On the Language Encoder of Contrastive Cross-modal Models

Anonymous ACL submission

Abstract

Contrastive cross-modal models such as CLIP and CLAP aid various vision-language (VL) and audio-language (AL) tasks. However, there has been limited investigation of and improvement in their language encoder – the central component of encoding natural language descriptions of image/audio into vector representations. We extensively evaluate how unsupervised and supervised sentence embedding training affect language encoder quality and cross-modal task performance. In VL pretraining, we found that sentence embedding training enhances language encoder quality and aids in cross-modal tasks, improving contrastive VL models such as CyCLIP. Sentence embedding training benefits AL tasks when the amount of training data is large. We analyze the representation spaces to understand the strengths of sentence embedding training, and find that it improves text-space uniformity, at the cost of decreased cross-modal alignment [1].

1 Introduction

Significant progress have been made in pretraining large-scale cross-modal models, such as CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021), for various vision-language (**VL**) applications such as retrieval and zero-shot image classification. These models are often pretrained with large amounts of data, e.g., OpenAI leverages ≈400M caption-image pairs to train CLIP while LAION-AI scaled up this number to 5B (Schuhmann et al., 2022; Cherti et al., 2023). Such a large amount of multi-modal pretraining data contains text captions at the same scale as the pretraining corpora of large language models (**LLMs**) such as BERT, which is pretrained on 3.3B words (Devlin et al., 2019).

The success of VL pretraining encourages research on contrastive learning models for other modalities like audio. Pretrained audio-language (**AL**) models such as AudioCLIP (Guzhov et al., 2022) and CLAP (Wu et al., 2023; Elizalde et al., 2023) show promising results on AL retrieval and zero-shot audio classification tasks.

It is clear that the language encoder in cross-modal contrastive models plays a central role when scaling-up pretraining of a specific modality and/or the amount of modalities. Therefore, analyzing and improving the language encoder become increasingly crucial. CLIP's language encoder (**CLIP LM**) – a decoder-only language model similar to GPT-2 (Radford et al., 2019) – has been investigated. Yan et al. (2022) showed that the CLIP LM outperforms BERT (Devlin et al., 2019) in clustering entities with prompting. Wolfe and Caliskan (2022) probed the CLIP LM, showing its word representations are less anisotropic (Ethayarajh, 2019), i.e., more uniformly distributed with respect to direction, than GPT-2. Complementary to research on CLIP LM, we focus on pretraining. CLIP-like models are often pretrained with cross-modal contrastive learning. *We measure – during pretraining – the effectiveness of systematically modeling the image captions with sentence embedding training* (Reimers and Gurevych, 2019; Gao et al., 2021), which is a natural fit to the captions.

We pretrain CLIP and one of its new variants CyCLIP (Goel et al., 2022) with sentence embedding training, as well as the conventional cross-modal contrastive learning. In addition to CLIP's cross-modal contrastive learning objective, CyCLIP explicitly optimizes for geometry consistency between the text and image representation spaces, making it a suitable model for validating the effectiveness of NLP methods. We evaluate pretrained models on an array of tasks involving one or two modalities such as SentEval, zero-shot VL retrieval, and image classification. We find that unsupervised sentence embedding training improves the language encoder quality and VL tasks. Supervised

[1] Code and model will be open upon publication.

sentence embedding training improves language encoder quality, but the benefit does not necessarily transfer to VL tasks. We analyze the learned representation spaces and find that sentence embedding training improves text-space uniformity (Wang and Isola, 2020) and reduces anisotropy (Ethayarajh, 2019; Wolfe and Caliskan, 2022).

We also investigate AL contrastive models such as CLAP (Wu et al., 2023). In contrast to VL pretraining, AL pretraining suffers from data scarcity and often leverages pretrained LLMs and audio encoders. We determine the effectiveness of sentence embedding training in both scenarios: continued pretraining with LLMs and audio encoders, and pretraining from scratch. We find that sentence embedding training improves AL tasks when the amount of data is large, while the benefits become less noticeable on small datasets. To the best of our knowledge, this is the first study on investigating and trying to improve the language encoder of AL contrastive learning, and we expect our results will encourage more research in this direction.

In summary, our **contributions** are as follows: (i) We extensively evaluate how unsupervised and supervised sentence embedding trainings affect VL and AL contrastive pretraining. Experimental results indicate improved VL performance. AL tasks see improvements when the amount of training data is large, while the benefits become less noticeable on small datasets. (ii) We show that unsupervised sentence embedding training improves the language encoder of CyCLIP (Goel et al., 2022), hence improves performance of cross-modal tasks. (iii) We conduct a comprehensive analysis on the alignment and uniformity of learned representation spaces following Wang and Isola (2020), and show that sentence embedding training improves uniformity of the text representation space, but at the cost of decreased cross-modal alignment.

2 Related work

CLIP LM. Research has focused on the language encoder of OpenAI CLIP (Radford et al., 2021). The model consists of a language encoder (CLIP LM) and an image encoder that are jointly trained on Web-scale caption-image pairs. Yan et al. (2022) stressed the importance of CLIP LM, showing that it outperforms BERT (Devlin et al., 2019) in tasks, such as entity clustering, through prompting. Bielawski et al. (2022) showed that CLIP LM outperforms BERT in "human-centric" tasks such as genre classification on books or movies. Santurkar et al. (2023) highlighted the importance of text captions for representation learning of CLIP by comparing it with SimCLR (Chen et al., 2020) in which no language supervisions is present. Training signals from language are shown to be detrimental, worthing any number of images in a sufficiently large dataset. Our work follows this direction, with a focus on determining how supervised or unsupervised sentence embedding trainings affect CLIP LM and VL contrastive learning.

Goel et al. (2022) introduced CyCLIP, incorporating extra training objectives than cross-modal contrastive learning such that the geometry consistency between the text and image spaces is improved. One of CyCLIP's training objectives is computing similarities between captions; this motivates us to determine how systematically modeling the captions through supervised or unsupervised sentence embedding training affects CyCLIP/CLIP.

Contrastive audio-language pretraining models have also been proposed. Guzhov et al. (2022) extended CLIP to audio tasks by adding an extra module and continued training on audio datasets. Similar distillation methods such as Wav2CLIP (Wu et al., 2022), have also been proposed. Elizalde et al. (2023) and Wu et al. (2023) independently proposed CLAP, in which a language encoder and an audio encoder are jointly trained on AL datasets, which resembles CLIP. *We focused on the language encoder in the AL models, and demonstrated the impact of sentence embedding training.* To the best of our knowledge, this is the first step in this direction.

Sentence embedding is an extensively investigated NLP topic. Methods ranging from bag-of-word averaging non-contextualized embeddings (Mikolov et al., 2013; Pennington et al., 2014) to training LSTMs (Hochreiter and Schmidhuber, 1997), e.g., SkipThought (Kiros et al., 2015) and InferSent (Conneau et al., 2017), have been proposed to effectively compose individual tokens to meaningful sentence representations. Methods that leverage post-hoc transforming (Li et al., 2020; Su et al., 2021) or finetuning the pretrained BERT in supervised (Reimers and Gurevych, 2019) or unsupervised scenarios (Gao et al., 2021) are also introduced. Zhang et al. (2022) show that grounding sentence embedding learning to images improves semantic textual similarity tasks. We present a focused investigation on the LM in CLIP/CyCLIP. We pretrained from scratch an LM and ResNet-50

(He et al., 2016) with cross-modal contrastive learning, as well as unsupervised or supervised sentence embedding training with image captions. Verifying the effectiveness of sentence embeddings – a critical component for retrieval and clustering (Reimers and Gurevych, 2019; Gao et al., 2021; Wang et al., 2021; Thakur et al., 2021; Geigle et al., 2022) – is of great importance because retrieval has been one of the main applications of CLIP-like models.

3 Method

Cross-modal contrastive learning plays a key role in training models such as CLIP/CLAP. We take image modality as an example for introducing this method. Consider a caption-image dataset $\{(I_i, T_i)\}_{i=1}^N$ that includes N caption-image pairs, and denote I^e and T^e as the output representations from an image and langauge encoder, respectively. The cross-modal contrastive loss (Radford et al., 2021) is defined as

$$\mathcal{L}_{\text{contra.}}(\tau) = -\sum_{j=1}^{N} \log \frac{\exp\left(\langle I_j^e, T_j^e \rangle / \tau\right)}{\sum_k \exp\left(\langle I_j^e, T_k^e \rangle / \tau\right)} - \sum_{k=1}^{N} \log \frac{\exp\left(\langle I_k^e, T_k^e \rangle / \tau\right)}{\sum_j \exp\left(\langle I_j^e, T_k^e \rangle / \tau\right)},$$

where τ is a trainable temperature parameter initialized to 0.07 and $\langle \cdot, \cdot \rangle$ computes cosine similarity.

CLIP's training objective solely stresses the alignment between the two modalities. CyCLIP (Goel et al., 2022) has improved CLIP (Radford et al., 2021) by additionally optimizing for improved representation space geometry, such that the image and text spaces are more consistent with each other. Concretely, CyCLIP explicitly optimizes two additional objectives for cross-modal and in-modal consistency as well as $\mathcal{L}_{\text{contra.}}$:

$$\mathcal{L}_{\text{C-cyclic}} = \sum_j \sum_k \left(\langle I_j^e, T_k^e \rangle - \langle I_k^e, T_j^e \rangle\right)^2,$$

$$\mathcal{L}_{\text{I-cyclic}} = \sum_j \sum_k \left(\langle I_j^e, I_k^e \rangle - \langle T_k^e, T_j^e \rangle\right)^2.$$

Intuitively, decreasing the cross-modal consistency loss $\mathcal{L}_{\text{C-cyclic}}$ makes the cross-modal similarity matrix more symmetric, as shown in Figure 1. Note that solely optimizing $\mathcal{L}_{\text{contra.}}$ is expected to symmetrize the cross-modal similarity matrix because the similarity of non-diagonal pairs are trained to be zero. Goel et al. (2022) showed that this scenario does not occur in practice and explicitly optimizing $L_{\text{C-cyclic}}$ is beneficial.

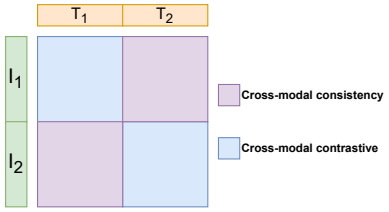


Figure 1: Cross-modal similarity matrix. Diagonal elements refer to cosine similarity between aligned caption-image pairs while non-diagonal elements refer to mismatched caption-image pairs.

Optimizing $\mathcal{L}_{\text{I-cyclic}}$, however, reduces the inconsistency between the overall geometry of the text and image spaces. Computing $\mathcal{L}_{\text{I-cyclic}}$ requires calculating the similarity between two text captions, i.e., $\langle T_k^e, T_j^e \rangle$. It is thus reasonable to hypothesize that *accurately computing caption similarities is beneficial for optimizing* $\mathcal{L}_{\text{I-cyclic}}$. We use unsupervised and supervised sentence embedding training methods and test the hypothesis (§4).

We use the widely used SimCSE (Gao et al., 2021) method for **unsupervise learning** caption representations. SimCSE also uses contrastive learning: a caption is input to a language encoder twice to obtain two vectors T^e and T_+^e. With dropout (Srivastava et al., 2014) enabled, T^e and T_+^e are generally different. These paired vectors serve as the positive training pairs for contrastive learning, while mismatched captions form negative pairs. We denote the unsupervised SimCSE loss as

$$\mathcal{L}_{\text{s}}(\tau) = -\sum_j \log \frac{\exp\left(\langle T_j^e, T_{j,+}^e \rangle / \tau\right)}{\sum_k \exp\left(\langle T_j^e, T_{k,+}^e \rangle / \tau\right)},$$

τ is fixed to 0.05 following Gao et al. (2021).

Another direction of sentence embedding training is **supervised training** (Reimers and Gurevych, 2019) on natural language inference (NLI) datasets, e.g., SNLI and MNLI (Bowman et al., 2015; Williams et al., 2018). We denote the objective as $\mathcal{L}_{\text{n}}$ and follow Gao et al. (2021) in using entailment pairs as T^e and T_+^e and the contradiction sentence as a hard negative.

Table 1 lists various models we use in our experiments as well as their training objectives. During the experiments, we sum up the objectives but weight them with different hyperparameters (λ) depending on the combinations, which are shown in §4. We add a suffix "s" to the name of models trained with $\mathcal{L}_{\text{s}}$ and "n" to models trained with $\mathcal{L}_{\text{n}}$.

	$\mathcal{L}_{\text{contra.}}$	$\mathcal{L}_{\text{C-cyclic}}$	$\mathcal{L}_{\text{I-cyclic}}$	$\mathcal{L}_{\text{s}}$	$\mathcal{L}_{\text{n}}$
CLI (A) P	✓	-	-	-	-
CLI (A) Ps	✓	-	-	✓	-
CLI (A) Pn	✓	-	-	-	✓
CyCLI (A) P	✓	✓	✓	-	-
CyCLI (A) Ps	✓	✓	✓	✓	-
CyCLI (A) Pn	✓	✓	✓	-	(✓)

Table 1: List of training objectives. We follow Radford et al. (2021) for CLIP, Wu et al. (2023) for CLAP, and Goel et al. (2022) for CyCLIP.

	Dataset	Pretraining	Retrieval	ZS	Size
VL	CC3M	✓	-	-	2,806,641
	MSCOCO	-	✓	-	5,000
	Flickr30K	-	✓	-	1,000
	CIFAR10	-	-	✓	10,000
	CIFAR100	-	-	✓	10,000
	ImageNet1K	-	-	✓	50,000
	ImageNetV2	-	-	✓	10,000
	ImageNetSketch	-	-	✓	5,0889
	ImageNet-O	-	-	✓	2,000
	ImageNet-A	-	-	✓	7,500
	ImageNet-R	-	-	✓	30,000
AL	Clotho	✓	✓	-	5,929
	AudioCaps	✓	✓	-	50,725
	FreeSound	✓	✓	-	194,895
	ESC50	-	-	✓	400
	US8K	-	-	✓	8,732

Table 2: Datasets and their amount of examples. We report amount of images for VL datasets and of waveform files for AL datasets. "ZS": zero-shot classification.

4 Experiments

4.1 Datasets

To *pretrain VL models* such as CLIP and CyCLIP, we follow Bugliarello et al. (2021); Goel et al. (2022) and use the Conceptual Captions dataset, which consists of approximately[2] 3M caption-image pairs (CC3M; Sharma et al. (2018)). CC3M has a reasonable size for pretraining and contains a broad coverage of Web content, making it a good option for learning generic VL representations (Bugliarello et al., 2021).

To *evaluate the trained VL models*, we follow (Radford et al., 2021) and conduct evaluations with zero-shot image-text retrieval on the Karpathy (Karpathy and Fei-Fei, 2015) test splits of Flickr30K (Plummer et al., 2015) and MSCOCO (Chen et al., 2015). We skip the evaluation on Flickr30K when supervised sentence embedding training is used, i.e., when $\mathcal{L}_{\text{n}}$ is considered in training. This is because Flickr30K captions are the premises in the SNLI dataset (Bowman et al., 2015), overlapping with the supervised sentence embedding training data. For zero-shot image classification, we use the standard benchmarks CIFAR10, CIFAR100 (Krizhevsky et al., 2009), and ImageNet1K (Russakovsky et al., 2015). Zero-shot image classification with domain shift, out-of-domain, and adversarial examples are also considered: ImageNetV2 (Recht et al., 2019), ImageNet-Sketch (Wang et al., 2019), ImageNet-O, ImageNet-A, and ImageNet-R (Hendrycks et al., 2021b,a).

To *pretrain AL models*, e.g., CLAP and CyCLAP, we conduct experiments with Clotho (Drossos et al., 2020) consisting of ≈6K caption-audio pairs, AudioCaps consisting of ≈50K[3] caption-audio pairs (Kim et al., 2019), and FreeSound (Fonseca et al., 2017; Mei et al., 2023) consisting of ≈195K caption-audio pairs. In contrast to the VL scenario, AL pretraining is known to be challenging due to data scarcity (Wu et al., 2023). We explore the effectiveness of sentence embedding training for AL with datasets with various scales of size.

To *evaluate pretrained AL models*, we conduct cross-modal retrieval and zero-shot audio classification tasks. For Clotho, we train the models on the training split and report retrieval results on the validation split. For AudioCaps and FreeSound, we select the best-performing checkpoint on the validation split and report test split results. We conduct zero-shot classification on the Environmental Sound Classification dataset (ESC50; Piczak (2015)) and UrbanSound8K (US8K; Salamon et al. (2014)), which have been widely used (Wu et al., 2023; Elizalde et al., 2023). ESC50 contains short audio clips containing the sound of different common events such as cats meowing and dogs barking; the clips are categorized into 50 classes, and US8K contains audio clips of urban event sounds such as drilling and street music; the clips are categorized into ten classes.

Table 2 lists all the cross-modal datasets and their usage. We follow Goel et al. (2022) in processing the VL datasets and Wu et al. (2023) in processing the AL datasets; the detailed steps of these processes are shown in Appendix §A.1.

4.2 Experiment settings

To *pretrain the VL models*, we use the same model architecture as CyCLIP (Goel et al., 2022), i.e., a

[2]CC3M images need to be downloaded by users. Due to broken URLs, the exact amount of data varies from time to time; Table 2 shows the exact number of images.

[3]Similar to CC3M, AudioCaps only provides audio captions while users need to download corresponding YouTube videos and convert their audio to waveforms and Table 2 shows the exact amount of waveforms we used.

ResNet-50 as the image encoder and Transformer (Vaswani et al., 2017) as the language encoder. We pretrain the model from scratch and largely reuse CyCLIP's hyperparameters. To weight different training objectives (Table 1) in CyCLIP, we set $\lambda_{\text{I-cyclic}}$ and $\lambda_{\text{C-cyclic}}$ to 0.25, $\lambda_{\text{contra.}}$ is set to 1.0, and we empirically set λ_{s} and λ_{n} to 0.1. We use a batch size of 80, and each pretraining trial is run for 64 epochs, taking four days with four A100 GPUs. We enable dropout in the language encoder and use dropout rate of 0.1. Appendix §A.1 lists the details of the hyperparameters.

The VL pretraining dataset CC3M has a validation split consisting of ≈15K text-image pairs. We select the best-performing checkpoint on this validation split for downstream task evaluations.

Due to data scarcity, *AL pretraining* often leverages pretrained language and audio encoders. We use the same model architecture as LAION-CLAP (Wu et al., 2023): pretrained RoBERTa-base (Liu et al., 2020) as the language encoder and pretrained Hierarchical Token-Semantic Audio Transformer (HTSAT; Chen et al. (2022)) as the audio encoder. HTSAT has shown to outperform CNNs in various audio tasks (Chen et al., 2022). We use the `HTSAT-tiny` variant with 31M parameters. Due to smaller dataset size, each pretraining experiment takes less than one day on a single A100 GPU. We use the default dropout rate of 0.1 of RoBERTa-base in our experiments when the unsupervised sentence embedding training objective is used. Table 9 lists other hyperparameters used in the experiments.

For weighting the training objectives, due to the small model size and dataset size, we grid search the optimal $\lambda_{\text{I-cyclic}}$, $\lambda_{\text{C-cyclic}}$, and λ_{s} from [0.1, 0.25, 0.5]; $\lambda_{\text{contra.}}$ is set to 1.0. Supervised sentence embedding training (λ_{n}) is not considered due to the small size of the AL datasets.

For the AL datasets Clotho, AudioCaps, and FreeSound, we select the best-performing checkpoint on the validation splits then conduct evaluation on downstream tasks.

5 Results and analyses

5.1 Results

Table 3 lists the **zero-shot VL retrieval** results on MSCOCO and Flickr30K, following the settings in Radford et al. (2021). We first utilize the pretrained image and language encoders in a VL model to encode images and captions into vectors. In text retrieval, we then input the image vector to retrieve the aligned captions and vice-versa for image retrieval. We report Recall@N for N in [1, 5, 10].

	Text Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
CLIP	15.70	37.22	49.06	12.48	31.10	42.23
CLIPs	**17.78**	**38.92**	**50.10**	**13.46**	**32.93**	**44.09**
CLIPn	15.74	35.66	47.38	13.12	31.46	42.55
CyCLIP	18.92	41.46	54.00	15.40	35.61	46.95
CyCLIPs	**21.30**	**44.34**	**56.54**	**16.69**	**37.75**	**49.24**
CyCLIPn	16.32	36.76	48.16	14.53	34.07	45.52

	Text Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
CLIP	31.80	62.10	72.90	25.50	52.28	**64.34**
CLIPs	**35.20**	**63.20**	**75.30**	**26.70**	**52.34**	64.32
CyCLIP	37.30	66.10	76.40	30.22	56.70	67.40
CyCLIPs	**40.00**	**69.30**	**79.70**	**31.74**	**58.02**	**69.46**

Table 3: Zero-shot **VL retrieval results** (%) on MSCOCO (top) and Flickr30K (bottom). Integrating unsupervised sentence embedding learning (CLIPs and CyCLIPs) noticeably improves zero-shot retrieval; supervised embedding training (CLIPn and CyCLIPn) has neutral to negative impacts.

Comparing CyCLIP and CLIP variants. We see that CyCLIP clearly outperforms CLIP across the board for both datasets, highlighting the significant value of incorporating $\mathcal{L}_{\text{C-cyclic}}$ and $\mathcal{L}_{\text{I-cyclic}}$ in optimizing for consistent geometry of the text and image representation spaces (Goel et al., 2022).

When **comparing CyCLIPs/CLIPs to CyCLIP/CLIP**, we observe the effectiveness of improving the language encoder with unsupervised sentence embedding training $\mathcal{L}_{\text{s}}$. CyCLIPs/CLIPs clearly surpass CyCLIP/CLIP in all configurations except Flickr30K-CLIPs-ImageRetrieval-R@10. This suggests that, *we improve CyCLIP on zero-shot vision-language retrieval tasks through learning better representations of the captions*. We also observe more gains in text retrieval than in image retrieval. For example, on Flickr30K@1, CyCLIPs outperforms CyCLIP by 2.70% (absolute) in text retrieval and by 1.52% (absolute) in image retrieval. This observation reflects the effectiveness of improved caption representations.

In contrast to the unsupervised embedding training scenario ($\mathcal{L}_{\text{s}}$ and CyCLIPs/CLIPs), supervised sentence embedding training ($\mathcal{L}_{\text{n}}$ and CyCLIPn/CLIPn) results in a neutral to negative impact on the overall retrieval results. Investigating the text and image representation spaces, we find that $\mathcal{L}_{\text{n}}$ extensively enforces a uniform text repre-

	Text Retrieval			Audio Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
CLAP	5.54	18.11	26.81	5.71	18.24	26.94
CLAPs	**5.97**	**18.74**	**27.54**	**6.09**	**19.10**	**27.53**
CyCLAP	5.69	18.94	27.91	5.95	19.11	27.97
CyCLAPs	**6.05**	**19.36**	**28.52**	**6.29**	**19.62**	**28.02**

	Text Retrieval			Audio Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
CLAP	**13.88**	34.16	48.90	11.67	**33.80**	**47.10**
CLAPs	13.49	**35.60**	**49.00**	**11.92**	32.54	45.47
CyCLAP	14.74	35.50	48.52	11.90	**34.95**	**48.61**
CyCLAPs	**14.93**	**36.84**	**51.00**	**12.08**	34.09	46.76

	Text Retrieval			Audio Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
CLAP	**44.10**	**76.80**	**87.67**	**34.82**	**70.62**	82.93
CLAPs	42.73	75.44	87.57	34.69	69.80	**82.99**
CyCLAP	**40.65**	74.19	**86.52**	34.13	69.24	82.30
CyCLAPs	39.81	**74.40**	85.79	**34.23**	**70.24**	**82.74**

Table 4: **Text-audio retrieval results** (%) on FreeSound (top), Clotho (mid), and AudioCaps (bottom). Adding unsupervised sentence embedding training consistently improves performance of CLAP/CyCLAP on the large dataset FreeSound, while improvements on the two small datasets are less noticeable.

	CLIP	CLIPn	CLIPs	CyCLIP	CyCLIPn	CyCLIPs
CIFAR10	28.31	**44.06**	36.80	38.67	41.16	**44.97**
CIFAR100	13.23	**17.93**	10.72	17.44	19.82	**22.05**
ImageNet1K	14.94	15.97	**16.01**	20.99	18.13	**22.13**
ImageNetV2	12.85	13.41	**14.09**	17.77	15.65	**18.68**
ImageNet-Sk.	7.72	7.75	**8.14**	11.67	9.93	**12.85**
ImageNet-O	20.75	**21.95**	21.30	27.05	24.45	**29.55**
ImageNet-A	3.59	3.41	**3.95**	5.03	4.45	**5.19**
ImageNet-R	18.39	**18.51**	18.24	24.37	23.07	**26.72**

Table 5: Zero-shot image classification (R@1 in %) on standard datasets (top) and datasets with distribution shift or adversarial examples (bottom).

sentation space such that the alignment between the text and image spaces is negatively affected; we provide more in-depth analyses in §5.2.

AL retrieval results on Clotho, AudioCaps, and FreeSound are listed in Table 4. Supervised sentence embedding training objective $\mathcal{L}_{\text{n}}$ is not considered because NLI datasets are much larger than AL datasets (e.g., MNLI: 433K; Clotho: 6K); subsampling introduces extra random factors that are difficult to control.

On FreeSound, we observe improvements when comparing CyCLAP to CLAP, demonstrating that explicitly optimizing for the consistency between the audio and text spaces as in CyCLIP (Goel et al., 2022) is also promising for improving AL retrieval tasks. We also observe consistent improvements of CLAPs/CyCLAPs to CLAP/CyCLAP, this shows the benefits of integrating unsupervised sentence embedding training objective during AL contrastive learning.

On Clotho, we observe overall improvements when comparing CyCLAP to CLAP. We see only one exception on text-retrieval-R@10 (i.e., 48.90% for CLAP and 48.52% for CyCLAP); When comparing CyCLAPs/CLAPs with CyCLAP/CLAP, we see clear improvements on text retrieval. However, this comes with a decreased performance on audio retrieval results of R@5 and R@10.

On AudioCaps, CyCLAP falls behind CLAP, showing that optimizing for geometry consistency brings no improvements on AudioCaps. The HTSAT audio encoder has already been pretrained with audio classification tasks on AudioSet (Gemmeke et al., 2017), from which AudioCaps is derived. This may contribute to the noisy results. Similarly, LAION-CLAP (Wu et al., 2023) reported that adding additional 630K AL pairs largely boosts AL retrieval performance on Clotho, but hurts on AudioCaps. We observe similar results when comparing CyCLAPs/CLAPs with CyCLAP/CLAP. We further conduct in-depth analyses (c.f. §5.2) on the audio caption properties of different datasets, and find that AudioCaps captions have a small vocabulary and the language use has very small variations, which likely limits the effectiveness of sentence embedding training.

Comparing VL and AL retrieval results in Table 3 and Table 4, we observe that (1) CyCLIP noticeably improves over CLIP than CyCLAP over CLAP; (2) improving the language encoder with sentence embedding training is more beneficial to VL than AL. We hypothesize that this is because AL pretraining starts with pretrained encoders, which have geometry that is difficult to alter due to the small AL dataset size. We further conduct AL pretraining from scratch (Appendix §A.4), where the language and audio encoders are randomly reinitialized. We observe that sentence embedding training brings more consistent and noticeable results, especially on FreeSound and Clotho. However, not utilizing the pretrained encoders leads to inferior absolute performances due to the small dataset sizes. We believe that resolving the data scarcity issue is still a critical step for future work in AL pretraining.

For **zero-shot image classification**, Table 5 lists

		CLAP	CLAPs	CyCLAP	CyCLAPs
FreeSound	ESC50	91.00	**91.75**	**92.25**	91.25
	US8K	82.02	**82.56**	**82.95**	82.65
Clotho	ESC50	72.25	**74.00**	77.00	**77.50**
	US8K	69.84	**70.58**	**71.99**	69.14
AudioCaps	ESC50	**80.75**	76.00	79.00	79.00
	US8K	**71.66**	66.30	69.06	**69.31**

Table 6: Zero-shot audio classification (R@1 in %) on ESC50 and US8K of models pretrained on FreeSound, Clotho, and AudioCaps.

the Top1 accuracy on standard image classification datasets (top) and datasets with distribution shifts or adversarial examples (bottom). We follow Radford et al. (2021) and use their prompts for zero-shot classification. For an image to be classified, we compute the cosine similarity between its vector and the encoded vector of all classes. Each of the classes is reformulated with various prompts. E.g., the ImageNet class "plane" is reformulated with 80 templates [4] such as "a photo of a " and "a blurry photo of a ", resulting in prompts "a photo of a plane" and "a blurry photo of a plane" (Radford et al., 2021). The vectors of encoded prompts of a class are averaged; we select the class with the maximum cosine similarity with the image vector.

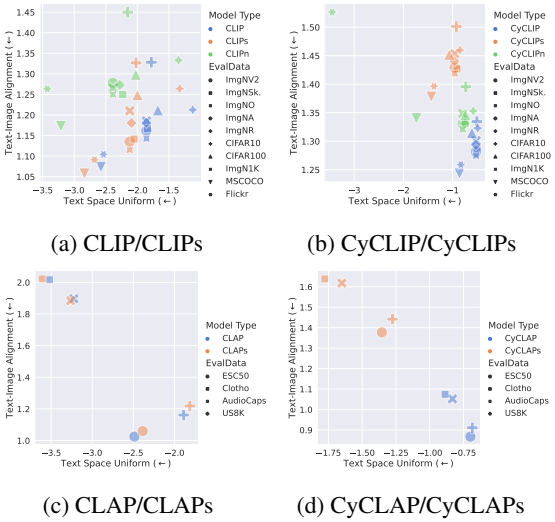
Similar trends as in the retrieval tasks are observed. CyCLIP variants outperform their CLIP counterparts; unsupervised sentence embedding training benefits both CyCLIP/CLIP while supervised sentence embedding training does not result in consistent improvement or deterioration.

For **zero-shot audio classification**, we follow the VL scenario to write several prompts (c.f. Appendix §A.3) such as "a sound of dog barking" and conduct similar experiments. Table 6 list the Top1 accuracy on ESC50 and US8K of models pretrained on FreeSound, Clotho, and AudioCaps respectively. CyCLAP generally outperforms CLAP, except for AudioCaps-US8K. We observe mixed performance when comparing different model configurations, and CyCLAPs/CLAPs perform on par with CyCLAP/CLAP across different datasets. Prompting inherently leads to performance with large variances (Zhao et al., 2021); we leave the extensive "prompt engineering" of designing more prompts for future work.

5.2 Analyses

Alignment and uniformity of representation spaces. In this section, we take a closer look at

[4] OpenAI templates are public on GitHub link.



(a) CLIP/CLIPs (b) CyCLIP/CyCLIPs

(c) CLAP/CLAPs (d) CyCLAP/CyCLAPs

Figure 2: Visualizing cross-modal alignment w.r.t. text space uniformity of trained VL and AL models. To visualize AL results, we use models pretrained on AudioCaps. We observe that sentence embedding training trades cross-modal alignment for text space uniformity.

the learned representation spaces. Following Wang and Isola (2020), we inspect the **alignment** and **uniformity** on the hypersphere of learned spaces. Considering a caption-image dataset $\{(I_i, T_i)\}_{i=1}^N$, we can compute the alignment and uniformity scores defined as:

$$\mathcal{L}_{\text{align}} \triangleq \mathop{\mathbb{E}}_{(I,T) \sim p_{\text{pos}}} \|I^e - T^e\|_2^2,$$

$$\mathcal{L}_{\text{T,uniform}} \triangleq \log \mathop{\mathbb{E}}_{T_i, T_j \overset{\text{i.i.d.}}{\sim} p_{\text{data}}} e^{-2\|T_i^e - T_j^e\|_2^2},$$

$$\mathcal{L}_{\text{I,uniform}} \triangleq \log \mathop{\mathbb{E}}_{I_i, I_j \overset{\text{i.i.d.}}{\sim} p_{\text{data}}} e^{-2\|I_i^e - I_j^e\|_2^2},$$

where $(I, T) \sim p_{\text{pos}}$ refers to aligned text-images pairs, $(T_i, T_j) \sim p_{\text{data}}$ refers to independent and identically distributed (IID) sampled text pairs, $(I_i, I_j) \sim p_{\text{data}}$ refers to IID sampled image pairs, and I^e, T^e respectively refer to the encoded image and text vectors. Recall that the trained models output ℓ-2 normalized vectors residing on the unit ball. Intuitively, we want vectors of aligned pairs of two modalities to be well aligned in the representation space, such that $\mathcal{L}_{\text{align}}$ is close to zero. However, we want the space of a single modality to be more uniform than anisotropic (Ethayarajh, 2019; Wolfe and Caliskan, 2022), such that the overall representation space capacity is well used. This results in a near minus infinite $\mathcal{L}_{\text{uniform}}$.

Figure 2 illustrates the results. We only show $\mathcal{L}_{\text{align}}$ w.r.t. $\mathcal{L}_{\text{T,uniform}}$ since our main focus is the language encoder. We see that unsupervised sen-

tence embedding training trades cross-modal alignment for improving the text space uniformity. For VL pretraining, supervised sentence embedding training (CLIPn/CyCLIPn) overly focuses on text space uniformity while the VL space alignment deteriorates, as evidenced when visualizing them using the Flickr30K (*) dataset in which the captions largely overlap with the dataset for supervised sentence embedding training (§4.1).

Another interesting observation is that the text encoder of CyCLIP/CyCLAP outputs representation space less uniform than that of CLIP/CLAP, as shown in Figure 2 (x-axis). Liang et al. (2022) show that randomly initialized encoders output vectors residing in different cones. The in-modal cyclic loss $\mathcal{L}_{\text{I-cyclic}}$ (§3) stresses consistency between cones; it is thus expected to be challenging to learn a uniform space while simultaneously preserving consistency between two spaces[5]. Sentence embedding training provides extra training signals.

Audio dataset analyses. Table 4 shows that sentence embedding training consistently improves on the largest AL dataset FreeSound; the benefits diminish on Clotho and AudioCaps. Besides dataset size, we further investigate properties of the AL datasets. We firstly compute word frequency, and then normalized it by the total number of words in the captions. Then we sort the words in decreasing order. Figure 3 shows the results in log scale. AudioCaps has the smallest vocabulary size, and there are little variations on the word use (i.e., a few words dominant the captions). This property could hinder improving the uniformity of the text space. Clotho, ≈9 times smaller than AudioCaps, has a larger vocabulary and a more uniform word frequency distribution of its captions. FreeSound has the largest number of and diverse captions among AL datasets; integrating sentence embedding training consistently improves the AL tasks.

LM quality. We evaluate the language encoder quality of pretrained VL models. Our motivation is two-fold. First, as a sanity check, we want to verify that incorporating sentence embedding training in VL contrastive learning still improves the language encoder's ability of representing general sentences. Second, the evaluation results help us measure the compatibility and possible interferences among the various training objectives (Pfeiffer et al., 2023). To this goal, we use the sentence embedding benchmark SentEval (Conneau and Kiela, 2018). Default SentEval configurations are used in all experiments, and we conduct both intrinsic (e.g., semantic textual similarity) and extrinsic tasks (e.g., sentiment analysis). Table 7 lists the averaged results on SentEval (Appendix §A.7 shows individual results). We observe that unsupervised sentence embedding training is generally beneficial for CyCLIP/CLIP on both intrinsic and extrinsic tasks. Supervised sentence embedding training results in more significant improvements[6], however, it negatively affects CyCLIPn on the sensitive intrinsic tasks.

[5]We find that improving text space uniformity also benefits the image space for CyCLIP. More discussions are presented in Appendix §A.2.

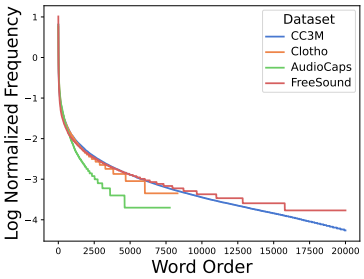


Figure 3: AudioCaps has the smallest vocabulary size; a small amount of frequent words dominants the captions, leading to small variations on the word use. We cut the vocabulary size to 20K for better visualization.

	CLIP	CLIPn	CLIPs	CyCLIP	CyCLIPn	CyCLIPs
Intrin.	55.45	**64.50**	57.58	53.72	49.70	**55.67**
Extrin.	67.03	**68.68**	67.64	65.82	**69.49**	67.95

Table 7: Averaged intrinsic and extrinsic SentEval task results of the language encoder in VL models.

6 Conclusion

We extensively investigate the effectiveness of sentence embedding training for pretraining contrastive vision-language and audio-language models. We show that it improves vision-language pretraining, resulting in a better CyCLIP. Sentence embedding training also improves audio-language pretraining on large datasets, while the benefits diminish on small datasets. We conduct comprehensive analyses and show that sentence embedding training increases text space uniformity, but with a cost of reduced cross-modal alignment.

[6]Improvements of supervised sentence embedding training may due to the observation that NLI datasets have similar domains and language use as SentEval tasks. In Appendix §A.6, we show that the improvements are indeed from supervised training, rather than domain similarity.

Limitations

We restrict our scope to cross-modal contrastive models with three most common modalities: language, image, and audio. While contrastive learning has been successfully extended to other modalities such as music, incorporating music poses additional challenges, particularly regarding licensing and the heterogeneity of music sources. Downloading music from the internet and mining reliable music-language pairs are time-consuming tasks, which we did not consider in detail in this study. Nevertheless, we conducted initial experiments on the music modality using MusicCaps (Agostinelli et al., 2023) and show promising results in §A.8.

In our audio-language pretraining experiments, we explored both pretraining from scratch and pretraining from publicly available language and audio encoders. We believe that a promising direction for future work would involve adapting the pretrained language encoder to the audio domain by performing additional pretraining (Gururangan et al., 2020) on audio descriptions before engaging in cross-modal contrastive learning. Nevertheless, we chose to follow the current methods in the literature to ensure consistent evaluations and facilitate meaningful comparisons.

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Inigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, et al. 2015. Semeval-2015 task 2: Semantic textual similarity, english, spanish and pilot on interpretability. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 252–263.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel M Cer, Mona T Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. Semeval-2014 task 10: Multilingual semantic textual similarity. In *SemEval@ COLING*, pages 81–91.

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez Agirre, Rada Mihalcea, German Rigau Claramunt, and Janyce Wiebe. 2016. Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *SemEval-2016. 10th International Workshop on Semantic Evaluation; 2016 Jun 16-17; San Diego, CA. Stroudsburg (PA): ACL; 2016. p. 497-511.* ACL (Association for Computational Linguistics).

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. Semeval-2012 task 6: A pilot on semantic textual similarity. In ** SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 385–393.

Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. *SEM 2013 shared task: Semantic textual similarity. In *Second joint conference on lexical and computational semantics (* SEM), volume 1: proceedings of the Main conference and the shared task: semantic textual similarity*, pages 32–43.

Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al. 2023. MusicLM: Generating music from text. *arXiv preprint arXiv:2301.11325*.

Romain Bielawski, Benjamin Devillers, Tim Van De Cruys, and Rufin Vanrullen. 2022. When does CLIP generalize better than unimodal models? when judging human-centric concepts. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pages 29–38, Dublin, Ireland. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Emanuele Bugliarello, Ryan Cotterell, Naoaki Okazaki, and Desmond Elliott. 2021. Multimodal Pretraining Unmasked: A Meta-Analysis and a Unified Framework of Vision-and-Language BERTs. *Transactions of the Association for Computational Linguistics*, 9:978–994.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Ke Chen, Xingjian Du, Bilei Zhu, Zejun Ma, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2022. HTS-AT: A hierarchical token-semantic audio transformer for sound classification and detection. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 646–650. IEEE.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. 2023. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829.

Alexis Conneau and Douwe Kiela. 2018. SentEval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 350–356, Geneva, Switzerland. COLING.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. 2020. Clotho: an audio captioning dataset. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 736–740.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023. Clap learning audio concepts from natural language supervision. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65. Association for Computational Linguistics.

Eduardo Fonseca, Jordi Pons Puig, Xavier Favory, Frederic Font Corbera, Dmitry Bogdanov, Andres Ferraro, Sergio Oramas, Alastair Porter, and Xavier Serra. 2017. Freesound datasets: a platform for the creation of open audio datasets. In *Hu X, Cunningham SJ, Turnbull D, Duan Z, editors. Proceedings of the 18th ISMIR Conference; 2017 oct 23-27; Suzhou, China.[Canada]: International Society for Music Information Retrieval; 2017. p. 486-93.* International Society for Music Information Retrieval (ISMIR).

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Gregor Geigle, Jonas Pfeiffer, Nils Reimers, Ivan Vulić, and Iryna Gurevych. 2022. Retrieve Fast, Rerank Smart: Cooperative and Joint Approaches for Improved Cross-Modal Retrieval. *Transactions of the Association for Computational Linguistics*, 10:503–521.

Jort F Gemmeke, Daniel PW 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 *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE.

Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan A. Rossi, Vishwa Vinay, and Aditya Grover. 2022. CyCLIP: Cyclic contrastive language-image pretraining. In *Advances in Neural Information Processing Systems*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. 2022. Audioclip: Extending clip to image, text and audio. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 976–980.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. 2021a. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. 2021b. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, page 168–177, New York, NY, USA. Association for Computing Machinery.

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.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. AudioCaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 119–132, Minneapolis, Minnesota. Association for Computational Linguistics.

Ryan Kiros, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. *Advances in neural information processing systems*, 28.

Alex Krizhevsky, Geoffrey Hinton, et al. 2009. Learning multiple layers of features from tiny images.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9119–9130, Online. Association for Computational Linguistics.

Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. 2022. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *Advances in Neural Information Processing Systems*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Ro{bert}a: A robustly optimized {bert} pretraining approach.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223, Reykjavik, Iceland. European Language Resources Association (ELRA).

Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D. Plumbley, Yuexian Zou, and Wenwu Wang. 2023. WavCaps: A ChatGPT-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *arXiv preprint arXiv:2303.17395*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 271–278, Barcelona, Spain.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 115–124, Ann Arbor, Michigan. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Jonas Pfeiffer, Sebastian Ruder, Ivan Vulić, and Edoardo Maria Ponti. 2023. Modular deep learning.

Karol J. Piczak. 2015. ESC: Dataset for Environmental Sound Classification. In *Proceedings of the 23rd Annual ACM Conference on Multimedia*, pages 1015–1018. ACM Press.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2641–2649.

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

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2019. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992. Association for Computational Linguistics.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252.

Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. 2014. A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM International Conference on Multimedia*, MM '14, page 1041–1044, New York, NY, USA. Association for Computing Machinery.

Shibani Santurkar, Yann Dubois, Rohan Taori, Percy Liang, and Tatsunori Hashimoto. 2023. Is a caption worth a thousand images? a study on representation learning. In *International Conference on Learning Representations*.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. 2021. Whitening sentence representations for better semantics and faster retrieval. *arXiv preprint arXiv:2103.15316*.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Ellen M. Voorhees and Dawn M. Tice. 2000. Building a question answering test collection. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '00, page 200–207, New York, NY, USA. Association for Computing Machinery.

Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. 2019. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, pages 10506–10518.

Kexin Wang, Nils Reimers, and Iryna Gurevych. 2021. TSDAE: Using transformer-based sequential denoising auto-encoderfor unsupervised sentence embedding learning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 671–688, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39:165–210.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Robert Wolfe and Aylin Caliskan. 2022. Contrastive visual semantic pretraining magnifies the semantics of natural language representations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3050–3061, Dublin, Ireland. Association for Computational Linguistics.

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.

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

An Yan, Jiacheng Li, Wanrong Zhu, Yujie Lu, William Yang Wang, and Julian McAuley. 2022. CLIP also understands text: Prompting CLIP for phrase understanding.

Miaoran Zhang, Marius Mosbach, David Adelani, Michael Hedderich, and Dietrich Klakow. 2022. MCSE: Multimodal contrastive learning of sentence embeddings. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5959–5969, Seattle, United States. Association for Computational Linguistics.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pages 12697–12706. PMLR.

A Appendix

A.1 Datasets and hyperparameters

For VL pretraining, our experiments largely follow those for CyCLIP (Goel et al., 2022) and for CLIP (Radford et al., 2021). We also directly reuse CyCLIP training hyperparameters but with a smaller batch size, as listed in Table 8.

For AL pretraining, our experiments largely follow that of for LAION-CLAP (Wu et al., 2023). To process the audio data, we sample the wavefiles at a rate of 48kHz and then convert them to FLAC format using `FFmpeg`[7]. We then use a hop size of 480, window size of 1024, and 64 mel-bins for computing Short-time Fourier transform (STFT) and mel-spectrograms. The audio encoder input thus has a dimension of 1024 for time steps and 64 for frequency bins. We list the hyperparameters in Table 9.

Hyperparameter	Value
Logit scale range	0 to 4.6052
Epochs	64
Batch size	80
Learning rate	0.0005
Optimizer	Adam
Scheduler	Cosine
Learning rate warmup steps	10000
Language encoder dropout	0.1

Table 8: Hyperparameters used for training VL models.

Hyperparameter	Value
Logit scale range	0 to 4.6052
Epochs	90
Batch size	80
Learning rate	0.00009
Optimizer	AdamW
Scheduler	Cosine
Learning rate warmup steps	9600
Language encoder dropout	0.1

Table 9: Hyperparameters used for training AL models. For FreeSound, we train for 30 epochs and use warmup steps 3200 due to its larger size.

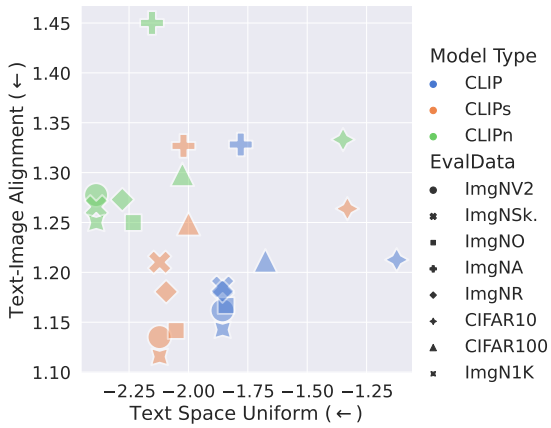
A.2 Text and image space consistency

CyCLIP variants ensure cross-modal consistency, such that improving the uniformity of the text space with sentence embedding training also benefits image space uniformity as shown in Figure 4. As expected, this observation does not hold for CLIP.

A.3 Zero-shot audio classification with prompts

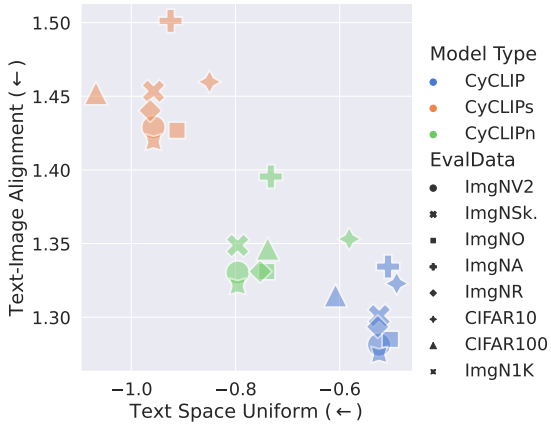
To resemble zero-shot image classification in VL experiments, we write several prompts for zero-shot audio classification, as listed in Table 10.

[7] https://ffmpeg.org/

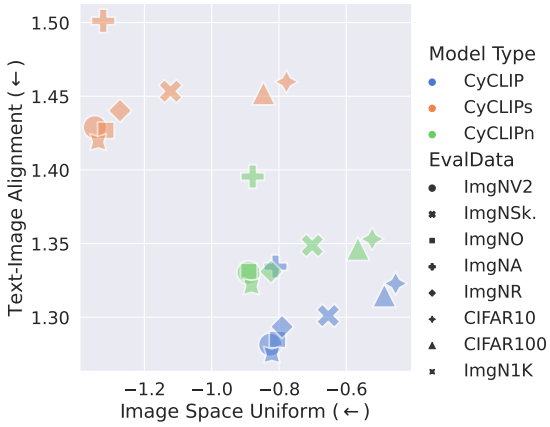


(a) CLIP: alignment vs. text space uniformity

(b) CLIP: alignment vs. image space uniformity



(c) CyCLIP: alignment vs. text space uniformity



(d) CyCLIP: alignment vs. image space uniformity

Figure 4: Comparing the text and image space consistency between CLIP and CyCLIP variants. Improving uniformity of the text space also benefits image space in CyCLIP.

A sound of `label`.
a sound of `label`.
The sound of `label`.
the sound of `label`.
A constant sound of `label`.
a constant sound of `label`.
A big sound of `label`.
a big sound of `label`.
A small sound of `label`.
A small sound of `label`.
A `label` is making a sound.
a `label` is making a sound.
An `label` is making a sound.
an `label` is making a sound.
A sound of `label` followed by a sound of `label`.
A sound of `label` followed by `label`.
A `label`.
An `label`.
`label`.
`label` and `label`.
A `label` is running.
A `label` is happening.

Table 10: We write several prompts for zero-shot audio classification resembling the VL prompts. "`label`" refers to the audio class label.

	Text Retrieval			Audio Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
CLAP	1.42	5.09	8.32	1.54	5.30	8.42
CLAPs	**1.63**	**5.17**	**8.40**	**1.69**	**5.46**	**8.89**
CyCLAP	1.21	5.16	8.24	1.49	5.38	8.60
CyCLAPs	**1.73**	**5.77**	**9.23**	**1.94**	**6.36**	**9.54**
	Text Retrieval			Audio Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
CLAP	**2.30**	7.85	13.88	2.28	7.94	13.47
CLAPs	2.11	**8.32**	**15.98**	**2.81**	**8.84**	**14.91**
CyCLAP	2.97	9.09	14.07	2.07	7.67	13.11
CyCLAPs	**3.54**	**10.05**	**15.41**	**2.64**	**8.88**	**14.87**
	Text Retrieval			Audio Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
CLAP	16.30	43.78	**58.41**	**14.19**	**40.02**	**53.98**
CLAPs	**17.76**	**44.10**	57.99	13.81	38.60	52.02
CyCLAP	**18.81**	44.83	**61.23**	**14.96**	39.94	54.67
CyCLAPs	18.18	**47.23**	61.02	13.96	**40.56**	**54.84**

Table 11: Text and audio retrieval results (%) on FreeSound (top), Clotho (mid) and AudioCaps (bottom) when pretraining AL models from scratch.

A.4 Training audio-language models from scratch

In §5.1 we show that sentence embedding training brings more noticeable impacts in learning VL models than in AL models; we conjecture that this is resulted from the fact that AL pretraining often leverages pretrained language and audio encoders (Elizalde et al., 2023; Wu et al., 2023). As a result, we conduct the experiments of pretraining the AL model from scratch, i.e., the language and audio encoders re-initialized. Table 11 lists the retrieval results on Clotho and AudioCaps.

Compared with Table 4, we observe a significant performance drop since the encoders are pretrained from scratch. The results are still noisy. We consider that larger scale AL datasets are necessary to highlight the effectiveness of learning consistent representation spaces and sentence embedding training.

A.5 Extended results of audio-language models

We observe that the differences between Table 4 results are smaller than in the VL scenario. As a result, we repeat each experiment three times and then report mean and variance of the results in Table 12. Similar observations as in Table 4 are obtained.

A.6 Unsupervised sentence embedding training with NLI datasets

When evaluating the language encoder on SentEval tasks (§5.2), it is possible that the improvements brought by supervised sentence embedding training is due to the fact that NLI datasets have similar domain and language use as the SentEval tasks. We thus conduct a new type of training, where we use sentences in the NLI datasets for unsupervised sentence embedding training with SimCSE, in addition to VL contrastive learning. We name this new training scheme as CLIPe and CyCLIPe.

Table 16 shows that the new training schemes, CLIPe and CyCLIPe fall behind the supervised sentence embedding training counterparts CLIPn and CyCLIPn on SentEval. This confirms that the gains of supervised sentence embedding trainings is from the NLI task supervision, e.g., premise and hypothesis relations, instead of other factors such as domain. For completeness, we also report CLIPe/CyCLIPe performance on VL retrieval tasks in Table 13 and zero-shot image classification in Table 17.

	Text Retrieval			Audio Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
CLAP	(5.73, 0.04)	(18.08, 0.00)	(27.15, 0.09)	(5.78, 0.02)	(18.13, 0.01)	(26.96, 0.01)
CLAPs	(**5.95**, 0.01)	(**18.71**, 0.06)	(**27.45**, 0.07)	(**6.08**, 0.00)	(**18.79**, 0.12)	(**27.45**, 0.10)
CyCLAP	(5.84, 0.01)	(18.86, 0.01)	(27.81, 0.06)	(5.91, 0.01)	(18.82, 0.05)	(27.68, 0.05)
CyCLAPs	(**6.11**, 0.00)	(**19.37**, 0.00)	(**28.42**, 0.06)	(**6.11**, 0.02)	(**19.31**, 0.11)	(**28.22**, 0.05)

	Text Retrieval			Audio Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
CLAP	(**14.93**, 0.55)	(35.02, 0.41)	(48.45, 0.12)	(**12.25**, 0.21)	(**33.38**, 0.23)	(**46.72**, 0.59)
CLAPs	(14.29, 1.43)	(**35.66**, 1.04)	(**49.67**, 0.22)	(12.15, 0.12)	(32.88, 0.06)	(46.06, 0.18)
CyCLAP	(13.75, 0.49)	(36.07, 0.46)	(48.77, 2.00)	(11.94, 0.07)	(**34.37**, 0.29)	(**48.13**, 0.51)
CyCLAPs	(**14.96**, 0.01)	(**37.58**, 1.74)	(**50.59**, 0.10)	(**12.16**, 0.24)	(34.23, 0.08)	(47.27, 0.18)

	Text Retrieval			Audio Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
CLAP	(**42.81**, 0.85)	(**76.21**, 0.19)	(**86.87**, 0.41)	(34.99, 0.03)	(70.05, 0.16)	(82.44, 0.14)
CLAPs	(35.08, 0.30)	(69.89, 0.08)	(82.47, 0.25)	(**44.26**, 1.33)	(**75.65**, 0.14)	(**87.39**, 0.16)
CyCLAP	(**42.11**, 2.50)	(**74.15**, 0.09)	(**86.10**, 0.09)	(**34.36**, 0.05)	(69.88, 0.44)	(82.49, 0.34)
CyCLAPs	(41.17, 4.30)	(74.09, 0.20)	(85.89, 0.02)	(33.94, 0.06)	(**70.14**, 0.07)	(**82.80**, 0.19)

Table 12: **Extended text-audio retrieval results** (%) on FreeSound (top), Clotho (mid), and AudioCaps (bottom). We repeat each experiment three times by changing random seeds, and then report the results in format: (mean of performance, variance of performance). Similar observations as Table 4 can be obtained.

	Text Retrieval			Image Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
CLIP	15.70	37.22	49.06	12.48	31.10	42.23
CLIPs	**17.78**	**38.92**	**50.10**	**13.46**	**32.93**	**44.09**
CLIPn	15.74	35.66	47.38	13.12	31.46	42.55
CLIPe	16.90	38.40	49.92	13.21	31.21	42.26
CyCLIP	18.92	41.46	54.00	15.40	35.61	46.95
CyCLIPs	**21.30**	**44.34**	**56.54**	**16.69**	**37.75**	**49.24**
CyCLIPn	16.32	36.76	48.16	14.53	34.07	45.52
CLIPe	16.22	37.52	49.22	14.05	32.56	43.15

Table 13: Zero-shot **VL retrieval results** (%) on MSCOCO (top) and Flickr30K (bottom).

	CLIP	CLIPn	CLIPs	CyCLIP	CyCLIPn	CyCLIPs
STS12	46.14	54.25	50.31	37.84	45.60	40.42
STS13	50.24	59.67	48.44	52.35	37.82	54.90
STS14	48.70	59.26	51.73	46.58	40.55	49.46
STS15	64.90	73.81	66.09	63.25	59.62	67.01
STS16	51.94	63.08	55.62	50.96	46.80	52.87
STS-B	61.54	68.36	65.04	60.30	54.88	60.72
SICKR	64.70	73.09	65.82	64.78	62.62	64.34
Avg	55.45	**64.50**	57.58	53.72	49.70	**55.67**
MR	61.07	63.66	61.11	59.51	62.78	60.65
CR	67.63	71.07	68.03	67.02	73.67	66.12
SUBJ	76.39	78.09	77.52	74.24	78.90	77.36
MPQA	74.60	77.25	74.80	74.69	80.13	76.16
SST2	61.67	66.89	63.65	61.50	68.42	64.25
TREC	60.80	56.00	60.60	55.80	53.80	62.80
MRPC	67.07	67.77	67.77	68.00	68.75	68.29
Avg	67.03	**68.68**	67.64	65.82	**69.49**	67.95

Table 14: Intrinsic (top) and extrinsic (bottom) SentEval task performance of the language encoder in VL models.

A.7 Complete results on SentEval

Our intrinsic evaluation tasks are the semantic textual similarity tasks: STS12-STS16, STS-B, SICKR (Marelli et al., 2014; Cer et al., 2017; Agirre et al., 2012, 2013, 2014, 2015, 2016). Extrinsic evaluation tasks are movie review (MR; Pang and Lee (2005)) product review (CR; Hu and Liu (2004)) subjectivity status (SUBJ; Pang and Lee (2004)), opinion polarity (MPQA; Wiebe et al. (2005)), sentiment analysis on SST2 (Socher et al., 2013), question-type classification (TREC; Voorhees and Tice (2000)), and paraphrase detection (MRPC; Dolan et al. (2004)). Table 14 shows individual task performances.

A.8 Preliminary experiments on the music modality

We further conducted a new experiment with the music modality: music-text retrieval on the MusicCaps dataset introduced by MusicLM (Agostinelli et al., 2023). MusicCaps consists of 5521 music-caption pairs, of which 2858 pairs are for training and 2663 are for validation. Each music clip is associated with hand-curated English descriptions (including genre, mood, tempo, singer voices etc.) from expert musicians. We use MusicCaps because it is open-sourced and publicly available. Follow-

	Text Retrieval			Music Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
CLAP	**6.05**	18.42	28.33	5.46	18.45	**28.50**
CLAPs	5.99	**18.66**	**28.85**	**6.05**	**19.01**	28.43
CyCLAP	**6.37**	18.98	29.45	**6.69**	18.87	**29.45**
CyCLAPs	6.34	**20.06**	**30.08**	6.62	**19.36**	29.13

Table 15: Text and music retrieval results (%) on MusicCaps.

ing tables show the retrieval results (the same experiment configurations as the audio modality are used; cf. §A.1).

It can be observed from Table 15 that improving the text encoder with unsupervised sentence embedding training also helps music-text retrieval in the music modality, especially in the text retrieval scenario (CLAPs generally outperforms CLAP; CyCLAPs generally outperforms CyCLAP). These music modality results are consistent with our previous findings on the image and audio modalities, and we plan to explore more in this direction in future work.

	CLIP	CLIPn	**CLIPe**	CLIPs	CyCLIP	CyCLIPn	**CyCLIPe**	CyCLIPs
STS12	46.14	54.25	46.54	50.31	37.84	45.60	42.03	40.42
STS13	50.24	59.67	46.29	48.44	52.35	37.82	35.56	54.90
STS14	48.70	59.26	47.66	51.73	46.58	40.55	27.57	49.46
STS15	64.90	73.81	65.48	66.09	63.25	59.62	46.65	67.01
STS16	51.94	63.08	52.39	55.62	50.96	46.80	33.83	52.87
STS-B	61.54	68.36	60.99	65.04	60.30	54.88	44.63	60.72
SICKR	64.70	73.09	62.86	65.82	64.78	62.62	47.53	64.34
Avg	55.45	**64.50**	54.60	57.58	53.72	49.70	39.69	**55.67**
MR	61.07	63.66	59.91	61.11	59.51	62.78	59.60	60.65
CR	67.63	71.07	68.74	68.03	67.02	73.67	64.61	66.12
SUBJ	76.39	78.09	75.86	77.52	74.24	78.90	74.16	77.36
MPQA	74.60	77.25	73.54	74.80	74.69	80.13	73.95	76.16
SST2	61.67	66.89	60.19	63.65	61.50	68.42	60.46	64.25
TREC	60.80	56.00	56.60	60.60	55.80	53.80	57.60	62.80
MRPC	67.07	67.77	67.83	67.77	68.00	68.75	67.48	68.29
Avg	67.03	**68.68**	66.10	67.64	65.82	**69.49**	65.41	67.95

Table 16: Evaluating the language encoder of different VL models with intrinsic (top) and extrinsic (bottom) SentEval tasks.

	CLIP	CLIPn	**CLIPe**	CLIPs	CyCLIP	CyCLIPn	**CLIPe**	CyCLIPs
CIFAR10	28.31	**44.06**	33.97	36.80	38.67	41.16	**50.48**	44.97
CIFAR100	13.23	**17.93**	12.30	10.72	17.44	19.82	21.76	**22.05**
ImageNet1K	14.94	15.97	15.74	**16.01**	20.99	18.13	20.07	**22.13**
ImageNetV2	12.85	13.41	13.51	**14.09**	17.77	15.65	17.34	**18.68**
ImageNet-Sk.	7.72	7.75	6.36	**8.14**	11.67	9.93	11.54	**12.85**
ImageNet-O	20.75	**21.95**	20.45	21.30	27.05	24.45	27.20	**29.55**
ImageNet-A	3.59	3.41	3.59	**3.95**	5.03	4.45	4.93	**5.19**
ImageNet-R	18.39	**18.51**	18.25	18.24	24.37	23.07	24.36	**26.72**

Table 17: Zero-shot image classification (R@1 in %) on standard datasets (top) and datasets with distribution shift or adversarial examples (bottom).