UNSUPERVISED COMPOSABLE REPRESENTATIONS FOR AUDIO

Giovanni Bindi, Philippe Esling
Institut de Recherche et Coordination Acoustique-Musique (IRCAM)
CNRS UMR 9912, Sorbonne Université
{bindi, esling}@ircam.fr

ABSTRACT

Current generative models are able to generate high-quality artefacts but have been shown to struggle with compositional reasoning, which can be defined as the ability to generate complex structures from simpler elements. In this paper, we focus on the problem of compositional representation learning for music data, specifically targeting the fully-unsupervised setting. We propose a simple and extensible framework that leverages an explicit compositional inductive bias, defined by a flexible auto-encoding objective that can leverage any of the current state-of-art generative models. We demonstrate that our framework, used with diffusion models, naturally addresses the task of unsupervised audio source separation, showing that our model is able to perform high-quality separation. Our findings reveal that our proposal achieves comparable or superior performance with respect to other blind source separation methods and, furthermore, it even surpasses current state-of-art supervised baselines on signal-to-interference ratio metrics. Additionally, by learning an a-posteriori masking diffusion model in the space of composable representations, we achieve a system capable of seamlessly performing unsupervised source separation, unconditional generation, and variation generation. Finally, as our proposal works in the latent space of pre-trained neural audio codecs, it also provides a lower computational cost with respect to other neural baselines.

1. INTRODUCTION

Generative models recently became one of the most important topic in machine learning research. Their goal is to learn the underlying probability distribution of a given dataset in order to accomplish a variety of downstream tasks, such as sampling or density estimation. These models, relying on deep neural networks as their core architecture, have demonstrated unprecedented capabilities in capturing intricate patterns and generating complex and realistic data [1]. Although these systems are able to generate impressive results that go beyond the replication of training data, some doubts have recently been raised about their actual reasoning and extrapolation abilities [2, 3]. Notably, a critical question remains on their capacity to perform *compositional reasoning*. The principle of compositionality states that the meaning of a complex expression is dependent on the meanings of its individual components and the rules employed to combine them [4, 5]. This concept also plays a significant role in machine learning [6], with a particular emphasis in the fields of NLP and vision. Indeed, compositionality holds a strong significance in the *interpretability* of machine learning algorithms [7], ultimately providing a better understanding of the behaviour of such complex systems. In line with recent studies on compositional inductive biases [8, 9], taking a compositional approach would allow to build better representation learning and more effective generative models, but research on compositional learning for audio is still lacking.

In this work, we specifically focus on the problem of compositional representation learning for audio and propose a generic and simple framework that explicitly targets the learning of composable representations in a fully unsupervised way. Our idea is to learn a set of low-dimensional latent variables that encode semantic information which are then used by a generative model to reconstruct the input. While we build our approach upon recent diffusion models, we highlight that our framework can be implemented with any state-of-the-art generative system. Therefore, our proposal effectively combines diffusion models and auto-encoders and represents, to the best of our knowledge, one of the first contributions that explicitly target the learning of unsupervised compositional semantic representations for audio. Although being intrinsically modality-agnostic, we show that our system can be used to perform *unsupervised source separation* and we validate this claim by performing experiments on standard benchmarks, comparing against both unsupervised and supervised baselines. We show that our proposal outperforms all unsupervised methods, and even supervised methods on some metrics. Moreover, as we are able to effectively perform latent source separation, we complement our decomposition system with a prior model that performs *unconditional generation* and *variation generation* [10]. Hence, our method is able to take an audio mixture as input, and generate several high-quality variations for one of the instrumental part only, effectively allowing to control regeneration of a source audio material in multi-instrument setups. Furthermore, we train a masking diffusion model in the latent space of composable representation and show

Attribution: G. Bindi, and P. Esling, "Unsupervised Composable Representations for Audio", in *Proc. of the 25th Int. Society for Music Information Retrieval Conf.,* San Francisco, United States, 2024.

that our framework is able to handle both decomposition and generation in an effective way without any supervision. We provide audio examples, additional experiments and source code on a supporting webpage [1]

2. BACKGROUND

In this section, we review the fundamental components of our methodology. Hence, we briefly introduce the principles underlying diffusion models and a recent variation rooted in autoencoders, referred to as Diffusion Autoencoder [11], which serves as the basis for our formulation.

Notation. Throughout this paper, we suppose a dataset $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^n$ of *i.i.d.* data points $\mathbf{x}_i \in \mathbb{R}^d$ coming from an unknown distribution $p^*(\mathbf{x})$. We denote $\theta \in \Theta \subseteq \mathbb{R}^p$, $\phi \in \Phi \subseteq \mathbb{R}^q$ and $\psi \in \Psi \subseteq \mathbb{R}^r$ as the set of parameters learned through back-propagation [12].

2.1 Diffusion models

Diffusion models (DMs) are a recent class of generative models that can synthesize high-quality samples by learning to reverse a stochastic process that gradually adds noise to the data. DMs have been successfully applied across diverse domains, including computer vision [13], natural language processing [14], audio [15] and video generation [16]. These applications span tasks such as unconditional and conditional generation, editing, super-resolution and inpainting, often yielding state of the art results.

This model family has been introduced by [17] and has its roots in statistical physics, but there now exist many derivations with different formalisms that generalise the original formulation. At their core, DMs are composed of a *forward* and *reverse* Markov chain that respectively adds and removes Gaussian noise from data. Recently, [18] established a connection between DM and denoising score matching [19, 20], introducing simplifications to the original training objective and demonstrating strong experimental results. Intuitively, the authors propose to learn a function $\boldsymbol{\epsilon}_\theta$ that takes a noise-corrupted version of the input and predicts the noise $\boldsymbol{\epsilon}$ used to corrupt the data. Specifically, the *forward* process gradually adds Gaussian noise to the data $\mathbf{x} \to \mathbf{x}_t$ according to an increasing noise variance schedule $\beta_1, \ldots, \beta_T$, following the distribution

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \boldsymbol{I}), \quad (1)$$

with $T \in \mathbb{N}$ and $t \in \{1, \ldots, T\}$. Following the notation $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$, diffusion models approximate the *reverse* process by learning a function $\boldsymbol{\epsilon}_\theta : \mathbb{R}^d \times \mathbb{R} \to \mathbb{R}^d$ that predicts $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{\epsilon}, \mathbf{0}, \mathbf{I})$ by

$$\min_{\theta \in \Theta} \quad \mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon}}\big[\|\boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t) - \boldsymbol{\epsilon}\|\big], \quad (2)$$

with $\boldsymbol{\epsilon}_\theta$ usually implemented as a U-Net [21] and the step $t \sim \mathcal{U}[0, T]$.

Deterministic diffusion. More recently, [22] introduced Denoising Diffusion Implicit Models (DDIM), extending the diffusion formulation with non-Markovian modifications, thus enabling deterministic diffusion models and substantially increasing their sampling speed. They also established an equivalence between their objective function and the one from [18], highlighting the generality of their formulation. Finally, [23] further generalized this approach and proposed Iterative $\alpha-$(de)Blending (IADB), simplifying the theory of DDIM while removing the constraint for the target distribution to be Gaussian. In fact, given a base distribution [2] $p_n(\mathbf{x}_0)$, we corrupt the input data by linear interpolation $\mathbf{x}_\alpha = (1-\alpha)\mathbf{x}_0 + \alpha\mathbf{x}$ with $\mathbf{x}_0 \sim p_n(\mathbf{x}_0)$ and learn a U-Net $\boldsymbol{\epsilon}_\theta$ by optimizing, e.g.,

$$\min_{\theta \in \Theta} \quad \mathbb{E}_{\alpha,\mathbf{x},\mathbf{x}_0}\big[\|\boldsymbol{\epsilon}_\theta(\mathbf{x}_\alpha, \alpha) - \mathbf{x}\|_2^2\big], \quad (3)$$

with $\alpha \sim \mathcal{U}[0, 1]$. This is known as the c variant of IADB, which is the closest formulation to DDIM. In our implementation, we instead use the d variant of IADB, which has a slightly different formulation that we do not report for brevity. We experimented with both variants and did not find significant discrepancies in performances.

Diffusion Autoencoders. All the methods described in the preceding paragraph specifically target unconditional generation. However, in this work we are interested in conditional generation and, more specifically, in a conditional encoder-decoder architecture. For this reason, we build upon the recent work by [11] named Diffusion Autoencoder (DiffAE). The central concept in this approach involves employing a learnable encoder to discover high-level semantic information, while using a DM as the decoder to model the remaining stochastic variations. Therefore, the authors equip a DDIM model ϵ_ϕ with a semantic encoder $E_\theta : \mathbb{R}^d \to \mathbb{R}^s$ with $s \ll d$ that is responsible for compressing the high-level *semantic* information [3] into a latent variable $\mathbf{z} \in \mathbb{R}^s$ as $\mathbf{z} = E_\theta(\mathbf{x})$. The DDIM model is, therefore, conditioned on such semantic representation and trained to reconstruct the data via

$$\min_{\theta \in \Theta, \phi \in \Phi} \quad \mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon}}\big[\|\boldsymbol{\epsilon}_\phi(\sqrt{\alpha}\mathbf{x}_0 + \sqrt{1-\alpha}\boldsymbol{\epsilon}, \mathbf{z}, t) - \boldsymbol{\epsilon}\|\big] \quad (4)$$

with $\alpha = \prod_{s=1}^t (1-\beta_s)$ and β_i being the variance at the $i-$th step. Since the DiffAE represents the state of the art for encoder-decoder models based on diffusion, we build our compositional diffusion framework upon this formulation, which we describe in the following section.

3. PROPOSED APPROACH

In compositional representation learning, we hypothesize that the information can be deconstructed into specific, identifiable parts that collectively makes up the whole input. In this work, we posit these parts to be distinct instruments in music but we highlight that this choice is uniquely dependent on the target application. Due to the lack of a widely-accepted description of compositional representations, we formulate a simple yet comprehensive definition that can subsequently be specialized to address particular

[1] https://github.com/ismir-24-sub/unsupervised_compositional_representations

[2] For simplicity we assume $p_n(\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_0; \mathbf{0}, \boldsymbol{I})$.

[3] In the domain of vision this could be the identity of a person or the type of objects represented in an image.

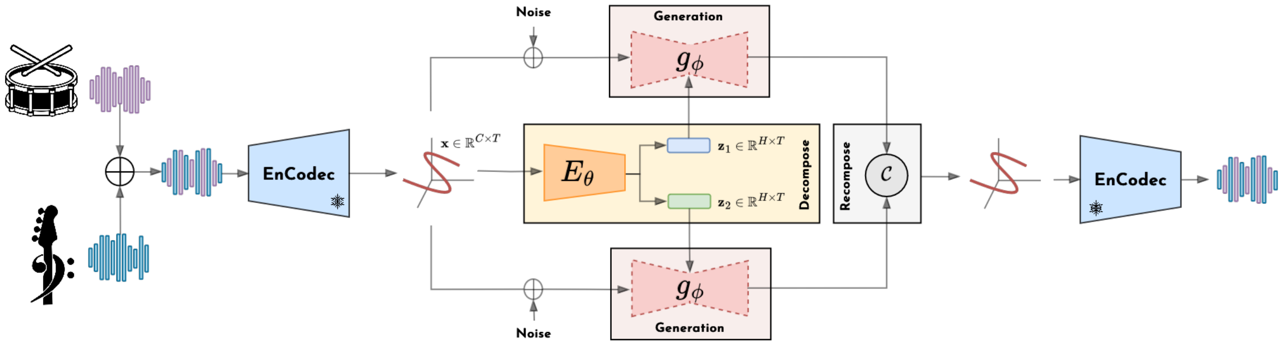
Noise
Generation
g_ϕ
EnCodec
$\mathbf{x} \in \mathbb{R}^{C\times T}$
E_θ
$\mathbf{z}_1 \in \mathbb{R}^{H\times T}$
$\mathbf{z}_2 \in \mathbb{R}^{H\times T}$
Decompose
Recompose
$\mathcal{C}$
EnCodec
g_ϕ
Generation
Noise

Figure 1. The overall architecture of our decomposition model. We first mix the sources, map the data $\mathbf{x}$ to the latent space through a frozen, pre-trained EnCodec model, and then decompose it into a set of latent variables (two shown here). These variables then condition a parameter-sharing diffusion model whose generation are then recomposed by an operator C.

cases [24, 25]. Specifically, we start from the assumption that observations $\mathbf{x} \in \mathbb{R}^d$ are realizations of an underlying latent variable model and that each concept is described by a corresponding latent $\mathbf{z}_i \in \mathcal{Z}_i$, where $i \in \{1, \dots, N\}$ with N being the total number of possible entities that compose our data. Then, we define a compositional representation of $\mathbf{x}$ as

$$\mathbf{x} = \mathcal{C}(\hat{\mathbf{z}}_1, \dots, \hat{\mathbf{z}}_N) = \mathcal{C}(f_1(\mathbf{z}_1), \dots, f_N(\mathbf{z}_N)), \qquad (5)$$

where $\mathcal{C} : \hat{\mathcal{Z}}_1 \times \hat{\mathcal{Z}}_2 \times \dots \hat{\mathcal{Z}}_N \to \mathbb{R}^d$ is a *composition operator* and each $f_i : \mathcal{Z}_i \to \hat{\mathcal{Z}}_i$ is a *processing function* that maps each latent variable to another *intermediate* space. By being intentionally broad, this definition does not impose any strong specific constraints a priori, such as the requirement for each subspace to be identical or the algebraic structure of the latent space itself. Hence, to implement this model, we rather need to consider careful intentional design choices and inductive biases. In this work, we constrain the intermediate space to be the data space itself, i.e. $\hat{\mathcal{Z}}_i = \mathbb{R}^d$ for all $i = 1, \dots, N$ and we focus on the learning of the latent variables and the processing functions. Finally, we set the composition operator to be a pre-defined function such as $mean$ or max and leave its learning to further investigations.

3.1 Decomposition

In this section, we detail our proposed model, as depicted in Figure 1. Globally, we follow an encoder-decoder paradigm, where we encode the data $\mathbf{x} \in \mathbb{R}^d$ into a set of latent representations $Z = \{\mathbf{z}_1, \dots, \mathbf{z}_N\}$, where $\mathbf{z}_i \in \mathcal{Z} \subseteq \mathbb{R}^h$ for each $i = 1, \dots, N$. This is done through an encoder network $E_\theta : \mathbb{R}^d \to \mathcal{Z} \times \dots \times \mathcal{Z}$ that maps the input $\mathbf{x}$ to the set of variables Z, i.e. $[\mathbf{z}_1, \dots, \mathbf{z}_N] = E_\theta(\mathbf{x})$. Each latent variable is then decoded separately through a parameter-shared diffusion model, which implements the *processing function* $f : \mathcal{Z} \to \mathbb{R}^d$ in Equation 5, mapping the latents to the data space. Finally, we reconstruct the input data $\mathbf{x}$ through the application of a *composition operator* $\mathcal{C}$ and train the system end-to-end through a vanilla iterative $\alpha-$(de)Blending (IADB) loss. Specifically, we learn a U-Net network $g_\phi : \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^h \to \mathbb{R}^d$ and a semantic encoder E_θ via the following objective

$$\min_{\theta\in\Theta, \phi\in\Phi} \quad \mathbb{E}_{\alpha,\mathbf{x},\mathbf{x}_0}\left[\|\hat{g}_\phi(\mathbf{x}_\alpha, \alpha) - \mathbf{x}\|_2^2\right], \qquad (6)$$

with $\alpha \sim \mathcal{U}[0, 1]$, $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{x}_0; \mathbf{0}, \boldsymbol{I})$ and

$$\hat{g}_\phi(\mathbf{x}_\alpha, \alpha) = \mathcal{C}(g_\phi(\mathbf{x}_\alpha, \alpha, \mathbf{z}_1), \dots, g_\phi(\mathbf{x}_\alpha, \alpha, \mathbf{z}_N)), \quad (7)$$

with $\mathbf{x}_\alpha = (1-\alpha)\mathbf{x}_0 + \alpha\mathbf{x}$ and $[\mathbf{z}_1, \dots, \mathbf{z}_N] = E_\theta(\mathbf{x})$. We chose the IADB paradigm due to its simplicity in implementation and intuitive nature, requiring minimal hyperparameter tuning.

At inference time, we reconstruct the input by progressively denoising an initial random sample coming from the prior distribution, conditioned on the components obtained through the semantic encoder.

A note on complexity. We found that using a single diffusion model proves effective instead of training N separate models for N latent variables. Consequently, we opt for training a parameter-sharing neural network g_ϕ. Nonetheless, the computational complexity of our framework is therefore N times that of a single DiffAE.

3.2 Recomposition

One of our primary objectives is to endow models with *compositional generation*, a concept we define as the ability to generate novel data examples by coherently recomposing distinct parts extracted from separate origins. This definition aligns with numerous related studies that posit compositional generalization as an essential requirement to bridge the gap between human reasoning and computational learning systems [26]. In this work, we allow for compositional generation by learning a prior model in the components' space. Specifically, once we have a well-trained decomposition model $D_{\theta,\phi} = (E_\theta, g_\phi)$ we learn a diffusion model in $\mathcal{Z}$ in order to obtain a full generative system. We define $\mathbf{z} = [\mathbf{z}_1, \dots, \mathbf{z}_N] = E_\theta(\mathbf{x})$ and train a IADB model to recover $\mathbf{z}$ from a masked view $\tilde{\mathbf{z}}$. At training time, with probability p_{mask}, we mask each latent variable $\mathbf{z}_i$ with a mask $\mathbf{m}_i \in \{0, 1\}^{dim(\mathcal{Z})}$ and optimize the diffusion model ϵ_ψ by solving

$$\min_{\psi\in\Psi} \mathbb{E}_{\alpha,\mathbf{z},\mathbf{z}_0,\mathbf{m}}[\|\mathbf{z} - \epsilon_\psi(\mathbf{z}_\alpha, \alpha, \mathbf{m})\|^2], \qquad (8)$$

Algorithm 1 Training prior model

```
  Input: dataset $\mathcal{D}$, U-Net $\epsilon_\psi$, pre-trained semantic encoder $E_\theta$, masking probability $p_{mask}$, learning rate $\gamma$.
  while not converged do
    for $\mathbf{x}$ in $\mathcal{D}$ do
      $\mathbf{z} = [\mathbf{z}_1, \dots, \mathbf{z}_N] = E_\theta(\mathbf{x})$.
      Sample $\alpha \sim \mathcal{U}[0,1]$ and $\mathbf{z}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.
      $\tilde{\mathbf{z}}_\alpha = (1-\alpha)\mathbf{z}_0 + \alpha\mathbf{z}$
      Draw $\mathbf{m} \in \{0,1\}^{dim(\mathcal{Z}) \times \cdots \times dim(\mathcal{Z})}$
      $\mathbf{z}_\alpha = \tilde{\mathbf{z}}_\alpha \odot \mathbf{m} + (1-\mathbf{m}) \odot \mathbf{z}$
      $\mathcal{L}(\psi, \mathbf{z}, \alpha, \mathbf{m}) = \|\mathbf{z} - \epsilon_\psi(\mathbf{z}_\alpha, \alpha, \mathbf{m})\|^2$
      Update $\psi \leftarrow \psi - \gamma \nabla_\psi \mathcal{L}(\psi, \mathbf{z}, \alpha, \mathbf{m})$
    end for
  end while
  Return: $\epsilon_\psi$
```

where $\mathbf{z}_\alpha = \tilde{\mathbf{z}}_\alpha \odot \mathbf{m} + (1-\mathbf{m}) \odot \mathbf{z}$ and $\tilde{\mathbf{z}}_\alpha = (1-\alpha)\mathbf{z}_0 + \alpha\mathbf{z}$. Here, $\mathbf{z}_0 \sim \mathcal{N}(\mathbf{z}_0; \mathbf{0}, \mathbf{I})$ and $\tilde{\mathbf{z}}_\alpha$ denotes the α-blended source $\mathbf{z}$. At each training iteration we randomly mask $\tilde{\mathbf{z}}_\alpha$ via $\mathbf{m}$ and train the diffusion model ϵ_ψ to recover the masked elements given the unmasked view $\mathbf{z}$. Our masking strategy allows for dropping each latent separately as well as all the latents simultaneously, effectively leading to a model that is able to perform both conditional and unconditional generation at the same time. In our application case, the conditional generation task reduces to the problem of generating variations. As our decomposition model proves to be effective in separating the stems of a given mixture, we obtain a system that is able to generate missing stems given the masked elements. Hence, this also addresses the accompaniment generation task. Algorithm 1 resumes the training process of the prior model.

4. EXPERIMENTS AND RESULTS

This section provides an overview of the experiments aimed at assessing the performance of our proposal in both decomposition (section 4.1) and recomposition (section 4.2) scenarios. Prior to diving into the specifics of each experiment, we provide a brief overview of the shared elements across our experiments, including data, evaluation metrics, and neural network architectures.

Data. We rely on the Slakh2100 dataset [27], a widely recognized benchmark in source separation, comprising 2100 tracks automatically mixed with separate stems. We selected this dataset because of its large-scale nature and the availability of ground truth separated tracks. Following recent approaches in generative models [28, 29], we rely on a pre-trained neural codec to map the audio data to an intermediate latent space, where we apply our approach. Specifically, we employ the EnCodec model [30], a Vector Quantized-VAE (VQ-VAE) model [31] that incorporates Residual Vector Quantization [32] to achieve state-of-the-art performances in neural audio encoding. We take 24 kHz mixtures from the Slakh2100 dataset, which we then feed to the pre-trained EnCodec model to extract the continuous representation obtained by decoding the discrete codes. EnCodec maps raw audio to latent trajectories with a sampling rate of 75 Hz. Specifically, we take audio crops of approximately $7s$ ($6.82s$), which are mapped via EnCodec to a latent code $\mathbf{x} \in \mathbb{R}^{128\times512}$.

MS-STFT	FAD (LC-A)	FAD (LC-M)
4.7	0.05	0.04

Table 1. EnCodec reconstruction quality, measured in terms of MS-STFT and FAD and computed following the procedure descried in section 4.

Evaluation metrics. Throughout this section, we report quantitative *reconstruction* metrics in terms of both Mean Squared Error (MSE) and Multi-Scale Short-Time Fourier Transform (MS-STFT) [33, 34] for latent and audio data, respectively. We perform the MS-STFT evaluation using five STFT with window sizes $\{2048, 1024, 512, 256, 128\}$ following the implementation of [34]. In order to evaluate the quality of the generated samples and the adherence to the training distribution, we also compute Fréchet Audio Distance (FAD) [35, 36] scores. Specifically, we obtain the FAD scores via the `fadtk` library [36], employing both the LAION-CLAP-Audio (LC-A) and LAION-CLAP-Music (LC-M) models [37], as it was shown in [36] that these embedding models correlate well with perceptual tests measuring subjective quality of pop music. In assessing FAD scores, we utilize the complete test set of Slakh2100, while for MSE and MS-STFT values, we randomly select 512 samples of $7s$ ($\sim$ 1 hour) from the same test set and report their mean and standard deviation. Finally, in order to provide the reader a reference value, we report in Table 1 the reconstruction metrics for the pre-trained EnCodec.

When assessing the effectiveness of *source separation* models, we adhere to common practice by relying on the `museval` Python library [38] to compute standard separation metrics: Source-to-Interference Ratio (SIR), Source-to-Artifact Ratio (SAR), and Source-to-Distortion Ratio (SDR) [39]. These metrics are widely accepted for evaluating source separation models, where SDR reflects sound quality, SIR indicates the presence of other sources, and SAR evaluates the presence of artifacts in a source. Specifically, following [39] we compute their scale-invariant (SI) versions and, hence, provide our results in terms of SI-SDR, SI-SIR and SI-SAR. The values shown are expressed in terms of mean μ and standard deviation σ computed on 512 samples of $\sim 7s$ from the Slakh2100 test set.

Architectures. We use a standard U-Net [21] with 1D convolution and an encoder-decoder architecture with skip connections. Each processing unit is a ResNet block [40] with group normalization [41]. Following [42], we feed the noise level information through Positional Encoding [43], conditioning each layer with the AdaGN mechanism. We also add multi-head self-attention [43] in the bottleneck layers of the U-Net. The semantic encoder mirrors the U-Net encoder block without the attention mechanism and maps the data $\mathbf{x} \in \mathbb{R}^{128\times512}$ to a set of variables $\mathbf{z} = [\mathbf{z}_1 \dots \mathbf{z}_i \dots \mathbf{z}_N]$ whose dimensionality is $\mathbf{z}_i \in \mathbb{R}^{1\times512}$.

Finally, these univariate latent variables condition the U-Net via a simple concatenation, which proved to be a sufficiently effective conditioning mechanism for the model to converge. We use the same U-Net architecture for both the decomposition and recomposition diffusion models.

4.1 Decomposition

In order to show the effectiveness of our decomposition method described in section 3.1, we perform multiple experiments on Slakh2100. Throughout this section, we fix the number of training epochs to 250 and use the AdamW optimizer [44] with a fixed learning rate of 10^{-4} as our optimization strategy. The U-Net and semantic encoder have 13 and 8 million trainable parameters, respectively. Finally, we use 100 sampling steps at inference time.

First, we show in Table 2 that our model can be used to perform unsupervised latent source separation and compare it against several non-neural baselines [45–49], as well as a recent study that explicitly targets neural latent blind source separation [50]. We also report the results obtained by Demucs [51], which is the current top performing fully-supervised state-of-the-art method in audio source separation. As the only non-neural baseline, LASS, has been trained and evaluated on the Drums + Bass subset, we perform our analysis on this split and subsequently perform an ablation study over the other sources.

Model	SI-SDR (↑)	SI-SIR (↑)	SI-SAR (↑)
rPCA [45]	-2.8 (4.8)	5.2 (7.3)	5.6 (4.6)
REPET [48]	-0.5 (4.8)	6.8 (7.0)	3.0 (5.2)
FT2D [49]	-0.2 (4.7)	5.1 (7.0)	3.1 (4.7)
NMF [46]	1.4 (5.0)	8.9 (7.6)	2.9 (4.5)
HPSS [47]	2.3 (4.8)	9.9 (7.5)	5.1 (4.6)
LASS [50]	-3.3 (10.8)	17.7 (11.6)	-1.6 (11.2)
Ours	5.5 (4.6)	**41.7** (9.3)	5.6 (4.6)
Demucs [51]	**11.9** (5.0)	37.6 (8.7)	**12.0** (5.0)

Table 2. Blind source separation results for the *Drums* + *Bass* subset. Our model is trained with the *mean* composition operator. The results are expressed in dB as the mean (standard deviation) across 512 elements randomly sampled from the test set of Slakh2100.

As we can see, our model outperforms the other baselines in terms of SI-SDR and SI-SIR and performs on par with respect to SI-SAR. Interestingly, our model outperforms the Demucs supervised baseline in terms of SI-SIR, which is usually interpreted as the amount of other sources that can be heard in a source estimate. In order to test LASS performances, we used their open source checkpoint which is trained on the Slakh2100 dataset, and followed their evaluation strategy. Unfortunately, we were not able to reproduce their results in terms of SDR but we found that their model performs well in terms of SI-SIR, which they did not measure in the original paper. Moreover, as LASS comprises training one transformer model per source, we found their inference phase to be more computationally demanding than ours. Finally, among non-neural baselines, we see that the HPSS model outperforms the others. This seems reasonable as HPSS is specifically built for separating percussive and harmonic sources and hence naturally fits this evaluation context.

Operator	MSE (↓) $\times 10^4$	MS-STFT (↓)
Sum	1.87820 (0.13418)	3.6 (0.1)
Mean	1.87020 (0.13183)	3.6 (0.1)
Min	2.54182 (0.17714)	4.5 (0.1)
Max	2.43302 (0.17510)	4.3 (0.1)

Table 3. Reconstruction quality in latent space (MSE) and audio (MS-STFT) of our decomposition-recomposition model for different recomposition operators for the *Drums* + *Bass* subset.

Moreover, in order to show the robustness of our approach against different sources and number of latent variables, we train multiple models on different subset of the Slakh2100 dataset, namely *Drums* + *Bass, Piano* + *Bass* and *Drums* + *Bass* + *Piano*. The interested reader can refer to our supplementary material and listen to the separation results.

Subsequently, we show that our objective in Equation 6 is robust across different composition operators. We show that, for simple functions such as *sum, min, max* and *mean* our model is able to effectively converge and provide accurate reconstructions. Again, we provide this analysis by training our model on the *Drums* + *Bass* subset of Slakh2100, fixing the number of components to 2. We report quantitative results in terms of two reconstruction metrics, the Mean Squared Error (MSE) and Multi-Scale STFT distance (MS-STFT) in Table 3. As we can see, *sum* and *mean* operators provided the best results, while *min* and *max* proved to be less effective. Nonetheless, the audio reconstruction quality measured in terms of MS-STFT provided reconstruction scores that are lower or comparable with respect to those obtained by evaluating EnCodec performances.

4.2 Recomposition

As detailed in section 3.2, once we are able to decompose our data into a set of composable representations we can then learn a prior model for generation from this new space. Since our decomposition model is able to compress meaningful information through the semantic encoder, we can learn a second latent diffusion model on this compressed representation to obtain a full generative model able to both decompose and generate data.

Here, we validate our claims by training a masked diffusion model for the *Drums* + *Bass* split of the Slakh2100 dataset. In Table 4, we show that our model can indeed produce good-quality unconditional generations by comparing it against a fully unconditional model. We measure the generation quality in terms of FAD scores computed against both the original as well as the encoded test data. Here, by original data we mean the audio coming

	Original		Encoded	
	FAD (LC-A) (↓)	FAD (LC-M) (↓)	FAD (LC-A) (↓)	FAD (LC-M) (↓)
Unconditional	0.09	0.09	0.06	0.06
$p_{mask} = 0.8$	0.12	0.11	0.08	0.07
Bass	0.03	0.03	0.01	0.01
Drums	0.09	0.08	0.05	0.05

Table 4. Audio quality of unconditional generations by our generative model. We demonstrate that we can jointly learn an unconditional and conditional model by showing that the FAD scores of $p_{mask} = 0.8$ are comparable to those of an unconditional latent diffusion model.

	Type	MSE $\times 10^3$	MS-STFT
Real	Drums	2.3259 (0.1287)	13.6 (0.4)
	Bass	1.4393 (0.0874)	9.38 (0.2)
Rand	Drums	4.8170 (0.1136)	20.5 (0.6)
	Bass	4.8814 (0.1157)	21.7 (0.7)

Table 5. Diversity of variations generated by our prior model, measured via the MSE and MS-STFT distances against ground truth and random components.

from the test split of Slakh2100, while the encoded data represents the same elements reconstructed with our decomposition algorithm. As we train on the representations obtained through the semantic encoder, the natural benchmark for the unconditional generation is given by the reconstructions that we can obtain through our decomposition model, which represents the bottleneck in terms of quality. Nonetheless, we show that the FAD scores do not drop substantially when comparing against the original audio, showing that we can indeed achieve a good generation quality. In the same table, we report the partial generation FAD scores. Instead of generating both components unconditionally, we generate the Bass (Drums) given the Drums (Bass), and measure the FAD against the original and the encoded test data, as done for the unconditional case. Given the presence of a ground-truth element, the FAD scores are lower, which is to be expected. Specifically, we can see that the drums generation is a more complex task with respect to the bass generation, as the model needs to synthesize more elements such as the kick, snare and hi-hats, matching the timing of a given bassline.

Lastly, as we strive for high-quality generations, we also aim to enhance diversity within our generations. Table 5 shows the diversity scores for partial generations obtained with our model. We measure diversity in terms of MSE and MS-STFT scores computed, respectively, in the latent and audio space. We compare our partial generations against real and random components, in order to provide the lower and upper bound for generation diversity. Specifically, given the Drums (Bass) we generate the Bass (Drums) and we compute both MSE and MS-STFT scores against the ground truth (Real) and random elements (Rand) coming from the test set of Slakh2100. From the values reported in Table 5, we can deduce that our model produces meaningful variations. We invite the interested readers to listen to our results on our support website.

5. DISCUSSION AND FURTHER WORKS

While our model proves to be effective for compositional representation learning, it still has shortcomings. Here, we briefly list the weaknesses of our proposal and highlight potential avenues for future investigations.

Factors of convergence. In this paper, we used EnCodec which already provides some disentanglement and acts as a sort of initialization strategy for our method. We argue that this property, jointly with the low dimensionality of the latent space enforced by our encoder leads our decomposition model to converge efficiently, not requiring further inductive biases towards source separation.

Limitations. First, there is no theoretical guarantee that the learned latent variables are bound to encode meaningful information. Exploring more refined approaches, as proposed by [52], could be interesting in order to incorporate a more principled method for learning disentangled latent representations. Furthermore, we observed that the dimensionality of the latent space significantly influences the representation content. A larger dimensionality allows the model to encode all the information in each latent, hindering the learning of distinct factors. Conversely, a smaller dimensionality may lead to under-performance, preventing the model to correctly converge. It could be interesting to investigate strategies such as Information Bottleneck [53] to introduce a mechanism to explicitly trade off expressivity with compression. Finally, using more complex functions as well as learnable operators is an interesting research direction for studying the interpretability of learned representations.

6. CONCLUSIONS

In this work, we focus on the problem of learning unsupervised compositional representations for audio. We build upon recent state-of-the-art diffusion generative models to design an encoder-decoder framework with an explicit inductive bias towards compositionality. We validate our approach on audio data, showing that our method can be used to perform latent source separation. Despite the theoretical shortcomings, we believe that our proposal can serve as a useful framework for conducting research on the topics of unsupervised compositional representation learning.

7. REFERENCES

[1] S. Bond-Taylor, A. Leach, Y. Long, and C. G. Willcocks, "Deep generative modelling: A comparative review of vaes, gans, normalizing flows, energy-based and autoregressive models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 7327–7347, 2022.

[2] A. Srivastava, A. Rastogi, A. Rao, A. A. M. Shoeb, A. Abid, A. Fisch, A. R. Brown, A. Santoro, A. Gupta, A. Garriga-Alonso, A. Kluska, A. Lewkowycz, A. Agarwal, A. Power, A. Ray, A. Warstadt, A. W. Kocurek, A. Safaya, A. Tazarv, A. Xiang, A. Parrish, A. Nie, and et al., "Beyond the imitation game: Quantifying and extrapolating the capabilities of language models," *Transactions on Machine Learning Research*, 2023. [Online]. Available: https://openreview.net/forum?id=uyTL5Bvosj

[3] M. Yuksekgonul, F. Bianchi, P. Kalluri, D. Jurafsky, and J. Zou, "When and why vision-language models behave like bags-of-words, and what to do about it?" in *International Conference on Learning Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=KRLUvxh8uaX

[4] P. Pagin and D. Westerståhl, "Compositionality i: Definitions and variants," *Philosophy Compass*, vol. 5, no. 3, pp. 250–264, 2010. [Online]. Available: https://compass.onlinelibrary.wiley.com/doi/abs/10.1111/j.1747-9991.2009.00228.x

[5] T. Janssen, "19 Compositionality: Its Historic Context," in *The Oxford Handbook of Compositionality*. Oxford University Press, 02 2012. [Online]. Available: https://doi.org/10.1093/oxfordhb/9780199541072.013.0001

[6] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman, "Building machines that learn and think like people," *Behavioral and Brain Sciences*, vol. 40, p. e253, 2017.

[7] J. Mu and J. Andreas, "Compositional explanations of neurons," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS'20. Red Hook, NY, USA: Curran Associates Inc., 2020.

[8] G. Hinton, "How to Represent Part-Whole Hierarchies in a Neural Network," *Neural Computation*, vol. 35, no. 3, pp. 413–452, 02 2023. [Online]. Available: https://doi.org/10.1162/neco_a_01557

[9] B. Lake and M. Baroni, "Human-like systematic generalization through a meta-learning neural network," *Nature*, vol. 623, no. 7985, pp. 115–121, Nov. 2023, publisher Copyright: © 2023, The Author(s).

[10] G. Mariani, I. Tallini, E. Postolache, M. Mancusi, L. Cosmo, and E. Rodolà, "Multi-source diffusion models for simultaneous music generation and separation," in *The Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=h922Qhkmx1

[11] K. Preechakul, N. Chatthee, S. Wizadwongsa, and S. Suwajanakorn, "Diffusion autoencoders: Toward a meaningful and decodable representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 619–10 629.

[12] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors." *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[13] T. Karras, M. Aittala, T. Aila, and S. Laine, "Elucidating the design space of diffusion-based generative models," in *Proc. NeurIPS*, 2022.

[14] Z. He, T. Sun, Q. Tang, K. Wang, X. Huang, and X. Qiu, "DiffusionBERT: Improving generative masked language models with diffusion models," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Toronto, Canada: Association for Computational Linguistics, Jul. 2023, pp. 4521–4534. [Online]. Available: https://aclanthology.org/2023.acl-long.248

[15] H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, W. Wang, and M. D. Plumbley, "AudioLDM: Text-to-audio generation with latent diffusion models," *Proceedings of the International Conference on Machine Learning*, 2023.

[16] J. Ho, T. Salimans, A. A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet, "Video diffusion models," in *ICLR Workshop on Deep Generative Models for Highly Structured Data*, 2022. [Online]. Available: https://openreview.net/forum?id=BBelR2NdDZ5

[17] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International conference on machine learning*. PMLR, 2015, pp. 2256–2265.

[18] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[19] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/file/3001ef257407d5a371a96dcd947c7d93-Paper.pdf

[20] P. Vincent, "A connection between score matching and denoising autoencoders," *Neural computation*, vol. 23, no. 7, pp. 1661–1674, 2011.

[21] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.

[22] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=St1giarCHLP

[23] E. Heitz, L. Belcour, and T. Chambon, "Iterative α -(de)blending: A minimalist deterministic diffusion model," in *ACM SIGGRAPH 2023 Conference Proceedings*, ser. SIGGRAPH '23. New York, NY, USA: Association for Computing Machinery, 2023. [Online]. Available: https://doi.org/10.1145/3588432.3591540

[24] J. Andreas, "Measuring compositionality in representation learning," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=HJz05o0qK7

[25] T. Wiedemer, P. Mayilvahanan, M. Bethge, and W. Brendel, "Compositional generalization from first principles," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. [Online]. Available: https://openreview.net/forum?id=LqOQ1uJmSx

[26] J. A. Fodor and Z. W. Pylyshyn, "Connectionism and cognitive architecture: A critical analysis," *Cognition*, vol. 28, no. 1-2, pp. 3–71, 1988.

[27] E. Manilow, G. Wichern, P. Seetharaman, and J. Le Roux, "Cutting music source separation some Slakh: A dataset to study the impact of training data quality and quantity," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2019.

[28] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," 2021.

[29] F. Schneider, O. Kamal, Z. Jin, and B. Schölkopf, "Moûsai: Text-to-music generation with long-context latent diffusion," 2023.

[30] A. Défossez, J. Copet, G. Synnaeve, and Y. Adi, "High fidelity neural audio compression," *arXiv preprint arXiv:2210.13438*, 2022.

[31] A. van den Oord, O. Vinyals, and K. Kavukcuoglu, "Neural discrete representation learning," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 6309–6318.

[32] N. Zeghidour, A. Luebs, A. Omran, J. Skoglund, and M. Tagliasacchi, "Soundstream: An end-to-end neural audio codec," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 30, p. 495–507, nov 2021. [Online]. Available: https://doi.org/10.1109/TASLP.2021.3129994

[33] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6199–6203.

[34] A. Caillon and P. Esling, "Rave: A variational autoencoder for fast and high-quality neural audio synthesis," 2021.

[35] K. Kilgour, M. Zuluaga, D. Roblek, and M. Sharifi, "Fréchet audio distance: A metric for evaluating music enhancement algorithms," 2019.

[36] A. Gui, H. Gamper, S. Braun, and D. Emmanouilidou, "Adapting frechet audio distance for generative music evaluation," in *Proc. IEEE ICASSP 2024*, 2024. [Online]. Available: https://arxiv.org/abs/2311.01616

[37] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2023.

[38] F.-R. Stöter, A. Liutkus, and N. Ito, "The 2018 signal separation evaluation campaign," in *Latent Variable Analysis and Signal Separation: 14th International Conference, LVA/ICA 2018, Surrey, UK*, 2018, pp. 293–305.

[39] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr - half-baked or well done?" 2018.

[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[41] Y. Wu and K. He, "Group normalization," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[42] P. Dhariwal and A. Q. Nichol, "Diffusion models beat GANs on image synthesis," in *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021. [Online]. Available: https://openreview.net/forum?id=AAWuCvzaVt

[43] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[44] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7

[45] P.-S. Huang, S. D. Chen, P. Smaragdis, and M. Hasegawa-Johnson, "Singing-voice separation from monaural recordings using robust principal component analysis," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 57–60.

[46] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[47] D. Fitzgerald, "Harmonic/percussive separation using median filtering," *13th International Conference on Digital Audio Effects (DAFx-10)*, 01 2010.

[48] Z. Rafii and B. Pardo, "Music/voice separation using the similarity matrix," in *Proceedings of the 13th International Society for Music Information Retrieval Conference, ISMIR 2012*, ser. Proceedings of the 13th International Society for Music Information Retrieval Conference, ISMIR 2012, 2012, pp. 583–588, 13th International Society for Music Information Retrieval Conference, ISMIR 2012 ; Conference date: 08-10-2012 Through 12-10-2012.

[49] P. Seetharaman, F. Pishdadian, and B. Pardo, "Music/voice separation using the 2d fourier transform," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017, pp. 36–40.

[50] E. Postolache, G. Mariani, M. Mancusi, A. Santilli, L. Cosmo, and E. Rodolà, "Latent autoregressive source separation," in *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, ser. AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023. [Online]. Available: https://doi.org/10.1609/aaai.v37i8.26131

[51] A. Défossez, N. Usunier, L. Bottou, and F. Bach, "Music source separation in the waveform domain," 2020. [Online]. Available: https://openreview.net/forum?id=HJx7uJStPH

[52] Y. Wang, Y. Schiff, A. Gokaslan, W. Pan, F. Wang, C. De Sa, and V. Kuleshov, "InfoDiffusion: Representation learning using information maximizing diffusion models," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 23–29 Jul 2023, pp. 36 336–36 354. [Online]. Available: https://proceedings.mlr.press/v202/wang23ah.html

[53] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," 2000.