Enhancing Language Model Performance through Self-Supervised Learning (SSL) Post Pre-Training with Text-Audio Pairs

Anshuman Sinha,[1,2] Aubin Rey,[1,2] and Camille Migozzi[1,2]

[1] *College of computing, Georgia Institute of Technology, Atlanta, GA*

[2] *All authors have equal contribution*

(Dated: 1st December 2023)

The field of research focusing on multi-modal contrastive learning strategies within the audio-text realm has rapidly gained intense interest. Contrastively trained Audio-Language Models (ALMs), such as CLAP, which establish a unified representation across audio and language modalities, have enhanced the efficacy in various subsequent tasks. These improvements are evident in areas like zero-shot audio classification and audio retrieval, among others. However, the ability of these models to understand natural language and temporal relations is still a largely unexplored and open field for research. In this paper, we propose to equip the multi-modal ALMs with language and temporal understanding without loosing their inherent prior capabilities of audio-language tasks. We propose a novel framework for post-pretraining the ALMs with a language adaptive technique while preserving its prior capabilities. We observe a encouraging performance gains on language and temporal understanding while still remain competitive on zero shot retrieval and classification tasks. We evaluate the model on various close and open-ended tasks.

I. INTRODUCTION

Audio, text, and images are among the most prevalent forms of information globally. In the past few years, the field of multi-modal contrastive pre-training has gained prominence because of its impressive performance across different subsequent uses. Among the field of multi-modal learning contrastive learning has emerged as an effective strategy for training models on extensive, less-structured internet-sourced data. One notable early model in this sphere is CLIP, developed by Radford *et al.* [1], which stands out as a groundbreaking vision-language model. The modelling technique capitalizes on this by learning the relationship between text and images, aligning them in a common latent domain. Additionally, encouraging results from CLIP facilitated essential related tasks such as formulating image captions, as discussed by Mokady *et al.* [2], and generating images from text, as explored by Rombach *et al.* [3]. Similar work also propagated into other multi-modal domains such as video-language as shown in Xu *et al.* [4], Fang *et al.* [5], Zhao *et al.* [6], Luo *et al.* [7], Cheng *et al.* [8], Ge *et al.* [9] and audio-language models such as shown in Elizalde *et al.* [10], Huang *et al.* [11], Guzhov *et al.* [12], Wu *et al.* [13], Deshmukh *et al.* [14], Wu *et al.* [15] Taking inspirations from CLIP, Elizalde *et al.* [10] in 2023 presented CLAP, a Contrastive Language Audio Pre-Training model. The model performs a similar contrastive pre-training with texts and audios, instead of texts and images. The authors showed unparalleled benchmarks in 16 downstream tasks, including but not limited to, innovative zero-shot audio categorization and audio retrieval based on textual prompts.

However previous authors have shown the possible limitations of these Audio-Language models in learning the actual natural language, while learning the relationship between texts and audio. Wu *et al.* [15] demonstrate that existing models primarily target nouns and verbs for retrieval, neglecting to make full use of the complete sentence. Wu *et al.* [15] conducted a study wherein they trained an ALM with captions stripped of everything except for nouns and verbs. Surprisingly, this refined ALM exhibited comparable or even superior performance to those trained with non-shuffled captions. This underscores the notion that ALMs can excel in such benchmarks without necessarily demonstrating compositional reasoning capabilities. Recent studies highlight certain limitations in models like CLIP's ability to understand compositional reasoning, despite their extensive training datasets [16–18] suggest that a key reason behind this shortfall is the

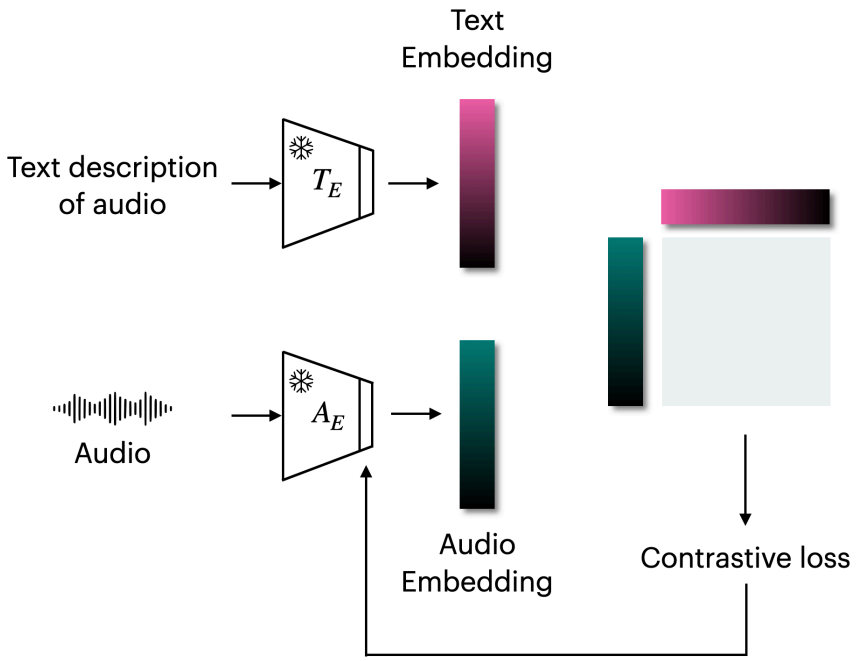

Figure 1: Contrastive learning paradigm [[Anshuman says: Update the figure, Remove the freeze]]

inherent focus of contrastive pre-training on retrieval tasks. This predisposition allows such models to excel in retrieval-based benchmarks, even in the absence of a robust compositional grasp. In a recent publication by Ghosh *et al.* [19] the authors critique existing audio-retrieval benchmarks, revealing ALMs' superficial success without true compositional understanding. In response, the authors introduce CompA, benchmarks gauging ALMs' compositional acumen, and debut CompA-CLAP, an evolved audio-language model enhanced with innovative contrastive training techniques.

In the light of making ALMs more robust we propose a novel self-supervised contrastive larning paradigm for post-pretraining the model. Taking inspirations from their work by Bagad *et al.* [20] on instilling time in the Video-Language models, we have tried to make the ALMs the model get a better relationship of modalities. In addition to the conventional contrastive loss, which derives negatives from different samples within the batch, we utilize time-order inversions within and across samples to produce extra negatives for both audio and text. Furthermore, we have expanded the contrastive loss to cater to video and text with time-order reversals, emphasizing reverse consistency. The following are the main outlines of our contribution in this work: [[Anshuman says: Include the reason for contrastive learning loss not performing well, as they just include the simillarity between text and audio without any attention to language.]]

- We show the current model fails to get a capture the correct relation between audio and texts ...
- We post-pretrain a multi-modal ALM ...
- We test on 5 tests against the benchmark result ...

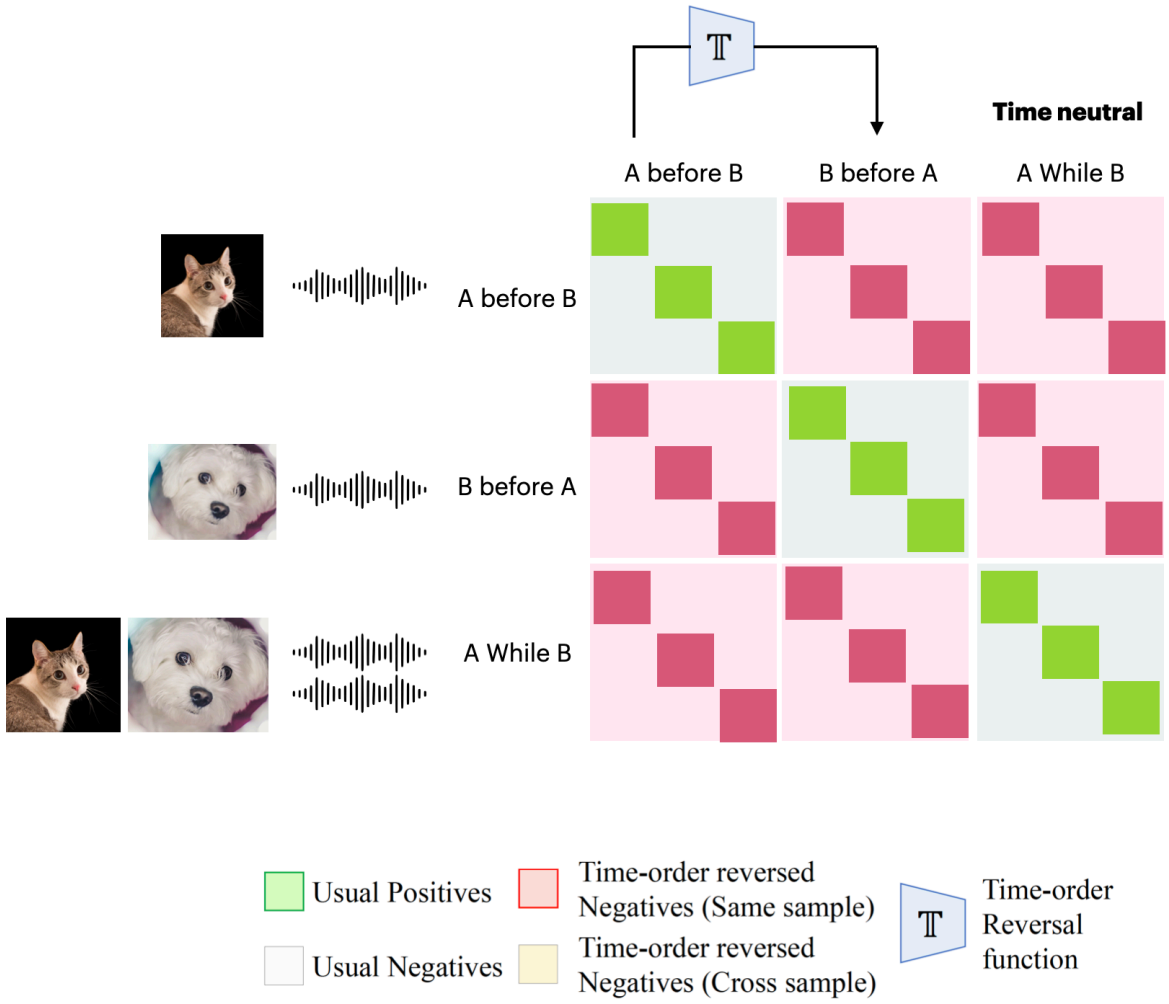

Figure 2: Loss formulation [[Anshuman says: Update the captions and the legends.]]

II. BACKGROUND AND RELATED WORK

A. Foundation models and Multi-modal text-audio learning

The swift advancement of Pretrained Foundation Models (PFMs) has extended beyond the realms of text, imagery, and graphical data, delving into extensive studies on auditory, visual, combined text-image, and multi-data formats. Additionally, the exploration of integrated PFMs that encompass the previously stated modalities has commenced. Hence, we present an overview of various sophisticated and comprehensive PFMs in this segment. Models that contrast audio and visual data have been employed for pinpointing the source of sounds in images [21, 22], retrieving information across different modalities [23], and classifying data in a zero-shot manner [12, 24]. Additionally, there's an increasing focus on models that pair audio with text, as demonstrated by the DCASE competition focused on retrieving audio based on language cues [25]. These models have been effectively utilized in categorizing and tagging musical audio by genre [26], identifying environmental sounds through language descriptors [27–30], and performing classification tasks without prior

exposure to the data [10, 27–30].

Current trends are shifting towards the use of linguistic interactions within auditory systems, showcasing a spike in interest for applications that intertwine language and sound. Innovations such as converting written text into audio outputs (as seen in the work of [31–33]) alongside the transformation of text into musical compositions by Agostinelli *et al.* [34] are becoming increasingly prominent. Additional endeavors include using textual prompts to differentiate sound sources (explored by Liu *et al.* [35]) and the generation of audio descriptions as shown by [14, 36] as fusion of linguistic algorithms with sound encoding techniques, redefining all auditory-based assignments as text-creation challenges. Their model, named Pengi [14], has set a new benchmark in 22 different tasks, underscoring the potential of integrating verbal modalities to bolster the functionality of audio systems. Predominantly, these innovations are powered by either a textual or an auditory encoder to fulfill their designated functions. Meanwhile, CLAP [10] distinguishes itself by mastering a unified conceptual framework that bridges the auditory and linguistic domains, demonstrating extraordinary capabilities in executing tasks without prior specific training, thus emerging as an influential archetype for interdisciplinary comprehension and interactions.

B. Self Supervised learning and Post pre-training

In recent times, the field of natural language processing (NLP) has witnessed considerable advances, particularly with the advent of pre-trained language models. Initiatives like BERT, introduced by Devlin *et al.* [37], and GPT, launched by Radford *et al.* [38], have set a new standard wherein pre-training followed by specialized fine-tuning has remained a widely adopted approach. This methodology has extended its impact into other areas of study, such as image processing [39, 40] and life sciences[41]. Additionally, new pre-training methods [42–45] have been explored to either boost the efficiency of the learning process or to amplify the generative capacities of these systems. [[Anshuman says: Aubin can you try work on this? Also we can have 1 image covering the whole background and SSL section]] [[Aubin says: The evolution of pre-training approaches has seen the emergence and growing significance of Self-Supervised Learning (SSL) in the NLP landscape. SSL methods use unlabeled data to create auxiliary tasks, enabling models to learn from the inherent structure and patterns present in the data itself. This paradigm shift has proven effective in overcoming challenges associated with limited labeled data. Furthermore, recent research about post pre-training strategies focused on optimizing models after the initial pre-training phase. Techniques such as domain adaptation, continual learning, and model distillation have been explored to improve the adaptability of pre-trained models to specific domains or tasks. These post pre-training strategies add versatility to the pre-trained models making them more flexible for diverse application cases.]]

C. Cross Attention

D. Zero-shot learning

III. METHODOLOGY

In this section, we describe the process to make the contrastive learning objective have better sense of language.

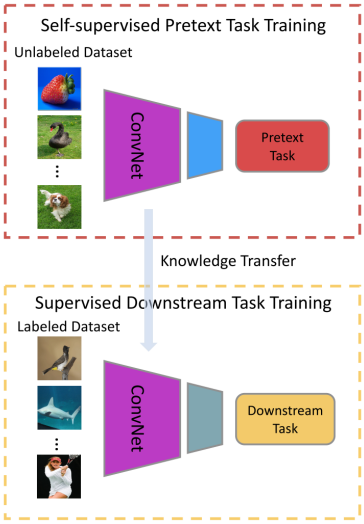

Figure 3: Self supervised learning architecture

A. Preliminaries

Define the variables, encoder, decoder, time and process adaption Introduction to Fundamentals. Consider $\mathcal{A}$ as the domain of audio recordings and $\mathcal{C}$ as the set of their corresponding textual transcripts. For any two discrete and non-overlapping audio clips a_i, a_j within $\mathcal{A}$, let their relevant transcripts be c_i, c_j found in $\mathcal{C}$. We introduce σ to represent a sequential relationship, where σ can be either *preceding* or *succeeding*.

We then define an integrated segment that respects the sequential order as (a_{ij}, c_{ij}), with a_{ij} constructed by the operation $[a_i; a_j]$, which merges the two audio clips, and c_{ij} represented by $[c_i; \sigma; c_j]$, which merges the transcripts in a manner that reflects the sequential relation σ. It should be noted that the arrangement of a_i and a_j within a_{ij} might vary depending on the value of σ. In the interest of brevity, we will refer to the composite audio-text pair as (a, c) except where additional specificity is required.

B. Data-processing

Define the process with which we make the augmentations and the process. We also need to figure out here [[Anshuman says: Camille can you try this for starters; I have given some starting points. We need to change this for our case.]]

Sequential Inversion Approach. In conventional contrastive learning frameworks for audio-text alignment, the usual practice is to synchronize audio segments a_i with their textual transcriptions c_i and to set these against distinct transcription pairs c_j that describe unrelated audio. Such an approach often overlooks nuanced sequential comprehension, as contrasting the content typically involves focusing on distinctive elements or themes within the audio. This limitation is apparent in methods akin to bag-of-words that are effective in contrastive scenarios, applied to both acoustic features and textual data.

We posit that without negative examples that share common thematic elements or soundscapes, there is no impetus for models to develop an understanding of chronological sequence. To address this, we propose an elementary technique to generate negative samples that encourage the model to prioritize sequential discernment.

A sequential inversion function Π is defined, which is applicable to concatenated audio or transcription pairs, effectuating a temporal reordering of the components:

$$\Pi(a) = \Pi([a_i; a_j]) := [a_j; a_i],$$

$$\Pi(c) = \Pi([c_i; \sigma; c_j]) := [c_j; \sigma; c_i].$$

The function Π is graphically represented, and it is essential to recognize that Π does not literally reverse time within the audio tracksrather, it rearranges the sequence of events within the compiled segments. Our goal is to cultivate a model capable of distinguishing an original audio-transcript pair (a, c) from its sequentially inverted counterpart $(a, \Pi(c))$, and also $(\Pi(a), c)$.

C. Post-pretraining with SSL

Preliminaries of Audio-Based Contrastive Pretraining

Consider the preprocessed audio sequence X_a such that $X_a \in \mathbb{R}^{F \times T}$, where 'F' indicates the frequency components (for instance, the Mel frequency bins) and 'T' signifies the quantity of temporal segments. We denote the corresponding textual data as X_t. Within a given batch containing 'N' instances, each audio and corresponding text are symbolized as $\{X_a, X_t\}_i$, with 'i' spanning from 0 to N. Dropping the index 'i' for brevity, we henceforth reference $\{X_a, X_t\}$ to symbolize a collection of 'N' audio-text pairs.

Each audio segment and its corresponding text description are processed through dedicated encoders. Denote $f_a(.)$ as the function characterizing the audio encoder, and $f_t(.)$ as that for the text encoder. For an ensemble of size N, we have:

$$\hat{X}_a = f_a(X_a); \quad \hat{X}_t = f_t(X_t)$$

Here, $\hat{X}_a$ embodies the audio feature space with a dimension 'V', while $\hat{X}_t$ represents the text feature space with a dimension 'U'.

To integrate the audio and text features, $\hat{X}_a$ and $\hat{X}_t$, into a shared multimodal feature space of dimension 'd', we employ trainable linear mappings:

$$E_a = L_a(\hat{X}_a); \quad E_t = L_t(\hat{X}_t)$$

where E_a and E_t exist in $\mathbb{R}^{N \times d}$, with L_a and L_t serving as the transformation matrices for audio and text modalities, respectively.

Subsequently, to assess the congruence between the embeddings E_a and E_t, we calculate their similarity:

$$C = \tau \cdot (E_t E_a^\top)$$

In this context, τ acts as a scaling constant that modulates the logarithmic scale. The similarity matrix C in $\mathbb{R}^{N \times N}$ is composed of 'N' compatible pairs along its diagonal and $N^2 - N$ non-compatible pairs elsewhere.

The objective function $\mathcal{L}$ is established as:

$$\mathcal{L} = 0.5 \cdot (\ell_{text}(C) + \ell_{audio}(C))$$

where $\ell_k = \frac{1}{N}\sum_{i=0}^{N} \log(\text{diag}(\text{softmax}(C)))$ computed across the textual and auditory dimensions respectively. This symmetric cross-entropy loss facilitates the joint optimization of the audio and text encoders alongside their corresponding linear transformations.

D. Sentence structure and loss function

Text encoder → text-embedding → T_e.

Audio encoder → audio-embedding → A_e.

$T_e.A_e$ for a batch x batch matrix. With positive and negative pairs.

Logits → Softmax →Probabilities → Cross-entropy(Probabilities, True labels).

We take the Loss = Loss texts ($\sum_i \text{row}_i$) + Loss audio ($\sum_i \text{col}_i$). The Text loss represents the loss term which is contributed by the text encoder, as we try to find the best texts among the $'n'$ texts produced by the T_e. Hence, this particular loss will try to improve the text-encoder. Both loss functions are necessary as we want both the T_e and A_e to be trained symmetrically and become better progressively.

[[Anshuman says: How to take into account the negative terms? Starting with Piyush et al.'s loss function, but we need to do a study on this]] The conventional contrastive loss function doesn't penalise the true negatives, it preferentially promotes the true positive and the rest are equally weighted negatively.

Objective Formulation. Our methodology leverages a pre-existing video-language model equipped with a visual encoder f_θ and a textual encoder g_ϕ. We derive video representations $z_u = f_\theta(u)$ and textual representations $z_t = g_\phi(t)$, both in the same dimensional space. The intention is to fine-tune $\Theta = \{\theta, \phi\}$ post-pretraining, enhancing temporal acuity while preserving its baseline retrieval capabilities. With only a limited dataset, we selectively refine parts of Θ, like its concluding layers.

For temporal alignment, we adopt a tailored adaptation of the InfoNCE loss function. This is used to discern the temporal sequence of video-text pairs. Taking a time-aligned video-text pair (u, t), we formulate a loss function that maintains the chronological order in the pair:

$$L_f = \sum_{(u,t)\in B} (TNCE_t(z_u, z_t) + TNCE(z_t, z_u)) + TNCE_o(z_t, z_u))$$

Here, TNCE stands for Temporal Noise Contrastive Estimation, a variant of the NCE loss tailored for temporal learning, and is calculated as:

$$TNCE(z_u, z_t) := -\log \frac{\exp(z_u \cdot z_t)}{\sum_{t'\in B_t} \exp(z_u \cdot z_{t'}) + C^{time} + C^{overlap}}$$

$$TNCE_o(z_u, z_t) := -\log \frac{\exp(z_u \cdot z_t)}{\sum_{t'\in B_t} \exp(z_u \cdot z_{t'}) + C^{time} + C^{overlap}}$$

In this expression, B represents the batch of (u, t) pairs, and B_t is the set of text samples within the batch that serve as temporal negatives. C^{time} is an accumulation of negatives fashioned via time-reversal, and is expressed as:

$$C^{time} = \alpha_{same} \exp(z_u \cdot z_{\Pi(t)}) + \alpha_{cross} \sum_{t'\in B_t \setminus \{t\}} \exp(z_u \cdot z_{\Pi(t')})$$

$$C^{overlap} = \alpha_{same_o} \exp(z_u \cdot z_{\Pi(t)}) + \alpha_{cross_o} \sum_{t' \in B_t \setminus \{o\}} \exp(z_u \cdot z_{\Pi(t')})$$

IV. EXPERIMENTS

A. Base model

Which model did we choose, and why did we choose this model to post-pretrain?

B. Training Dataset

Explain a bit about the original training dataset and then about our choice of the current dataset.

C. Downstream Tasks

Explain the list of downstream tasks along with the dataset which we want to use. Explain the relevance of the tasks as well, as in what it means to your model and methodology. [[Aubin says: Zero shot changing prompts: i can hear [class label], this is an audio of [class label],'[class label],this is [class label],this is a sound of [class label] and comparing our accuracy with the CLAP]]

D. Evaluation metrics

The evaluation metrics are easy, just explain the code in mathematical details.

V. RESULTS

Results section should cover all the downstream tasks. Along with some explanation of the results.

Table I: Effect of different prompts on ESC50 (ZS). [[Anshuman says: Not updated]]

Prompt	**ESC50 (acc)**
'i can hear [class label]'	0.786
'this is an audio of [class label]'	0.8005
'[class label]'	0.812
'this is [class label]'	0.8135
'this is a sound of [class label]'	0.826

Table II: Classification performance [[Anshuman says: Not updated]]

Models	**Task 1**	**Task 2**	**Task 3**	**Task 4**	**Task 5**
CLAP	78.14	49.7, 48.4	25.8	2.1	3.7, 11.7
T-CLAP	73.87	48.9	62.8	40.38	28.7
T-CLAP ($\alpha_s = 0.0$)	71.2	46.4, 49.1	-	-	-
T-CLAP ($\alpha_c = 0.0$)	-	-	-	-	-

Table III: Retrieval performance [[Anshuman says: Not updated]]

	ESC-50			**Clotho**		
Method	**R@1**	**R@5**	**R@10**	**R@1**	**R@5**	**R@10**
CLAP	99.34	62.28	69.89	0.0259.71	0.2563	0.1593
T-CLAP	51.65	-	-	0.0217	-	-

VI. CONCLUSION

VII. REFERENCES

$$\text{Cosine similarity}(A_i, T_i) = \frac{A_i \cdot T_i}{\|A_i\| + \|T_i\|} \tag{1}$$

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, Learning transferable visual models from natural language supervision, in *International conference on machine learning* (PMLR, 2021) pp. 8748–8763.

[2] R. Mokady, A. Hertz, and A. H. Bermano, Clipcap: Clip prefix for image captioning, arXiv preprint arXiv:2111.09734 (2021).

[3] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, High-resolution image synthesis with latent diffusion models, in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2022) pp. 10684–10695.

[4] H. Xu, G. Ghosh, P.-Y. Huang, D. Okhonko, A. Aghajanyan, F. Metze, L. Zettlemoyer, and C. Feichtenhofer, Videoclip: Contrastive pre-training for zero-shot video-text understanding, arXiv preprint arXiv:2109.14084 (2021).

[5] H. Fang, P. Xiong, L. Xu, and Y. Chen, Clip2video: Mastering video-text retrieval via image clip, arXiv preprint arXiv:2106.11097 (2021).

[6] S. Zhao, L. Zhu, X. Wang, and Y. Yang, Centerclip: Token clustering for efficient text-video retrieval, in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2022) pp. 970–981.

[7] H. Luo, L. Ji, M. Zhong, Y. Chen, W. Lei, N. Duan, and T. Li, Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning, Neurocomputing **508**, 293 (2022).

[8] F. Cheng, X. Wang, J. Lei, D. Crandall, M. Bansal, and G. Bertasius, Vindlu: A recipe for effective video-and-language pretraining, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023) pp. 10739–10750.

[9] Y. Ge, Y. Ge, X. Liu, D. Li, Y. Shan, X. Qie, and P. Luo, Bridging video-text retrieval with multiple choice questions, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022) pp.

16167–16176.

[10] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, Clap learning audio concepts from natural language supervision, in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2023) pp. 1–5.

[11] Q. Huang, A. Jansen, J. Lee, R. Ganti, J. Y. Li, and D. P. Ellis, Mulan: A joint embedding of music audio and natural language, arXiv preprint arXiv:2208.12415 (2022).

[12] A. Guzhov, F. Raue, J. Hees, and A. Dengel, Audioclip: Extending clip to image, text and audio, in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2022) pp. 976–980.

[13] 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 *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2023) pp. 1–5.

[14] S. Deshmukh, B. Elizalde, R. Singh, and H. Wang, Pengi: An audio language model for audio tasks, arXiv preprint arXiv:2305.11834 (2023).

[15] H.-H. Wu, O. Nieto, J. P. Bello, and J. Salomon, Audio-text models do not yet leverage natural language, in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2023) pp. 1–5.

[16] T. Thrush, R. Jiang, M. Bartolo, A. Singh, A. Williams, D. Kiela, and C. Ross, Winoground: Probing vision and language models for visio-linguistic compositionality, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022) pp. 5238–5248.

[17] Z. Ma, J. Hong, M. O. Gul, M. Gandhi, I. Gao, and R. Krishna, Crepe: Can vision-language foundation models reason compositionally?, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023) pp. 10910–10921.

[18] 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 *The Eleventh International Conference on Learning Representations* (2022).

[19] S. Ghosh, A. Seth, S. Kumar, U. Tyagi, C. K. Evuru, S. Ramaneswaran, S. Sakshi, O. Nieto, R. Duraiswami, and D. Manocha, Compa: Addressing the gap in compositional reasoning in audio-language models, arXiv preprint arXiv:2310.08753 (2023).

[20] P. Bagad, M. Tapaswi, and C. G. Snoek, Test of time: Instilling video-language models with a sense of time, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023) pp. 2503–2516.

[21] H. Chen, W. Xie, T. Afouras, A. Nagrani, A. Vedaldi, and A. Zisserman, Localizing visual sounds the hard way, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021) pp. 16867–16876.

[22] H.-H. Wu, M. Fuentes, P. Seetharaman, and J. P. Bello, How to listen? rethinking visual sound localization, arXiv preprint arXiv:2204.05156 (2022).

[23] D. Surís, C. Vondrick, B. Russell, and J. Salamon, It's time for artistic correspondence in music and video, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022) pp. 10564–10574.

[24] H.-H. Wu, P. Seetharaman, K. Kumar, and J. P. Bello, Wav2clip: Learning robust audio representations from clip, in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2022) pp. 4563–4567.

[25] H. Xie, S. Lipping, and T. Virtanen, Language-based audio retrieval task in dcase 2022 challenge, arXiv preprint arXiv:2206.06108 (2022).

[26] I. Manco, E. Benetos, E. Quinton, and G. Fazekas, Contrastive audio-language learning for music, arXiv preprint arXiv:2208.12208 (2022).

[27] Y. Zhao, J. Hessel, Y. Yu, X. Lu, R. Zellers, and Y. Choi, Connecting the dots between audio and text without parallel data through visual knowledge transfer, arXiv preprint arXiv:2112.08995 (2021).

[28] S. Lou, X. Xu, M. Wu, and K. Yu, Audio-text retrieval in context, in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2022) pp. 4793–4797.

[29] X. Mei, X. Liu, J. Sun, M. D. Plumbley, and W. Wang, On metric learning for audio-text cross-modal retrieval, arXiv preprint arXiv:2203.15537 (2022).

[30] A. S. Koepke, A.-M. Oncescu, J. Henriques, Z. Akata, and S. Albanie, Audio retrieval with natural language queries: A benchmark study, IEEE Transactions on Multimedia (2022).

[31] D. Ghosal, N. Majumder, A. Mehrish, and S. Poria, Text-to-audio generation using instruction-tuned llm and latent diffusion model, arXiv preprint arXiv:2304.13731 (2023).

[32] 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, arXiv preprint arXiv:2301.12503 (2023).

[33] R. Huang, J. Huang, D. Yang, Y. Ren, L. Liu, M. Li, Z. Ye, J. Liu, X. Yin, and Z. Zhao, Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models, arXiv preprint arXiv:2301.12661 (2023).

[34] A. Agostinelli, T. I. Denk, Z. Borsos, J. Engel, M. Verzetti, A. Caillon, Q. Huang, A. Jansen, A. Roberts, M. Tagliasacchi, *et al.*, Musiclm: Generating music from text, arXiv preprint arXiv:2301.11325 (2023).

[35] X. Liu, Q. Kong, Y. Zhao, H. Liu, Y. Yuan, Y. Liu, R. Xia, Y. Wang, M. D. Plumbley, and W. Wang, Separate anything you describe, arXiv preprint arXiv:2308.05037 (2023).

[36] S. Ghosh, S. Kumar, C. K. R. Evuru, R. Duraiswami, and D. Manocha, Recap: Retrieval-augmented audio captioning, arXiv preprint arXiv:2309.09836 (2023).

[37] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[38] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, *et al.*, Improving language understanding by generative pre-training, (2018).

[39] P. He, X. Liu, J. Gao, and W. Chen, Deberta: Decoding-enhanced bert with disentangled attention, arXiv preprint arXiv:2006.03654 (2020).

[40] H. Bao, L. Dong, S. Piao, and F. Wei, Beit: Bert pre-training of image transformers, arXiv preprint arXiv:2106.08254 (2021).

[41] A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C. L. Zitnick, J. Ma, *et al.*, Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences, Proceedings of the National Academy of Sciences **118**, e2016239118 (2021).

[42] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, arXiv preprint arXiv:1910.13461 (2019).

[43] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, The Journal of Machine Learning Research **21**, 5485 (2020).

[44] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, Electra: Pre-training text encoders as discriminators rather than generators, arXiv preprint arXiv:2003.10555 (2020).

[45] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).