Understanding Co-speech Gestures in-the-wild

Sindhu B Hegde*, K R Prajwal*, Taein Kwon, Andrew Zisserman
Visual Geometry Group, Dept. of Engineering Science, University of Oxford
{sindhu, prajwal, taein, az}@robots.ox.ac.uk
https://www.robots.ox.ac.uk/~vgg/research/jegal

Abstract

Co-speech gestures play a vital role in non-verbal communication. In this paper, we introduce a new framework for co-speech gesture understanding in the wild. Specifically, we propose three new tasks and benchmarks to evaluate a model's capability to comprehend gesture-speech-text associations: (i) gesture based retrieval, (ii) gestured word spotting, and (iii) active speaker detection using gestures. We present a new approach that learns a tri-modal video-gesture-speech-text representation to solve these tasks. By leveraging a combination of global phrase contrastive loss and local gesture-word coupling loss, we demonstrate that a strong gesture representation can be learned in a weakly supervised manner from videos in the wild. Our learned representations outperform previous methods, including large vision-language models (VLMs). Further analysis reveals that speech and text modalities capture distinct gesture related signals, underscoring the advantages of learning a shared tri-modal embedding space. The dataset, model, and code are available at: https://www.robots.ox.ac.uk/~vgg/research/jegal.

1. Introduction

Humans gesture when they talk – gesturing is an integral part of human communication, together with speech and facial expressions. Gestures can vary from *beats* – two phase hand movements (up/down, left/right etc) that emphasize particular words or phrases and match the rhythm of the speech, but do not carry semantic content – to *iconic* and *deictic* gestures that are representational and illustrate the *content* of the speech [6, 31]. For example, hands and arms moving apart can accompany a speech segment indicating that something is "huge", or as illustrated in Fig 1, an inward pointing gesture to depict the uttered word "my".

Non-verbal communication accounts for 55% of overall communication[1], highlighting the need for machines to understand non-verbal gestural elements in order to have a holistic understanding of human communication. A clear application is enriching human-computer interaction (HCI) through gestures, and this requires machines to comprehend the semantics of the user's hand gestures. Another application is to detect if a person is speaking based on their gestures, or spot specific words or phrases in a video based on gestures alone. More generally, being able to recognize gestures and determine their semantic and temporal alignment with speech enables human communication to be studied at scale [23].

In this paper, our objective is to learn and evaluate co-speech gesture representations. To this end, we propose three tasks and evaluation benchmarks that act as a proxy for assessing real world applications: (1) gesture based cross-modal retrieval, (2) gesture word spotting, and (3) active speaker detection via gestures. We perform large-scale training featuring $\approx$ 7000 speakers and evaluate on in-the-wild videos from the AVSpeech dataset [17].

Figure 1. Co-speech gestures supplement the spoken language – we show examples for six phrases here, with common words. Learning to associate gestures with the uttered phrases is essential for a holistic understanding of human communication.

All three tasks require models to learn to associate gesture clips with speech segments or their corresponding textual transcripts. For this, we propose a model termed **J**oint **E**mbedding space for **G**estures, **A**udio, and **L**anguage (**JEGAL**)[2] that facilitates matching gestures to words and phrases in the accompanying speech. The matches can be based on the style of the speech (intonation, stress, prosody) or the semantic content of the phrase

*equal contribution

[1] https://online.utpb.edu/about-us/articles/communication/how-much-of-communication-is-nonverbal/

[2] JEGAL, known as Zhuge Liang in Chinese, was a prominent historical figure from China's Three Kingdoms period and is regarded as a symbol of wisdom.

or particular words. However, learning a rich joint gesture-audio-language embedding space is a very challenging task. The associations between gestures and speech are typically sparse and ambiguous, with a high degree of variability across speakers. Usually, only a few of the spoken words are clearly gestured. Additionally, the same sentence can be gestured very differently in different contexts and by different people. Gestures also depend on the speaker's emotion, culture, and social scenario (formality, private vs. public, with friends or strangers, etc.). Furthermore, some types of gestures, such as beat gestures, carry no semantic information, resulting in no direct mapping from the gesture to words. The sparse and weak cross-modal correlations makes gesture representation learning a very unique research problem.

We make three key design choices in our approach that result in a strong tri-modal gesture representation. To start with, we learn gesture video representations from large-scale *weak* cross-modal supervision. The supervision is weak because we only use phrase-level speech audio and transcripts – since we do not have any information on which words in the speech are gestured for videos collected 'in-the-wild'. Second, we obtain cross-modal supervision in the form of both audio and the corresponding text transcript (in Section 6.1, we demonstrate that speech and text modalities capture complementary gesture-related signals). Third, we introduce a new gesture-word alignment and spotting loss that explicitly encourages learning of word-level correspondences.

To summarize, we make the following contributions: (i) we propose a new framework for co-speech gesture understanding with three new tasks and evaluation benchmarks; (ii) we learn a joint tri-modal embedding space in a weakly-supervised manner with a combination of global phrase-level objective, and a local word-level gesture coupling loss; (iii) we demonstrate that the learned JEGAL representation performs on par with vision-language foundation models on three gesture-centric tasks and is useful for practical applications.

2. Related Work

Spectrum of Human Gestures. Gestures can be classified broadly into four major classes [31]. Emblematics convey a clear symbolic meaning, e.g. a "thumbs-up". Iconic gestures are used to convey meaning co-occurring with the articulated speech, e.g. "revolving door" is accompanied with the hands moving in a circular motion. Deictic gestures are pointing gestures using the index finger. The most common are the "beat" gestures which are co-speech gestures that are temporally aligned with the prosodic characters of human speech. They are prominent on lexically stressed syllables. Past works [23, 53] have studied how these gesture classes relate to the speech. Several works have attempted to recognize hand, head, and facial emblematic gestures [15, 19]. Recognizing deictic gestures can help in identifying the referring object being pointed to in a conversation, an essential part of human-robot interaction [30]. The other two gesture classes, i.e., beat and iconic, are quite diverse and are difficult to associate with a well-defined set of words or hand movements. This work takes a data-driven approach to learn gesture representations with the help of speech and natural language.

Co-speech Gesture Understanding. Spoken discourse consists of multiple streams of information: language, lip movements, facial expressions, and hand gestures. As opposed to the advancements in the other streams [9, 37, 42], co-speech gesture understanding is a relatively under-explored area. One possible reason for this could be that gestures are only sparsely correlated with the speech. As a result, some of the works have resorted to developing models for specific cases where the gestures are clear, e.g. weather narration [24, 47]. Some works [20, 33, 34] have tried to detect and recognize gestures in laboratory settings while using multiple stereo and IR cameras. Recently, GestSync [22] learns gesture representations by solving for cross-modal synchronization with speech. This objective can lead to the model capturing low-level associations rather than high-level semantics, leading to poor performance on tasks like retrieval, and word spotting. Our work is the first one to learn gesture representations which capture the semantics, style and also learn word-level associations.

Understanding Gestures in Sign Language. Sign language understanding and recognition is another body of work where models need to understand and associate gestures to words and phrases to solve tasks like sign recognition [5, 11, 32, 38, 44, 62, 64], sign language retrieval [12, 16], sign language translation [10, 11, 51, 58] and sign language production [45, 46, 49]. Sign language gesture understanding is quite different compared to co-speech gesture understanding. In sign language, text transcription is a *translation* of what is being signed/gestured. Thus, the words in the text can be a summary or paraphrasing of what is being gestured, with even a mismatch in the temporal ordering. In co-speech gestures, the speaker is the gesturer, and the gestures are being made by the speaker to directly accompany each word he (she) utters. Hence, these two tasks require very different approaches.

Co-speech Gesture Generation. Several works have focused on generating natural gestures that match a given speech segment. This task has the advantage of being "freely supervised" – large-scale datasets can be curated for this task with almost no manual effort as it only requires unlabeled videos of people talking. Speech2Gesture [21] trains speaker-specific models to generate hand skeleton motion for a given speech segment. Recent papers [7, 59, 61] have moved towards more speaker-independent approaches while also using text to obtain strong semantic supervision. In particular, GestureDiffuCLIP [7] learns a joint gesture-text embedding to improve gesture generation. As will be seen in the results, one clear distinction from the work presented in this paper is that [7] does not learn word-level correspondences, which makes a significant difference to the gesture understanding tasks that we evaluate.

Gesture Recognition Datasets. ChaLearn ConGD and IsoGD [54] are two gesture recognition datasets, providing

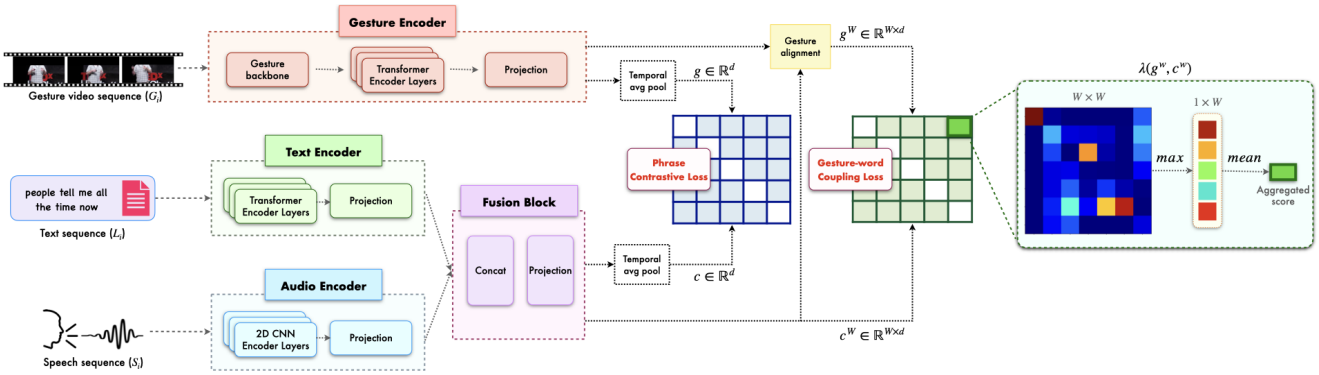

Figure 2. The **JEGAL** architecture. The three input modalities (video, text, speech) are each encoded with a modality-specific encoder, followed by a fusion block to merge speech and text representations. The encoder outputs are average-pooled to obtain global (phrase-level) gesture and speech-text embeddings. During training, these provide the inputs for the global 'phrase contrastive loss'. The gesture alignment module aggregates the relevant video frames to obtain a local word-level gesture embedding for each speech-text word. During training, these provide the inputs for the local 'gesture-word coupling loss'. The two losses encourage the learning of global and local correspondences between the three modalities.

benchmarks for the ChaLearn challenges. However, these datasets are of people using gestures for a task, (e.g. playing a game or controlling an appliance), and are not suitable for learning or evaluating co-speech gestures. Montalbano II [18] is another dataset covering gestures from a vocabulary of 20 Italian sign gesture categories. Again, this is not suitable for our task.

Learning Video Representations. Representation learning in videos [36, 40, 52, 55] has gained significant attention driven by the availability of large-scale video datasets. Learning video representations from text offers a promising advantage by incorporating interpretability via language. Recent works on vision language representation learning [4, 26, 41, 50, 56] highlight this potential. On the other hand, since audio is naturally paired with video, other studies [25, 48] have explored learning video representations from audio. More recently, multimodal approaches have emerged for video representation learning. LanguageBind [63] utilizes depth, infrared, audio, and video to enhance video representations, while Video-LLaMA [60] learns video representations from free-form text and audio. Following these successes, our work aims to leverage multimodal data - video, audio and text - to advance the understanding of gestures.

3. Method

Our goal is to learn co-speech gesture representations from speech and text supervision. Given a dataset *G, S, L* of gesture clips, the accompanying speech segments and their corresponding transcriptions, our goal is to learn gesture representations that capture the rich semantics (from text) and utterance style (from speech) of what is being spoken.

3.1. Overview

JEGAL learns gesture representations by solving two multimodal contrastive objectives between the gesture video and the two other modalities, i.e., speech and text. Each of the three modalities are first encoded using separate encoders $\mathbb{G}$, $\mathbb{S}$, $\mathbb{L}$ to get modality-specific embeddings $\mathbf{g^T} \in \mathbb{R}^{T \times d}$, $\mathbf{s^w} \in \mathbb{R}^{W \times d/2}$, $\mathbf{l^w} \in \mathbb{R}^{W \times d/2}$ to obtain frame-level (T) and word-level (W) representations. The speech and text embeddings are fused into joint speech-text embeddings $\mathbf{c^w} \in \mathbb{R}^{W \times d}$ as depicted in Fig 2.

We learn these representations using (1) a *global phrase-level contrastive objective*, and (2) a *local gesture-word coupling loss*. The first one encourages the model to learn global semantics to match a gesture clip to a speech/text segment. The second objective enforces the model to find the strongest word-level matches between the gesture clip and the other two modalities. We describe our architecture and loss functions in detail below.

3.2. Gesture Encoder

Gesture backbone. Given a gesture clip $G \in (T,h,w,3)$ of T frames, we encode it using a stack of 3D convolution layers, similar to previous audio-visual networks [3, 37, 38], where the first layer has a temporal receptive field of 5 frames to capture the motion information. We obtain a sequence of T visual feature vectors of dimension d. These feature vectors are further encoded using a stack of Transformer encoder layers to get $\mathbf{f_g} \in \mathbb{R}^{T \times d}$. We initialize the backbone weights from GestSync [22] and keep it frozen.

Gesture head. We use a Transformer encoder followed by a projection layer to get the gesture embeddings, $\mathbf{g^T} \in \mathbb{R}^{T \times d}$.

3.3. Text Encoder

Text backbone. Given a text transcription L corresponding to the gesture video clip, we use the final layer outputs from a pre-trained bi-directional language model, multilingual Roberta XLM-Base [14] to obtain text representations. The output of the text backbone is a sequence of sub-word feature vectors, f_l.

Text head. The text head is similar to the gesture head that uses a stack of Transformer layers. The sub-word embeddings f_l are encoded and projected to get the final sub-word embeddings l^{sw} of feature dimension $d/2$ each. We aggregate these sub-word

tokens to word-level tokens later in the fusion block.

3.4. Audio Encoder

Given a speech waveform S, we convert it into melspectrograms which is encoded using a stack of 2D-CNN layers following previous works [3, 13]. The output of the audio encoder is a sequence of T' speech feature vectors, $\mathbf{s} \in \mathbb{R}^{T' \times d/2}$.

3.5. Fusion Block

Before we fuse the speech and text embeddings, we aggregate the text and audio embeddings to obtain word-level feature vectors. We average the sub-word embeddings for each word to get word-level text embeddings $\mathbf{l^w} \in \mathbb{R}^{W \times d/2}$. Using the start and end times of each word, we average the speech features for each word to get word-level features $\mathbf{s^w} \in \mathbb{R}^{W \times d/2}$. We fuse the speech and text features by concatenating along the feature dimension to get joint word-level representations, $\mathbf{c^w} \in \mathbb{R}^{W \times d}$.

3.6. Gesture-Word Alignment

The word boundaries are aligned with the speech, but not necessarily with the gestures in the video. The gestures can be longer or shorter than the window in which the word is uttered, and can also be offset. To handle this discrepancy, we propose an attention-based pooling mechanism to obtain the gesture embedding corresponding to the word. We first pad the speech-based word start-end times with $p=10$ video frames on either side. Let S,E be the start and end frames for the padded window for the word c^{w_i}. We obtain the word-level gesture embedding, g^{w_i} by using the word embedding c^{w_i} for attention-pooling over the extended temporal interval of the word:

$$g^{w_i} = \sum_{j=S}^{E} \left(\frac{\exp(\gamma \cdot g^{T_j} \cdot c^{w_i})}{\sum_{j=S}^{E} \exp(\gamma \cdot g^{T_j} \cdot c^{w_i})} \right) \cdot g^{T_j} \tag{1}$$

3.7. Training Objective

We only have reliable supervision at the phrase level — meaning we know that a specific text or speech segment corresponds to a gesture clip. However, we do not know which individual words are gestured. Keeping this in mind, we employ two loss functions.

Global Phrase Contrastive Loss. To obtain the global phrase embeddings, we average pool the speech-text embeddings and the video frame embeddings to give us $\mathbf{c}$ and $\mathbf{g}$ respectively. Given a batch of N samples, we employ the contrastive InfoNCE loss [35] to encourage similarity between the N positive triplets, and dissimilarity between the N^2-N negative triplets:

$$\mathcal{L}_{seq} = -\frac{1}{N} \sum_{i=1}^{N} \left(\log \frac{\exp(\gamma \cdot cos(g_i, c_i))}{\sum_{j=1}^{n} \exp(\gamma \cdot cos(g_i, c_j))} \right) \tag{2}$$

where γ is the temperature and *cos* is the cosine similarity.

Local Gesture-word Coupling Loss. Pooling the word-level representations to compute the global phrase loss can lead to weak gesture-word associations (Table 6, row 1). However, directly training to match gestures to words is not possible – very few words are gestured in a given phrase, and we do not know which of them are. Thus, we devise a new strategy to learn word-level correspondences using phrase-level supervision. Given a pair of word-level gesture and speech-text ($\mathbf{g^w}$,$\mathbf{c^w}$) embeddings, we first find the closest gesture g^{w_j} for each speech-text word c^{w_i}. Our hypothesis is that a matching ($\mathbf{g^w}$,$\mathbf{c^w}$) will have a higher number of strong word couplings than a non-matching pair. With this idea, we define the following scoring function and the gesture-word coupling loss:

$$\lambda(g_n^w, c_n^w) = \frac{1}{W} \sum_{i=1}^{W} \max_{j=1,2..W} cos(g_n^{w_i}, c_n^{w_j}) \tag{3}$$

$$\mathcal{L}_{couple} = -\frac{1}{N} \sum_{i=1}^{N} \left(\log \frac{\exp(\gamma \cdot \lambda(g_i^w, c_i^w))}{\sum_{j=1}^{N} \exp(\gamma \cdot \lambda(g_i^w, c_j^w))} \right) \tag{4}$$

The gesture-word coupling loss simply maximizes λ for matching gesture-speech-text samples while minimizing λ for the negative ones. In other words, the model is encouraged to find more strong word-level couplings for positive gesture-speech-text phrases in the batch.

Our final loss function is a weighted sum of the two losses:

$$\mathbb{L} = \beta \cdot \mathcal{L}_{seq} + (1-\beta) \cdot \mathcal{L}_{couple} \tag{5}$$

3.8. Implementation Details

We now describe the essential implementation details, more details can be found in the supplementary material.

Training data. We train our model, JEGAL, on triplets of gesture clips, speech segments, and text transcriptions. For the gesture frame inputs, we resize the frames to 270×480 pixels. We extract melspectrograms with a hop length of 10ms. The word-aligned text transcriptions are tokenized into wordpiece tokens. Using the start-end time of the word boundaries, we randomly sample a video clip between $2-10$ seconds in length.

Modality drop. In order to encourage the model to learn both speech and text representations equally well, we randomly set one of these modality inputs to zero 50% of the time. This is commonly done in audio-visual speech recognition models [2, 48]. This also allows us to use only one modality (speech or text) during inference, if necessary.

Model hyper-parameters. For the text and gesture heads, we set the number of Transformer layers to 3 and 6 respectively. The Transformer uses a hidden dimension of 512 and a feed-forward dimension of 2048 with 8 attention heads.

Training hyper-parameters. We use the AdamW optimizer [27] with a learning rate of $5e^{-5}$, weight decay of $1e^{-4}$ and betas $(0.9, 0.98)$. We reduce the learning rate by a factor of 5 when the validation performance does not improve for 2 epochs.

3.9. Training Datasets

We train our model using the following datasets: (i) PATS [21], and (ii) a subset of the MultiVSR dataset [39]. The dataset

Table 1. We train and evaluate on multiple datasets consisting of 720 hours of gesture clips comprising 7000+ speakers. For evaluation, we curate task-specific benchmarks from the publicly available AVSpeech [17] dataset.

Dataset	split	# hours	# spk.	avg. clip duration (s)	# videos
PATS [21]	train	162.3	24	11.37	51390
MultiVSR [39]	train	556.1	6934	15.31	130510
Combined	train	718.4	6958	14.2	181900
AVS-Ret	test	0.31	404	2.27	500
AVS-Spot	test	0.38	384	2.76	500
AVS-Asd	test	0.44	398	3.15	500

specifics are outlined in Table 1. PATS [21] is a publicly available video dataset from 25 speakers sourced from diverse platforms such as lectures, talk-shows, YouTube, and televangelists. The subset from the MultiVSR dataset is composed of 556 hours of interviews, narrations, and talks spanning a broad spectrum of speakers and a rich vocabulary.

Pre-processing: We resample all videos to 25 FPS, and the speech to 16kHz. We leverage WhisperX [8] in cases where datasets lack word-aligned text transcripts. Additionally, using the $L2$ distance between consecutive frame body keypoints, we filter out samples with minimal gesture activity. We also make sure to mask out the face region to avoid leakage from lip movements. Table 1 presents the final statistics of all the datasets.

4. Downstream Tasks and Evaluation

We describe our newly curated evaluation benchmarks and the different downstream tasks to evaluate the quality of our learned gesture representations. The first is cross-modal retrieval, the second is spotting gestured words, and the third is active speaker detection. Note that in all the tasks, while we use the joint speech-text embedding, we can obtain uni-modal scores by inputting zeros to omit a modality during inference.

4.1. Cross-modal Retrieval

Given a gallery of gesture-speech-text samples, the task is to retrieve a gesture clip given a speech segment and/or text and vice-versa. Concretely, given a speech or text as query, we obtain a speech-text embedding, $c \in \mathbb{R}^d$ and rank the gesture embeddings $g \in \mathbb{R}^d$ in the gallery by cosine similarity, highest being at the top. We do the same process for the gesture to speech-text retrieval as well.

Retrieving relevant gestures for a text or speech segment enables several practical applications. For digital avatars, we can retrieve most plausible hand gesture clips to accompany what the avatar is speaking, leading to a more immersive and engaging experience. In gaming applications, given a database of gesture sequences, the developer can automatically select the most relevant gestures to go with the in-game dialogues. Gestures can assist in language learning [29] by improving word-level memory retention (e.g. eat, kick, clap). Language teaching apps will be able to retrieve gesture clips for sentences to improve the speed of foreign language learning.

4.2. Gesture Word Spotting

Given a gesture clip with the accompanying speech/text segment and a word of choice from this segment, the goal is to localize the word in the gesture clip. Concretely, we obtain word-level speech-text (c^w) embeddings and frame-level gesture embeddings, g^T. To localize the i-th word, c^{w_i}, we compute the cosine similarity of the word embedding with all the gesture frame embeddings. The localization of the word in the video is simply obtained by keeping only the locations with similarity scores $\geq \delta = 0.5$.

Spotting can be useful to enhance transcriptions by supplementing the plain words with stress and emotion labels. Another application would be to create word-level gesture databases, e.g. a thousand different ways the word "big" is gestured by people all over the world, which will be useful for language and communication analysis.

4.3. Active Speaker Detection

Given gesture clips of P different speakers, and a speech (S) and/or text segment (T), the goal is to predict the active speaker A who is uttering the queried speech/text. To do this, we extract the sequence-aggregated gesture features $g_i \in \mathbf{R}^d, i \in 1,2,...,P$ for each of the P clips. Given the query speech or text, we obtain the speech-text feature, c. The active speaker A is the one whose gesture and speech-text cosine similarity is maximum:

$$A = \underset{i \in 1,2,...,S}{\operatorname{argmax}} \; cos(c, g_i) \tag{6}$$

The majority of audio-visual models, encompassing tasks like speech recognition, generation, and translation, primarily operate on inputs containing a single speaker. Thus, there arises a necessity to identify the speaker within a video segment. To determine the active speaker in a multi-speaker scenario, previous works [1, 43] have shown the benefits of resorting to the face for lip-sync with the audio, and text subtitles when the audio is corrupted. We extend this thread even further. What happens if the lip region is occluded or unclear? Another important use-case is privacy preserving [57] active speaker detection: what if the active speaker detection needs to be done without leaking the face identity of the speaker? We show that we can successfully do this – with very little identity information, i.e. by only using the hand gestures, we can determine who is speaking.

4.4. AVSpeech Test Benchmarks

Using the AVSpeech official test set [17], we manually curate three separate evaluation benchmarks for the three downstream gesture tasks. The statistics for the evaluation test sets are summarized in Table 1.

AVS-Ret. We create a new cross-modal retrieval benchmark containing diverse gesture clips of hundreds of unique speakers.

Table 2. Cross-modal retrieval performance on the AVS-Ret benchmark (Sec 4.4). JEGAL outperforms the baselines by a large margin.

Method	Mod.		Speech-text to Gesture retrieval					Gesture to Speech-text retrieval				
	T	A	R@5 ↑	R@10↑	R@25↑	R@50↑	MR ↓	R@5 ↑	R@10↑	R@25↑	R@50↑	MR ↓
Random	✓	✓	1.00	2.00	5.00	10.00	250	1.00	2.00	5.00	10.00	250
Zero-shot												
Clip4Clip [28]	✓	✗	7.40	11.00	17.60	25.80	139.0	4.59	7.39	13.57	22.75	167.0
Language-Bind [63]	✓	✗	2.60	4.60	9.00	17.20	190.5	2.20	4.20	8.20	15.80	204.5
GestSync [22]	✗	✓	3.60	5.60	13.20	19.80	212.5	3.20	6.60	18.40	29.80	127.5
Fine-tuned												
Clip4Clip [28]	✓	✗	8.00	12.60	17.60	26.40	132.0	3.60	7.00	19.20	30.20	125.0
Language-Bind [63]	✓	✗	5.80	10.80	14.00	20.40	140.5	4.80	8.00	12.60	24.40	180.0
GestSync [22]	✗	✓	10.00	18.20	27.40	41.20	70.5	11.60	16.60	27.40	40.00	82.5
GestureDiffuClip [7]	✓	✗	7.90	12.80	21.20	30.60	112.0	7.80	10.40	19.00	29.20	128.5
Ours												
JEGAL	✓	✗	13.40	20.60	35.80	48.60	57.5	14.40	27.00	37.20	49.20	51.0
JEGAL	✗	✓	11.00	20.00	34.20	46.20	59.0	12.20	20.60	37.00	45.60	60.5
JEGAL	✓	✓	**18.80**	**30.80**	**46.40**	**62.00**	**31.0**	**18.20**	**20.20**	**51.40**	**70.20**	**24.5**

We choose a gallery of 500 clips, which also contain isolated clean speech and accurate text transcriptions. We verify that the clips contain reasonable gesture activity and transcripts with at least two nouns or verbs or adjectives. For evaluation, we use the standard metrics used in other video-text retrieval works [56, 63], i.e. Recall@K and Median Rank. We evaluate both gesture (g) to content (speech-text c) retrieval and vice-versa and show both unimodal and multimodal retrieval performance.

AVS-Spot. To quantitatively evaluate the gesture spotting task, we manually curate a new test dataset where we search and annotate clips that clearly contain a word that is gestured. We obtain 500 such clips, each containing a target word that is clearly gestured. The manual annotation process removes all kinds of label noise in the test set, allowing for a faithful evaluation of our newly defined gesture spotting task. Additionally, we also manually annotate these target words with binary "stress/emphasis" labels, which can have important cues about the gesture (Table 5). We provide additional annotations of the AVS-Spot test set and more results with it in the supplementary.

AVS-Asd. To build the evaluation dataset for active speaker detection, we first choose 500 "target" clips. For these target clips, we create three evaluation subsets, where we choose $P-1$ clips from different speakers, where $P=2,4,6$. We report the accuracy of detecting the correct target speaker out of the P different speakers.

5. Results

5.1. Baselines

For baselines, we report performance of zero-shot pre-trained video-language models [56, 63] and pre-trained GestSync [22], which learns gesture-audio correspondences by solving for audio-visual synchronization. We also report scores after fine-tuning all these models further on our training data for a fair comparison. In addition, we compare with the semantic encoder of GestureDiffuCLIP [7], by training it on our dataset.

5.2. Cross-modal Retrieval

In Table 2, we compare the performance of JEGAL against other baselines on the cross-modal retrieval task. Zero-shot evaluation of foundational vision-language models like LanguageBind [63] and Clip4Clip [28] leads to higher than chance performance. These models are designed to capture different kinds of features: they cannot handle a large number of frames, and learn non-gesture attributes like identity and scene. Fine-tuning these models improves their performance on the task, but it is still far from the performance of JEGAL . GestSync [22] clearly performs better than the foundational vision-language models post-finetuning. However, since this network is trained to detect synchronization offsets in speech and video, its representations perform poorly for global semantic tasks like retrieval. This is also partly true for our model when we turn off the global phrase loss (Table 6 row 2 vs row 3). GestureDiffuCLIP's semantic encoder [7] performs only second best among the baselines. The lack of local word-level semantic supervision leads to an inferior performance compared to JEGAL for both GestSync and GestureDiffuCLIP.

Furthermore, none of the baseline approaches ingest and fuse multi-modal speech-text inputs. Our JEGAL model outperforms previous methods by a large margin. JEGAL can retrieve gestures from speech or text queries with similar performance. The opposite direction is also true, i.e. retrieving speech or text for a query gesture clip. Finally, retrieving with the fused speech-text representation is clearly better than the unimodal variants, showing that the speech and text embeddings each encode information that is not present in the other modality.

Table 3. Gesture-based word spotting performance on the AVS-Spot benchmark (Sec 4.4).

Method	Mod.		Accuracy ↑
	T	A	
Zero-shot			
Clip4Clip [28]	✓	✗	9.20
Language-Bind [63]	✓	✗	9.40
GestSync [22]	✗	✓	15.63
Fine-tuned			
Clip4Clip [28]	✓	✗	17.59
Language-Bind [63]	✓	✗	15.80
GestSync [22]	✗	✓	21.04
GestureDiffuClip [7]	✓	✗	19.50
Ours			
JEGAL	✓	✗	61.00
JEGAL	✗	✓	41.80
JEGAL	✓	✓	**63.60**

5.3. Gesture Word Spotting

Table 3 compares the spotting accuracy of different methods on the AVS-Spot benchmark. Unlike JEGAL that uses the word-level L_{couple} loss, all the baseline methods, including the semantic encoder of GestureDiffuCLIP [7], use only a phrase-level loss and hence, struggle to learn fine-grained word-level associations. Through this task, we also see the big advantage of using text modality in addition to speech – text-based gesture spotting is more accurate than using audio. This is expected, since word-level semantic correspondences are easier to learn in text space. Having said that, using speech alongside text still gives a clear improvement even in the gesture spotting task.

In Fig 3, we show two examples of spotting gestured words. The heatmaps show the similarity of the word (red is higher) along the video frames. In the first example, the lady gestures "beautiful" with her fingers, and in the second example, the speaker clenches the fist to show "energy". We can see that not all words get a high similarity score, only ones with distinctive gestures are spotted. Furthermore, the spotting does not exactly align with the speech-based word boundary (indicated by yellow vertical lines). Our alignment layer (Sec 3.6) allows the model to look beyond the speech-based boundary and find the exact frames where the word is gestured.

5.4. Active Speaker Detection

In Table 4, we show the accuracy of identifying the target speaker for a given text and/or speech segment. This task is different from our other tasks – it does not need a strong holistic understanding of the gesture sequence like retrieval, nor does it need semantic word-level understanding. It can simply be solved by checking for frame-level synchronization, which is exactly why we see GestSync [22] performs the best on this task. None of the other models are trained with strong frame-level video-speech supervision, and hence, perform worse. JEGAL comes at a close second after GestSync. We also see that speech information is more useful here compared to text.

Table 4. Performance of active speaker detection on AVS-Asd benchmark. We report the mean class accuracy of predicting the active speaker among a set of S speakers, where $S = 2,4,6$.

Method	Mod.		2 spk.	4 spk.	6 spk.
	T	A			
Random	✓	✓	50.0	25.0	16.7
Zero-shot					
Clip4Clip [28]	✓	✗	62.6	39.6	31.8
Language-Bind [63]	✓	✗	51.4	32.2	23.8
GestSync [22]	✗	✓	54.2	31.6	23.8
Fine-tuned					
Clip4Clip [28]	✓	✗	63.8	42.3	32.8
Language-Bind [63]	✓	✗	56.8	38.7	30.1
GestSync [22]	✗	✓	81.2	64.8	54.4
GestureDiffuClip [7]	✓	✗	61.8	39.4	28.6
Ours					
JEGAL	✓	✗	65.6	44.4	34.6
JEGAL	✗	✓	74.4	50.0	40.4
JEGAL	✓	✓	**76.8**	**57.8**	**48.0**

6. Insights and Ablations

In this section, we provide additional insights into the gesture signals learned from the speech and text modalities, the impact of the two loss functions on the downstream tasks, and the choice of speech-text fusion.

6.1. Speech v/s Text Modalities

We have already seen how our model can flexibly leverage speech or text modality to solve three different kinds of tasks. In Fig 4, we show an example where audio-based gesture word spotting is successful, but text-based spotting is not. We see that the uttered word "action" has been emphasized (using pitch graph). This example leads us to do a deeper analysis of gesture word spotting based on stress cues. We divide the binary stress labels to split the AVS-Spot test set (500 samples) into two subsets, one containing only samples with stressed/emphasized words (100 samples) and the other subset containing the remaining words (400 samples). In Table 5, we report the same spotting accuracy metric on these subsets separately. While we saw previously in Table 3 that speech-based gesture spotting is clearly inferior compared to text-based gesture spotting, it is not always the case. We see that the difference in spotting accuracy between the stressed and non-stressed words is higher for the speech modality. This indicates that speech modality pays more attention to emphasis of the word compared to the text modality.

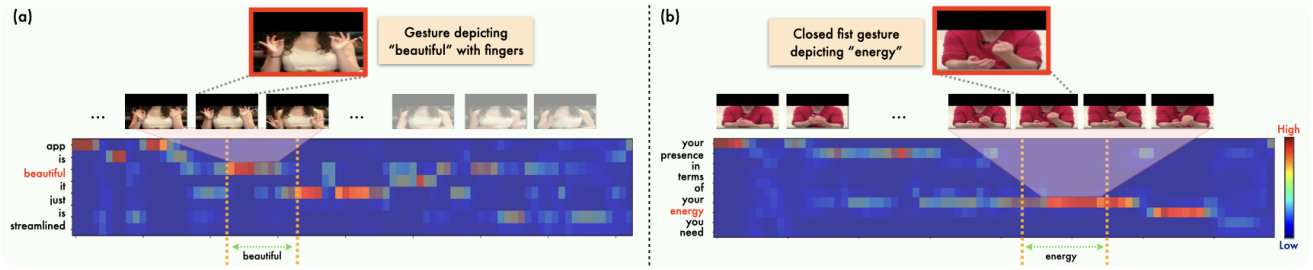

Figure 3. JEGAL can spot the gestured words in a video clip. Here, we show a similarity heatmap of words vs video frames. The vertical yellow lines indicate the speech-based word boundaries of 'beautiful' and 'energy'. The red triangles zoom into the corresponding frames where JEGAL detects the words, clearly aligning with the gestures. For the word "beautiful" the gestured segment is smaller than the spoken word boundary and for the word "energy" it extends well beyond to the right of the word boundary. The alignment layer (Sec 3.6) allows the model to look beyond just the speech-based boundaries. Note that our model learns to perform gestured-word spotting without using any training labels on which words are gestured.

Table 5. Stressed words are more likely to be spotted with speech-based spotting than non-stressed words. As seen in the column "Δ", the difference between stressed vs. non-stressed word spotting is higher for the speech modality.

Modality	With stress	W/o stress	All words	Δ
Text	68.00	59.25	61.00	14.8%
Speech	54.00	38.75	41.80	39.4%

6.2. Impact of loss functions

In Table 6, we study the impact of our two loss functions. In the first row, we only train with the phrase contrastive loss, which captures the global semantics. The second row is the variant trained only with gesture-word coupling loss, which captures local word-level semantics. Compared to the dual loss model, the results from these individual loss models are significantly worse. Specifically, the variant without the word coupling loss performs poorly on the spotting task and the one without the sequence contrastive loss performs poorly on retrieval and active speaker detection. The combination of the two losses performs the best across all the tasks, thus demonstrating the complementary nature of the two training objectives.

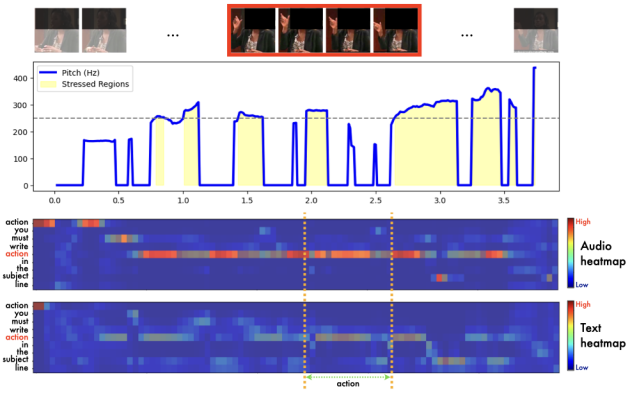

Figure 4. Audio and text heatmap examples when words are stressed.

Table 6. Each loss encourages feature learning at different temporal granularity, and combination of the two loss functions performs the best.

Loss	Retrieval			Spotting	ASD
	R@5 ↑	R@10 ↑	MR ↓	Acc. ↑	Acc. ↑
Seq. contrastive	12.20	23.60	45	20.83	44.2
Word coupling	8.50	14.60	76	52.46	14.8
Seq. + Word coupling	**18.80**	**30.80**	**31**	**63.60**	**48.0**

6.3. Fusion techniques

In Table 7, we ablate different ways to fuse the speech and text features. The first case is to not fuse at all and have two separate pairwise contrastive losses: gesture-audio and gesture-text. We can see in the first two rows that this is an inferior design. In fact, it is better to train with a single contrastive head after fusing the speech and text embeddings (rows 3, 4). The fusion strategy of choice would be to concatenate, rather than average.

Table 7. Ablation study on fusing speech and text modalities. Training without fusing is far worse, as the model cannot perform the tasks by using multiple information streams at the same time.

	Retrieval			Spotting	ASD
	R@5 ↑	R@10 ↑	MR ↓	Acc. ↑	Acc. ↑
Pairwise (with text)	9.39	15.58	70	34.31	29.6
Pairwise (with audio)	9.80	16.60	72	23.67	31.4
Late fusion (avg.)	17.00	26.40	40	56.04	41.2
Late fusion (concat.)	**18.80**	**30.80**	**31**	**63.60**	**48.0**

7. Conclusion

In this work, we learn a joint embedding space that captures cross-modal relationships with gestures, speech, and language. We show that we can learn such an embedding space with weak supervision using a careful design of two loss functions. We evaluate these new representations on three new downstream tasks and manually curated test sets. We observe that the two modalities, i.e., speech and text, learn complementary features that can be useful for different kinds of gesture-related tasks. One promising future direction would be to explore 2D and 3D keypoint-based inputs to make the network computationally lighter and less susceptible to distracting features.

Acknowledgements: The authors would like to thank Piyush Bagad, Ragav Sachdeva, Jaesung Hugh, Paul Engstler for their valuable discussions. The authors are further grateful to Alyosha Efros, Jitendra Malik, and Justine Cassell for their insightful inputs and suggestions. They also extend their thanks to David Pinto for setting up the data annotation tool and to Ashish Thandavan for his support with the infrastructure. This research is funded by EPSRC Programme Grant VisualAI EP/T028572/1, an SNSF Postdoc.Mobility Fellowship P500PT_225450 and a Royal Society Research Professorship RSRP\R\241003.

References

[1] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. The conversation: Deep audio-visual speech enhancement. In *INTERSPEECH*, 2018. 5

[2] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. Deep lip reading: a comparison of models and an online application. In *INTERSPEECH*, 2018. 4

[3] Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. Self-supervised learning of audio-visual objects from video. In *Proc. ECCV*, 2020. 3, 4

[4] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 3

[5] Samuel Albanie, Gül Varol, Liliane Momeni, Hannah Bull, Triantafyllos Afouras, Himel Chowdhury, Neil Fox, Bencie Woll, Rob Cooper, Andrew McParland, and Andrew Zisserman. BOBSL: BBC-Oxford British Sign Language Dataset. 2021. 2

[6] Michael Andric and Steven L Small. Gesture's neural language. *Frontiers in psychology*, 3:99, 2012. 1

[7] Tenglong Ao, Zeyi Zhang, and Libin Liu. Gesturediffuclip: Gesture diffusion model with clip latents. *ACM Transactions on Graphics (TOG)*, 42(4):1–18, 2023. 2, 6, 7

[8] Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. Whisperx: Time-accurate speech transcription of long-form audio. In *INTERSPEECH*, 2023. 5

[9] Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, et al. Seamless: Multilingual expressive and streaming speech translation. *arXiv preprint arXiv:2312.05187*, 2023. 2

[10] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7784–7793, 2018. 2

[11] Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10023–10033, 2020. 2

[12] Yiting Cheng, Fangyun Wei, Jianmin Bao, Dong Chen, and Wenqiang Zhang. Cico: Domain-aware sign language retrieval via cross-lingual contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19016–19026, 2023. 2

[13] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Workshop on Multi-view Lip-reading, ACCV*, 2016. 4

[14] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019. 3

[15] Trevor J Darrell, Irfan A Essa, and Alex P Pentland. Task-specific gesture analysis in real-time using interpolated views. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18 (12):1236–1242, 1996. 2

[16] Amanda Duarte, Samuel Albanie, Xavier Giró-i Nieto, and Gül Varol. Sign language video retrieval with free-form textual queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14094–14104, 2022. 2

[17] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T. Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation. *ACM Trans. Graph.*, 37, 2018. 1, 5

[18] Sergio Escalera, Jordi Gonzàlez, Xavier Baró, Miguel Reyes, Oscar Lopes, Isabelle Guyon, Vassilis Athitsos, and Hugo Escalante. Multi-modal gesture recognition challenge 2013: Dataset and results. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 445–452, 2013. 3

[19] William T Freeman and Michal Roth. Orientation histograms for hand gesture recognition. In *International workshop on automatic face and gesture recognition*, pages 296–301. Citeseer, 1995. 2

[20] Esam Ghaleb, Ilya Burenko, Marlou Rasenberg, Wim Pouw, Peter Uhrig, Judith Holler, Ivan Toni, Aslı Özyürek, and Raquel Fernández. Co-speech gesture detection through multi-phase sequence labeling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4007–4015, 2024. 2

[21] Shiry Ginosar, Amir Bar, Gefen Kohavi, Caroline Chan, Andrew Owens, and Jitendra Malik. Learning individual styles of conversational gesture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3497–3506, 2019. 2, 4, 5

[22] Sindhu B Hegde and Andrew Zisserman. Gestsync: Determining who is speaking without a talking head. In *Proc. BMVC*, 2023. 2, 3, 6, 7

[23] Adam Kendon. *Gesture: Visible action as utterance*. Cambridge University Press, 2004. 1, 2

[24] Sanshzar Kettebekov, Mohammed Yeasin, and Rajeev Sharma. Improving continuous gesture recognition with spoken prosody. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, pages I–I. IEEE, 2003. 2

[25] Sangho Lee, Jiwan Chung, Youngjae Yu, Gunhee Kim, Thomas Breuel, Gal Chechik, and Yale Song. Acav100m: Automatic curation of large-scale datasets for audio-visual video representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10274–10284, 2021. 3

[26] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 3

[27] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 4

[28] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomput.*, 508:293–304, 2022. 6, 7

[29] Manuela Macedonia, Karsten Müller, and Angela D Friederici. The impact of iconic gestures on foreign language word learning and its neural substrate. *Human brain mapping*, 32(6):982–998, 2011. 5

[30] Cynthia Matuszek, Liefeng Bo, Luke Zettlemoyer, and Dieter Fox. Learning from unscripted deictic gesture and language for human-robot interactions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2014. 2

[31] David McNeill. *Hand and mind: What gestures reveal about thought*. University of Chicago press, 1992. 1, 2

[32] Yuecong Min, Aiming Hao, Xiujuan Chai, and Xilin Chen. Visual alignment constraint for continuous sign language recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11542–11551, 2021. 2

[33] Pavlo Molchanov, Xiaodong Yang, Shalini Gupta, Kihwan Kim, Stephen Tyree, and Jan Kautz. Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4207–4215, 2016. 2

[34] Louis-Philippe Morency, Ariadna Quattoni, and Trevor Darrell. Latent-dynamic discriminative models for continuous gesture recognition. In *2007 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2007. 2

[35] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 4

[36] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 3

[37] K R Prajwal, Triantafyllos Afouras, and Andrew Zisserman. Sub-word level lip reading with visual attention. In *Proc. CVPR*, 2022. 2, 3

[38] K R Prajwal, Hannah Bull, Liliane Momeni, Samuel Albanie, Gül Varol, and Andrew Zisserman. Weakly-supervised fingerspelling recognition in british sign language videos. In *Proc. BMVC*, 2022. 2, 3

[39] K R Prajwal, Sindhu Hegde, and Andrew Zisserman. Scaling multilingual visual speech recognition, 2025. 4, 5

[40] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6964–6974, 2021. 3

[41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[42] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR, 2023. 2

[43] Akam Rahimi, Triantafyllos Afouras, and Andrew Zisserman. Reading to listen at the cocktail party: Multi-modal speech separation. In *Proc. CVPR*, 2022. 5

[44] Charles Raude, K R Prajwal, Liliane Momeni, Hannah Bull, Samuel Albanie, Andrew Zisserman, and Gül Varol. A tale of two languages: Large-vocabulary continuous sign language recognition from spoken language supervision. *arXiv*, 2024. 2

[45] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. Progressive transformers for end-to-end sign language production. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 687–705. Springer, 2020. 2

[46] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. Signing at scale: Learning to co-articulate signs for large-scale photo-realistic sign language production. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5141–5151, 2022. 2

[47] Rajeev Sharma, Jiongyu Cai, Srivat Chakravarthy, Indrajit Poddar, and Yogesh Sethi. Exploiting speech/gesture co-occurrence for improving continuous gesture recognition in weather narration. In *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*, pages 422–427. IEEE, 2000. 2

[48] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. Learning audio-visual speech representation by masked multimodal cluster prediction. *arXiv preprint arXiv:2201.02184*, 2022. 3, 4

[49] Stephanie Stoll, Necati Cihan Camgoz, Simon Hadfield, and Richard Bowden. Text2sign: towards sign language production using neural machine translation and generative adversarial networks. *International Journal of Computer Vision*, 128(4): 891–908, 2020. 2

[50] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7464–7473, 2019. 3

[51] Laia Tarrés, Gerard I Gállego, Amanda Duarte, Jordi Torres, and Xavier Giró-i Nieto. Sign language translation from instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5625–5635, 2023. 2

[52] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022. 3

[53] Petra Wagner, Zofia Malisz, and Stefan Kopp. Gesture and speech in interaction: An overview, 2014. 2

[54] J. Wan, S. Z. Li, Y. Zhao, S. Zhou, I. Guyon, and S. Escalera. ChaLearn looking at people RGB-D isolated and continuous datasets for gesture recognition. In *2016 IEEE Conference on*

Computer Vision and Pattern Recognition Workshops (CVPRW), pages 761–769, 2016. 2

[55] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14549–14560, 2023. 3

[56] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, Sen Xing, Guo Chen, Junting Pan, Jiashuo Yu, Yali Wang, Limin Wang, and Yu Qiao. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022. 3, 6

[57] Runhua Xu, Nathalie Baracaldo, and James Joshi. Privacy-preserving machine learning: Methods, challenges and directions. *arXiv preprint arXiv:2108.04417*, 2021. 5

[58] Aoxiong Yin, Tianyun Zhong, Li Tang, Weike Jin, Tao Jin, and Zhou Zhao. Gloss attention for gloss-free sign language translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2551–2562, 2023. 2

[59] Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Transactions on Graphics (TOG)*, 39(6):1–16, 2020. 2

[60] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. 3

[61] Yihao Zhi, Xiaodong Cun, Xuelin Chen, Xi Shen, Wen Guo, Shaoli Huang, and Shenghua Gao. Livelyspeaker: Towards semantic-aware co-speech gesture generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20807–20817, 2023. 2

[62] Hao Zhou, Wengang Zhou, Yun Zhou, and Houqiang Li. Spatial-temporal multi-cue network for continuous sign language recognition. In *Proceedings of the AAAI conference on artificial intelligence*, pages 13009–13016, 2020. 2

[63] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, et al. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. *arXiv preprint arXiv:2310.01852*, 2023. 3, 6, 7

[64] Ronglai Zuo and Brian Mak. C2slr: Consistency-enhanced continuous sign language recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5131–5140, 2022. 2