Interactive Machine Learning for Generative Models

Junichi Shimizu, Ireti Olowe, Terence Broad,
Gabriel Vigliensoni, Prashanth Thattai, Rebecca Fiebrink
Creative Computing Institute
University of the Arts London
London, UK
{j.shimizu, i.olowe, t.broad}@arts.ac.uk
{g.vigliensoni, p.thattairavikumar, r.fiebrink}@arts.ac.uk

Abstract

Effective control of generative media models remains a challenge for specialised generation tasks, including where no suitable dataset to train a contrastive language model exists. We describe a new approach that enables users to interactively create bespoke text-to-media mappings for arbitrary media generation models, using a small number of examples. This approach—very distinct from contrastive language pretraining approaches—facilitates new strategies for using language to drive media creation in creative contexts not well served by existing methods.

1 Motivation and Background

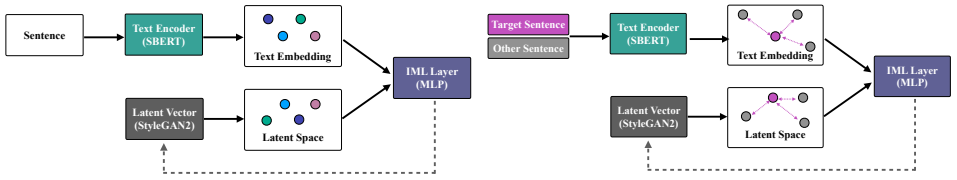
Text-to-media models using contrastive language pretraining approaches, such as CLIP [11] and CLAP [15], offer useful ways to drive media generation. However, they depend on the large amounts of relevant text-media data, which is not available for more specialised creative generation tasks—including less common generation modalities (e.g., choreography, textile weaving) and generation using specialised or personalised models (e.g., models trained on an artist's own works, or models targeting professional-quality generation from a narrow range of sound effects). Controlling such models instead generally depends on methods for manipulating latent vectors. When latent space is low-dimensional, this can be sufficient for meaningful control (e.g., [14]). Some current research considers techniques for improving usability in higher-dimensional spaces; e.g., [3] demonstrates a method to force dimensions of a latent space to map to musically meaningful attributes.

2 Our Approach

Our complementary approach enables people to dynamically build and adjust text-to-media mappings for arbitrary generative models. Here, a user can first quickly define a basic relationship between natural language and a model's latent space by labeling a few media exemplars (e.g., initially chosen through random generation or mouse selection on a 3D latent space projection) with short text descriptions. A simple model (e.g., a 1-hidden-layer multilayer perceptron) is trained on these to map from language embeddings to generative model latent vectors (Figure 1a). A user can then type a new text description to jump to a new location in latent space. Crucially, a user can also employ an interactive machine learning (IML) [6] workflow, adding and deleting training examples to iteratively improve the mapping or dynamically adjust it to focus on new areas of interest in the latent space. IML has been shown to be helpful to creators designing bespoke mappings in contexts including control of sound synthesis [7] and game engine [5] parameters, and [13] recently demonstrated its use in designing instruments in which gesture controls generative sound models.

Our approach also allows people to use text to generate local variations on a given point in latent space. This is inspired by StyleCLIP's approach to global directions that manipulate direction by giving

37th Conference on Neural Information Processing Systems (NeurIPS 2023).



(a) Text-to-media mapping method (b) Local variation method

Figure 1: System Explanation

a pair of images from CLIP(text) embedding and Style code $s \in S$ as proposed by Patashnik et al. [9]. StyleCLIP distinguishes between the manifolds of image embeddings and text embeddings in CLIP's joint embedding space by utilizing manipulation directions ($\Delta\mathbf{s}$) determined by the relevance of each style channel to the target attribute. In our approach, we compute the desired direction of variation by subtracting target embedding and other embeddings as well as those latent vector(Figure 1b). The IML layer learn relative distance between target and other data to make a variation of target data.Further offer a user interface slider to interactively adjust the magnitude of variation.

We have illustrated this approach using a StyleGAN2-based model for drum sample and drum loop sound generation [8], as well as Sentence-BERT [12] for text embeddings. In StyleGAN2, performing image manipulation in W or W^+ space can show better disentanglement rather than Z space which is randomly sampled, so we use the latent vector from W. However, alternatives are possible, e.g., using [16]. In our implementation, we use a 1-hidden-layer MLP to map from sentence embedding $\mathbf{s} \in \mathbb{R}^{384}$ to a drum sound latent vector $\mathbf{w} \in \mathbb{R}^{512}$, though this could be adjusted for other contexts. Our user interface (Appendix A) supports an iterative design of the text-sound model and sound generation and variation. This interface also includes 8 sliders for manually varying sounds along latent space dimensions identified by applying PCA to pre-computed $\mathbf{w}$ vectors, as well as a list of words commonly used to describe percussive sounds [1] which we hypothesise (but have not yet tested) may help new users efficiently identify and inspire useful labels. We emphasise that this approach is applicable to any model generating media from a latent vector, and Appendix B illustrates some results on images, generated using a variant of this user interface.

3 Novel Interaction Strategies for Text-Driven Media Generation

The accompanying video `https://vimeo.com/867836056/d2d2e5b9d5` illustrates some strategies for using the proposed approach. Being trained on small numbers of examples, this approach is ill-suited to long, detailed text instructions characteristic of CLIP/CLAP-based models. Rather, this approach supports distinctive interaction strategies, for instance: (1) A creator can dynamically choose and change which characteristics are relevant to them, for instance, initially providing labels related to the rhythmic density of a drum loop, and then, after identifying a region of latent space with desired density, adding labels to describe and navigate relative "grooviness" of drum patterns in this region. (2) A creator can use words that have personally specific meanings, or that are only defined relative to the current local context (e.g., "squishy,", "weird", "boring"). (3) Text-driven navigation offers interesting alternatives to slider-based interpolation between known latent vectors. For instance, if one drum loop is labeled "straight" and another "groovy," a prompt like "between straight and groovy" can interpolate, "extremely groovy" can extrapolate, and prompts using other words (e.g., "funky and a bit noisy") can explore the vicinity with more nuance and even explicit or implicit reference to other labeled examples. (4) The mapping can be refined to better represent one's expectations within an area of interest, for instance tweaking the "funky and a bit noisy" generated sound with on-screen variation sliders until one is happier with the outcome, then adding a new training example for the mapping pairing this description with the varied sound.

This approach will not always produce satisfactory mappings for all text-sound pairs, given the high dimensionality of the input and output latent spaces, and the possibility for users to provide inconsistent labelings. However, the interactive ability to explore and change mappings means that users can adapt their expectations and strategies through experimentation.

4 Ethical Implications

Popular text-driven models that use contrastive language pre-training often present ethical problems regarding sourcing their large training sets without consent or proper IP consideration, as well as bias in what types of media are included or excluded by nature of relying on very large, available datasets. In contrast, our research provides creative practitioners with new approaches to controlling generative media models that may be trained on smaller datasets and/or for more bespoke tasks. In these cases, it may often be easier to ethically and legally source training data (e.g., from an artist's own collection of work). Further, the relationships between text and media are defined by the artist themselves; while certain biases are still potentially present in any text embedding model [10], this approach is not likely to result in the same types of systematic biases that have been observed in text-generated media in large-scale contrastive pre-training models [4]. Our approach thus potentially presents ethical advantages compared to existing popular methods.

References

[1] Robert Bell. *PAL: The Percussive Audio Lexicon*. PhD thesis, Swinburne University of Technology, Melbourne, Australia, 2015.

[2] Terence Broad, Frederic Fol Leymarie, and Mick Grierson. Network bending: Expressive manipulation of deep generative models. In Juan Romero, Tiago Martins, and Nereida Rodríguez-Fernández, editors, *Artificial Intelligence in Music, Sound, Art and Design*, 2021.

[3] Nick Bryan-Kinns, Berker Banar, Corey Ford, Courtney N Reed, Yixiao Zhang, Simon Colton, and Jack Armitage. Exploring xai for the arts: Explaining latent space in generative music. *arXiv preprint arXiv:2308.05496*, 2023.

[4] Nassim Dehouche. Implicit stereotypes in pre-trained classifiers. *IEEE Access*, 9:167936–167947, 2021.

[5] Carlos Gonzalez Diaz, Phoenix Perry, and Rebecca Fiebrink. Interactive machine learning for more expressive game interactions. In *2019 IEEE Conference on Games (CoG)*, pages 1–2. IEEE, 2019.

[6] Jerry Alan Fails and Dan R. Olsen. Interactive machine learning. In *Proceedings of the 8th International Conference on Intelligent User Interfaces*, IUI '03, page 39–45, New York, NY, USA, 2003. URL `https://doi.org/10.1145/604045.604056`.

[7] Rebecca Fiebrink, Perry R. Cook, and Dan Trueman. Human model evaluation in interactive supervised learning. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, page 147–156, New York, NY, USA, 2011. URL `https://doi.org/10.1145/1978942.1978965`.

[8] Tun-Min Hung, Bo-Yu Chen, Yen-Tung Yeh, and Yi-Hsuan Yang. A benchmarking initiative for audio-domain music generation using the FreeSound Loop Dataset. In *Proceedings of the International Society for Music Information Retrieval Conference*, 2021. URL `https://github.com/allenhung1025/LoopTest`.

[9] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2085–2094, October 2021.

[10] Davor Petreski and Ibrahim C Hashim. Word embeddings are biased. but whose bias are they reflecting? *AI & SOCIETY*, 38(2):975–982, 2023.

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[12] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL `https://arxiv.org/abs/1908.10084`.

[13] Gabriel Vigliensoni and Rebecca Fiebrink. Steering latent audio models through interactive machine learning. In *Proceedings of the 14th International Conference on Computational Creativity (ICCC'23)*, page 19–23, 6 2023.

[14] Gabriel Vigliensoni, Louis McCallum, Esteban Maestre, and Rebecca Fiebrink. R-VAE: Live latent space drum rhythm generation from minimal-size datasets. *Journal of Creative Music Systems 1(1)*, 2022. URL https://doi.org/10.5920/jcms.902.

[15] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[16] Zongze Wu, Dani Lischinski, and Eli Shechtman. Stylespace analysis: Disentangled controls for stylegan image generation, 2020.

Appendix A: Prototype User Interface Images

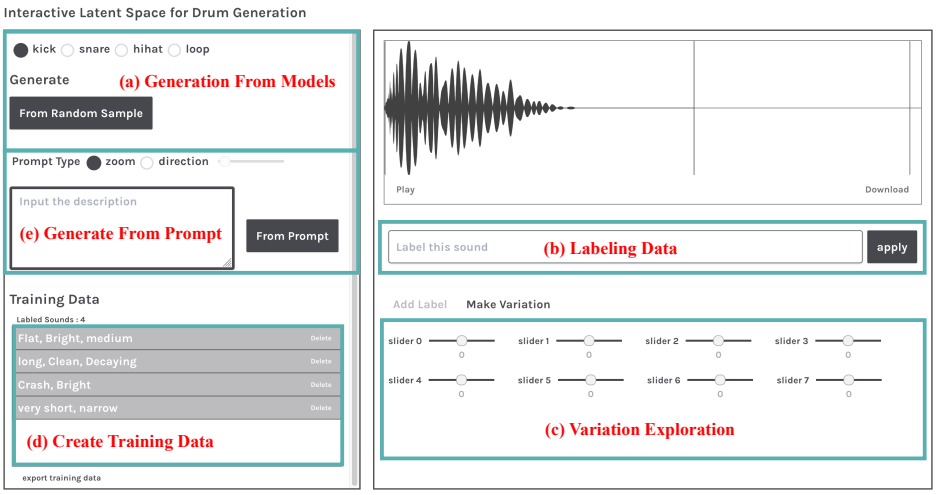


Figure 2: The user interface for applying this method to drum sample and loop generation. In (a), a user can generate a new sound from the selected model (kick, snare, hihat, or percussion loop) by randomly choosing a latent space vector. The generated sound will then appear at top right and be playable. A user can repeat this until they identify a sound of interest, at which point they can use the text field (b) to assign a text label (one word or a longer description) to the sound. Once a sound is assigned a label, it appears in the list of training data (d); sounds in this list can be selected, at which point their waveform appears in top right and they can be played or have their label edited. Sounds in the training data can also be deleted. Section (c) of the user interface, "Make Variation", allows the user to transform any selected sound by applying sliders. As sliders are manipulated, the sound at top right is replaced with the variation.

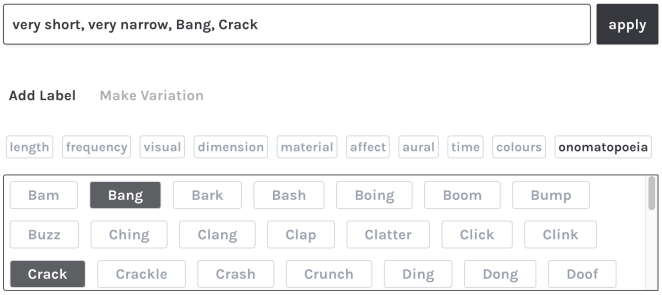


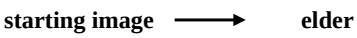
Figure 3: When working with any of the single drum-sound sample models (kick, snare, hihat), the "Add Label" pane (appearing just below the data labelling text box (b) in Figure 2) can be used. This box provides a list of words commonly used to describe percussive sounds [1]. Clicking on any of these words adds text to the data labeling box above it, where they can be edited if desired. Our intention in providing this functionality is to help new users identify words that might make for useful labels; future research will test its usefulness.

Appendix B: Illustration of Approach Applied to Image Generation

In the image domain example below, we used the StyleGAN2 FFHQ 1024 model, specifically utilized in [2].

(a) Three training examples are first defined.

(b) After these (and only these) examples are defined, generating new images using the prompts "child", "man", "elder" and "youth" results in these images.



(c) When the image at left is used as the starting image, the prompt "elder" at slider values of 30%, 60%, and 90% results in the three variations at right. Ideally, it should control only the specific attribute ("elder"). However, when the slider value is increased, this affects other attributes slightly (e.g., smiling). Future work could explore how to modify the proposed system regulate such tiny style attributes through a dedicated style channel.

Figure 4