The Thirty-Seventh AAAI Conference on Artificial Intelligence (AAAI-23)

Accommodating Audio Modality in CLIP for Multimodal Processing

Ludan Ruan, Anwen Hu, Yuqing Song, Liang Zhang, Sipeng Zheng, Qin Jin*

School of Information, Renmin University of China
{ruanld,anwenhu,syuqing,zhangliang00,zhengsipeng,qjin}@ruc.edu.cn

Abstract

Multimodal processing has attracted much attention lately especially with the success of pre-training. However, the exploration has mainly focused on vision-language pre-training, as introducing more modalities can greatly complicate model design and optimization. In this paper, we extend the state-of-the-art Vision-Language model CLIP to accommodate the audio modality for Vision-Language-Audio multimodal processing. Specifically, we apply inter-modal and intra-modal contrastive learning to explore the correlation between audio and other modalities in addition to the inner characteristics of the audio modality. Moreover, we further design an audio type token to dynamically learn different audio information type for different scenarios, as both verbal and nonverbal heterogeneous information is conveyed in general audios. Our proposed CLIP4VLA model is validated in different downstream tasks including video retrieval and video captioning, and achieves the state-of-the-art performance on the benchmark datasets of MSR-VTT, VATEX, and Audiocaps.The corresponding code and checkpoints will be released at *https://github.com/ludanruan/CLIP4VLA*.

Introduction

Multimodal processing (Portillo-Quintero, Ortiz-Bayliss, and Terashima-Marín 2021; Carion et al. 2020; Sun et al. 2019) aims to learn the general knowledge across multiple modalities of our daily perception, such as text, vision and audio. Due to the high complexity and high training cost of multimodal alignment, most works focus on the processing of two modalities such as text and vision. However, only visual and textual information may be insufficient to comprehensively understand a realistic scenario. For example, in sports program, the sound of the race start gun and the cheers of the crowd can describe the intensity of the competition even more than the picture, and the narrator's commentary helps the general audience with less sports knowledge to better understand the progress of the game. Therefore, it is necessary to equip the video pre-training models with audio modality modeling.

In recent years, pre-training has achieved great success in multimodal processing. For example, Vision-Language (VL) pre-training models (Carion et al. 2020; Sun et al. 2019; Portillo-Quintero, Ortiz-Bayliss, and Terashima-Marín 2021) have shown superior performance for understanding tasks such as text-visual retrieval and flexible scalability for generation tasks such as video captioning. Audio pre-training models (Gong, Chung, and Glass 2021; Baevski et al. 2020; Chen et al. 2022) can represent complex audio information. As learning general correlations of vision, text and audio via pre-training from scratch is highly computation costly (e.g. 768 TPU days for VATT (Akbari et al. 2021)), one straight-forward idea is to combine the state-of-the-art VL models with the pre-trained audio backbones. However, it faces two main challenges. First, the text, vision and audio backbones usually have different model structures, which makes it hard to combine via a unified training strategy. For example, the audio pre-training models for Automatic Speech Recognition (ASR) normally process audio at the phoneme level, whose parameters are too heavy compared with the VL models. Second, there is currently no single audio backbone that can fully handle rich and different types of information conveyed in general audios, which can be roughly categorized as verbal information and nonverbal information. The verbal information refers to the human speech in the video, which delivers linguistic semantics of the video. The nonverbal information refers to ambient sounds which can reflect natural events occurring in the video, such as raining. Due to the heterogeneity of these two types of information, the existing audio models usually focus on handling only one type. However, both types of information are indispensable for the comprehensive video understanding. It is naive and cumbersome to apply multiple audio backbones to encode the two types of audio information respectively.

To tackle the above two challenges, in this paper, we propose **CLIP4VLA** (**CLIP** for **V**ision, **L**anguage and **A**udio), which extends CLIP to accommodate the audio modality with unified tri-encoder structure for multimodal processing. Specifically, we employ the state-of-the-art VL model CLIP (Portillo-Quintero, Ortiz-Bayliss, and Terashima-Marín 2021) as the vision and text encoders, and propose an audio encoder with the same architecture as the vision encoder to ensure the training consistency and efficiency. To simultaneously encode both verbal information and nonverbal information from the audio track of videos, we design an audio type token to dynamically con-

*Corresponding Author.

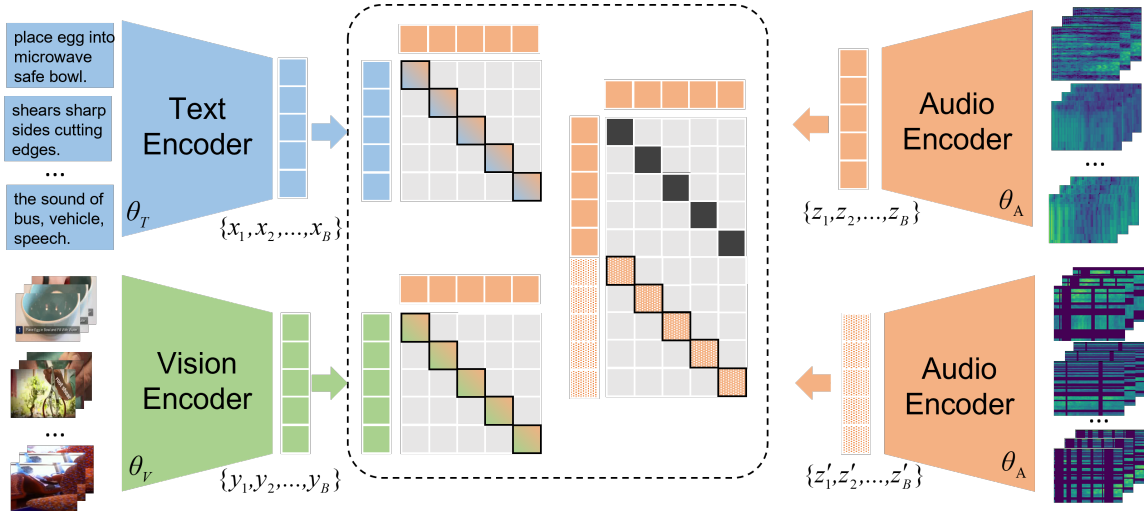

Figure 1: An overview of our CLIP4VLA model, which consists of three backbones: Text Encoder, Vision Encoder, and Audio Encoder. After encoding a batch of text features $\{x_1, x_2, \cdots, x_B\}$, vision features $\{y_1, y_2, \cdots, y_B\}$ and audio features $\{z_1, z_2, \cdots, z_B\}$, we pre-train the model with four kinds of contrastive learning objectives for text-audio, video-audio, augmented_audio-original_audio respectively. The black squares are not included in the calculation

trol the learned information type. During pre-training, we apply both inter-modal and intra-modal contrastive learning to learn the correlation across audio modality with other modalities and the inner characteristics of the audio modality. To better utilize the multimodal representations learned by CLIP4VLA, we further explore different modality fusion methods for video-text downstream tasks on various datasets. CLIP4VLA is demonstrated to be effective on both retrieval and captioning tasks, requiring much less hardware resource and training time.

Our contributions can be summarized as follows:

- We propose CLIP4VLA for learning correlation across textual, visual and audio information in videos by accommodating the audio encoder in CLIP.
- To fully exploit the rich audio information in videos, we propose to explicitly encode both verbal information and nonverbal information with audio type tokens.
- We design intra-modal and inter-modal contrastive learning for pre-training CLIP4VLA and explore multiple modality fusion methods for video downstream tasks.
- Our model achieves the state-of-the-art performance in retrieval and captioning tasks on the benchmark datasets of MSR-VTT, VATEX, and Audiocaps.

Related Work

Audio Pre-training

Audio pre-training works aim to well represent nonverbal information in ambient sound (Gemmeke et al. 2017; Gong, Chung, and Glass 2021; Guzhov et al. 2022, 2020; Wu et al. 2022) or verbal information in human speech (Liu, Li, and Lee 2021; Tang, Lei, and Bansal 2021; Chung et al. 2019; Baevski et al. 2020; Hsu et al. 2021; Chen et al. 2022) . For nonverbal information encoding, recent works (Guzhov et al. 2022, 2020; Wu et al. 2022; Gong, Chung, and Glass 2021) prove that audio representation learning can benefit from other modalities (i.e. images) by transfer learning. To encode verbal information, self-supervised methods are always utilized to learn inherent characteristic, ranging from auto-regressive learning (Chung et al. 2019; Liu, Chung, and Glass 2021; Liu, Li, and Lee 2021) to contrastive learning (Baevski et al. 2020; van den Oord, Li, and Vinyals 2018; Ling et al. 2020). Furthermore, wav2vec2.0 (Baevski et al. 2020), HuBERT (Hsu et al. 2021), WavLM (Chen et al. 2022) demonstrate that self-supervised learning with a large amount of unlabeled data could boost the model's performance on semantic related tasks (i.e. ASR) and decrease the demand of labeled data. Our CLIP4VLA has two changes compared with the previous audio pre-training works: Firstly, previous works mainly focus on audio encoding and ignore cross-modality understanding, while CLIP4VLA enhances audio representation by both self-supervised learning and cross-modal alignment. Secondly, previous works only focus on one specific type of audios while CLIP4VLA extract both verbal and nonverbal information for general video understanding.

Video-Text Pre-training

Most video-text pre-training works (Sun et al. 2019; Tang, Lei, and Bansal 2021; Xu et al. 2021; Lei et al. 2021; Sun et al. 2020; Zhu and Yang 2020; Luo et al. 2020) focus on the vision-text alignment in videos. VideoBERT (Sun et al. 2019) and CBT (Sun et al. 2020) are pioneering works to explore Video-Language representation by self-supervised learning. For fine-grained multimodal understanding, HERO (Li et al. 2020) designs a temporal-specific proxy task and UniVL (Luo et al. 2020) designs a generation proxy task. ClipBERT (Lei et al. 2021) further explores

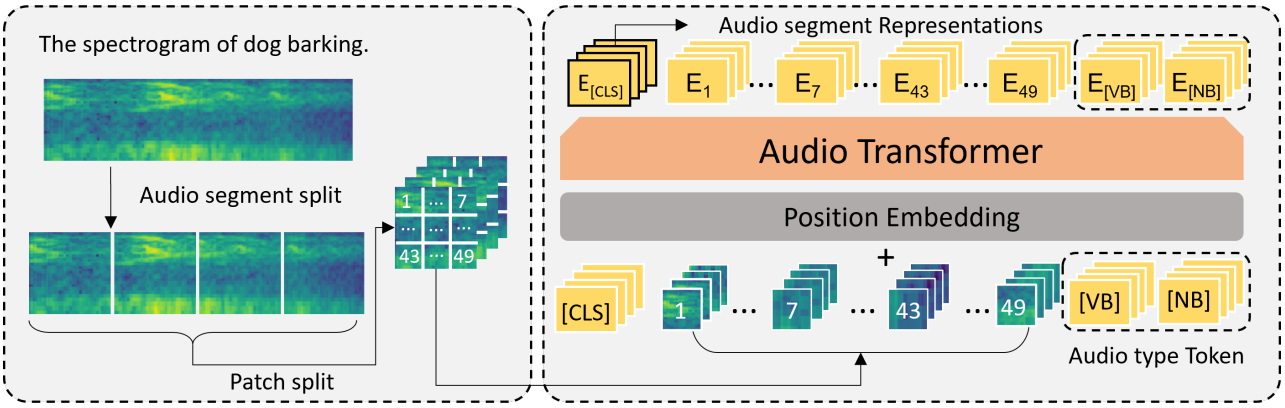

Figure 2: An overview of the audio encoding. The audio spectrogram is first split into segments along temporal dimension without overlap, the spectrogram of each audio segment is then split into a patch sequence of 7 × 7 without overlap. After flattening, A [CLS] token and an audio type embedding are added to the start and the end of the patch sequence respectively. Each patch embedding is added with a learnable positional embedding and then fed into the Audio encoder, which keeps the same structure as the visual encoder. The output of [CLS] is used as the final audio segment representation

an end-to-end manner by inputting sparse sampled frames from video clips rather than extracted video features from pre-trained backbones (Xie et al. 2018). These works well explore the correlation between vision and text modalities but ignore audio information in videos.

Recently, some works (Alayrac et al. 2020; Akbari et al. 2021; Liu et al. 2021) try to incorporate audio modality during pretraining for tri-modal understanding. OPT (Liu et al. 2021) focuses on the speech of image descriptions, which is greatly different from the audio in general videos. To encode general videos, VATT (Akbari et al. 2021) explores representing all three modalities with one modality-agnostic encoder. Our work also focuses on general videos for broader application and there are major two differences. Firstly, VATT is trained from scratch with heavy computation load while our model learns triple-modal correlation based on existing VL pe-trained model. Secondly, VATT does not distinguish verbal and nonverbal information in audios while CLIP4VLA respectively learns their correlation with other two modalities from different types of videos.

Method

In this section, we describe the proposed CLIP4VLA model and the multimodal contrastive learning objectives for pretraining in details. Given a batch of videos and their corresponding descriptions, we first extract audios from the videos and formulate the video batch, audio batch and text batch as V, A and T respectively. The target of our CLIP4VLA model is to learn rich semantic representations for the three modalities, so that the corresponding video, audio and text with similar semantics can be embedded close to each other though in different modalities, while those with different semantics be embedded further away.

With the multimodal representations fully learned, we adapt the model on different downstream tasks including cross-modal retrieval and multimodal captioning to verify the effectiveness of our CLIP4VLA model.

Model Structure

As illustrated in Figure 1, our proposed CLIP4VLA model consists of three backbones to handle the textual, visual and audio signals respectively. The details of the audio processing and audio backbone structure are illustrated in Figure 2.

Text & Vision Encoder. We employ CLIP (Portillo-Quintero, Ortiz-Bayliss, and Terashima-Marín 2021) as our text and vision encoders to encode text input T and vision input V. Each $t_i \in T$ is first tokenized into a token sequence and then added with a start token [SOS] and an end token [EOS], denoted as $\{t_i^1, t_i^2, \cdots, t_i^{L_T}\}$. After text encoding, outputs of each token are collected as word-level representations $\{x_i^1, x_i^2, \cdots, x_i^{L_T}\}$. Following CLIP, we choose the output of [EOS] token as the global text representation of t_i, denoted as x_i^g.

For visual information in videos, we uniformly sample L_V frames from $v_i \in V$ in the temporal dimension as the vision sequence $\{v_i^1, v_i^2, \cdots, v_i^{L_V}\}$. Specifically, each frame is split into a sequence of patches without overlap and then added with a [CLS] token. During vision encoding, patch sequence of each frame is independently fed into the vision encoder to model the spatial relationship between patches. The final output of the [CLS] token is chosen as the vision representation of each video frame. Finally, for the vision sequence $\{v_i^1, v_i^2, \cdots, v_i^{L_V}\}$, we acquire frame-level vision representations $y_i = \{y_i^1, y_i^2, \cdots, y_i^{L_V}\}$. By average pooling of y_i, we get a global vision embedding, denoted as y_i^g.

Audio Encoder Well-trained specialists could infer ambient events or human voice by watching spectrograms. Thus it is also possible for machine to encode audio information with visual spectrograms as inputs. To keep architecture consistency across different modalities, we design our audio encoder with the same model structure as the vision encoder. To process audios similarly as the visual signals, the first thing to do is to transfer the 1-dimensional long audio $a_i \in A$ into the image format, a matrix in the shape

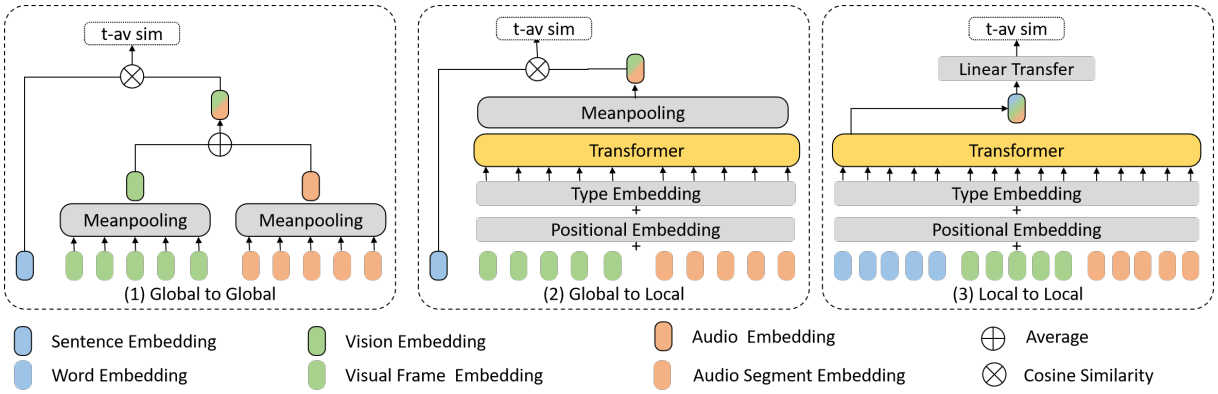

Figure 3: An overview of different modal fusion methods. We explore 3 methods for video retrieval, which includes (1) Global to Global, (2) Global to Local, (3) Local to Local

of 224×224×3 in this paper. To be specific, we convert the audio waveform into 224-dimensional log Mel filterbank (fbank) features with 32ms Hamming window every 8ms. In this way, a t-second audio stream will be transferred into a spectrogram in the shape of 125t × 224. We cut the spectrogram into k×224×224 along the temporal dimension without overlap, and pad with zeros if the last part is less than 224. Therefore, a t-second audio will be finally transferred into $\lceil \frac{t}{1.792} \rceil$ frames $\{a_i^1, a_i^2, \cdots, a_i^{L_A}\}$ with $a_i^j \in R^{224\times224\times3}$. Then the normalized segment frames can be encoded similarly with frame images. The final sequence of audio segment representations is denoted as $z_i = \{z_i^1, z_i^2, \cdots, z_i^{L_A}\}$. We also get a global audio embedding by average pooling of z_i, denoted as z_i^g.

Audio Type Token As we consider roughly two types of information in the audios of common videos (**Ver**bal information and **Nonver**bal information), we design audio type tokens to effectively control which type of features the audio encoder tends to generate. To be specific, after flattening patches of each audio segment, an audio type token [VB]/[NB] is added at the end of the patch sequence according to different application scenarios. For example, for dialogue or commentary, the audio type token [VB] could be used to encode the verbal information. While for natural activities/events, where the nonverbal information is more important, the [NB] token can be used as the control signal to extract the audio features from the nonverbal aspect. Furthermore, for the complex scenarios where both verbal information and nonverbal information are crucial, these two types of embeddings can also be combined for better video understanding. During pre-training, we set the audio type token according to the characteristics of audio pre-training datasets. During fine-tuning or testing, we add both audio type tokens at the end of the flatten patch sequences to flexibly extract both verbal and nonverbal information.

Pre-training

In this section, we introduce pre-training objectives of our CLIP4VLA model. To learn semantic representations of text, vision and audio, we explore contrastive learning from two perspectives: inter-modal and intra-modal. The inter-modal contrastive learning is designed to learn the correlation between audio modality and text/vision modality. The intra-modal contrastive learning aims to learn the inherent characteristics of the audio modality. We choose the NCE loss (Józefowicz et al. 2016) for both inter-modal and intra-modal contrastive learning.

Inter-modal Learning During inter-modal learning, which learns cross-modal alignments between text, vision and audio, positive pairs of cross-modal representations should be closer than negative ones. In this work, we construct negative pairs of cross-modal representations within a mini-batch. With global embeddings of text, vision and audio modalities, we compute the cosine similarity matrix in $B \times B$ for text-audio pairs and vision-audio pairs within a mini-batch, where B is the batch size. Since the vision and text encoders have been well pre-trained to learn vision-language alignment, we mainly train the audio encoder by maximizing the cosine similarity of B positive pairs while minimizing the cosine similarity of the $B^2 - B$ negative pairs. The symmetric cross entropy loss is calculated as follows:

$$\text{NCE}_{at} = \frac{1}{B}\sum_{i}^{B} \log \frac{\exp(z_i^g \cdot x_i^g)}{\sum_j^B \exp(z_i^g \cdot x_j^g)} \quad (1)$$

$$\text{NCE}_{av} = \frac{1}{B}\sum_{i}^{B} \log \frac{\exp(z_i^g \cdot y_i^g)}{\sum_j^B \exp(z_i^g \cdot y_j^g)} \quad (2)$$

where x^g, y^g, z^g refer to global embbeddings of text, vision and audio modalities.

Intra-modal Learning To enhance the information representation ability of audio encoder, we further optimize it with intra-modal self-supervised learning. We first augment the audio a_i to $\hat{a}_i$ by randomly masking the audio spectrograms along both channel and temporal dimension (Liu, Li, and Lee 2021). To be specific, we randomly sample the start step along the channel step and the time step with probability of 5% and 15% respectively, then we mask the subsequent 10 consecutive steps from the start step. Overlap is allowed in the masking. The original audio a_i and its augmented version $\hat{a}_i$ then can be seen as a positive pair for the contrastive

Model	MSR-VTT				VATEX			
	R@1	R@5	R@10	MedianR	R@1	R@5	R@10	MedianR
W2VV++ (Li et al. 2019)	18.9	45.3	57.5	-	34.3	73.6	83.7	-
CE (Liu et al. 2019)	20.9	48.8	62.4	5.0	47.9	84.2	91.3	2.0
MMT (Gabeur et al. 2020)	26.6	57.1	69.6	-	-	-	-	-
HGR (Chen et al. 2020)	-	-	-	-	35.1	73.5	83.5	2.0
SSB (Patrick et al. 2021)	30.1	58.5	69.3	3.0	45.9	82.4	90.4	1.0
UniVL (Luo et al. 2020)	20.6	49.1	62.9	6.0	-	-	-	-
ClipBERT (Lei et al. 2021)	22.0	46.8	59.9	6.0	-	-	-	-
VLM (Xu et al. 2021)	28.1	55.5	57.4	4.0	-	-	-	-
CLIP	31.2	53.7	64.2	4.0	39.7	72.3	82.2	-
CLIP-FRL (Chen et al. 2021)	38.2	66.0	75.7	-	47.1	82.3	90.6	-
CLIP4Clip (Luo et al. 2021)	44.5	71.4	81.6	2.0	55.9	89.2	95.0	1.0
CLIP2Video (Fang et al. 2021)	45.6	72.5	81.7	2.0	57.3	90.0	95.5	1.0
CLIP4VLA	**46.2**	**73.5**	**83.5**	**2.0**	**63.5**	**91.5**	**95.9**	**1.0**

Table 1: Video Retrieval Performance on MSR-VTT-1kA and VATEX

Model	Modality	R@1	R@5	R@10	MedianR
VGGish	A	18.5	-	62.0	-
VGGSound	A	22.4	-	69.2	-
MoEE	A	22.5	-	69.5	-
CE	A	23.1	56.2	70.7	4.0
CLIP4VLA	A	**28.4**	**60.9**	**76.2**	4.0
CE	AV	28.0	-	80.4	-
CLIP4VLA	AV	**33.6**	**68.1**	**82.3**	3.0

Table 2: Retrieval Performance Comparison on Audiocaps

learning. Similar to the inter-modal NCE loss, other masked audios within a mini-batch are negative samples for a_i. The symmetric cross entropy loss is calculated as follows:

$$\mathrm{NCE}_{a\hat{a}} = \frac{1}{B}\sum_{i}^{B}\log\frac{\exp(z_i^g \cdot \hat{z_i}^g)}{S} \tag{3}$$

$$S = \sum_{j}^{B}\exp(z_i^g \cdot \hat{z_j}^g) + \sum_{k\neq i}^{B}\exp(z_i^g \cdot {z_k}^g) \tag{4}$$

where $\hat{z_i}^g$ is the global embedding of masked audio $\hat{a_i}$.

The final pre-training loss for CLIP4VLA is the sum of inter-modal NCE and intra-modal NCE objectives:

$$\mathcal{L} = \mathrm{NCE}_{at} + \mathrm{NCE}_{av} + \mathrm{NCE}_{a\hat{a}} \tag{5}$$

Fine-tuning

To verify the effectiveness of the learned representations for text, vision and audio, we fine-tune the CLIP4VLA model for multiple downstream tasks.

Fine-tuning for Video Retrieval Video Retrieval aims to search the target video based on a video caption as the retrieval query. Without encoding audio information, existing video retrieval works (Liu et al. 2019; Miech, Laptev, and Sivic 2018; Chen et al. 2020) only focus on the matching between text and vision modality. Benefiting from the tri-modality encoding ability of CLIP4VLA, we fully explore both vision and audio information in the video for text-video retrieval. Since there are three modalities involved in this task, effective multimodal fusion is important. In this paper, we explore three multimodal fusion approaches for text-video retrieval, including (1) *Global to Global*, (2) *Global to Local*, and (3) *Local to Local*. As illustrated in Figure 3, the *Global to Global* approach directly calculates similarity based on the global embeddings of vision and audio modalities via mean pooling. For the *Global to Local* approach, we apply a Video Temporal Encoding Module (N-layer transformer) to encode temporal relevance of vision and audio modalities, and calculate the similarity between text feature and fused video feature. For the *Local to Local* approach, we apply a Fine-grained Cross-modality Fusion Module (N-layer transformer) to further exploit the fine-grained correlation of text to vision and audio modalities. We analyze these multimodal fusion methods in the supplementary material.

Fine-tuning for Video Captioning Besides the video retrieval task, Video Captioning (Zhang et al. 2020; Lin, Gan, and Wang 2021; Wang et al. 2022) is another challenging task on video understanding, which aims to generate fluent natural language description of video contents. To conduct sentence generation, we introduce a Multimodal Caption Generator (N-layer transformer encoder) upon CLIP4VLA. At the t^{th} decoding step, we feed previous generated words, vision frames and audio segments into CLIP4VLA. After intra-model encoding with three encoders, we concatenate the fine-grained features to construct multimodal sequence U_i. Input the sequence into Multimodal Caption Generator, the t^{th} word is predicted as follows:

$$H_i = \mathrm{MCG}(U_i), \tag{6}$$

$$p_i^t = \mathrm{softmax}(f(h_i^t)), \quad h_i^t \in H_i, \tag{7}$$

where MCG refers to the Multimodal Caption Generator, $f(\cdot)$ is the linear output layer, p_i^t is the predicted probabilities over the whole vocabulary size.

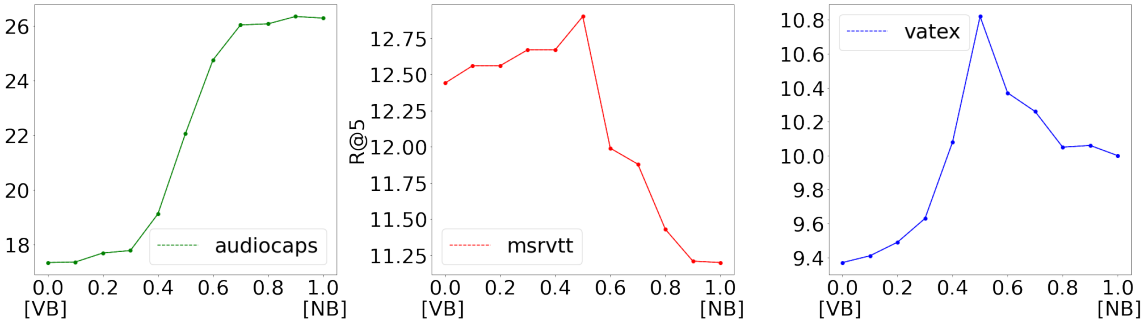

Figure 4: The zero-shot text-audio retrieval performance on different datasets with different audio type token. The x-axis represents the mixing ratio of [VB] and [NB] type embeddings, and the y-axis represents the retrieval performance

Model	BLUE4	METEOR	ROUGE	CIDER
ORG-TRAL	43.6	28.8	62.1	50.9
SemSynAN	44.3	28.8	62.5	50.1
APML	41.9	29.9	62.6	49.8
UniVL	42.2	28.8	61.2	49.9
CMG	43.7	29.4	62.8	55.9
Clip4Caption	46.1	30.7	**64.8**	57.7
CLIP4VLA	**46.7**	**31.1**	64.4	**58.0**

Table 3: Captioning Performance on MSR-VTT

Model	BLUE4	METEOR	ROUGE	CIDER
Shared E	28.4	21.7	47.0	45.1
Shared E-D	27.9	21.6	46.8	44.2
ORG-TRAL	32.1	22.2	48.9	49.7
SCST-C-B-F	33.3	22.8	49.6	54.6
CLIP4VLA	**36.4**	**25.0**	**54.7**	**59.7**

Table 4: Captioning Performance on VATEX

Experiments

Experiment Settings

We first pre-train our proposed CLIP4VLA on large scale datasets including Howto100M (Miech et al. 2019) and Audioset (Gemmeke et al. 2017), then fine-tune it for the retrieval and captioning tasks on three datasets: MSR-VTT (Xu et al. 2016), VATEX (Wang et al. 2019), Audiocaps (Kim et al. 2019). The evaluation metrics are Recall@n (R@n) and Median R for retrieval tasks, and BLUE-n, METEOR, ROUGE, CIDER for captioning tasks.

Pre-training Datasets Our pre-training data includes instructional video dataset Howto100M (Miech et al. 2019) and event video dataset Audioset (Gemmeke et al. 2017). To enable our audio encoder to distinguish verbal and nonverbal audio information, we choose [VB] as the audio type token for Howto100M and [NB] for Audioset, respectively. More details of data processing can be found in the supplementary material.

Fine-tuning Datasets We evaluate the pre-trained CLIP4VLA on retrieval and captioning benchmarks, including MSR-VTT (Xu et al. 2016), VATEX (Wang et al. 2019) and Audiocaps (Kim et al. 2019). After filtering out the silent videos, MSR-VTT remains 7867 and 884 videos for training and testing on the retrieval task, and 5867, 448, and 2617 videos for training, validation, and testing on the captioning task. VATEX remains 24667, 1427, and 1421 videos for training, validation, and testing on retrieval, and 24667, 2845, and 5698 videos for training, validation, and testing on captioning. Audiocaps keeps 49712, 495, and 967 videos for training, validation, and testing on the retrieval task. For training cost comparison with previous work, we further measure our model on event classification datasets of UCF101 (Soomro, Zamir, and Shah 2012) and ESC50 (Piczak 2015). The former one contains 13K videos of 101 action classes, and the latter one contains 2K audio clips of 50 classes.

Comparison with the State-of-the-arts

Video Retrieval To demonstrate the effectiveness of our proposed CLIP4VLA, we first evaluate it for video retrieval on three benchmarks. Baselines could be grouped into three categories, corresponding to the three blocks in the table: 1) *Classical Retrieval Methods*: W2VV++ (Li et al. 2019), CE (Liu et al. 2019), HGR (Chen et al. 2020), MMT (Gabeur et al. 2020), SSB (Patrick et al. 2021), MoEE (Miech, Laptev, and Sivic 2018); 2) *Pre-training based Methods*: UniVL (Luo et al. 2020), ClipBERT (Lei et al. 2021), VLM (Xu et al. 2021); 3) *CLIP-based Methods*: CLIP (Portillo-Quintero, Ortiz-Bayliss, and Terashima-Marín 2021), CLIP-FRL (Chen et al. 2021), CLIP4Clip (Luo et al. 2021), CLIP2Video (Fang et al. 2021). Our model understands videos according to both vision and audio information. However, on visual-centric video datasets MSR-VTT and VATEX, not all videos contain audio information. To deal with the missing modality problem for each silent video during testing, we pair it with the audio of the most similar video, which is chosen from corresponding training set according to cosine similarity of global vision features. As shown in Table 1, firstly, our model CLIP4VLA achieves state-of-the-art performance on both MSR-VTT and VATEX datasets. Secondly, CLIP-based methods significantly outperform other baselines, which shows video understanding can benefit a lot from large-scale image-text pre-training.

	Components	MSR-VTT				VATEX			
		R@1	R@5	R@10	MedianR	R@1	R@5	R@10	MedianR
1	Scratch	2.5	7.8	11.1	124.0	1.9	6.0	9.4	156.0
2	+Initial	3.6	11.6	16.5	136.5	3.7	12.5	19.3	72.0
3	+Inter-modal NCE	6.8	15.4	21.6	81.5	7.0	19.7	27.5	40.0
4	+Intra-modal NCE	10.5	25.4	37.5	21.0	9.3	24.9	34.2	**26.0**
5	+Audio Type Token	**10.6**	**26.5**	**38.0**	**19.0**	**9.9**	**25.4**	**34.3**	29.0

Table 5: Impact of key components for audio retrieval on MSR-VTT & VATEX

Model	Training Cost	Batch Size	Training Param	ESC50(A)	UCF101(V& A)
VATT-Medium	512∼768 TPU days	2048	264M	84.7	89.6
CLIP4VLA	48 V100 days	256	88M	86.8	91.9

Table 6: The Comparison of Training Cost and event classification performance of VATT and CLIP4VLA on ESC50 and UCF101 (audio features of ESC50, vision and audio features of UCF101)

Thirdly, with audio content as extra input, our CLIP4VLA achieves better performance than other CLIP-based methods. This indicates that our model could well encode the correlation across text, vision and audio modality.

Besides visual-centric datasets, we also evaluate our model on audio-centric video dataset Audiocaps. As shown in Table 2, either with only audio representations or both audio and vision representations of videos, our model achieves state-of-the-art video retrieval performance on Audiocaps. What's more, CLIP4VLA with both audio and vision information outperforms the one with only audio information. This indicates that our model could better understand audio-centric videos by leveraging vision information.

Video Captioning We further validate the adaptability of CLIP4VLA to video captioning task on MSRVTT and VATEXT. As shown in Table 3 and Table 4, our model achieves state-of-the-art captioning performance on both datasets as well. This indicates that our model also possesses good caption generation capability by leveraging well-aligned multimodal representations.

Ablation Study

Audio Type Token To verify the validity of our proposed audio type token for different kinds of audio information encoding, we conduct the experiment to compare the audio retrieval performance when adjusting the mixing ratio of the two type embeddings of [NB] and [VB]. As shown in Figure 4, with the mixing ratio of [NB] embedding increased, the audio retrieval result on the Audiocaps dataset is significantly improved, because most of the audios in Audiocaps dataset are ambient sound. However, for the video datasets MSR-VTT and VATEX, the best results are yielded when the [NB] and [VB] embeddings are mixed with a ratio of 1:1, which further demonstrates that videos usually contain complex audios with both verbal and nonverbal information, while previous multimodal pre-training works have not specifically considered handling them simultaneously. The results on the three datasets show that our audio type token can effectively control the information aspect of encoded audio features for different application scenarios.

Key Components Table 5 ablates the contributions from key components of our model. The text-audio retrieval results on MSR-VTT and VATEX datasets consistently demonstrate the effectiveness of each proposed component. Especially, compared with row 1, directly initializing the audio backbone with vision backbone (row 2) has brought obvious gains, which further demonstrates that the audio information learning can benefit from existing visual knowledge.

Training Cost Fully exploiting the existing vision-text knowledge for audio pre-training can not only help the audio representation learning, but also reduce the training cost. In this section we compare the training cost and the classification performance on ECS50 and UCF101 with VATT (Akbari et al. 2021), which is a vision-text-audio model pre-trained from scratch. For fair comparison, we follow the VATT to train a linear classifier on top of the frozen multimodal backbones, and report the mean accuracy over official splits (5-fold and 3-fold cross validation for ESC50 and UCF101 respectively) . As the results shown in Table 6, our CLIP4VLA model achieves better downstream results with much less training cost, which demonstrates the advantages of learning audio from the existing visual-text knowledge.

Conclusion

We propose CLIP4VLA for Vision-Language-Audio processing by extending the VL pre-training model CLIP to accommodate the audio modality in a unified and economic way, which incorporates an audio encoder with the same structure as the vision backbone for training consistency and efficiency. To take full advantage of multimodal training data, we propose the contrastive learning from both inter- and intra-modal perspectives. Considering both verbal information and nonverbal information contained in general audios, we further propose an audio type token to explicitly encode these two types of information. CLIP4VLA is validated by the video retrieval and video captioning tasks on MSR-VTT, VATEX, and Audiocaps benchmark datasets and achieves the state-of-the-art performance.

Acknowledgments

This work was partially supported by National Key R&D Program of China (No. 2020AAA0108600) and National Natural Science Foundation of China (No. 62072462).

References

Akbari, H.; Yuan, L.; Qian, R.; Chuang, W.; Chang, S.; Cui, Y.; and Gong, B. 2021. VATT: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text. In *NeurIPS*.

Alayrac, J.; Recasens, A.; Schneider, R.; Arandjelovic, R.; Ramapuram, J.; Fauw, J. D.; Smaira, L.; Dieleman, S.; and Zisserman, A. 2020. Self-Supervised MultiModal Versatile Networks. In *NeurIPS*.

Baevski, A.; Zhou, Y.; Mohamed, A.; and Auli, M. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *NeurIPS*.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-End Object Detection with Transformers. In *ECCV*.

Chen, A.; Hu, F.; Wang, Z.; Zhou, F.; and Li, X. 2021. What Matters for Ad-hoc Video Search? A Large-scale Evaluation on TRECVID. In *ICCV*.

Chen, S.; Wang, C.; Chen, Z.; Wu, Y.; Liu, S.; Chen, Z.; Li, J.; Kanda, N.; Yoshioka, T.; Xiao, X.; Wu, J.; Zhou, L.; Ren, S.; Qian, Y.; Qian, Y.; Wu, J.; Zeng, M.; and Wei, F. 2022. WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. *IEEE J. Sel. Top. Signal Process.*

Chen, S.; Zhao, Y.; Jin, Q.; and Wu, Q. 2020. Fine-Grained Video-Text Retrieval With Hierarchical Graph Reasoning. In *CVPR*.

Chung, Y.; Hsu, W.; Tang, H.; and Glass, J. R. 2019. An Unsupervised Autoregressive Model for Speech Representation Learning. In *Interspeech*.

Fang, H.; Xiong, P.; Xu, L.; and Chen, Y. 2021. CLIP2Video: Mastering Video-Text Retrieval via Image CLIP. *CoRR*.

Gabeur, V.; Sun, C.; Alahari, K.; and Schmid, C. 2020. Multi-modal Transformer for Video Retrieval. In *ECCV*.

Gemmeke, J. F.; Ellis, D. P. W.; Freedman, D.; Jansen, A.; Lawrence, W.; Moore, R. C.; Plakal, M.; and Ritter, M. 2017. Audio Set: An ontology and human-labeled dataset for audio events. In *ICASSP*.

Gong, Y.; Chung, Y.; and Glass, J. R. 2021. AST: Audio Spectrogram Transformer. In *Interspeech*.

Guzhov, A.; Raue, F.; Hees, J.; and Dengel, A. 2020. ESResNet: Environmental Sound Classification Based on Visual Domain Models. In *ICPR*.

Guzhov, A.; Raue, F.; Hees, J.; and Dengel, A. 2022. Audioclip: Extending Clip to Image, Text and Audio. In *ICASSP*.

Hsu, W.; Bolte, B.; Tsai, Y. H.; Lakhotia, K.; Salakhutdinov, R.; and Mohamed, A. 2021. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE ACM Trans. Audio Speech Lang. Process.*

Józefowicz, R.; Vinyals, O.; Schuster, M.; Shazeer, N.; and Wu, Y. 2016. Exploring the Limits of Language Modeling. *CoRR*.

Kim, C. D.; Kim, B.; Lee, H.; and Kim, G. 2019. AudioCaps: Generating Captions for Audios in The Wild. In *NAACL-HLT*.

Lei, J.; Li, L.; Zhou, L.; Gan, Z.; Berg, T. L.; Bansal, M.; and Liu, J. 2021. Less is More: ClipBERT for Video-and-Language Learning via Sparse Sampling. In *CVPR*.

Li, L.; Chen, Y.; Cheng, Y.; Gan, Z.; Yu, L.; and Liu, J. 2020. HERO: Hierarchical Encoder for Video+Language Omni-representation Pre-training. In *EMNLP*.

Li, X.; Xu, C.; Yang, G.; Chen, Z.; and Dong, J. 2019. W2VV++: Fully Deep Learning for Ad-hoc Video Search. In *ACM MM*.

Lin, K.; Gan, Z.; and Wang, L. 2021. Augmented Partial Mutual Learning with Frame Masking for Video Captioning. In *AAAI*.

Ling, S.; Liu, Y.; Salazar, J.; and Kirchhoff, K. 2020. Deep Contextualized Acoustic Representations for Semi-Supervised Speech Recognition. In *ICASSP*.

Liu, A. H.; Chung, Y.; and Glass, J. R. 2021. Non-Autoregressive Predictive Coding for Learning Speech Representations from Local Dependencies. In *Interspeech*.

Liu, A. T.; Li, S.; and Lee, H. 2021. TERA: Self-Supervised Learning of Transformer Encoder Representation for Speech. *IEEE ACM Trans. Audio Speech Lang. Process.*

Liu, J.; Zhu, X.; Liu, F.; Guo, L.; Zhao, Z.; Sun, M.; Wang, W.; Lu, H.; Zhou, S.; Zhang, J.; and Wang, J. 2021. OPT: Omni-Perception Pre-Trainer for Cross-Modal Understanding and Generation. *CoRR*.

Liu, Y.; Albanie, S.; Nagrani, A.; and Zisserman, A. 2019. Use What You Have: Video retrieval using representations from collaborative experts. In *BMVC*.

Luo, H.; Ji, L.; Shi, B.; Huang, H.; Duan, N.; Li, T.; Li, J.; Bharti, T.; and Zhou, M. 2020. UniVL: A Unified Video and Language Pre-Training Model for Multimodal Understanding and Generation. arXiv:2002.06353.

Luo, H.; Ji, L.; Zhong, M.; Chen, Y.; Lei, W.; Duan, N.; and Li, T. 2021. CLIP4Clip: An Empirical Study of CLIP for End to End Video Clip Retrieval. *CoRR*.

Miech, A.; Laptev, I.; and Sivic, J. 2018. Learning a Text-Video Embedding from Incomplete and Heterogeneous Data. *CoRR*.

Miech, A.; Zhukov, D.; Alayrac, J.; Tapaswi, M.; Laptev, I.; and Sivic, J. 2019. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *ICCV*.

Patrick, M.; Huang, P.; Asano, Y. M.; Metze, F.; Hauptmann, A. G.; Henriques, J. F.; and Vedaldi, A. 2021. Support-set bottlenecks for video-text representation learning. In *ICLR*.

Piczak, K. J. 2015. ESC: Dataset for Environmental Sound Classification. In *ACM MM*.

Portillo-Quintero, J. A.; Ortiz-Bayliss, J. C.; and Terashima-Marín, H. 2021. A Straightforward Framework for Video Retrieval Using CLIP. In *MCPR*.

Soomro, K.; Zamir, A. R.; and Shah, M. 2012. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. *CoRR*.

Sun, C.; Baradel, F.; Murphy, K.; and Schmid, C. 2020. Learning Video Representations using Contrastive Bidirectional Transformer. In *ECCV*.

Sun, C.; Myers, A.; Vondrick, C.; Murphy, K.; and Schmid, C. 2019. VideoBERT: A Joint Model for Video and Language Representation Learning. In *ICCV*.

Tang, Z.; Lei, J.; and Bansal, M. 2021. DeCEMBERT: Learning from Noisy Instructional Videos via Dense Captions and Entropy Minimization. In *NAACL-HLT*.

van den Oord, A.; Li, Y.; and Vinyals, O. 2018. Representation Learning with Contrastive Predictive Coding. *CoRR*.

Wang, H.; Lin, G.; Hoi, S. C. H.; and Miao, C. 2022. Cross-Modal Graph With Meta Concepts for Video Captioning. *IEEE Trans. Image Process.*

Wang, X.; Wu, J.; Chen, J.; Li, L.; Wang, Y.; and Wang, W. Y. 2019. VaTeX: A Large-Scale, High-Quality Multilingual Dataset for Video-and-Language Research. In *ICCV*.

Wu, H.; Seetharaman, P.; Kumar, K.; and Bello, J. P. 2022. Wav2CLIP: Learning Robust Audio Representations from Clip. In *ICASSP*.

Xie, S.; Sun, C.; Huang, J.; Tu, Z.; and Murphy, K. 2018. Rethinking Spatiotemporal Feature Learning For Video Understanding. In *ECCV*.

Xu, H.; Ghosh, G.; Huang, P.; Arora, P.; Aminzadeh, M.; Feichtenhofer, C.; Metze, F.; and Zettlemoyer, L. 2021. VLM: Task-agnostic Video-Language Model Pre-training for Video Understanding. In *ACL*.

Xu, J.; Mei, T.; Yao, T.; and Rui, Y. 2016. MSR-VTT: A Large Video Description Dataset for Bridging Video and Language. In *CVPR*.

Zhang, Z.; Shi, Y.; Yuan, C.; Li, B.; Wang, P.; Hu, W.; and Zha, Z. 2020. Object Relational Graph With Teacher-Recommended Learning for Video Captioning. In *CVPR*.

Zhu, L.; and Yang, Y. 2020. ActBERT: Learning Global-Local Video-Text Representations. In *CVPR*.