UniBriVL: Robust Universal Representation and Generation of Audio Driven Diffusion Models

Sen Fang[1], **Bowen Gao**[2,*], **Yangjian Wu**[3], **Teik Toe Teoh**[4]
[1,2]Victoria University, [3]Hainan University, [4]Nanyang Technological University
{sen.fang, bowen.gao}@live.vu.edu.au, yangjian.wu@hainanu.edu.cn
ttteoh@ntu.edu.sg

Abstract

Multimodal large models have been recognized for their advantages in various performance and downstream tasks. The development of these models is crucial towards achieving general artificial intelligence in the future. In this paper, we propose a novel universal language representation learning method called UniBriVL, which is based on Bridging-Vision-and-Language (BriVL). **Uni**versal **BriVL** embeds audio, image, and text into a shared space, enabling the realization of various multimodal applications. Our approach addresses major challenges in robust language (both text and audio) representation learning and effectively captures the correlation between audio and image. Additionally, we demonstrate the qualitative evaluation of the generated images from UniBriVL, which serves to highlight the potential of our approach in creating images from audio. Overall, our experimental results demonstrate the efficacy of UniBriVL in downstream tasks and its ability to choose appropriate images from audio. The proposed approach has the potential for various applications such as speech recognition, music signal processing, and captioning systems.

1 Introduction

Sound and vision affect people's core cognition in many areas, such as feeling, information processing and communication. Sound and vision are closely related. However, most of the existing methods only have a single cognitive ability, and some only study text-vision, text-voice, etc. Recent studies have shown that leveraging large-scale Internet data for self-supervised pre-training of models offers better results than relying on high-quality or manually labeled data sets (Pan et al., 2022), such as the recently popular chatGPT. Moreover, multiple studies demonstrate the effectiveness of multimodal models over single or bimodal models in several fields and tasks (Chen et al., 2022a), such as Microsoft's latest BEiT3 (Wang et al., 2022), Meta's ImageBind (Girdhar et al., 2023), etc.

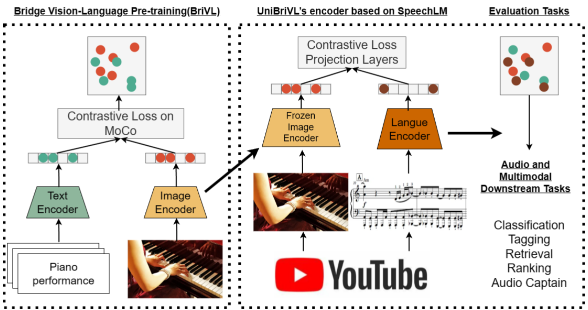


Fig. 1: Our UniBriVL architecture and training flow, we train in conjunction with a SpeechLM encoder, enabling a unified text and audio entry.

Data volume is the basic element for training large-scale language models. Since BERT of Devlin et al. (2018) (perhaps even earlier (Ma and Zhang, 2015)), the pre-training model of NLP has been benefiting from large-scale corpora. According to the theory of Kaplan et al. (2020), the language model gradually reflects a scaling law (the rule that the model capacity increases with the model volume). Manual annotation of large amounts of data in supervised learning is very expensive, so self-supervised learning is valued for large model training. In order to expand the boundary of the research field and break the limitation of the lack of relevant resources (Hsu et al., 2021), we explore a new multimodal self-monitoring model based on the latest excellent work: **Bridging-Vision-and-Language** (Fei et al., 2022). It's a new effort similar to OpenAI CLIP (Radford et al., 2021) and Google ALIGN (Jia et al., 2021). Like CLIP, BriVL can rearrange images based on how well they match text images to find the best match. BriVL[1] model has excellent effect on image and text retrieval tasks, surpassing other common multimodal pre-training models in the same period.

* Collaborator Author.

[1]https://github.com/BAAI-WuDao/BriVL

Proceedings of the The 3rd Workshop on Multi-lingual Representation Learning (MRL), pages 1–11
December 7, 2023

In this work, we propose UniBriVL, an audio-visual correspondence model that extracts training from the BriVL model. As shown in Figure 1, the principle of UniBriVL is to freeze the BriVL visual model, run video on the visual stream of the model, and train a new model to predict BriVL embedding independently from the audio stream. The entry point for our selection of the new language modality is Microsoft's latest developed model, SpeechLM (Zhang et al., 2023), which is a fusion model of text and audio. It is capable of outputting text and audio as the same representation. This allows us to input text, audio, or both when using the model. Consequently, this significantly enhances the adaptability of the model to various tasks, such as audio-text retrieval, image retrieval, audio recognition, image captioning, and even theoretically enables better perception of real-life scenarios through simultaneous processing of live speech and text. We conducted a comprehensive evaluation of our model in the aforementioned tasks. The experimental results demonstrate its strong generalizability and excellent performance in the main experiments.

Finally, we use UniBriVL to guide the generation of model Stable Diffusion[2] (Rombach et al., 2022) output images, and intuitively verify that the embedded space is meaningful. Experimental results show that this method can effectively choose appropriate images from audio. This is a significant contribution to the field of multimodal learning, as prior methods mainly focused on generating images from text or image inputs, rather than audio inputs. In addition, compared with other fully supervised models, UniBriVL theoretically requires less data to obtain competitive performance in downstream tasks, that is, it performs pre-training more effectively than competitive methods, because it does not need to completely re learn the visual model, only needs to train the audio model. It is a reproducible and potential application model, and we will provide our model and more code information after publication.

2 Related Works

The impetus for our research is the considerable progress noticed in multimodal learning, specifically during the early part of 2022. The comparison of BriVL's performance with CLIP (Radford et al., 2021) indicates noteworthy improvements across various benchmarks. Likewise, Microsoft's SpeechLM (Zhang et al., 2023) outshines the former Wav2Vec (Baevski et al., 2020) in several dimensions. We posit that fusing the strengths of BriVL and SpeechLM could indeed result in an enhancement over Wav2CLIP[3]. Crucially, the field is presently underexplored in terms of pioneering endeavors concerning the use of audio-guided diffusion models for image generation.

2.1 Audio dependent multimodal models

There have been many multimodal works that have taken audio into account before, and some have replaced text with audio as the main object for matching with images (Ilharco et al., 2019; Chrupała, 2022). In addition to AudioCLIP (Guzhov et al., 2021) and other similar but actually different work, the most similar to us is Wav2CLIP (Wu et al., 2022). For CLIP, the BriVL we use has the following differences and advantages: Firstly,

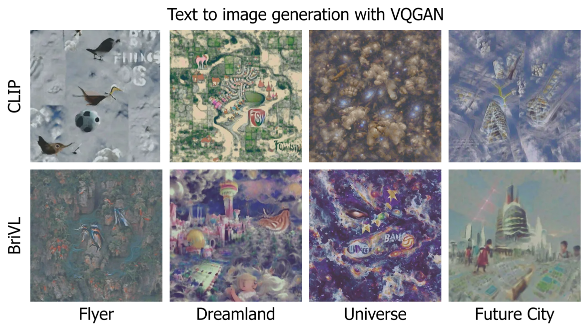


Fig. 2: Examples of CLIP (top) and BriVL (bottom) to image generation from text, BriVL's labels in x-axis are translated.

BriVL has more weak semantic relevance, so our model is more imaginative (We also use naturally distributed weak semantic data.). For example, here are two groups of graphs in Figure 2 generated by using CLIP and BriVL respectively using GAN for comparison and understanding in the field of text-guided generation. Secondly, for our network architecture, because there is not necessarily a fine-grained area match between the image and audio, we lost the time-consuming target detector and adopted a simple and more efficient dual tower architecture. Thirdly, BriVL designed a cross modal comparative learning algorithm based on the single modal comparative learning method MoCo (He et al., 2020), which has different advantages than CLIP.

[2] https://github.com/CompVis/stable-diffusion

[3] https://github.com/descriptinc/lyrebird-wav2clip

2.2 Audio driven image generation

For many years, people have been trying to give AI people multimodal perception and thinking, and one of the main ideas is to simulate people's impressions of different external inputs, namely image generation. The pursuit of applications and methods for generating different images is the direction of researchers' efforts. With the emergence of different generation models, such as Goodfellow introduced GAN in 2014, there has been a lot of excellent work in the field of GAN-based image generation (Karras et al., 2017; Cudeiro et al., 2019; Yi et al., 2020; Zhang et al., 2021a; Song et al., 2022; Zhang et al., 2021b,c; Wu et al., 2021; Lahiri et al., 2021; Richard et al., 2021; Thies et al., 2020; Wen et al., 2020; Chen et al., 2020b). Then, from single mode to multi-mode, from text guidance about 15 years later to audio guidance (Qiu and Kataoka, 2018) 20 years later (of course, there are more and earlier attempts and exceptions), several impressive works appeared (Xu et al., 2018; Zhu et al., 2021; Hessel et al., 2021; Saharia et al., 2022b,a). At a time when diffusion models have achieved success in many fields, exploring based on this work is meaningful.

2.3 Background information

SpeechLM (Zhang et al., 2023) is a neural network model that combines speech and text information to perform language modeling. It consists of two parts: a Speech Transformer and a Shared Transformer, which are enhanced with a random swapping mechanism. The Speech Transformer uses a standard Transformer with relative position embedding to process the speech waveform into speech features, which are then masked and further processed by the Speech Transformer to obtain higher-level representations. A speech waveform $\boldsymbol{S}$ is first processed into a sequence of speech features $\boldsymbol{X} = (x_1, x_2, \ldots, x_M)$ by a stack of 1-D convolutional layers. They follow HuBERT to mask the speech feature $\boldsymbol{X}$ with the mask probability of 8% and the mask length of 10. Then the masked features, $\hat{\boldsymbol{X}}$, are fed into the Speech Transformer for higher level representations $\boldsymbol{H}^l = \text{Transformer}(\boldsymbol{H}^{l-1})$, where l means the layer and $\boldsymbol{H}^0 = \hat{\boldsymbol{X}}$ indicates the input. The Shared Transformer has the same architecture, but takes in both the encoded speech representations and the embeddings derived from tokenized text units. To better align the speech and text representations in the same latent space, they introduce a random swapping mechanism that randomly replaces some speech features with corresponding text embeddings. They randomly select some positions from the unmasked region of speech and replace the lower representations $h_i^{L/2}$ with the corresponding unit embeddings u_i, where the units are extracted from the input speech sample. In this way, the speech and text modalities can be shuffled into one sequence and treated equally. This is also one of the advantages of our model, we can use it for tasks that require text-image matching as well as voice-image matching, which is very convenient.

3 Methodology And Experiments

BriVL is a model trained on 650 million text image weak semantic datasets. They designed a cross modal comparison learning algorithm based on the monomodal comparison learning method MoCo (He et al., 2020), and maintained the negative sample queue in different training batches through a mechanism called Memory Bank, so as to obtain a large number of negative samples for use in the comparison learning method. In simple terms, it does not incorporate momentum encoders or negative sample queues, instead relying on computing the InfoNCE loss (Oord et al., 2018) within each batch. Specifically, the number of negative samples for each positive image-text pair is determined by the mini-batch size, affording greater flexibility and efficiency in training. It also shows the SOTA results in such scenes as image annotation, image zero sample classification, and input features of other downstream multimodal tasks. Even the guidance generation model has excellent performance.

As mentioned in the introduction, UniBriVL replaces the text encoder with the audio/shared encoder encoder by model of BriVL (In fact, as mentioned in the background information, SpeechLM's feature extraction is shared across text and audio types. The model is retrained after changing the BriVL code, and then fine-tuned together with SpeechLM.), runs the image through it, and trains the new model to predict that only the matching image embedded content is obtained from the audio. We refer to the exclusive multilayer perceptron of BriVL, which can not only enhance performance but also prepare for possible downstream tasks. After the audio encoder is fine-tuned, we freeze it and use it in the UniBriVL image generation task as a qualitative evaluation of our experimental results.

3.1 Dataset for performance test

We select diverse set of data ranging from various number of clips, number of categories, and perform diverse tasks including classification, retrieval, and generation. For evaluation, we use relevant metrics detailed in Table 1 for each task.

3.2 Dataset for training

To train audio-image correspondence, we use the files of the AudioSet (Gemmeke et al., 2017) video datasets as the audio input for our rearrangement of the generated images. AudioSet comprises a growing ontology that encompasses 632 distinct audio event classes and a comprehensive corpus of 2.1 million videos. These clips are annotated by human experts and extracted from YouTube videos, each lasting ten seconds. The ontology is structured as a hierarchical graph of event categories, encompassing a diverse spectrum of human and animal sounds, musical genres and instruments, as well as everyday environmental sounds. We randomly select one image from each sample video, cut them into squares, and sample them down to 64 × 64. The audio sampling rate is 16,000Hz. We use it to train the model, which helps to increase the applicability of the model. In total, we randomly selected 200,000 segments for training and then selected some additional audio for our image generation task.

3.3 Feature extraction processing methods

For image and audio encoders, we use EfficientNet-B7 (Tan and Le, 2019) as the CNN in the image encoder, and the backbone SpeechLM (Zhang et al., 2023) as the basic transformer in the audio encoder. The self concerned block is composed of 4 Transformer encoder layers and MLP block respectively, with two fully connected layers and one ReLU activation layer. For all models, we use grid search to find the best hyperparameter. For other hyperparameters (such as batch size, training steps, learning rate, etc.), we directly use the suggested values in the original papers. Note that for per-instance perturbation, we adopt the appropriate quantity compared to the original epochs.

Picture Encoding. The technique employed by BriVL utilizes random grayscale conversion for the input picture, along with random color jitter for data enrichment. A 720P resolution is utilized for all videos in the dataset, with non-compliant ones being converted to 480P. The pictures are then trimmed to 360 × 360 pixels. Patch features from the picture are captured via a Transformer, followed by employing an average pooling layer for feature integration. To further refine the extraction and depiction of interrelations among the picture patch features, a self-attention (SA) block containing multiple Transformer encoder layers is employed by the BriVL team[4]. Each Transformer encoder layer encompasses a multi-head attention (MHA) layer and a feed-forward network (FFN) layer (Fei et al., 2022):

$$\mathbf{T}' = \text{LayerNorm}(\mathbf{T} + \text{MHA}(\mathbf{T})) \quad (1)$$

$$\mathbf{T} = \text{LayerNorm}(\mathbf{T}' + \text{FFN}(\mathbf{T}')) \quad (2)$$

Post this, they make use of an average pooling layer to amalgamate the extracted patch features:

$$\mathbf{q}^{(i)} = \frac{1}{N_p} \sum_{j=1}^{N_p} \mathbf{T}_j \in \mathbb{R}^c \quad (3)$$

wherein $\mathbf{T}_j$ stands for the j-th column of $\mathbf{T}$. To project $\mathbf{q}^{(i)}$ to the joint cross-modal embedding space, a two-layer MLP block equipped with a ReLU activation layer is used. This results in generating the ultimate d-dimensional picture embedding $\mathbf{y}^{(i)} \in \mathbb{R}^d$.

Audio Encoder. For audio input, we first convert the original audio waveform (1D) into a spectrum (2D) as the input of SpeechLM, and pool the entire 512 dimensional audio sequence to output an embedding. The SpeechLM embedding is computed by the weighted average of outputs from all transformer layers. The SpeechLM[5] model inspired by HuBERT (Hsu et al., 2021) consists of a Speech Transformer and a Shared Transformer, which are enhanced with the random swapping mechanism. The Transformer is optimized to predict the discrete target sequence **z**, in which each $z_t \in [C]$ is a C-class categorical variable. The distribution over the classes is parameterized with

$$p(c|\mathbf{n}_t) = \frac{\exp(\text{sim}(\mathbf{K}^P \mathbf{n}_t^L, \mathbf{e}_c)/\tau)}{\sum_{c'=1}^{C} \exp(\text{sim}(\mathbf{K}^P \mathbf{n}_t^L, \mathbf{e}_{c'})/\tau)} \quad (4)$$

where $\mathbf{K}^P$ is a projection matrix, $\mathbf{n}_t^L$ is the output hidden state for step t, $\mathbf{e}_c$ is the embedding for class c, $\text{sim}(a, b)$ means the cosine similarity between a and b, and $\tau = 0.1$ scales the logit (Chen

[4] https://github.com/BAAI-WuDao/BriVL
[5] https://aka.ms/SpeechLM

Dataset	Task	Clip (Split)	Class	Metric
ESC-50 (Piczak, 2015)	MC/ZS	2k (5 folds)	50	ACC
UrbanSound8K (Salamon et al., 2014)	MC/ZS	8k (10 folds)	10	ACC
VGGSound (Chen et al., 2020a)	MC/ZS	185k	309	mAP
DESED (Turpault et al., 2019)	AR	2.5k (valid)	10	F1
VGGSound (Chen et al., 2020a)	CMR	15k (test)	309	MRR
Clotho (Drossos et al., 2020)	AC	5k (evaluation)		COCO

Table 1: Downstream tasks, including 1. classification: multi-class (MC), zero-shot (ZS), 2. retrieval: audio (AR) and cross-modal retrieval (CMR), and 3. audio captioning (AC) task, with various of clips, classes, and common metrics.

et al., 2022b). The SpeechLM embedding is calculated by the weighted average of all transformer layer outputs of SpeechLM, where the weights are learned during fine tuning. In the process of fine-tuning, we either update or freeze the parameters of SpeechLM.

3.4 Training process

Adhering to BriVL's method, we employ a similar cross modal comparative loss delineated upon the concept of MoCo (He et al., 2020), a mechanism that facilitates dynamic sample queue formation for contrastive learning. Our approach, with two negative queues, enables a larger negative sample size without equivalent mini-batch size, thereby economizing GPU resources. The cross projection loss function, $CXLoss = L(f(Image), Language) + L(Image, g(Language))$ (f, g: projection functions and L: contrastive loss). For all models, we use grid search to find the best hyperparameter. For other hyperparameters (such as batch size, training steps, learning rate, etc.), we directly use the suggested values in the original papers. Note that for per-instance perturbation, we adopt the appropriate quantity compared to the original epochs. The topk parameter is set to 1, which indicates that we only consider the top-scoring prediction for each input instance. The queue_size parameter is set to 9600, which controls the number of instances that can be processed in parallel. We use a momentum value of 0.99 to stabilize the learning process and prevent oscillations during training. The temperature parameter is set to 0.07, which scales the logits output of the model to control the softness of the predicted probability distribution. Finally, we use a grid_size of 4 to divide the input image into a grid of smaller sub-regions for object detection tasks.

4 Task 1: UniBriVL Performance Test

We begin by discussing the training, development, and evaluation process of the UniBriVL model. We use publicly available datasets of varying sizes and tasks, including classification, retrieval, and audio captioning tasks. We compare UniBriVL with some widely used as strong benchmarks in this field and evaluate its performance in these tasks. Additionally, we investigate the effect of sound volume on the generated images. We hypothesize that the volume of sounds can influence the generated images. Hence, we explore the influence of sound volume on image features extracted from the sound using the sound correlation model. We also perform quantitative image analysis to evaluate the performance of UniBriVL compared to previous work, such as S2I and Pedersoli et al. We test model with five categories from VEGAS (Zhou et al., 2018) and compare its performance with other methods in terms of generating visually plausible images.

4.1 Training, development, and evaluation

We selected publicly available audio classification data of different sizes, which are generally used for evaluation (Cramer et al., 2019), and also included some audio tasks/data, as shown in table 1, including classification, retrieval and audio captioning. ESC-50 (Piczak, 2015) is a simple data set with only 2 thousand samples, while UrbanSound8K (Salamon et al., 2014) is a large environmental data set with 10 categories. VGGSound (Chen et al., 2020a) is a huge set of audio and video materials as we said before, including the widest and most diverse range of audio molds. DESED is used again as an audio extraction (AR) job because DESED can perform sound extraction at the fragment level. Finally, Clotho (Drossos et al., 2020) is a unique set of audio subtitles.

Model	Classification			Retrieval		
	ESC-50	UrbanSound8K	VGGSound	DESED (AR)	VGGSound (CMR)	
	ACC	ACC	mAP	F1	A→I (MRR)	I→A (MRR)
Supervise	0.5200	0.6179	0.4331			
OpenL3	0.733	0.7588	0.3487	0.1170	0.0169	0.0162
Wav2CLIP	0.8595	0.8101	0.4663	0.3955	0.0566	0.0678
UniBriVL	**0.9307**	**0.8722**	**0.4885**	0.4111	**0.0641**	**0.0612**
SOTA	0.959	0.8949	0.544			
UniBriVL (ZS)	0.412	0.4024	0.1001			

Table 2: In the subsequent classification and acquisition work, there will be supervised training, other audio representation modes, OpenL3, and the latest SOTA (Guzhov et al., 2021; Kazakos et al., 2021). ZS is based on UniBriVL as a zero sample size model, some of which are derived from the original literature.

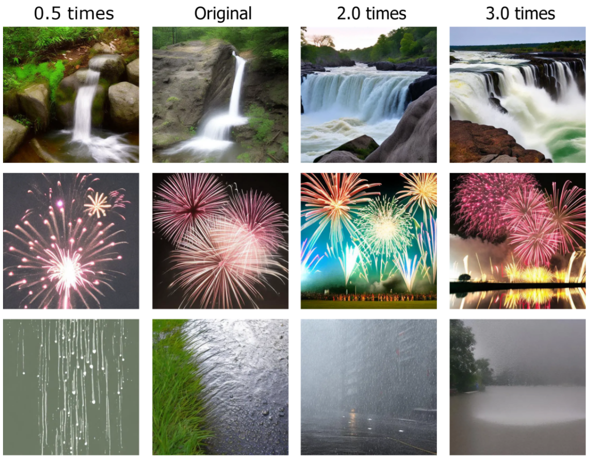


Fig. 3: Generated images by inputting different volumes of sounds. The numbers in the table is the relative loudness to the original sound.

For multi-class (MC) classification problems, an MLP-based classifier is employed, with a corresponding number of classes as output. In DESED, we use the way of simulating UniBriVL and sed_eval[6] to realize audio retrieval (AR). At the same time, we also explore the performance of ours when dealing with multimodal tasks, and how to transfer zero samples to other modalities.

4.2 Sound volume

To establish the reliability of our method's capability to learn the connection between sound and vision, we analyzed the influence of sound volume on generated images. Specifically, we explored how changes in sound volume may affect the generated image. To achieve this, we adjusted the sound volume levels during testing and extracted features for the corresponding sound files. These modified sound features were then input into our pre-trained generator, which was trained on a standard volume scale. The final three sets of images can prove our hypothesis that the magnitude of different volume levels is usually positively correlated with the effects and meanings displayed in the images.

	Method	VEGAS (5 classes)		
		R@1	FID (↓)	IS (↑)
(A)	Pedersoli et al.	23.10	118.68	1.19
(B)	S2I	39.19	114.84	1.45
(C)	S2V	77.58	34.68	4.01
(D)	Ours	**81.31**	**31.48**	**5.42**

Table 3: **Comparison to the baseline:** Pedersoli et al. (2022) and **existing sound-to-image/video method: S2I and S2V** (Fanzeres and Nadeu, 2021; Sung-Bin et al., 2023). Our method outperforms the others both qualitatively and quantitatively in the VEGAS dataset.

4.3 Quantitative image analysis

We conducted a comparative analysis of our proposed model against publicly available prior works S2I[7] (Fanzeres and Nadeu, 2021; Sung-Bin et al., 2023) and Pedersoli et al. (2022). It should be noted that while the latter is not primarily designed for sound-to-image conversion, it employs a VQVAE-based model to generate sound-to-depth or segmentation. We trained our model and Pedersoli et al. using the same training setup as S2I, including five categories in VEGAS, to ensure a fair comparison. As shown in Table 3, our proposed model outperforms all other models while generating visually compelling and recognizable images. We assert that this superior performance can be attributed to the combination of visually enriched audio embeddings and a powerful image generator.

[6] https://github.com/TUT-ARG/sed_eval

[7] https://github.com/leofanzeres/s2i

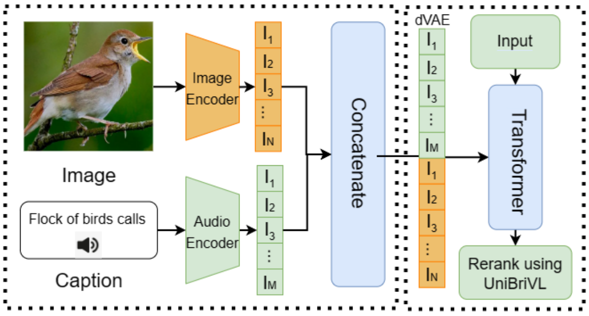
Model	B1	B4	M	RL	Cr
Baseline	0.389	0.015	0.084	0.262	0.074
Wav2CLIP	0.393	0.054	0.104	0.271	0.100
UniBriVL	**0.434**	**0.107**	**0.115**	**0.268**	**0.126**

Table 4: Results of audio captionin, ASR, compared with baseline (Drossos et al., 2020). We tested some tasks on the test tools we worked on previously[8] and we exclude Bleu2/3, list Bleu1/4 (B1/4), METEOR (M), ROUGEL (RL), CIDEr (Cr).

4.4 Downstream task result analysis

As shown in Tables 2 and 4, in training, we monitor the benchmark by training from scratch on each downlink (with random initialization of the encoder weights). Next, we compare UniBriVL with other publicly available OpenL3 (Cramer et al., 2019) pre-trained on different pretext tasks in OpenL3. OpenL3 multimodal self-monitoring training with AudioSet. It serves as a strong benchmark for different audio tasks, such as audio classification and retrieval. We extract features from OpenL3 (512 dim) and UniBriVL (512 dim) and apply the same training scheme to all downstream classification and retrieval tasks. In the chart, we can see that in the retrieval of classification, we are slightly better than our previous work, with an average increase of about 0.04, and only some deficiencies in AR. But it's only about 0.02. We approach or slightly outperform our previous work in retrieval tasks. On tasks such as BLEU and audio captioning, we have some advantages over the baseline, which to our knowledge are not state-of-the-art, but are sufficient to prove their effectiveness.

In sumary, our model has good effects in both data sets of audio retrieval classification, for the source of our strengths: In the Classification tasks, on the four datasets, three of us achieved good results close to or exceeding SOTA. one of reason may be related to our data, and the other may be the effect of BriVL. As for the lack of excellent performance in AR tasks, it may be due to the excessive divergence of the BriVL dataset. If we retrain the basic model on a large scale, we may achieve better results. In the Retrieva tasks, such mrr tasks from A to I, from I to A we have also achieved excellent results, which mainly comes from the excellent training effect of the previous two towers model and the pre-training model. In addition, we believe that increasing the amount of data has the potential to further improve performance on audio tasks.



(a) Stage One (b) Stage Two

Fig. 4: UniBriVL controls the concept map of the stable diffusion model after the model matches the image features through the input language.

5 Task2: Speech Generation Picture Based on Diffusion Model

Our method uses the UniBriVL model to guide the generation of Stable Diffusion. This process utilizes meaningful embedding in the embedding space, by calculating the matching score between audio and image to rearrange the image, and this rearrangement idea is like CLIP. Our code is improved from the official model code and similarity calculation tools[9]. In the reasoning stage, as shown in Figure 4, the matching score of the audio and the generated image can be calculated through the pre-trained UniBriVL, ultimately achieving the effect of guiding the generation of the most matched image. The rearranged images are all provided by selecting from the 100th epoch of the same 20 text inputs. We found that this method can generate images that are appropriate for a given audio input, as confirmed by feedback from related experiments.

5.1 Correlation between sounds and images

This section aims to investigate whether the proposed method generates graphs that are also relevant to humans. Because simply proving authenticity is not enough to prove the deep connection between sound and image, to demonstrate the connection between the two, we conducted a test similar to previous work (Ilharco et al., 2019; Wan et al., 2019). Participants were presented with two images, each with different sound categories as input and the image closest to the given sound. We conducted three tests and obtained a series of option values. By collecting participants' options, we aim to evaluate the effectiveness of the model in generating images related to different sound categories.

[9] https://github.com/BAAI-WuDao/BriVL

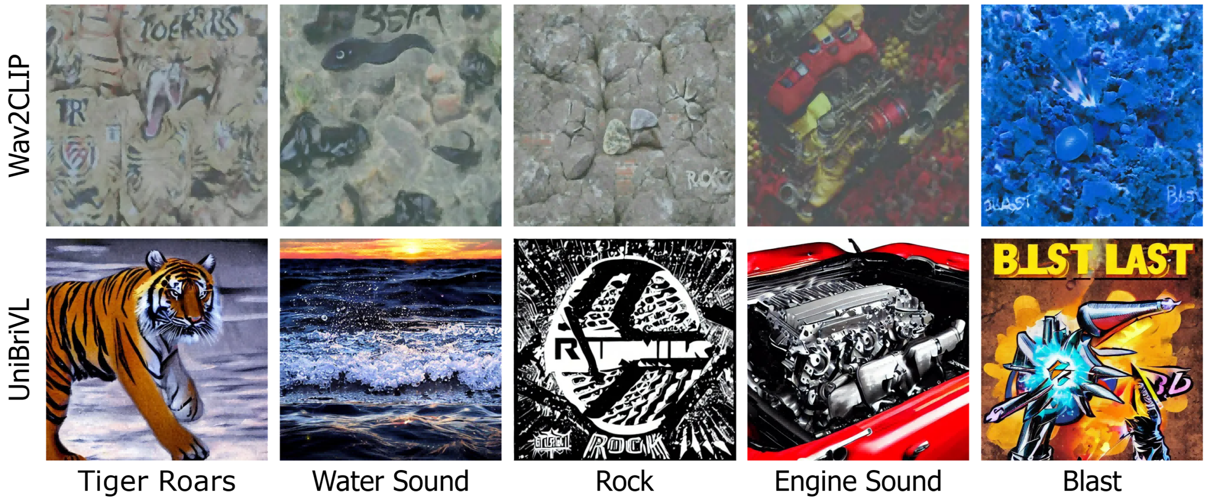


Fig. 5: Images generated from five-piece audio in AudioSet (Gemmeke et al., 2017). Top: Wav2CLIP, Bottom: UniBriVL - corresponding audio input labels in x-axis. Experiments have shown that our tools are effective.

Options	Positive	Negative	Neither
Wav2CLIP	75%	13%	12%
UniBriVL	79%	10%	11%

Table 5: Human scores on correlation between sounds and images, Wav2CLIP works for comparison

The experimental results are shown in Table 5, which collected participants' reactions and classified them as positive, negative, or neutral. A positive option indicates that participants have chosen images generated from input sound, while a negative option indicates their preference for images generated from different categories of sound. Participants who believe that neither of these images represents the sound they hear are considered neutral. Our research results indicate that the majority of participants believe that the generated images are related to the input sound, thus verifying our method's ability to generate images related to a given sound, and it was a good match.

5.2 Comparison with previous work

In previous work, Wav2CLIP also tried to generate text/audio maps. Here are two sets of pictures for comparison with our work. Figure 2 shows the text output image of CLIP and BriVL. Figure 5 shows another group of pictures generated by Wav2CLIP and UniBriVL using audio.

However, in general, they all generated appropriate images, and they have their own characteristics: for example, in their understanding of "Tiger Roads", UniBriVL is more realistic, and WavCLIP is more abstract. When they faced the input of "Water Sound", our work generated a small stream, WavCLIP generated symbolic images similar to fish fossils, and the other images have similar features. Even considering the characteristics of the GAN model, this result can further prove the superiority of our work, which also indicates that our exploration and attempt to generate images using a universal audio guided diffusion model is meaningful; For the generation of audio, they exhibit two characteristics of convergence and divergence between the two models, as we can see, convergence still corresponds to the image. Divergence is reflected in Figure 5 generated by audio, which is more imaginative than Figure 2 generated by text. This is because our BriVL weak semantic text image dataset has strong imagination, and another reason is that audio itself has strong divergence ability, which will enhance the associative ability of audio driven models.

6 Summary & Conclusion

This article introduces a UniBriVL method for generating generic representations. The results show that UniBriVL is able to output general, robust sound representations, and that UniBriVL can be easily transferred to multimodal jobs, such as audio classification, audio retrieval, audio captioning and audio image generation. In future research, we will explore a number of interpretable machine learning methods, consider extending to 6 modalities to our work, just like ImageBind (Girdhar et al., 2023). We will also consider exploring more efficient presentation and using the Consistency Models (Song et al., 2023) and the NeRF (Mildenhall et al., 2020) as the next version of the work and method.

Limitations

We fine-tune the language encoder on SpeechLM-large model, but are limited by the fact that we use part of the AudioSet data, which is a bit less than the original Microsoft training data, perhaps making performance limited. Lastly, it is essential to consider the potential influence of external factors such as background noise, reverberation, or speaker variability on the performance of the speech recognition system. These factors were not extensively addressed in our study, and their impact on the model's performance may be a subject for further investigation.

In summary, our study is subject to limitations concerning the representativeness of the training data, potential language and accent bias, and the focus solely on the language encoder component. These limitations should be taken into account when interpreting our results and considering the application of the model in real-world scenarios. Further research, incorporating diverse datasets and investigating other components of the speech recognition system, would be valuable to overcome these limitations and enhance the overall performance of speech recognition technology.

Ethics Statement

All datasets we train actively exclude harmful, pornographic, and private content, and are only used for research purposes. The participants we recruited, except for some who volunteered, received satisfactory compensation for the rest. The academic tools and human assessment related tests used in this article comply with all regulations or relevant permits.

Biases & Content Acknowledgment Although our ability to generate images through audio is impressive, it should be noted that this model may be influenced by human factors to output content that enhances or exacerbates social biases. In addition, we note a parallel work called WavBriVL, but they are based on simple representation matching, while we use the latest text-audio fusion feature extraction methods and train them with the help of a novel loss. They use Gans to generate images, and we use diffusion models to generate images. Our submission time and their appearance are within three months, so there is no need to compare it to their model or data.

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460.

Hang Chen, Hengshun Zhou, Jun Du, Chin-Hui Lee, Jingdong Chen, Shinji Watanabe, Sabato Marco Siniscalchi, Odette Scharenborg, Di-Yuan Liu, Bao-Cai Yin, et al. 2022a. The first multimodal information based speech processing (misp) challenge: Data, tasks, baselines and results. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 9266–9270. IEEE.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. 2020a. Vggsound: A large-scale audio-visual dataset. In *ICASSP*, pages 721–725. IEEE.

Lele Chen, Guofeng Cui, Celong Liu, Zhong Li, Ziyi Kou, Yi Xu, and Chenliang Xu. 2020b. Talking-head generation with rhythmic head motion. In *European Conference on Computer Vision*, pages 35–51. Springer.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. 2022b. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*.

Grzegorz Chrupała. 2022. Visually grounded models of spoken language: A survey of datasets, architectures and evaluation techniques. *Journal of Artificial Intelligence Research*, 73:673–707.

Jason Cramer, Ho-Hsiang Wu, Justin Salamon, and Juan Pablo Bello. 2019. Look, listen, and learn more: Design choices for deep audio embeddings. In *ICASSP*, pages 3852–3856. IEEE.

Daniel Cudeiro, Timo Bolkart, Cassidy Laidlaw, Anurag Ranjan, and Michael J Black. 2019. Capture, learning, and synthesis of 3d speaking styles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10101–10111.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. 2020. Clotho: An audio captioning dataset. In *ICASSP*.

Leonardo A Fanzeres and Climent Nadeu. 2021. Sound-to-imagination: Unsupervised crossmodal translation using deep dense network architecture. *arXiv preprint arXiv:2106.01266*.

Nanyi Fei, Zhiwu Lu, Yizhao Gao, Guoxing Yang, Yuqi Huo, Jingyuan Wen, Haoyu Lu, Ruihua Song, Xin Gao, Tao Xiang, et al. 2022. Towards artificial general intelligence via a multimodal foundation model. *Nature Communications*, 13(1):1–13.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE ICASSP 2017*, New Orleans, LA.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind: One embedding space to bind them all.

Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. 2021. Audioclip: Extending clip to image, text and audio. *arXiv preprint arXiv:2106.13043*.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Gabriel Ilharco, Yuan Zhang, and Jason Baldridge. 2019. Large-scale representation learning from visually grounded untranscribed speech. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Tero Karras, Timo Aila, Samuli Laine, Antti Herva, and Jaakko Lehtinen. 2017. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics (TOG)*, 36(4):1–12.

Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. 2021. Slow-fast auditory streams for audio recognition. In *ICASSP*, pages 855–859.

Avisek Lahiri, Vivek Kwatra, Christian Frueh, John Lewis, and Chris Bregler. 2021. Lipsync3d: Data-efficient learning of personalized 3d talking faces from video using pose and lighting normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2755–2764.

Long Ma and Yanqing Zhang. 2015. Using word2vec to process big text data. In *2015 IEEE International Conference on Big Data (Big Data)*, pages 2895–2897. IEEE.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2020. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Xichen Pan, Peiyu Chen, Yichen Gong, Helong Zhou, Xinbing Wang, and Zhouhan Lin. 2022. Leveraging unimodal self-supervised learning for multimodal audio-visual speech recognition. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4491–4503, Dublin, Ireland. Association for Computational Linguistics.

Fabrizio Pedersoli, Dryden Wiebe, Amin Banitalebi, Yong Zhang, and Kwang Moo Yi. 2022. Estimating visual information from audio through manifold learning. *arXiv preprint arXiv:2208.02337*.

Karol J. Piczak. 2015. ESC: Dataset for Environmental Sound Classification. In *ACM Multimedia*, page 1015. ACM Press.

Yue Qiu and Hirokatsu Kataoka. 2018. Image generation associated with music data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2510–2513.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, et al. 2021. Learning transferable visual models from natural language supervision. *ICML*.

Alexander Richard, Michael Zollhöfer, Yandong Wen, Fernando De la Torre, and Yaser Sheikh. 2021. Meshtalk: 3d face animation from speech using cross-modality disentanglement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1173–1182.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on*

Computer Vision and Pattern Recognition (CVPR), pages 10684–10695.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. 2022a. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*, volume 35, pages 36479–36494. Curran Associates, Inc.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022b. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494.

J. Salamon, C. Jacoby, and J. P. Bello. 2014. A dataset and taxonomy for urban sound research. In *ACM Multimedia*, pages 1041–1044, Orlando, FL, USA.

Linsen Song, Wayne Wu, Chen Qian, Ran He, and Chen Change Loy. 2022. Everybody's talkin': Let me talk as you want. *IEEE Transactions on Information Forensics and Security*, 17:585–598.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. 2023. Consistency models.

Kim Sung-Bin, Arda Senocak, Hyunwoo Ha, Andrew Owens, and Tae-Hyun Oh. 2023. Sound to visual scene generation by audio-to-visual latent alignment.

Mingxing Tan and Quoc Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR.

Justus Thies, Mohamed Elgharib, Ayush Tewari, Christian Theobalt, and Matthias Nießner. 2020. Neural voice puppetry: Audio-driven facial reenactment. In *European conference on computer vision*, pages 716–731. Springer.

Nicolas Turpault, Romain Serizel, Ankit Parag Shah, and Justin Salamon. 2019. Sound event detection in domestic environments with weakly labeled data and soundscape synthesis. In *DCASE*, New York City, United States.

Chia-Hung Wan, Shun-Po Chuang, and Hung-Yi Lee. 2019. Towards audio to scene image synthesis using generative adversarial network. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 496–500.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. 2022. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*.

Xin Wen, Miao Wang, Christian Richardt, Ze-Yin Chen, and Shi-Min Hu. 2020. Photorealistic audio-driven video portraits. *IEEE Transactions on Visualization and Computer Graphics*, 26(12):3457–3466.

Haozhe Wu, Jia Jia, Haoyu Wang, Yishun Dou, Chao Duan, and Qingshan Deng. 2021. Imitating arbitrary talking style for realistic audio-driven talking face synthesis. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 1478–1486.

Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. 2022. Wav2clip: Learning robust audio representations from clip. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4563–4567. IEEE.

Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. 2018. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324.

Ran Yi, Zipeng Ye, Juyong Zhang, Hujun Bao, and Yong-Jin Liu. 2020. Audio-driven talking face video generation with learning-based personalized head pose. *arXiv preprint arXiv:2002.10137*.

Chenxu Zhang, Saifeng Ni, Zhipeng Fan, Hongbo Li, Ming Zeng, Madhukar Budagavi, and Xiaohu Guo. 2021a. 3d talking face with personalized pose dynamics. *IEEE Transactions on Visualization and Computer Graphics*.

Chenxu Zhang, Yifan Zhao, Yifei Huang, Ming Zeng, Saifeng Ni, Madhukar Budagavi, and Xiaohu Guo. 2021b. Facial: Synthesizing dynamic talking face with implicit attribute learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3867–3876.

Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. 2021c. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3661–3670.

Ziqiang Zhang, Sanyuan Chen, Long Zhou, Yu Wu, Shuo Ren, Shujie Liu, Zhuoyuan Yao, Xun Gong, Lirong Dai, Jinyu Li, and Furu Wei. 2023. Speechlm: Enhanced speech pre-training with unpaired textual data.

Yipin Zhou, Zhaowen Wang, Chen Fang, Trung Bui, and Tamara L Berg. 2018. Visual to sound: Generating natural sound for videos in the wild. In *CVPR*.

Hao Zhu, Man-Di Luo, Rui Wang, Ai-Hua Zheng, and Ran He. 2021. Deep audio-visual learning: A survey. *International Journal of Automation and Computing*, 18(3):351–376.