution on $S^{d-1}_{\mathrm{shared}}$ and calculate the distances between their distributions. For a single-modal query, we calculate the cosine similarity between the mean (i.e., Fréchet mean [35, 36]) over samples obtained from the distribution of a query and that of each content item in a dataset. For a multimodal query, we calculate the Fréchet mean over all samples obtained from the distribution of each query and use it like a single-modal query. When the similarity score between a pair of content items is high, it indicates that they are matched. We thus sort the similarity scores in descending order and retrieve the content item in the dataset that scored higher with respect to the query.

4. EXPERIMENTS AND RESULTS

This section describes comparison experiments to quantitatively evaluate how closely the probability distributions of positive instances were located on $S^{d-1}_{\mathrm{shared}}$, and a qualitative analysis of the proposed method to further investigate the nature of the learned representation of each content item.

4.1 Experimental Setup

4.1.1 Dataset

For the experiments, we used the following two benchmark datasets with different characteristics. We determined the size of each test set by following the setup in [1, 37].

YT8M-MusicVideo dataset [1] is a subset of the YouTube-8M dataset [38], comprising videos tagged as "music video." We collected 73,113 triplets consisting of music audio (average length of 4 min with a 48 kHz sampling rate), its thumbnail image (an RGB image with an aspect ratio of 16:9), and its metadata including title, channel name, and upload date from 60,785 YouTube channels. We randomly split the dataset into training (64,001 songs), validation (7,112 songs), and test (2,000 songs) sets with no YouTube channels overlapping across these sets. For evaluation, we conducted our experiments three times with different seed values when training the encoders.

AS5M dataset (Album Songs 5 Million dataset) is a private dataset that contains triplets of a music audio excerpt (a 30 s audio preview for trial listening, with a 44.1 kHz sampling rate), its cover image (a square RGB image), and its metadata including song title, artist name, collection name, music genre, and release date. The dataset contains 5,920,828 audio excerpts and their metadata by 174,629 artists, and 1,115,668 cover images. Because multiple excerpts from a music album are associated with a single cover image, each image corresponds to about 5.3 excerpts on average. The songs encompass a variety of music genres (over 250). We randomly split the dataset into training, validation, and test sets with an eight-one-one ratio and with no artists or images overlapping across these sets. For evaluation, we constructed ten folds of test subsets by randomly selecting 2,000 triplets of an audio excerpt, a cover image, and a text prompt for each fold from the test set.

4.1.2 Implementation Details

Encoder architecture: We used an audio model of contrastive language-audio pretraining (CLAP) [20] as the backbone network for the audio encoder, and used image and text models of contrastive language-image pretraining (CLIP) [19] as the backbone network for the image and text encoders. Before training, we set the parameters of the pre-trained models available at Transformers [39] (i.e., "laion/clap-htsat-fused" for CLAP (audio model) and "vit_base_patch16_224" for CLIP (vision and text models)) to the encoders. During training, we updated the projection layers of the encoders.

Audio: The music audio of each song was converted to a mel spectrogram through a CLAP feature extractor available at Transformers [39], and the audio encoder was trained using the spectrogram as input. In training the audio encoder, we applied a masking technique including frequency masking and time masking [40] and a random crop

technique regarding the time domain to the spectrogram for data augmentation [41].

Image: We used an RGB image resized to $224\,\text{px} \times 224\,\text{px}$ as the input of the image encoder. In training the image encoder, we applied a random resized crop (scale=[0.08, 1.0], ratio=[0.75, 1.33]), random horizontal flip (probability=0.5), and random erasing (probability=0.2) [42] to all images for data augmentation.

Text: We tokenized text generated by using a keyword-to-caption augmentation technique [20] [2] with a maximum length of 77, which is the same setup as CLIP [19]. In training the text encoder, words corresponding to metadata are randomly dropped [43] at a ratio of 0.05 for each metadata.

Training: We used 16 NVIDIA A100 GPUs under each experimental condition, and each GPU computed 64 triplets of audio, images, and text per iteration. Our implementation was based on PyTorch [44]. In training the encoders, we used the Adam optimizer [45] with a learning rate of 1.0×10^{-4}. We used $d = 512$ (dimensions of latent variables) following the setup in [5]. For the vMF distribution, we set $\kappa(\cdot) \in (64, 128)$ to obtain a clear distribution following the setup of [16]. We empirically set the number of samples L to 16. For $\mathcal{L}_C$, we set the temperature-scaling value (Equation (10)) to $\tau = 0.07$, which was originally used in MoCo [26]. For $\mathcal{L}_S$, we calculated the SSW_1 distance 100 times for a set of random U, following [22] (i.e., 16 samples from ζ_j and 16 samples from η_j were projected onto 100 different great circles to match distributional shapes from 100 different views). On the basis of preliminary studies, we set the weight λ_S to 1.0.

4.1.3 Ranking-Based Evaluation Metrics

We used three standard evaluation metrics for retrieval tasks: the mean reciprocal rank (MRR) [46], the recall@k (R@k), and the median rank (MR) [1]. MRR is a statistic measure utilized to evaluate the quality of retrieval results. Given a set of queries, MRR calculates the average of reciprocal ranks of the first correct (i.e., original) content item. A higher MRR value indicates a more accurate and efficient retrieval method. R@k evaluates how correctly content items are retrieved in the top results. For retrieval tasks, a higher R@k means that the retrieval method is more practical. We set $k = 1$ for the R@k and displayed R@1 as a percentage. MR represents the median value of the ranks of the retrieved correct content item. In our context, a lower MR is desirable because it indicates that the correct content item is ranked closer to the top of the retrieval results.

4.2 Conditions

We compared our method (**Proposed** based on $\mathcal{L}$) with two competitive methods that utilize probabilistic representation learning for text-image retrieval, **PCME** [4] and **MPC** [5].

[2] Since the text prompt generation using a template sentence with metadata is known to be effective for retrieval tasks [19], for the YT8M-MusicVideo dataset, we generated a text prompt using: "title" is a music video uploaded by "channel name" on "upload date." For the AS5M dataset, we generated a text prompt using: "song title" is a(n) "music genre" song by "artist name", released on "release date." "song title" is collected to "collection name."

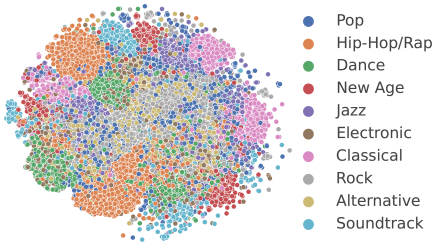

Figure 3. Visualization of the learned representations of audio, images, and text in the test subsets of the AS5M dataset with respect to music genre tags using t-SNE [51].

For music audio and other modalities, probabilistic representation learning for multimodal MIR has not yet been investigated, so we solely used the multimodal probabilistic contrastive loss $\mathcal{L}_C$ (Section 3.1) as a baseline method (**Baseline**) in order to investigate the effectiveness of $\mathcal{L}_S$.

4.3 Results

As shown in Tables 1–6, our method outperformed PCME [4] and MPC [5], which are competitive methods for text-image retrieval, in all the retrieval tasks on both datasets. Likewise, our method was superior to the baseline method based on the modified MCInfoNCE [16] in nearly all retrieval tasks. We thus confirmed that $\mathcal{L}_S$ was effective in achieving better performances. The results also showed that a multimodal query outperformed a single-modal query for most tasks. Our method can seamlessly create multimodal queries from multiple probability distributions, bringing benefits to multimodal MIR.

The performance differences between the datasets can be partly explained by their sizes since our method uses transformer models as the encoders. Several studies have shown that the performance of transformer models follows a scaling law [47–50]. This scaling law has been confirmed in experiments with data from various modalities [47–49] and in transfer learning [50]. In practice, the YT8M-MusicVideo dataset is two orders of magnitude smaller than the AS5M dataset, resulting in a decrease in performance. The performance differences between the tasks, as well as between the datasets, can also be explained on the basis of the scaling law. In our experiments, we used the CLAP audio model, which was trained on the LAION-Audio-630K dataset [20]. This dataset is several orders of magnitude smaller than the one used for training the CLIP models, which can lead to the decreased performance in audio-related retrieval tasks.

We provide additional comparison experiments that demonstrate the effectiveness of our proposed method in our supplementary materials [1].

4.4 Qualitative Analysis

We investigated the nature of the learned representations of music audio, images, and text by visualizing them regarding music genres. We utilized music audio, images, and text for 12,180 songs for the top 10 most popular genres in test subsets of the AS5M dataset. We calculated the Fréchet mean over all samples obtained from the distribution of each content item and mapped each of them to a two-dimensional

Table 1. Comparison on YT8M-MusicVideo dataset for multimodal image retrieval.

	Audio → Image			Text → Image			Audio & Text → Image		
Method	MRR ↑	R@1 (%) ↑	MR ↓	MRR ↑	R@1 (%) ↑	MR ↓	MRR ↑	R@1 (%) ↑	MR ↓
PCME	–	–	–	0.025 ± 0.003	0.73 ± 0.08	369	–	–	–
MPC	–	–	–	0.014 ± 0.001	0.2 ± 0.11	425	–	–	–
Baseline	0.024 ± 0.001	$\mathbf{0.73 \pm 0.09}$	272	0.048 ± 0.001	1.92 ± 0.12	166	0.044 ± 0.001	1.55 ± 0.11	166
Proposed	$\mathbf{0.028 \pm 0.001}$	0.65 ± 0.08	**247**	$\mathbf{0.115 \pm 0.0}$	$\mathbf{6.68 \pm 0.1}$	**92**	$\mathbf{0.119 \pm 0.002}$	$\mathbf{6.8 \pm 0.29}$	**72**

Table 2. Comparison on YT8M-MusicVideo dataset for multimodal text retrieval.

	Audio → Text			Image → Text			Audio & Image → Text		
Method	MRR ↑	R@1 (%) ↑	MR ↓	MRR ↑	R@1 (%) ↑	MR ↓	MRR ↑	R@1 (%) ↑	MR ↓
PCME	–	–	–	0.023 ± 0.002	0.73 ± 0.16	372	–	–	–
MPC	–	–	–	0.013 ± 0.001	0.13 ± 0.05	427	–	–	–
Baseline	0.026 ± 0.001	0.6 ± 0.18	226	0.046 ± 0.001	1.47 ± 0.1	167	0.054 ± 0.002	1.83 ± 0.3	131
Proposed	$\mathbf{0.039 \pm 0.001}$	$\mathbf{1.17 \pm 0.09}$	**180**	$\mathbf{0.118 \pm 0.002}$	$\mathbf{6.87 \pm 0.21}$	**89**	$\mathbf{0.139 \pm 0.002}$	$\mathbf{7.97 \pm 0.46}$	**55**

Table 3. Comparison on YT8M-MusicVideo dataset for multimodal audio retrieval.

	Image → Audio			Text → Audio			Image & Text → Audio		
Method	MRR ↑	R@1 (%) ↑	MR ↓	MRR ↑	R@1 (%) ↑	MR ↓	MRR ↑	R@1 (%) ↑	MR ↓
Baseline	0.021 ± 0.001	0.52 ± 0.05	263	0.028 ± 0.001	0.68 ± 0.08	219	0.032 ± 0.002	0.83 ± 0.26	191
Proposed	$\mathbf{0.027 \pm 0.001}$	$\mathbf{0.58 \pm 0.06}$	**235**	$\mathbf{0.041 \pm 0.003}$	$\mathbf{1.25 \pm 0.37}$	**173**	$\mathbf{0.05 \pm 0.002}$	$\mathbf{1.75 \pm 0.25}$	**141**

Table 4. Comparison on AS5M dataset for multimodal image retrieval.

	Audio → Image			Text → Image			Audio & Text → Image		
Method	MRR ↑	R@1 (%) ↑	MR ↓	MRR ↑	R@1 (%) ↑	MR ↓	MRR ↑	R@1 (%) ↑	MR ↓
PCME	–	–	–	0.069 ± 0.004	2.82 ± 0.34	131	–	–	–
MPC	–	–	–	0.026 ± 0.002	0.62 ± 0.15	240	–	–	–
Baseline	0.046 ± 0.002	1.37 ± 0.19	141	0.125 ± 0.005	6.21 ± 0.56	50	0.1 ± 0.004	4.39 ± 0.53	60
Proposed	$\mathbf{0.074 \pm 0.004}$	$\mathbf{2.94 \pm 0.46}$	**94**	$\mathbf{0.539 \pm 0.005}$	$\mathbf{45.37 \pm 0.65}$	**2**	$\mathbf{0.508 \pm 0.008}$	$\mathbf{41.35 \pm 1.12}$	**2**

Table 5. Comparison on AS5M dataset for multimodal text retrieval.

	Audio → Text			Image → Text			Audio & Image → Text		
Method	MRR ↑	R@1 (%) ↑	MR ↓	MRR ↑	R@1 (%) ↑	MR ↓	MRR ↑	R@1 (%) ↑	MR ↓
PCME	–	–	–	0.067 ± 0.003	2.73 ± 0.27	131	–	–	–
MPC	–	–	–	0.025 ± 0.002	0.57 ± 0.13	239	–	–	–
Baseline	0.062 ± 0.002	1.93 ± 0.27	82	0.126 ± 0.006	5.99 ± 0.59	47	0.146 ± 0.007	6.96 ± 0.76	30
Proposed	$\mathbf{0.113 \pm 0.004}$	$\mathbf{4.99 \pm 0.37}$	**46**	$\mathbf{0.541 \pm 0.007}$	$\mathbf{44.21 \pm 0.99}$	**2**	$\mathbf{0.58 \pm 0.009}$	$\mathbf{47.75 \pm 1.19}$	**2**

Table 6. Comparison on AS5M dataset for multimodal audio retrieval.

	Image → Audio			Text → Audio			Image & Text → Audio		
Method	MRR ↑	R@1 (%) ↑	MR ↓	MRR ↑	R@1 (%) ↑	MR ↓	MRR ↑	R@1 (%) ↑	MR ↓
Baseline	0.045 ± 0.002	1.32 ± 0.2	138	0.067 ± 0.003	2.11 ± 0.24	77	0.069 ± 0.003	2.43 ± 0.32	74
Proposed	$\mathbf{0.072 \pm 0.004}$	$\mathbf{2.62 \pm 0.33}$	**92**	$\mathbf{0.115 \pm 0.005}$	$\mathbf{4.86 \pm 0.47}$	**44**	$\mathbf{0.126 \pm 0.006}$	$\mathbf{5.54 \pm 0.62}$	**37**

space using t-SNE [51]. Figure 3 shows that their learned representations form clusters regarding music genres. That is, audio, images, and text in each of these genres are closely associated with each other.

5. CONCLUSION

We proposed a method for multimodal MIR that leverages the probabilistic representations of content items. Our contributions can be summarized as follows. First, we leveraged the von Mises-Fisher (vMF) distribution, which has been used for single-modal tasks [14–16] but has not been used for multimodal retrieval tasks. In addition, the recently-invented spherical sliced-Wasserstein (SSW) [22] p-distance for optimal transport is surprisingly computationally efficient and useful, but has not yet been used in the MIR community. Moreover, we designed the two novel loss functions, $\mathcal{L}_C$ and $\mathcal{L}_S$, using both probabilistic contrastive learning and optimal transport to facilitate probabilistic multimodal representation learning. To our knowledge, this is the first work to utilize these reusable insights for probabilistic representation learning. Second, we confirmed the effectiveness of integrating the contrastive loss function $\mathcal{L}_C$ with the loss function $\mathcal{L}_S$ based on the optimal transport distance through quantitative evaluations, and showed that the proposed method can retrieve more appropriate content items for single-modal and multimodal queries. Third, we conducted a qualitative analysis, showing that music audio, images, and text for the same music style are located close to each other on S^{d-1}_{shared}. These results demonstrated that the proposed method is effective for multimodal MIR.

The underlying principles of the proposed method can work for any retrieval tasks regardless of modalities, which will lead to a broader scope of application. As such, we believe that the proposed method will shed light on other challenging retrieval tasks and usher in practical solutions.

6. ACKNOWLEDGMENTS

This work was supported in part by JST CREST Grant Number JPMJCR20D4 and JSPS KAKENHI Grant Number 22K18017, Japan.

7. REFERENCES

[1] D. Surís, C. Vondrick, B. Russell, and J. Salamon, "It's time for artistic correspondence in music and video," in *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition Conference (CVPR)*, 2022, pp. 10 564–10 574.

[2] Q. Huang, A. Jansen, J. Lee, R. Ganti, J. Y. Li, and D. P. Ellis, "MuLan: A joint embedding of music audio and natural language," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, 2022, pp. 559–566.

[3] H. Li, J. Song, L. Gao, P. Zeng, H. Zhang, and G. Li, "A differentiable semantic metric approximation in probabilistic embedding for cross-modal retrieval," in *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, 2022, pp. 11 934–11 946.

[4] S. Chun, S. J. Oh, R. S. De Rezende, Y. Kalantidis, and D. Larlus, "Probabilistic embeddings for cross-modal retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 8415–8424.

[5] A. Neculai, Y. Chen, and Z. Akata, "Probabilistic compositional embeddings for multimodal image retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 4547–4557.

[6] T. Jebara, R. Kondor, and A. Howard, "Probability product kernels," *J. Mach. Learn. Res.*, vol. 5, pp. 819–844, 2004.

[7] L. Vilnis and A. McCallum, "Word representations via gaussian embedding," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014, pp. 1–12.

[8] Y. Shi and A. K. Jain, "Probabilistic face embeddings," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 6902–6911.

[9] M. Kirchhof, K. Roth, Z. Akata, and E. Kasneci, "A non-isotropic probabilistic take on proxy-based deep metric learning," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022, pp. 435–454.

[10] S. J. Oh, K. Murphy, J. Pan, J. Roth, F. Schroff, and A. Gallagher, "Modeling uncertainty with hedged instance embedding," *arXiv preprint arXiv:1810.00319*, 2018.

[11] J. Park, J. Lee, I.-J. Kim, and K. Sohn, "Probabilistic representations for video contrastive learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 14 711–14 721.

[12] J. Chang, Z. Lan, C. Cheng, and Y. Wei, "Data uncertainty learning in face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 5710–5719.

[13] B. D. Roads and B. C. Love, "Enriching imagenet with human similarity judgments and psychological embeddings," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 3547–3557.

[14] S. Li, J. Xu, X. Xu, P. Shen, S. Li, and B. Hooi, "Spherical confidence learning for face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 15 629–15 637.

[15] T. R. Scott, A. C. Gallagher, and M. C. Mozer, "von mises-fisher loss: An exploration of embedding geometries for supervised learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 10 612–10 622.

[16] M. Kirchhof, E. Kasneci, and S. J. Oh, "Probabilistic contrastive learning recovers the correct aleatoric uncertainty of ambiguous inputs," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2023, pp. 17 085–17 104.

[17] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proceeding of the International Conference on Learning Representations (ICLR)*, 2014.

[18] T. R. Davidson, L. Falorsi, N. De Cao, T. Kipf, and J. M. Tomczak, "Hyperspherical variational auto-encoders," in *Proceeding of the Conference on Uncertainty in Artificial Intelligence (UAI)*, 2018, pp. 856–865.

[19] 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 *Proceedings of the International Conference on Machine Learning (ICML)*, 2021, pp. 8748–8763.

[20] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[21] C. Villani, *Topics in optimal transportation*. American Mathematical Soc., 2021, vol. 58.

[22] C. Bonet, P. Berg, N. Courty, F. Septier, L. Drumetz, and M.-T. Pham, "Spherical sliced-wasserstein," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

[23] P. H. Le-Khac, G. Healy, and A. F. Smeaton, "Contrastive representation learning: A framework and review," *IEEE Access*, vol. 8, pp. 193 907–193 934, 2020.

[24] K. Sohn, "Improved deep metric learning with multi-class N-pair loss objective," in *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS)*, vol. 29, 2016, pp. 1857–1865.

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

[26] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 9729–9738.

[27] C. Frogner, C. Zhang, H. Mobahi, M. Araya, and T. A. Poggio, "Learning with a wasserstein loss," in *Proceeding of the Advances in Neural Information Processing Systems (NIPS)*, vol. 28, 2015.

[28] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2017, pp. 214–223.

[29] D. Garg, Y. Wang, B. Hariharan, M. Campbell, K. Q. Weinberger, and W.-L. Chao, "Wasserstein distances for stereo disparity estimation," in *Proceeding of the Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 22 517–22 529.

[30] S. Kolouri, K. Nadjahi, U. Simsekli, R. Badeau, and G. Rohde, "Generalized sliced wasserstein distances," in *Proceeding of the Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019.

[31] T. Bendokat, R. Zimmermann, and P.-A. Absil, "A grassmann manifold handbook: Basic geometry and computational aspects," *arXiv preprint arXiv:2011.13699*, 2020.

[32] J. Delon, J. Salomon, and A. Sobolevski, "Fast transport optimization for monge costs on the circle," *SIAM J. Appl. Math.*, vol. 70, no. 7, pp. 2239–2258, 2010.

[33] J. Rabin, J. Delon, and Y. Gousseau, "Transportation distances on the circle," *J. Math. Imaging Vis.*, vol. 41, no. 1, pp. 147–167, 2011.

[34] S. Hundrieser, M. Klatt, and A. Munk, *The Statistics of Circular Optimal Transport*. Springer Nature Singapore, 2022, pp. 57–82.

[35] M. Fréchet, "Les éléments aléatoires de nature quelconque dans un espace distancié," *Annales de l'institut Henri Poincaré*, vol. 10, no. 4, pp. 215–310, 1948.

[36] H. Karcher, "Riemannian center of mass and mollifier smoothing," *Commun. Pure Appl. Math.*, vol. 30, no. 5, pp. 509–541, 1977.

[37] L. Prétet, G. Richard, and G. Peeters, "Cross-modal music-video recommendation: A study of design choices," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, 2021, pp. 1–9.

[38] 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.

[39] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Transformers: State-of-the-art natural language processing," in *Proceedings of the Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP-SD)*, 2020, pp. 38–45.

[40] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, 2019.

[41] R. Takahashi, T. Matsubara, and K. Uehara, "Data augmentation using random image cropping and patching for deep CNNs," *IEEE Trans. Circuits. Syst. Video Technol.*, vol. 30, no. 9, pp. 2917–2931, 2019.

[42] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2020, pp. 13 001–13 008.

[43] T. Sellam, D. Das, and A. P. Parikh, "BLEURT: Learning robust metrics for text generation," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020, pp. 7881–7892.

[44] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019, pp. 8024–8035.

[45] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

[46] N. Craswell, "Mean reciprocal rank," in *Encyclopedia of Database Systems*. Springer US, 2009.

[47] M. Cherti, R. Beaumont, R. Wightman, M. Wortsman, G. Ilharco, C. Gordon, C. Schuhmann, L. Schmidt, and J. Jitsev, "Reproducible scaling laws for contrastive language-image learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 2818–2829.

[48] J. Droppo and O. Elibol, "Scaling laws for acoustic models," in *Proceedings of Interspeech 2021*, 2021, pp. 2576–2580.

[49] D. Hernandez, J. Kaplan, T. Henighan, and S. McCandlish, "Scaling laws for transfer," *arXiv preprint arXiv:2102.01293*, 2021.

[50] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling laws for neural language models," *arXiv preprint arXiv:2001.08361*, 2020.

[51] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.