i-Code: An Integrative and Composable Multimodal Learning Framework

Ziyi Yang*, Yuwei Fang*, Chenguang Zhu, Reid Pryzant, Dongdong Chen, Yu Shi, Yichong Xu, Yao Qian, Mei Gao, Yi-Ling Chen, Liyang Lu, Yujia Xie, Robert Gmyr, Noel Codella, Naoyuki Kanda, Bin Xiao, Lu Yuan, Takuya Yoshioka, Michael Zeng, Xuedong Huang

Microsoft Azure Cognitive Services Research
{ziyiyang, yuwfan, chezhu}@microsoft.com

Abstract

Human intelligence is multimodal; we integrate visual, linguistic, and acoustic signals to maintain a holistic worldview. Most current pretraining methods, however, are limited to one or two modalities. We present i-Code, a self-supervised pretraining framework where users may flexibly combine the modalities of vision, speech, and language into unified and general-purpose vector representations. In this framework, data from each modality are first given to pretrained single-modality encoders. The encoder outputs are then integrated with a multimodal fusion network, which uses novel merge- and co-attention mechanisms to effectively combine information from the different modalities. The entire system is pretrained end-to-end with new objectives including masked modality unit modeling and cross-modality contrastive learning. Unlike previous research using only video for pretraining, the i-Code framework can dynamically process single, dual, and triple-modality data during training and inference, flexibly projecting different combinations of modalities into a single representation space. Experimental results demonstrate how i-Code can outperform state-of-the-art techniques on five multimodal understanding tasks and single-modality benchmarks, improving by as much as 11% and demonstrating the power of integrative multimodal pretraining.

1 Introduction

True humanlike intelligence incorporates information from a variety of signals and sensory organs (Schank and Abelson 1975). This implies that intelligent systems should be integrative, incorporating signals from all available modalities. In many practical data regimes this corresponds to the modalities of vision (V), language (L), and speech/audio (S). Although there has been tremendous progress in making models to understand one modality (Devlin et al. 2019; Hsu et al. 2021; Chen et al. 2022a) or two modalities (Su et al. 2019; Lu et al. 2019; Li et al. 2019; Radford et al. 2021; Jia et al. 2021; Yuan et al. 2021) through self-supervised and semi-supervised pretraining, it is a non-trivial task to extend these successes to a three-modality system which can simultaneously interpret vision (V), language (L) and speech (S).

A few previous and concurrent attempts are made for the three-modality pretraining (Akbari et al. 2021; Zellers et al. 2022), however, one important difficulty is the pretraining requires enormous amounts of three-modality data like captioned videos, which is often several orders of magnitude smaller than the available single- or dual-modality data. E.g., the largest annotated video dataset at the time of writing consists of 180M clips (Zellers et al. 2021), while the largest available image-caption dataset has 900M pairs (Yuan et al. 2021). Another challenge is to effectively featurize raw signals from different modalities and allow them to interact and fuse with one another, in an end-to-end framework.

To address these problems, we propose two solutions. First, in addition to three-modality videos, we leverage large-scale dual-modality data, e.g., images with captions (V+L), speech with transcripts (S+L) and video narrations (V+S). This greatly expands the size and diversity of pretraining data while covering all three target modalities. Second, instead of building a standalone model from scratch, we leverage state-of-the-art pretrained single-modality encoders to generate powerful representations from raw signals. We propose a fusing architecture that can integrate the outputs of the single-modality encoders and conducts cross-modality understanding to get a final prediction. To design the best fusing architecture, we experiment with variations on the self-attention mechanism inside the transformer architecture, including mechanisms that merge the attention scores of different modalities. To this end, we propose **i-Code**, an integrative and composable multimodal learning framework, where "i" stands for the multimodal integration.

i-Code is pretrained on double and triple-modality data using various self-supervision objectives, including: i) masked unit modeling, where all input signals are converted into discrete tokens, and the goal is to predict the tokens for the masked units of each modality; ii) contrastive learning, where two input modalities are provided and the model predicts whether the signals come from the same triple (or pair) in the training data. We evaluate i-Code on diverse multimodal and single-modal benchmarks. Experimental results demonstrate the effectiveness of the proposed multimodal pretraining framework: i-Code outperforms state-of-the-art algorithms across five multimodal datasets, as well as single modality tasks such as the GLUE NLP benchmark, improving over the previous best by as much as 11%.

Our novelties include: (1) To the best of our knowledge, i-Code is the first work using dual-modality data for

*Co-first authors.

three-modality pretraining. Using pretrained encoders for all three modalities and the flexibility of switching module also make i-Code unique of its kind. (2) How to effectively fuse three modalities was unclear in the field of language-speech-vision modeling. We systematically and thoroughly investigate several fusion mechanisms, i.e., co-attention, merge attention (Section 4.1) and Mixture of Experts (Section 6). (3) We demonstrate the effectiveness of unified masked units modeling for vision, language and speech (Section 4.2). (4) i-Code can dynamically switch between single-modality encoders, which is a previously unexplored topic in multimodal pretraining and makes i-Code a more agile and composable framework (Section 6). (5) In previous multimodal models, performance of uni-modal encoders deteriorated after multimodal training. Our multimodal framework can improve uni-modal encoders performance (Section 5.4).

2 Related Work

Jointly co-learning vision and language modalities is an active area of multimodal research. One category follows a two-tower architecture with independent encoders for two modalities (Radford et al. 2021; Jia et al. 2021; Yuan et al. 2021; Chung, Zhu, and Zeng 2021; Alayrac et al. 2022). Multimodality fusion is achieved via a projection layer which is added to the single-modality encoder. These models are pretrained on image-caption or speech-text data with contrastive learning loss. These models show outstanding cross-modality retrieval performance along with zero and few-shot prediction performance. Another body of research seeks to achieve cross-modality interaction via a shared encoder (Dosovitskiy et al. 2020; Lu et al. 2019; Su et al. 2019; Li et al. 2019). For example, in VL-BERT (Su et al. 2019), vision features and language token embeddings are inputted together into the encoder, and the task is to predict the masked tokens based on the detection features and language contexts. Video-language learning has also been an active research area (Zellers et al. 2021; Tang, Lei, and Bansal 2021; Miech et al. 2019; Xu et al. 2021), where models are trained on videos frames and automatic speech recognition transcripts.

One fundamental question in deep learning is can we develop one unified model to learn vision, language and speech modalities altogether (Kaiser et al. 2017; Baevski et al. 2022). To solve this question, there has been increasing research on modeling the multimodal components of video data: textual transcripts, video frames, and audio waveform, etc (Zellers et al. 2022; Akbari et al. 2021; Alayrac et al. 2020). E.g., VATT (Akbari et al. 2021) builds a transformer encoder on top of projections of raw input data from videos (3D RGB voxels, waveforms, and token embeddings) and does not use single-modality encoder. We demonstrate that leveraging state-of-the-art single modality encoders for multimodal learning can effectively boost the multimodal model performance. The i-Code framework also extends the pretraining data from videos to dual-modality data.

3 Large-Scale Multimodal Pretraining Data

To facilitate effective multimodal learning, we collect large-scale multimodal data for pretraining. We collect two types of data: three-modality video and dual-modality datasets.

Video is a large-scale data resource that contains all three modalities and is widely available on public streaming platforms. We choose the recently published video dataset YT-Temporal-180M (Zellers et al. 2021) because of its great diversity in video corpus topics, high quality filtering and selection, and large-scale quantity. We collect 180 million video clips in the YT-Temporal-180M dataset, using the provided videos IDs and time stamps. For each clip, we evenly sample 8 video frames as visual inputs. For speech, the raw waveforms of audios are extracted to be further processed by the downstream speech encoder. Each clip also comes with a textual script that has been carefully denoised from the original ASR transcripts. However, misalignment between the frames and transcripts is a concerning and common issue in video data (Tang, Lei, and Bansal 2021; Miech et al. 2019): narration transcripts can be irrelevant or temporally misaligned with the visual frames. To alleviate this issue, we generate the caption for the high-resolution mid-frame of each clip with the captioning API of Azure Cognitive Services, to augment the video dataset. More details on how we leverage the captions can be found in Section 4.2. 54k clips are held out as the validation set.

As high-quality three-modality videos are limited in size, we also resort to dual-modality datasets for pretraining, which have been widely used in dual-modality learning such as visual-language representation learning (Radford et al. 2021; Jia et al. 2021; Yuan et al. 2021), zero-shot cross-modality generation (Ramesh et al. 2021), automatic speech recognition (ASR).. i-Code leverages the following dual-modality datasets during pretraining:

1. Visual-Language. We use 72.8 million image-caption pairs from the pretraining data of the Florence computer vision foundation model (Yuan et al. 2021). Data are collected with a programmatic data curation pipeline from the Internet, then selected and post-filtered (Yuan et al. 2021). 25K pairs are held out as the validation set.

2. Language-Speech. We use internal 75k-hour transcribed English speech data. This dataset, containing 63.2 million transcript-speech pairs, is diverse in scenarios, including Cortana, far-field speech, and call center. Again, 25k pairs are kept as the validation set.

3. Visual-Speech. For visual and speech pair datasets, we leverage Spoken Moments in Time (SMiT), a video-narration dataset. SMiT comprises 500k spoken captions each of which depicts a broad range of different events in a short video (Monfort et al. 2021). The validation split contains 5k examples.

To the best of our knowledge, this is the first time that paired datasets have been used to train vision-language-speech models. In the experiment section, we compare the performance of models pretrained with paired and video datasets, respectively. To balance between dual-modality datasets with difference sizes, we perform Exponentially

Smoothed Weighting[1] when sampling from difference data resources. We discover that combining both types of datasets can further boost the model's performance.

4 The i-Code Multimodal Framework

In this section, we introduce the overall model architecture of i-Code and how we pretrain i-Code on the aforementioned large-scale multimodal datasets in a self-supervised manner.

4.1 Model Architecture

i-Code has three single-modality encoder and a multimodal fusion module (Figure 1). The raw input for each modality is fed into its corresponding single-modality encoder, then all encoded inputs are fed through a linear projection layer and integrated with the modality fusion network. Due to this architecture design, i-Code can process various kinds of inputs: single-modality inputs, any combination of two modalities, and all three modalities together.

Instead of training each single-modality encoder from scratch, we design our framework to be modular: any pretrained model can be swapped in to fill the role of a single-modality encoder. This provides the fusion network with high-quality contextual representations for more effective multimodal understanding. We opt to leverage state-of-the-art models for each modality:

Language Encoder. We use the recently published DeBERTa V3 base (He et al. 2020) as the language encoder. This pretrained language model with 183 million parameters has a disentangled attention mechanism that has helped it achieve record-breaking performance on the GLUE and SuperGLUE NLP benchmarks.

Vision Encoder. We adopt CoSwin transformer (Yuan et al. 2021) as the vision encoder. To enable i-Code to process both images and sequence of frames (video), we instantiate a video CoSwin transformer from a pretrained CoSwin transformer (Yuan et al. 2021) as the vision encoder, following the procedure in Liu et al. (2022). The video CoSwin transformer has 91 million parameters.

Speech Encoder. Recently, there has been significant advances in learning speech representations through diverse network architectures (Schneider et al. 2019; Baevski et al. 2020; Hsu et al. 2021; Chen et al. 2022a). To leverage these state-of-the-art techniques in speech representation learning, we use a pretrained WavLM-large model (Chen et al. 2022a) of 315 million parameters as the speech encoder. WavLM contains a temporal convolutional encoder to featurize the input speech waveform, followed by a transformer encoder.

Note other encoders can also be used besides these three mentioned above. For example, we experimented with another speech encoder HuBERT and include the results in Appendix E.1. We also show that for a already pretrained i-Code framework, it is possible to switch to other single modality encoders without pretraining from scratch, indicating the composable property of i-Code (Section 6).

Multimodal Fusion Module. Features extracted by each single-modality encoder are then projected to the fusion network's hidden dimension by a 1-layer feed-forward network. The projected features are input to the modality fusion network to generate integrative multimodal representations. Since positional information is already incorporated by single-modality encoders, we do not use positional embeddings in the fusion module.

The backbone of the fusion network is a transformer encoder, where each layer conducts cross-modality attention, forward projection, and layer normalization. To facilitate more effective cross-modality understanding, we explore two variations on the traditional attention mechanism: merge-attention and co-attention, as illustrated in Figure 1.

Merge-attention. Different modalities share the same attention parameters. To distinguish between different modalities, an identification embedding unique to each modality is added to the projected features (on all temporal and spatial dimensions). Projected features from different modalities are concatenated together (the temporal and spatial dimensions are flattened for visual inputs) and fed into the fusion network, where each layer is the same as the classical transformer encoder layer (Vaswani et al. 2017).

Co-attention. Each transformer layer first conducts self-attention among features of each individual modality, with modality-specific attention parameters (Lu et al. 2019). For example, let the language, vision and speech outputs from a previous transformer layer be X_L, X_V and X_S. Now, we can write a single attention head focusing on the language modality as:

$$X_L^{\text{self}} = \text{Self-Attention-Language}(Q_L, K_L, V_L),$$

where query $Q_L = X_L W_L^Q$, key $K_L = X_L W_L^K$, value $V_L = X_L W_L^V$; W_L^Q, W_L^K, and W_L^V are modality-specific attention matrices (in this case language). The self-attention sublayer (with the residual connection and layer normalization) is followed by a cross-modality attention:

$$X_L^{\text{cross}} = \text{Cross-Attention-Language}(Q_L^{\text{cross}}, K_L^{\text{cross}}, V_L^{\text{cross}}),$$

where $Q_L^{\text{cross}} = X_L^{\text{self}} W_{Lc}^Q$, $K_L^{\text{cross}} = [X_V^{\text{self}}, X_S^{\text{self}}] W_{Lc}^K$, $V_L^{\text{cross}} = [X_V^{\text{self}}, X_S^{\text{self}}] W_{Lc}^V$; W_{Lc}^Q, W_{Lc}^K, and W_{Lc}^V are cross-attention parameters of the language modality. Figure 1 illustrates the merge- and co-attention mechanisms. For a fusion network module with merge-attention, we use 6 transformer encoder layers with hidden size 768 and the fusion module has 154 million parameters. For the co-attention fusion module, to keep its model size close to the merge-attention one, we use 3 layers and the same hidden size, ending up with 163 million parameters. The parameters in the fusion module are randomly initialized in pretraining and are not instantiated from pretrained checkpoints.

Furthermore, we investigated the relationship between model scale and pretraining performance, replacing the dense transformer fusion encoder with a Mixture-of-Experts encoder and observing consistent performance gains (Section 6).

4.2 Pretraining i-Code

In this subsection, we introduce how we pretrain i-Code. We first discuss the multimodal pretraining objectives: masked units modeling and cross-modality contrastive learning. Then we introduce the optimization and training details.

[1] github.com/google-research/bert/blob/master/multilingual.md

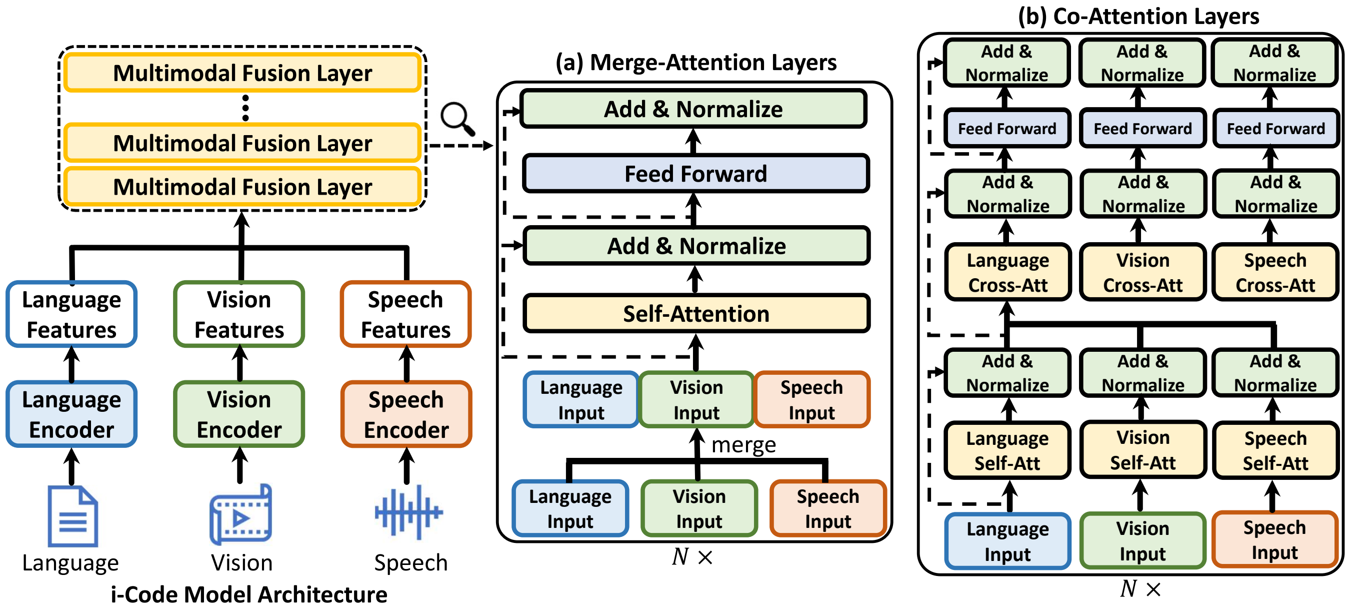

Figure 1: Left: i-Code model architecture. Right: the attention and feed-forward operation in a fusion network layer with (a) merge-attention and (b) co-attention. For simplicity, we only draw the residual connection of the language modality.

Masked Units Modeling: Masked Language Modeling (MLM). Masked Language Modeling (MLM) has achieved remarkable success in self-supervised learning for both language (Devlin et al. 2019) and vision-language pretraining (Dou et al. 2022). During pretraining, we mask out 30% of the text tokens[2]. The task is to predict the masked tokens and the loss $\mathcal{L}_{\text{MLM}}$ is the cross entropy between the ground-truth and the predicted token indices.

Masked Vision Modeling (MVM). For vision self-supervised learning, we adopt the consistent high-level strategy as masked language modeling. We convert vision inputs to discrete tokens, mask out regions in the inputs images, and maximize the cross-entropy between the prediction and ground-truth tokens of the masked out regions. Given a sequence of frames, we leverage PeCo (Dong et al. 2021), a state-of-the-art visual vector quantized variational Autoencoder (VQ-VAE), to discretize each frame to tokens. For masking, we adopt the 3D tube-masking strategy proposed in Wang et al. (2022) to mask image regions across the temporal dimension, where the masking patch ratio for one frame is 50%. We introduce the details in Appendix E.

Masked Span Modeling (MSM). We discretize the speech utterance into tokens with a speech quantizer model, i.e., the quantizer of wav2vec 2.0 (Baevski et al. 2020). We use the same masking strategy as in HuBERT (Hsu et al. 2021) and wav2vec 2.0 (Baevski et al. 2020), where p% of the time steps are randomly selected as start indices, and the next l-step span is masked (Hsu et al. 2021). We follow the default setting of pretraining WavLM and HuBERT by choosing $l = 10$ and $p = 8$. The MSM loss $\mathcal{L}_{\text{MSM}}$ is the cross-entropy between the prediction and labels.

Cross-Modality Contrastive Learning The second group of pretraining objectives are the cross-modality contrastive learning objectives. Each single modality input is first encoded by the corresponding encoder and then fed into the multimodal encoder **individually**. Next, we average each group of single-modality embeddings. For language and speech, the multimodal encoder outputs are averaged along the sequential/temporal dimension. For vision inputs, they are averaged along both temporal and spatial dimensions. We denote the l_2 normalized representations for vision, language and speech as $\boldsymbol{u}_v, \boldsymbol{u}_l, \boldsymbol{u}_s$ respectively.

Similar to previous works (Yuan et al. 2021; Jia et al. 2021; Radford et al. 2021), the vision-language contrastive loss $\mathcal{L}_{vl}$ for a minibatch $\mathcal{B}$ is then defined as:

$$\begin{aligned} \mathcal{L}_{vl} &= \mathcal{L}_{v2l} + \mathcal{L}_{l2v}, \\ \mathcal{L}_{v2l} &= -\frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\frac{\exp(\tau_{vl}\langle \boldsymbol{u}_v^{(i)}, \boldsymbol{u}_l^{(i)}\rangle)}{\sum_{j=1}^{|\mathcal{B}|}\exp(\tau_{vl}\langle \boldsymbol{u}_v^{(i)}, \boldsymbol{u}_l^{(j)}\rangle)}, \\ \mathcal{L}_{l2v} &= -\frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\frac{\exp(\tau_{vl}\langle \boldsymbol{u}_l^{(i)}, \boldsymbol{u}_v^{(i)}\rangle)}{\sum_{j=1}^{|\mathcal{B}|}\exp(\tau_{vl}\langle \boldsymbol{u}_l^{(i)}, \boldsymbol{u}_v^{(j)}\rangle)}. \end{aligned} \tag{1}$$

$\mathcal{L}_{v2l}$ and $\mathcal{L}_{l2v}$ denote the vision-to-language and language-to-vision contrastive learning objectives respectively; τ_{vl} is a learnable scaling parameters; $\langle\, ,\, \rangle$ denotes the inner product. We also define $\mathcal{L}_{vs}$ and $\mathcal{L}_{ls}$ for vision-speech and language-speech contrastive learning. To increase the effective batch size, we concatenate the batches across all GPUs and employ the recently proposed gradient-cache technique (Gao et al. 2021). This also ensures that when pretraining on dual data, the effective batch from steps accumulation can contain pairs from different dual datasets.

[2] Similarly to BERT pretraining, 10% of the masked tokens are replaced with random token, and 10% of the time we keep the tokens unchanged and 80% are replaced with the MASK token.

For pretraining with video, captions and ASR transcripts are concatenated as the language input to the visual-language contrastive learning and MLM. The final pretraining objective is the weighted sum of the masked units modeling and contrastive learning objectives:

$$\mathcal{L} = \alpha\mathcal{L}_{\mathrm{MLM}} + \beta\mathcal{L}_{\mathrm{MVM}} + \gamma\mathcal{L}_{\mathrm{MSM}} + \lambda(\mathcal{L}_{vl} + \mathcal{L}_{vs} + \mathcal{L}_{ls}). \tag{2}$$

We experiment with both fixed and learnable combination weights. We do not observe significant differences in downstream performance and empirically find that $\alpha = 0.5, \beta = 0.6, \gamma = 1, \lambda = 1$ works well. On either paired datasets or videos, we pretrain for 3 epochs on 72 A100 GPUs, with effective batch size 1728. The learning rate is 2×10^{-5} for the fusion module and 10^{-5} for modality encoders with 20000 warm-up steps, and the optimizer is AdamW.

5 Experiments

In this section, we evaluate i-Code and compare it with previous work on a variety of downstream tasks, including multimodal sentiment & emotion analysis, multimodal inference, video QA and single-modality tasks. We refer readers to the appendix for details on experiment settings, hyperparameters, and performance standard deviation for each downstream task due to space limit.

5.1 Multimodal Sentiment & Emotion Analysis

We test i-Code on the largest dataset of multimodal sentiment analysis and emotion recognition to date, CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) (Zadeh et al. 2018) of 23,453 videos. It has two tasks: sentiment analysis and emotion recognition. For sentiment analysis, given a video, the models need to predict the sentiment levels from "highly negative (-3)" to "highly positive (3)" (Zadeh et al. 2018). Evaluation metrics are mean average errors (MAE), correlation (Corr) between the predicted and ground-truth labels and F1. The dataset can also be evaluated as a binary classification task, by grouping sentiment -3 to 0 as one class and 1 to 3 as another [3].

We tested several configurations of i-Code, with results shown in Table 1. We compare with MulT (Tsai et al. 2019), ICCN (Sun et al. 2020), MISA (Hazarika, Zimmermann, and Poria 2020), ScaleVLAD (Luo et al. 2021), and Self-MM (Yu et al. 2021). i-Code sets the state-of-the-art on this task, e.g., improving 3% on the correlation. i-Code models trained on the dual dataset exhibit better performance than the one trained on video dataset, and the merge-attention outperforms the co-attention on this dataset. We also explore directly finetuning on the downstream task, without pretraining the fusion module ("No Pretrain").

For emotion recognition, videos are categorized into {happiness, sadness, anger, fear, disgust, surprise}. Evaluation metrics are accuracy, precision, recall and micro-F1. We evaluate on the unaligned version of the dataset since the alignment information is not always available in real-world scenarios. i-Code improves upon previous best models by 4.2% on accuracy and 3.3% on F1 (Table 2). The co-attention surpasses the merge-attention. Leveraging dual and video data together for pretraining achieves the best result. i-Code even surpasses previous models which had additional access to the alignment information (Table 17).

Model	Pretrain	Att.	MAE (↓)	Corr.	Acc-2	F1
MulT			0.591	69.4	81.6	81.6
ICCN			0.565	71.3	84.2	84.2
MISA			0.555	75.6	85.5	85.3
ScaleVLAD			0.527	78.1	86.4	86.3
Self-MM			0.530	76.5	85.2	85.3
i-Code	No Pretrain	Merge	0.529	79.6	86.8	86.5
	Dual	Merge	**0.502**	**81.1**	**87.5**	**87.4**
	No Pretrain	Co	0.510	80.6	87.3	87.5
	Dual	Co	0.525	80.9	87.1	87.0
	Video	Merge	0.519	80.8	87.3	87.1
	Dual+Video	Merge	0.507	80.7	87.3	87.2

Table 1: Experimental results on CMU MOSEI Sentiment Analysis dataset.

Model	Pretrain	Att.	Acc.	F1	Prec.	Recall
DFG (Zadeh et al. 2018)			38.6	49.4	53.4	45.6
MISA			39.8	45.0	37.1	57.1
RAVEN			40.3	51.1	63.3	42.9
MuIT			42.3	52.3	63.6	44.5
HHMPN (Zhang et al. 2021)			43.4	52.8	59.1	47.6
TAILOR (Zhang et al. 2022)			46.0	52.9	**63.9**	45.2
SIMM (Wu et al. 2019)			41.8	48.4	48.2	48.6
ML-GCN (Chen et al. 2019)			43.7	52.4	57.3	48.2
i-Code	No Pretrain	Merge	49.2	54.6	50.3	59.8
	Dual	Merge	49.4	55.4	49.4	**63.0**
	No Pretrain	Co	49.5	55.0	50.2	60.0
	Dual	Co	**50.2**	**56.2**	50.7	**63.0**
	Video	Merge	49.4	55.3	49.6	62.4
	Dual+Video	Merge	49.8	56.0	50.8	62.1

Table 2: Experimental results on MOSEI Emotion Recognition dataset.

We then test on a humor detection dataset, UR-FUNNY (Hasan et al. 2019). Given a video clip with subscripts, video frames and sound, the task is to predict whether this clip will lead to immediate laughter. Baseline models include those that also leverage three-modality inputs, e.g., Bi-Bimodal-Fusion Network (Han et al. 2021), Low-rank Matrix Fusion (LMF, Liu et al. (2018)), MultiBench (Liang et al. 2021) and Tensor Fusion Network (TFN, Zadeh et al. (2017)). Due to the space limit, the i-Code model pretrained with "dual datasets" is abbreviated as "D", "videos" as "V", "dual datasets+video" as "DV", "no pretraining" as "NP", the merge-attention fusion network as "M", and the co-attention fusion network as "C". E.g., the i-Code model trained on dual datasets with the co-attention is denoted as "**i-Code D+C**". Results in Table 3 show that i-Code outperforms the previous best model by 7.5% and video pretraining shows the best performance.

[3] We also present the result of grouping [-3, -1] as one class and [1, 3] as another in the appendix (Table 15).

Model	**i-Code** D+M	**i-Code** NP+M	**i-Code** D+C	**i-Code** NP+C	**i-Code** V+M	**i-Code** DV+M	MulT	MISA	MultiBench	BBFN	LMF	TFN
Acc.	76.94	73.16	77.00	73.6	79.17	**79.67**	70.55	70.61	66.7	71.68	67.53	68.57

Table 3: Results on the UR-FUNNY test set.

Model	**i-Code** D+M	**i-Code** NP+M	**i-Code** D+C	**i-Code** V+M	HERO	Craig.Starr	GVE (C3D)	GVE (ResNet)	DF-BERT
Acc.	**72.90**	71.60	72.09	72.61	68.59	69.43	68.15	68.39	67.84

Table 4: Results on the VIOLIN test set.

5.2 Multimodal Inference

To better assess how i-Code reasons across modalities, we evaluate i-Code on a multimodal inference dataset VIOLIN (Liu et al. 2020). The input is a video clip from a television show. This clip consists of frames V, aligned subtitles T and sound S. The clip is paired with a text hypothesis H and the task is to decide whether the hypothesis contradicts or entails the video clip. We append the video subtitles to the text hypothesis H, separated by the [SEP] token. The multimodal representation is the average of the outputs from the fusion network. A binary classifier is trained on the ensuing representation. Results are summarized in Table 4. Baselines include HERO, Craig.Starr, GVE (Chen and Kong 2021), and DF-BERT (Liu et al. 2020). i-Code improves upon the previous best model by 3.47%.

5.3 Video Question & Answering

We then test on video question answering (VQA) datasets. We concatenate the question, a candidate answer, and subtitles together (separated by [SEP]) as the text input. The text input, visual frames and speech waveforms are fed together to the i-Code model to average the outputs across modalities as the multimodal representation of the QA pair $\{q, a_i\}$ conditioned on the clip. A projection layer transforms the representation to the logit, followed by softmax.

How2QA dataset (Li et al. 2020) contains 31.7k video clips. Baseline models include HERO (Li et al. 2020), the 2021 ICCV VALUE winner Craig.Starr (Shin et al. 2021), DUKG (Li, He, and Feng 2021), CLIP (Radford et al. 2021), CLIP+SlowFast and ResNet+SlowFast (Li et al. 2021). Results on the public test split are listed in Table 5. KnowIT is a knowledge-based VQA dataset with 24,282 human-annotated question-answer pairs (Garcia et al. 2020). We compare i-Code with DiagSumQA (Engin et al. 2021), variants of knowledge based VQA models ROCK (Garcia et al. 2020) and ROLL (Garcia and Nakashima 2020). As shown in Table 6, i-Code sets the new state-of-the-art.

5.4 Single Modality Evaluations

In Table 7, we compare i-Code (D+M) against previously published multimodal models (FLAVA (Singh et al. 2022), UNITER (Chen et al. 2020)) on the language-only benchmark GLUE (Wang et al. 2018). i-Code has set a new state-of-the-art for multimodal models by a significant margin of 11% on the average performance. Even compared to language-only models, i-Code still shows stronger performance and outperforms DeBERTa v3 base (which the i-Code language encoder is initialized from) on 7 out of 8 tasks. As indicated in Table 7, previous multimodal models, especially V+L models, usually exhibit inferior performance compared to language models. The performance gap is typically attributed to the inferior quality of language data in multimodal datasets (e.g., image captions). We conjecture that the MLM objective and language-speech contrastive learning improve our encoder quality. i-Code pretrained with DeBERTa-v3 large also outperforms large-configuration language encoders (Table 12).

We also evaluate on vision-only action recognition dataset Kinetics-600 (Carreira et al. 2018). We test the vision encoder from i-Code pretraining (on video data with merge attention). As shown in Table 8, i-Code improves upon Florence-base, the model that i-Code vision encoder is initialized from, as well as previous models of similar or large model size, e.g., VATT (Akbari et al. 2021), X3D-XL (Feichtenhofer 2020), TimeSformer-L (Bertasius, Wang, and Torresani 2021), SlowFast-R101-NL (Feichtenhofer et al. 2019) and LGD-3D-101 (Qiu et al. 2019).

We then experiment on the speech benchmark SUPERB (Chen et al. 2022b), including Speaker Identification (SID), Keyword Spotting (KS), Automatic Speaker Verification (ASV). Results are in Table 9 and baselines include TERA (Liu, Li, and Lee 2021), wav2vec, HuBERT and WavLM. Recall that the speech encoder in i-Code is instantiated from WavLM-large. i-Code speech encoder outperforms WavLM-large on SID by 1.68% and ASV. It is on par with it on other tasks. Compared with other baselines, i-Code exhibits much stronger performance. Evaluation results on more SUPERB tasks can be found in Table 13 in the appendix.

6 Analysis

In this section, we present the exploration on the effectiveness of model size on i-Cod performance, modality effusiveness, how composable i-Code model design is, and the effectiveness of multimodal pretraining.

MoE Multimodal Fusion Encoder To further investigate the relationship between i-Code scale and performance, we pretrained a i-Code model using sparsely activated Mixture-

Model	**i-Code** D+M	**i-Code** NP+M	**i-Code** D+C	**i-Code** NP+C	**i-Code** V+M	**i-Code** DV+M	Craig.Starr	HERO	DUKG	CLIP	CLIP-SF	ResNet-SF
Acc.	75.41	74.41	75.52	74.76	**75.73**	75.21	74.74	74.32	73.92	69.34	72.87	74.32

Table 5: Results on the HOW2QA dataset.

Model	**i-Code** D+M	**i-Code** NP+M	**i-Code** D+C	**i-Code** NP+C
Acc.	**80.5**	78.1	**80.5**	79.8
Model	**i-Code** V+M	DiagSumQA	ROLL	ROCK concepts
Acc.	80.0	78.1	71.5	65.4
Model	ROCK image	ROCK facial	ROCK caption	-
Acc.	65.4	65.4	63.5	-

Table 6: Results on the KnowIT video Q&A dataset.

of-Experts (MoE) encoder for multimodal fusion (Jacobs et al. 1991; Shazeer et al. 2017). Similar to Fedus, Zoph, and Shazeer (2022), we replaced the final feed-forward layers of each transformer block in the merge-attention fusion encoder (Figure 1) with a routing block employing top-1 gating and 32 experts, each expert consisting of a pair of feed-forward layers separated by a GELU activation function. The MoE version of the fusion encoder has 483 million parameters, compared to the 154 million parameters of the dense model. As shown in Table 10, using MoE for multimodal fusion further improves i-Code performance.

Modality Effectiveness & Flexibility. We investigate how effective modalities are in multimodal datasets. E.g., in MOSEI Emotion Recognition (Table 11), we find the speech (S) modality to be the most effective. This observation is reasonable considering the emotional quality of human speech (Peerzade, Deshmukh, and Waghmare 2018). Leveraging dual modalities improves upon using the single-modality, and using all modalities produces the best overall performance. These competitive results also demonstrate that i-Code can infer with any combinations of modalities, e.g., text-less or speech-only emotion analysis.

Switching the Single-modality Encoder. After i-Code is pre-trained, one may want to switch the single-modality encoder with another one, e.g., a newly developed model. Can we leverage the pretrained i-Code model without having to pretrain from scratch? E.g., we first pretrain an i-Code model using DeBERTa as the language encoder, and then we switch the DeBERTa model with the BERT-base-uncased model and continue pretraining. Its performance on the UR-FUNNY development set is plotted with blue in Figure 2. We pretrain another i-Code model with the same BERT encoder from scratch, with performance plotted with orange in Figure 2. Utilizing previous checkpoints converges faster. This shows switching the single-modality encoder in a pretrained i-Code model without pretraining from scratch is feasible. More analysis are available in the appendix, e.g., Tables 18 and 19 on pretraining i-Code with different encoders.

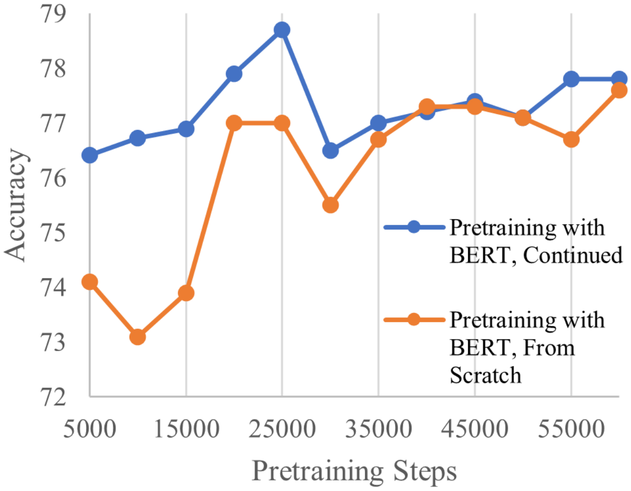

Figure 2: Experiment of switching the language encoder from DeBERTa to BERT after pretraining the i-Code model.

Pretraining Effectiveness. As shown in Tables 1 to 6, pretraining improves upon without pretraining (NP+M). We explore the impact of pretraining objectives, presented in Table 20. Training with either contrastive learning or masked units modeling yields competitive performance.

More analysis and experiment results are presented in the appendix. For example, we conduct video-language cross-modal retrieval experiments on MSR-VTT dataset. We find that using audio modality can further improve retrieval performance (Table 14). Moreover, using audio modality alone to retrieve video already exhibits reasonable performance. Appendix can be found at https://arxiv.org/abs/2205.01818.

7 Conclusion

We introduced i-Code, a multimodal pretraining framework that jointly learns representations for vision, language, and speech. i-Code is integrative and flexible, able to dynamically process one, two, or three modalities at a time for the construction of shared representation spaces. The model leverages novel attention mechanisms and loss functions to effectively combine information from these disparate modalities. We show that pretraining on dual-modality datasets can also yield competitive or even better performance than pretraining on videos, the data resource that previous three-modality models were restricted to. i-

Model	CoLA	SST-2	RTE	MRPC	QQP	MNLI	QNLI	STS-B	Avg.
Single-modality Language Models									
BERT	52.1	93.5	66.4	88.9	71.2	84.6	90.5	85.8	79.1
RoBERTa	63.6	94.8	78.7	90.2	91.9	87.6	92.8	91.2	86.4
DeBERTa V3	69.2	96.2	84.8	90.2	92.5	90.6	94.1	91.4	88.6
Multimodal Models									
UNITER	37.4	89.7	55.6	69.3	89.2	80.9	86.0	75.3	72.9
FLAVA	50.7	90.9	57.8	81.4	90.4	80.3	87.3	85.7	78.0
VisualBERT	38.6	89.4	56.6	71.9	89.4	81.6	87.0	81.8	74.5
VL-BERT	38.7	89.8	55.7	70.6	89.0	81.2	86.3	82.9	74.2
ViLBERT	36.1	90.4	53.7	69.0	88.6	79.9	83.8	77.9	72.4
CLIP	25.4	88.2	55.3	69.9	65.3	33.5	50.5	16.0	50.5
i-Code	**70.1**	**96.3**	**85.6**	**91.0**	**92.6**	**90.5**	**94.3**	**91.9**	**89.0**

Table 7: GLUE experiments. Best multimodal models are in bold and the highest of all baselines are underlined.

Model	#Params	TOP-1	TOP-5
AttentionNAS	-	79.8	94.4
LGD-3D-101	-	81.5	95.6
SlowFast-R101-NL	-	81.8	95.1
X3D-XL	11M	81.9	95.5
TimeSformer-L	121.4M	82.2	95.6
VATT-Base	87.9M	80.5	95.5
VATT-Medium	155M	82.4	**96.1**
Florence-Base	91M	82.5	95.9
i-Code	91M	**83.0**	**96.1**

Table 8: Results on Kinetics-600 action recognition dataset.

Tasks Metrics	SID Acc.	KS Acc.	ASV EER↓
TERA	57.57	89.48	15.98
vq-wav2vec	38.80	93.38	7.99
wav2vec2.0 Large	86.14	96.66	5.65
HuBERT Large	90.33	95.29	5.98
WavLM Large	95.46	**97.86**	3.77
i-Code	**97.14**	97.57	**3.73**

Table 9: Results on the speech benchmark SUPERB. The best results are in bold and the second best are underlined.

	MOSEI SA	MOSEI EM	KnowIT
Dense (152M)	85.4	49.4	80.0
MoE (483M)	**85.81**	**49.7**	**80.7**
	HOW2QA	UR-FUNNY	VIOLIN
Dense (152M)	75.73	79.17	72.61
MoE (483M)	**76.40**	**80.1**	**72.67**

Table 10: Comparison between i-Code dense and MoE fusion encoders, with merge attention and video pretraining.

V	L	S	Acc.	F1	Prec.	Recall
✓			45.0	50.0	45.9	54.9
	✓		46.3	52.6	44.5	64.3
		✓	47.3	52.7	46.4	60.1
✓	✓		49.0	54.8	49.2	61.8
✓		✓	48.0	53.3	47.5	60.7
	✓	✓	49.2	**56.1**	48.8	**65.8**
✓	✓	✓	**49.4**	55.4	**49.4**	63.0

Table 11: Vision (V), language (L) and speech (S) modality effectiveness on MOSEI Emotion Recognition.

Code sets new state-of-the-art on 5 video understanding tasks and single-modality benchmarks.

Acknowledgements

We thank anonymous reviewers for their feedback. We would also like to thank Linjie Li, Lijuan Wang, Junjian Xia, Min Gao for the help in data preparation.

References

Ahuja, K.; Dandapat, S.; Sitaram, S.; and Choudhury, M. 2022. Beyond Static models and test sets: Benchmarking the potential of pre-trained models across tasks and languages. In *Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP*, 64–74.

Akbari, H.; Yuan, L.; Qian, R.; Chuang, W.-H.; Chang, S.-F.; Cui, Y.; and Gong, B. 2021. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems*, 34.

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a Visual Language Model for Few-

Shot Learning. In *Advances in Neural Information Processing Systems*.

Alayrac, J.-B.; Recasens, A.; Schneider, R.; Arandjelović, R.; Ramapuram, J.; De Fauw, J.; Smaira, L.; Dieleman, S.; and Zisserman, A. 2020. Self-supervised multimodal versatile networks. *Advances in Neural Information Processing Systems*, 33: 25–37.

Baevski, A.; Hsu, W.-N.; Xu, Q.; Babu, A.; Gu, J.; and Auli, M. 2022. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *International Conference on Machine Learning*, 1298–1312. PMLR.

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

Bertasius, G.; Wang, H.; and Torresani, L. 2021. Is Space-Time Attention All You Need for Video Understanding? In *International Conference on Machine Learning*, 813–824. PMLR.

Carreira, J.; Noland, E.; Banki-Horvath, A.; Hillier, C.; and Zisserman, A. 2018. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*.

Chen, J.; and Kong, Y. 2021. Explainable Video Entailment with Grounded Visual Evidence. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Chen, S.; Wang, C.; Chen, Z.; Wu, Y.; Liu, S.; Chen, Z.; Li, J.; Kanda, N.; Yoshioka, T.; Xiao, X.; et al. 2022a. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6): 1505–1518.

Chen, S.; Wu, Y.; Wang, C.; Chen, Z.; Chen, Z.; Liu, S.; Wu, J.; Qian, Y.; Wei, F.; Li, J.; et al. 2022b. Unispeech-sat: Universal speech representation learning with speaker aware pre-training. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6152–6156. IEEE.

Chen, Y.-C.; Li, L.; Yu, L.; El Kholy, A.; Ahmed, F.; Gan, Z.; Cheng, Y.; and Liu, J. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*, 104–120. Springer.

Chen, Z.-M.; Wei, X.-S.; Wang, P.; and Guo, Y. 2019. Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5177–5186.

Chung, Y.-A.; Zhu, C.; and Zeng, M. 2021. SPLAT: Speech-Language Joint Pre-Training for Spoken Language Understanding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1897–1907.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 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)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Dong, X.; Bao, J.; Zhang, T.; Chen, D.; Zhang, W.; Yuan, L.; Chen, D.; Wen, F.; and Yu, N. 2021. PeCo: Perceptual Codebook for BERT Pre-training of Vision Transformers. *arXiv preprint arXiv:2111.12710*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

Dou, Z.-Y.; Xu, Y.; Gan, Z.; Wang, J.; Wang, S.; Wang, L.; Zhu, C.; Zhang, P.; Yuan, L.; Peng, N.; et al. 2022. An Empirical Study of Training End-to-End Vision-and-Language Transformers. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 18145–18155. IEEE Computer Society.

Engin, D.; Schnitzler, F.; Duong, N. Q.; and Avrithis, Y. 2021. On the hidden treasure of dialog in video question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2064–2073.

Fedus, W.; Zoph, B.; and Shazeer, N. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *The Journal of Machine Learning Research*, 23(1): 5232–5270.

Feichtenhofer, C. 2020. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 203–213.

Feichtenhofer, C.; Fan, H.; Malik, J.; and He, K. 2019. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6202–6211.

Gao, L.; Zhang, Y.; Han, J.; and Callan, J. 2021. Scaling Deep Contrastive Learning Batch Size under Memory Limited Setup. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, 316–321.

Garcia, N.; and Nakashima, Y. 2020. Knowledge-Based Video Question Answering with Unsupervised Scene Descriptions. In *Proceedings of the European Conference on Computer Vision*.

Garcia, N.; Otani, M.; Chu, C.; and Nakashima, Y. 2020. KnowIT VQA: Answering knowledge-based questions about videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 10826–10834.

Han, W.; Chen, H.; Gelbukh, A.; Zadeh, A.; Morency, L.-p.; and Poria, S. 2021. Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, 6–15.

Hasan, M. K.; Rahman, W.; Zadeh, A. B.; Zhong, J.; Tanveer, M. I.; Morency, L.-P.; and Hoque, M. E. 2019. UR-FUNNY: A Multimodal Language Dataset for Understanding Humor. 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), 2046–2056.

Hazarika, D.; Zimmermann, R.; and Poria, S. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, 1122–1131.

He, P.; Liu, X.; Gao, J.; and Chen, W. 2020. DEBERTA: DECODING-ENHANCED BERT WITH DISENTANGLED ATTENTION. In *International Conference on Learning Representations*.

Hsu, W.-N.; Bolte, B.; Tsai, Y.-H. H.; Lakhotia, K.; Salakhutdinov, R.; and Mohamed, A. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29: 3451–3460.

Jacobs, R. A.; Jordan, M. I.; Nowlan, S. J.; and Hinton, G. E. 1991. Adaptive mixtures of local experts. *Neural computation*, 3(1): 79–87.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.-H.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, 4904–4916. PMLR.

Kaiser, L.; Gomez, A. N.; Shazeer, N.; Vaswani, A.; Parmar, N.; Jones, L.; and Uszkoreit, J. 2017. One model to learn them all. *arXiv preprint arXiv:1706.05137*.

Li, G.; He, F.; and Feng, Z. 2021. A CLIP-Enhanced Method for Video-Language Understanding. *arXiv preprint arXiv:2110.07137*.

Li, L.; Chen, Y.-C.; Cheng, Y.; Gan, Z.; Yu, L.; and Liu, J. 2020. HERO: Hierarchical Encoder for Video+ Language Omni-representation Pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2046–2065.

Li, L.; Lei, J.; Gan, Z.; Yu, L.; Chen, Y.-C.; Pillai, R.; Cheng, Y.; Zhou, L.; Wang, X. E.; Wang, W. Y.; et al. 2021. VALUE: A Multi-Task Benchmark for Video-and-Language Understanding Evaluation. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Li, L. H.; Yatskar, M.; Yin, D.; Hsieh, C.-J.; and Chang, K.-W. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Liang, P. P.; Lyu, Y.; Fan, X.; Wu, Z.; Cheng, Y.; Wu, J.; Chen, L. Y.; Wu, P.; Lee, M. A.; Zhu, Y.; et al. 2021. MultiBench: Multiscale Benchmarks for Multimodal Representation Learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Liu, A. T.; Li, S.-W.; and Lee, H.-y. 2021. Tera: Self-supervised learning of transformer encoder representation for speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29: 2351–2366.

Liu, J.; Chen, W.; Cheng, Y.; Gan, Z.; Yu, L.; Yang, Y.; and Liu, J. 2020. Violin: A large-scale dataset for video-and-language inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10900–10910.

Liu, Z.; Ning, J.; Cao, Y.; Wei, Y.; Zhang, Z.; Lin, S.; and Hu, H. 2022. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3202–3211.

Liu, Z.; Shen, Y.; Lakshminarasimhan, V. B.; Liang, P. P.; Zadeh, A. B.; and Morency, L.-P. 2018. Efficient Low-rank Multimodal Fusion With Modality-Specific Factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2247–2256.

Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Luo, H.; Ji, L.; Huang, Y.; Wang, B.; Ji, S.; and Li, T. 2021. ScaleVLAD: Improving Multimodal Sentiment Analysis via Multi-Scale Fusion of Locally Descriptors. *arXiv preprint arXiv:2112.01368*.

Miech, A.; Alayrac, J.-B.; Smaira, L.; Laptev, I.; Sivic, J.; and Zisserman, A. 2020. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9879–9889.

Miech, A.; Zhukov, D.; Alayrac, J.-B.; Tapaswi, M.; Laptev, I.; and Sivic, J. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2630–2640.

Monfort, M.; Jin, S.; Liu, A.; Harwath, D.; Feris, R.; Glass, J.; and Oliva, A. 2021. Spoken Moments: Learning Joint Audio-Visual Representations From Video Descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14871–14881.

Peerzade, G. N.; Deshmukh, R.; and Waghmare, S. 2018. A review: Speech emotion recognition. *Int. J. Comput. Sci. Eng*, 6(3): 400–402.

Qiu, Z.; Yao, T.; Ngo, C.-W.; Tian, X.; and Mei, T. 2019. Learning spatio-temporal representation with local and global diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12056–12065.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763. PMLR.

Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, 8821–8831. PMLR.

Saunshi, N.; Plevrakis, O.; Arora, S.; Khodak, M.; and Khandeparkar, H. 2019. A theoretical analysis of contrastive unsupervised representation learning. In *International Conference on Machine Learning*, 5628–5637. PMLR.

Schank, R. C.; and Abelson, R. P. 1975. Scripts, plans, and knowledge. In *IJCAI*, volume 75, 151–157.

Schneider, S.; Baevski, A.; Collobert, R.; and Auli, M. 2019. wav2vec: Unsupervised Pre-Training for Speech Recognition. In *INTERSPEECH*.

Shazeer, N.; Mirhoseini, A.; Maziarz, K.; Davis, A.; Le, Q.; Hinton, G.; and Dean, J. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

Shin, M.; Mun, J.; On, K.-W.; Kang, W.-Y.; Han, G.; and Kim, E.-S. 2021. Winning the ICCV'2021 VALUE Challenge: Task-aware Ensemble and Transfer Learning with Visual Concepts. *arXiv preprint arXiv:2110.06476*.

Singh, A.; Hu, R.; Goswami, V.; Couairon, G.; Galuba, W.; Rohrbach, M.; and Kiela, D. 2022. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15638–15650.

Su, W.; Zhu, X.; Cao, Y.; Li, B.; Lu, L.; Wei, F.; and Dai, J. 2019. VL-BERT: Pre-training of Generic Visual-Linguistic Representations. In *International Conference on Learning Representations*.

Sun, Z.; Sarma, P.; Sethares, W.; and Liang, Y. 2020. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 8992–8999.

Tang, Z.; Lei, J.; and Bansal, M. 2021. Decembert: Learning from noisy instructional videos via dense captions and entropy minimization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2415–2426.

Tsai, Y.-H. H.; Bai, S.; Liang, P. P.; Kolter, J. Z.; Morency, L.-P.; and Salakhutdinov, R. 2019. Multimodal Transformer for Unaligned Multimodal Language Sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 6558–6569.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, A.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 353–355.

Wang, R.; Chen, D.; Wu, Z.; Chen, Y.; Dai, X.; Liu, M.; Jiang, Y.-G.; Zhou, L.; and Yuan, L. 2022. Bevt: Bert pretraining of video transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14733–14743.

Wang, Y.; Shen, Y.; Liu, Z.; Liang, P. P.; Zadeh, A.; and Morency, L.-P. 2019. Words can shift: Dynamically adjusting word representations using nonverbal behaviors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 7216–7223.

Wu, X.; Chen, Q.-G.; Hu, Y.; Wang, D.; Chang, X.; Wang, X.; and Zhang, M.-L. 2019. Multi-View Multi-Label Learning with View-Specific Information Extraction. In *IJCAI*, 3884–3890.

Xu, H.; Ghosh, G.; Huang, P.-Y.; Okhonko, D.; Aghajanyan, A.; Metze, F.; Zettlemoyer, L.; and Feichtenhofer, C. 2021. VideoCLIP: Contrastive Pre-training for Zero-shot Video-Text Understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 6787–6800.

Xu, J.; Mei, T.; Yao, T.; and Rui, Y. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5288–5296.

Yu, W.; Xu, H.; Yuan, Z.; and Wu, J. 2021. Learning Modality-Specific Representations with Self-Supervised Multi-Task Learning for Multimodal Sentiment Analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 10790–10797.

Yuan, L.; Chen, D.; Chen, Y.-L.; Codella, N.; Dai, X.; Gao, J.; Hu, H.; Huang, X.; Li, B.; Li, C.; et al. 2021. Florence: A New Foundation Model for Computer Vision. *arXiv preprint arXiv:2111.11432*.

Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; and Morency, L.-P. 2017. Tensor Fusion Network for Multimodal Sentiment Analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1103–1114.

Zadeh, A. B.; Liang, P. P.; Poria, S.; Cambria, E.; and Morency, L.-P. 2018. Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2236–2246.

Zellers, R.; Lu, J.; Lu, X.; Yu, Y.; Zhao, Y.; Salehi, M.; Kusupati, A.; Hessel, J.; Farhadi, A.; and Choi, Y. 2022. Merlot reserve: Neural script knowledge through vision and language and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16375–16387.

Zellers, R.; Lu, X.; Hessel, J.; Yu, Y.; Park, J. S.; Cao, J.; Farhadi, A.; and Choi, Y. 2021. Merlot: Multimodal neural script knowledge models. *Advances in Neural Information Processing Systems*, 34.

Zhang, D.; Ju, X.; Zhang, W.; Li, J.; Li, S.; Zhu, Q.; and Zhou, G. 2021. Multi-modal multi-label emotion recognition with heterogeneous hierarchical message passing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 1.

Zhang, Y.; Chen, M.; Shen, J.; and Wang, C. 2022. Tailor versatile multi-modal learning for multi-label emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 9100–9108.