Recognizing Co-Speech Gestures in-the-Wild

Sindhu B Hegde, K R Prajwal, and Andrew Zisserman

Visual Geometry Group, Dept. of Engineering Science, University of Oxford
{sindhu,prajwal,az}@robots.ox.ac.uk
https://www.robots.ox.ac.uk/~vgg/research/grw

Abstract. While humans naturally gesture during speech, only a sparse subset of these co-speech gestures are visually depictive and semantically linked to specific spoken words. In this paper, we introduce a large-scale dataset – *Gesture Recognition in the Wild* (GRW), comprising co-speech gestures corresponding to a diverse vocabulary of 155 words. GRW contains 140k manually annotated video clips where the word is spoken, with 17k instances of semantic co-speech gestures including their frame-level temporal boundaries. The video clips are collected 'in the wild' from public-facing discourse, including lectures, talk shows, and interviews, covering a diverse range of speakers and visual conditions.
We also introduce video models to: (a) classify gestures as semantic or not; (b) recognize the word corresponding to a co-speech gesture; and (c) temporally localize the gesture. These models are trained and evaluated on the GRW dataset and compared against a range of strong baselines, establishing benchmark results for all three tasks. The dataset, annotations, and trained models are publicly available on the project website.

Keywords: gesture recognition · gesture localization

1 Introduction

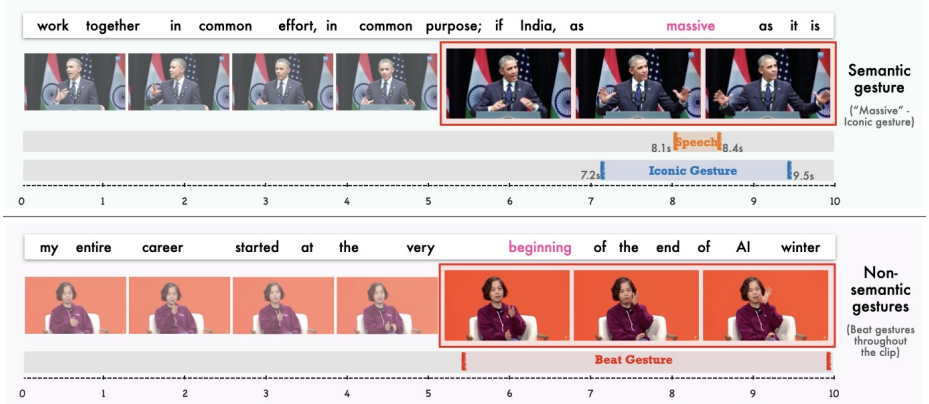

Fig. 1: Semantic vs. non-semantic co-speech gestures. **(Top)** Obama performs an iconic semantic gesture for the word "massive". Notice the significant temporal offset between the physical gesture and the spoken word. **(Bottom)** There are no semantic gestures around the word "beginning". This paper focuses on building dataset and models that can automatically recognize and localize semantic gestures in real-world clips.

Humans naturally move their hands and upper body while speaking, yet only a sparse subset of these movements are *meaning-bearing*. In this work, we focus on **semantic co-speech gestures**, i.e. instances where a speaker produces a clear, visually depictive (iconic, deictic, or metaphoric) movement directly attributable to a specific spoken word (e.g., tracing a *circle* while saying 'circle', or moving the hands *apart* while saying 'massive' in Fig 1). Humans also make other gestures while speaking, such as *beat* gestures (rhythmic emphasis) and incidental motions (fidgeting, resting hands), but these are considered *non-semantic*.

Unfortunately, no large-scale dataset currently exists that has annotations for these semantic co-speech gestures in unconstrained situations. Distinguishing these semantic gestures in-the-wild is challenging: they are relatively sparse in time, highly variable in form, and often require integrating both linguistic and visual context. Consequently, while recent multimodal foundation models for video understanding [26, 40, 42] excel at generic recognition and retrieval, they remain remarkably brittle when tasked with recognizing specific, fine-grained gestures. To address this critical gap, we make the following contributions:

Gesture Recognition in the Wild (GRW) Dataset: We introduce the first large-scale, unconstrained benchmark for semantic co-speech gestures. GRW features manually curated annotations for $\approx$ 140k video clips across 37k unique identities, indicating whether or not a semantic gesture is present. For $\approx$17k semantic instances, we provide word labels and frame-accurate temporal boundaries for both the speech and the corresponding gesture, explicitly accounting for the natural temporal misalignment between the two. The word labels are distributed across a vocabulary of 155 conceptual words.

Insights into dynamics of gesture and spoken words: Leveraging the diversity of our data, we provide a comprehensive analysis of how humans gesture in the wild. We reveal fundamental insights into the temporal envelope of gestures (is the word gestured before it is spoken or after?) and the relative likelihood of words to elicit a semantic gesture (which words are most often gestured?)

New models for three core gesture tasks: We formalize three core gesture understanding tasks: (i) semantic gesture classification – is the gesture semantic or not? (ii) word-level gesture recognition, and (iii) temporal gesture localization. We then develop and train models for these tasks using the GRW dataset. Our results demonstrate the importance of modeling extended temporal motion context for semantic gesture classification. Finally, we show that our models outperform existing state-of-the-art vision-language models, establishing a strong benchmark and foundation for future multimodal gesture research.

2 Related Work

Human gestures are studied across multimodal machine learning, HCI, graphics, and linguistics. Existing resources vary significantly in gesture type (isolated commands vs. co-speech vs. sign language), sensing modality (RGB/RGB-D, 2D/3D pose), and supervision granularity.

Gesture Datasets. Many datasets focus on predefined, isolated command gestures (e.g., swipes, thumbs-up) for HCI applications. Examples include Jester (27

Table 1: Comparison of gesture datasets. The Gesture Recognition in the Wild (GRW) dataset uniquely provides verified semantic word labels and precise temporal gesture boundaries for in-the-wild co-speech gestures while supporting multimodal inputs.

Dataset	Modality	# Identities	Hours	In-the-wild?	Gestured word labels	Temporal Bounds
Isolated / Command-Style						
Jester [30]	RGB	1,376	∼123	✗	✗	✓
EgoGesture [41]	RGB-D	50	24	✗	✗	✓
NVGesture [31]	RGB-D+IR	20	∼10	✗	✗	✓
ChaLearn LAP [39]	RGB-D	21	∼14	✗	✗	✓
Co-Speech Gestures						
Trinity [13]	Mocap+Audio	1	4	✗	✗	✗
TalkingWithHands [24]	Mocap+Audio	50	150	✗	✗	✗
BEAT [27]	Mocap+Audio	30	76	✗	✓	✗
Speech2Gesture [16]	2D+Audio	10	144	✓	✗	✗
PATS [2]	2D+Audio+Text	25	251	✓	✗	✗
AVS-Spot [20]	RGB+Audio+Text	384	0.4	✓	✓	✗
SeamlessInteraction [1]	RGB+Audio+Text	4,000	4,065	✗	✗	✗
Ours (GRW)	**RGB+Audio+Text**	37,409	173	✓	✓	✓

classes) [30], EgoGesture (83 classes) [41], NVGesture [31], SocialGesture [8] (4 deitic gestures in multi-person interaction settings), and the ChaLearn LAP challenges (IsoGD/ConGD) scaling to hundreds of categories [39]. While valuable, these datasets model gestures-as-commands and lack the spoken-word grounding and linguistic alignment necessary for co-speech semantic modeling.

Sign language datasets: Corpora like BOBSL [3], How2Sign [10], WLASL [25], and YouTube-ASL [38] benchmark sign language recognition and generation. Here, hand shape and motion are the primary communication channel rather than an auxiliary speech accompaniment. Although annotated at the word level, their discrete linguistic structure differs fundamentally from co-speech gestures, which are optional, highly variable, and only partially aligned with lexical items, preventing the direct transfer of methods.

Co-speech gesture datasets: High-fidelity 3D motion capture datasets, such as Trinity Speech-Gesture [13] and Talking With Hands [24], provide clean trajectories, but lack diverse speakers and contexts. The BEAT dataset [27] adds emotion and semantic-relevance annotations, categorizing beats, iconic, deictic, and metaphoric gestures. But, these datasets are limited by scale and diversity. A recent large-scale dataset, Seamless Interaction [1] is massive in scale (4000+ speakers, 4000+ hours) but it is recorded in laboratory settings where dialogues are enacted with instructions to the actors, and do not contain gestured word labels either. There are also "in-the-wild" datasets used in gesture synthesis works, that source videos from TED talks and TV shows. For example, Speech2Gesture [16] (144 hours, 10 speakers) and PATS [2] (251 hours, 25 speakers) aggregate auto-detected poses, audio, and transcripts. While critical for scaling generation models, these datasets lead to models that do not generate semantic gestures well, due to lack of specific annotations for the same [33].

Gesture Recognition and Understanding. Historically treated as an action recognition sub-problem, early gesture recognition employs 2D/3D CNNs and sequence models evaluated on isolated datasets [23, 35]. Analyzing co-speech

gestures, however, requires detecting their presence and segmenting temporal boundaries. A CRF-based sequence-labeling approach was proposed [15] to segment object-depicting gestures, however, this was in a constrained setting rather than in-the-wild co-speech gestures. Only recently, JEGAL [20] formalized tasks such as gesture-based retrieval and word spotting using tri-modal (video, speech, text) representations. While aligned with modeling semantics, these benchmarks focus on representation learning and retrieval rather than explicitly recognizing and segmenting semantic gestures with precise, word-level boundaries.

Gesture Synthesis. Gesture generation has historically included rule-based and procedural systems inspired by linguistics. In recent years, data-driven models dominate, enabled by large gesture data and progress in sequence modeling. Speech2Gesture [16] was a seminal work that learned personalized patterns from in-the-wild monologues using auto-extracted poses. Subsequent modern generative models like StyleGestures [4] and ExpressGesture [14] improved realism, control, and coverage. Audio-driven models capture prosody well but struggle with semantic gestures. Incorporating transcripts or text embeddings has been a step towards addressing this. Rhythmic Gesticulator [5] structures explicit rhythm and semantics, while ConvoFusion [32] uses diffusion to emphasize specific words. However, these models still predominantly rely on speech stress and prosody, noting that sparse semantic gestures remain difficult to capture due to a lack of non-beat annotations in training data.

3 The Gesture Recognition in the Wild (GRW) Dataset

Fig. 2: Samples from the **GRW** dataset. (a) Lexical and kinematic diversity of semantic gestures. (b) Even when gesturing the same conceptual word ("SPIRAL"), speakers employ different physical motions. The dataset captures these gestures across a wide variety of speakers, poses, and camera angles.

The GRW dataset comprises 139,503 manually annotated video clips curated from diverse, in-the-wild speaker environments. The core dataset features 17,340 precisely localized positive instances of semantic gestures, distributed across a broad vocabulary of 155 conceptual words. The dataset also comprises 122,163 verified negative samples where a target vocabulary word is spoken, but no corresponding semantic gesture is physically performed. Finally, we also include an automatically labeled set of 67,214 clips (refer to Sec 4.4) with word labels and approximate gesture boundaries that can be used for large-scale pre-training.

Table 2: Overview of the GRW dataset splits. It contains manually verified data for evaluation and finetuning, and large-scale pseudo-labeled data for pre-training.

Task / Split	Total Clips	Hours	Vocab	Semantic	Non-Semantic
Semantic Gesture Classification					
Train Split	135,503	149.26	155	15,340	120,163
Test Split	4,000	4.41	100	2,000	2,000
Total	**139,503**	**153.67**	**155**	**17,340**	**122,163**
Word Recognition & Localization					
Pre-train Split (noisy)	67,214	74.26	155	67,214	-
Train Split (clean)	15,340	16.94	155	15,340	-
Test Split (clean)	2,000	2.21	100	2,000	-
Total (clean)	**17,340**	**19.15**	**155**	**17,340**	-

In Fig 2, we show a few representative samples from the dataset. Unlike laboratory-recorded isolated gesture datasets, GRW captures gestures in in-the-wild conditions public-facing discourse with natural intra- and inter-class variations. Fig. 3 details the 155-word vocabulary across semantic domains ranging from literal actions ('Translational', Pointing') to complex concepts ('State/process', 'Physical'). Node sizes (reflecting frequency) confirm a robust distribution across common anchors (e.g., 'bye', 'together') and nuanced descriptors (e.g., 'spiral', expand'), establishing GRW as a comprehensive benchmark for real-world gesture understanding. Each sample provides video, audio, word-level transcripts, and a binary semantic gesture label; positive samples additionally include the lemmatized target word w and its precise temporal gesture boundary $\mathcal{B}$.

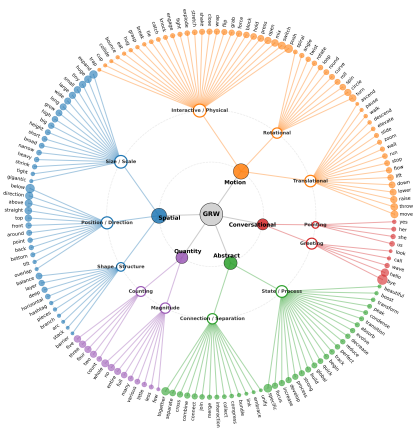

Fig. 3: Semantic Taxonomy of the GRW Dataset. We organize our 155-word vocabulary into a three-tiered hierarchy, radiating from five high-level semantic domains to specific target words (leaf nodes). Node size is strictly proportional to the frequency of annotated clips per word.

4 Dataset Curation Pipeline

In this section, we describe the pipeline used to curate the GRW word-level semantic gesture dataset and benchmark. The principal steps are: (i) large-scale automatic mining and filtering from a word-aligned speech corpus; and (ii) multi-stage human annotation with redundancy and verification. Table 3 shows the number of clips and the yield of semantic gestures at these stages of the pipeline.

4.1 Mining and filtering to obtain candidate video clips

The objective is to obtain a large set of word-aligned video clips potentially containing associated co-speech gestures.

Table 3: Overview of the GRW dataset curation pipeline. We detail the progression of our data from raw video filtering to final manual annotation. After rigorous quality checks, we obtain over 17k annotated semantic clips.

Pipeline Stage	Total Clips	Yield	Hours	Vocab	Non-Semantic	Notes
Raw videos	1,529,572	-	1,699.52	200	n/a	Sourced from MultiVSR
Mining and filtering	525,226	-	578.40	185	n/a	Keypoint filtering
Assigned for manual labeling	161,828	-	178.21	185	n/a	Output of Section 4.1
After recognition annotations	19,364	11.9%	21.20	155	122,163	Output of Section 4.2
After localization annotation	17,340	10.7%	19.15	155	n/a	Output of Section 4.3

Video data source. We use the English subset of MultiVSR [34], a large-scale multilingual visual speech recognition dataset with word-aligned transcripts, as our in-the-wild video pool. Sourced from YouTube, MultiVSR contains extended full-length versions of the AVSpeech [12] videos. The videos exhibit substantial real-world variability across lighting, viewpoints, backgrounds, and speaker demographics. We pre-process the videos to obtain stable upper-body gesture crops. Details on pre-processing are given in the arXiv version.

Lexicon of gesturable words. Given the MultiVSR English vocabulary, we target *plausibly gesturable words* for iconic, metaphoric, or deictic gestures (e.g., motion verbs, spatial relations, size/shape adjectives), excluding high-frequency function words that rarely elicit semantic depiction. To systematically construct this lexicon, we extract the most frequent lemmas from the dataset's transcripts and prompted an LLM (Gemini [17]) to score each word based on its likelihood to elicit a semantic gesture. The top-scoring candidates are then manually verified, yielding a final set of 200 English lemmas expected to induce distinctive gestures. For every match in the dataset to a lemma on this list, we extract a 4-second temporal window centered around the word timestamp. This yields a large collection of candidate clips (1.5M clips), potentially containing semantic gestures. However, this initial pool contains both true positives and substantial noise (e.g., no-gesture segments, beat-only motion, hands out of frame).

Visual filtering for gestures. To reduce annotation cost and improve positive sample yield, we apply automatic pose keypoint filtering [28]. For each candidate clip, we compute: (i) the fraction of frames with confident hand/wrist detections, and (ii) the mean temporal hand displacement (velocity), $|p_t - p_{t-1}|_2$, with p_t denoting the frame t hand centroid. Candidates failing minimum visibility or motion thresholds are discarded. This keypoint-based filtering retains approximately 525k candidates across 185 target words. As the mined distribution is highly long-tailed, naive sampling would over-represent high-frequency words. We therefore construct a semi-balanced $\approx$ 162k clip subset for manual annotation, sampling 100 to 1500 instances per word to ensure low-frequency lemma coverage and prevent high-frequency dominance.

4.2 Manual annotation for word recognition

The objective is to partition the 161,828 candidate clips into two sets: those containing a co-speech gesture corresponding to a target word, and those that don't (e.g., beat or random gestures). Both sets are retained for subsequent model training.

Stage-1: Annotating semantic gestures. The annotators assign a binary label to each of the candidate samples, indicating the presence of a corresponding semantic gesture for a target word. The annotators are trained to only mark a clip as 'Positive' if there is a clearly identifiable gesture that depicts the target word. We developed a custom web-based interface using (VIA) [11] (details in arXiv version). For each target word, annotators review a grid of associated video clips (with audio playback) and assign the binary label. This selection stage required approximately 1,000 annotation hours.

Stage-1: Quality control. Each sample is independently labeled by five annotators, screened using a held-out set of 500 gesture clips. To reduce mistakes post-annotation, we aggregate labels via majority voting, discarding words with high disagreement, yielding roughly 19k semantic videos (11.9% positive rate).

4.3 Manual annotation for word localization

Stage-2: Annotating temporal gesture boundaries. The input to this stage is the set of 19k output semantic gesture clips obtained from Section 4.2. For each clip, annotators mark the start-end timestamps of the gesture corresponding to the target word. This task is more demanding than binary recognition, as it requires precise temporal localization through scrubbing, frame stepping, and repeated viewing. The interface and detailed annotation guidelines are provided in the arXiv version. This stage required $\approx$ 500 hours of annotation efforts.

Stage-2: Quality control. To ensure gesture boundary reliability, we retain only those samples for which *at least three annotators* provide start-end timestamps that differ by less than 1-second. This filtering yields 17,340 samples containing semantically grounded gestures with reliable word-level temporal boundaries, which becomes our final GRW semantic dataset. This amounts to 19.15 hours of clean data.

4.4 Curating a pre-train set

To scale downstream word recognition and temporal localization without prohibitive manual annotation costs, we employ a pseudo-labeling strategy. By applying a strict confidence threshold to the classifier's predicted probability of a semantic gesture (≥ 0.9) to our trained semantic classifier (Section 6.2), we automatically mine 67,214 high-confidence positive clips from the $\approx$ 363k unannotated candidates (Table 3). For weak localization labels, we pad each clip's speech boundaries using word-specific average start-end frame offsets derived from the clean GRW train split. This weakly-labeled set is then used to pretrain our recognition and localization models.

5 Dataset Analysis

In this section, we present an initial analysis of the co-speech gestures in the GRW dataset. We are interested in three questions: (i) how likely a given word is to be accompanied by a semantically meaningful gesture; (ii) how the timing of the gesture segment relates to the spoken word segment; and (iii) how variable the gestures are for the same word across speakers and context. Prior

works [20, 27] have investigated some of these questions, but have been in laboratory settings or with limited scale and diversity of the data – potentially restricting the scope and generality of their findings.

5.1 How likely are the words to be gestured?

Fig 4 shows the proportion of gestured instances among all clips submitted for annotation. The likelihood that a word is accompanied by a semantically meaningful gesture varies dramatically across the lexicon. While some words are gestured very frequently, for example, 'bye' is gestured approximately 68%, while others are rarely depicted. At the extreme low end, 'look' is gestured in 1% of cases. This long-tailed distribution highlights the sparsity of semantic gestures in natural discourse and motivates the need for large-scale filtering and careful annotation to obtain high-quality gesture–word pairs.

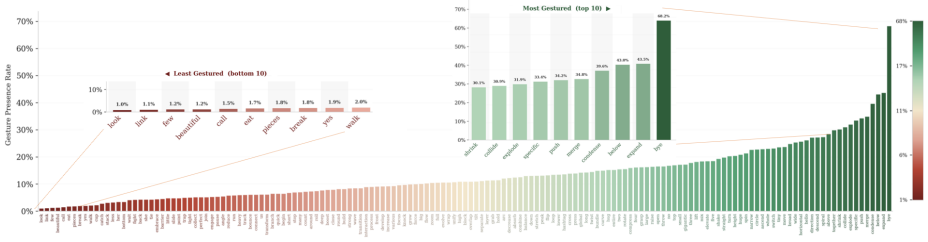

Fig. 4: Not all words are gestured equally: We plot the likelihood that a spoken word is accompanied by a depictive semantic gesture. Abstract concepts and common verbs (left inset) rarely elicit semantic gestures ($< 2\%$), whereas inherently spatial or physical actions like push, below, and bye (right inset) are gestured with high frequency.

5.2 Gesture vs. speech boundaries

We analyze the temporal relationship between spoken words and their semantic gestures on a larger scale and find that gesture–speech alignment is loose and highly variable. The Violin plot in Fig 5(a) shows that the average duration of spoken word segments is much shorter than their corresponding gesture segments across all annotated words. Fig 5 (b) quantifies this mismatch. Gestures start before spoken word (97.5%) and end after spoken word (85.7%). This also corroborates the findings in [37]. Finally, Fig 5 (c) visualizes averaged activation heatmaps for a few word classes, showing that gestures are often distributed asymmetrically around the speech center. Table 4 compares the average durations of speech and gestures for selected words. Together, these analyses highlight that gesture boundaries are loosely aligned with speech, making word-level gesture recognition and gesture localization particularly challenging.

Table 4: Average speech and gesture durations for a few words from the GRW dataset. Values are (speech | gesture) in seconds.

join	condense	two	interaction	knock	spiral	hello
0.31 | 1.81	0.50 | 1.30	0.22 | 1.34	0.60 | 2.07	0.30 | 1.56	0.46 | 2.17	0.34 | 1.01

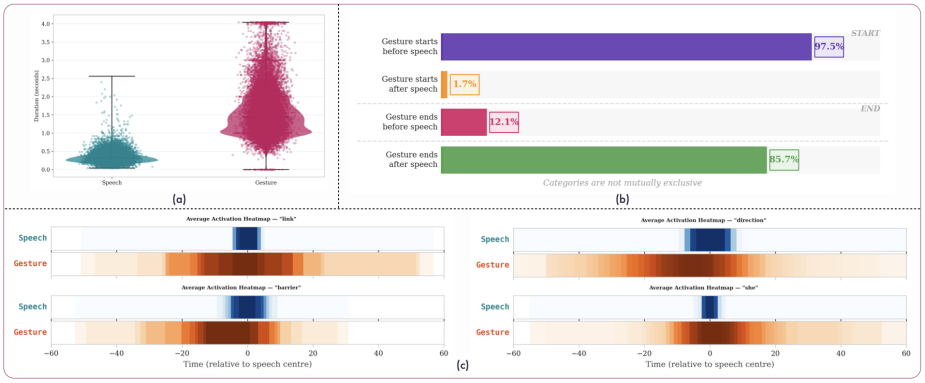

Fig. 5: Gesture-speech temporal alignment. Semantic gestures rarely align perfectly with spoken words. As shown in (a), gestures are significantly longer in duration than speech. (b) The vast majority of gestures start before (97.5%) and end after (85.7%) the target word is spoken. (c) Activation heatmaps (aggregated across all samples for a specific word) visually confirm this "envelope" effect.

5.3 Variations in gestures

Co-speech gestures vary substantially across speakers and contexts: a single word can elicit multiple distinct movements, and different words may share visually similar gestures. For example (Fig. 2), speakers depict 'spiral' using one or both hands, via vertical or horizontal motions. Similarly, 'zoom' elicits spreading palms, opening fingers, or moving hands apart. Conversely, 'shrink' is gestured by bringing hands inward, pinching fingers, or pressing palms together.

6 Recognizing Co-speech Gestures

We introduce two complementary co-speech gesture recognition models that are trained on GRW: (i) a **semantic gesture classification** model that predicts whether a short video clip contains a clear semantic (iconic, deictic, metaphoric) gesture or not, and (ii) a **gestured word recognition & localization** model that (a) predicts the gestured word class, and (b) localizes the gesture segment in time within a short clip. The architecture of the two models is illustrated in Fig 6. Both models are built on top of a frozen visual backbone pretrained to extract per-frame hand shape and position features. The features are then ingested by a series of transformer blocks for task-specific temporal modeling and prediction. In the following, we first describe the common visual backbone and then the two task specific transformer architectures.

6.1 Gesture backbone

Given a video clip $V \in \mathbb{R}^{T\times H\times W\times 3}$, we use SHuBERT [18] as our pretrained gesture backbone to extract frame-level features. Pretrained on large-scale sign-language data [38], SHuBERT provides robust visual features that inherently capture fine-grained hand shapes and semantically distinctive motions. We extract the hidden states from L layers of this Transformer-based model to construct a multi-layer sequence tensor $f_{\text{multi}} \in \mathbb{R}^{L\times T\times d_{in}}$. Here, $d_{in} = 768$ repre-

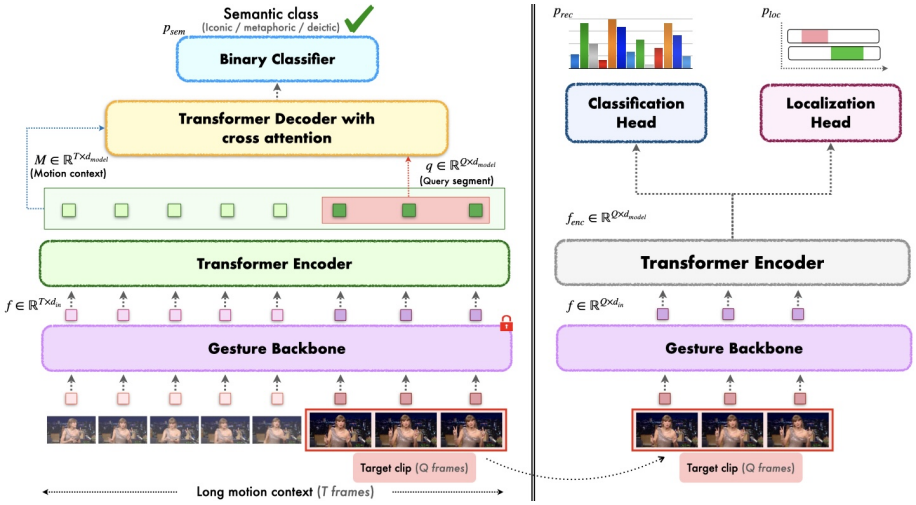

Fig. 6: Left: A binary semantic gesture classifier for a query clip, given a broad temporal motion context. The model uses cross-attention between a target query clip and the motion context to determine if the gesture within the query clip is semantic or not. **Right:** The word recognition and localization model operates on the query clip and is trained to classify the word class, and the precise temporal boundary of the gesture (per-frame binary labels).

sents the feature dimension, and T is the number of frames (corresponding to approximately 4-10 seconds of video).

Weighted layerwise feature aggregation. To optimally fuse SHuBERT's layers, an MLP processes f_{multi} to generate L scalar scores, which are softmax-normalized into mixing weights α. A weighted sum across the layer dimension then collapses this into a single sequence $f \in \mathbb{R}^{T \times d_{in}}$, serving as the direct input to our downstream task-specific models (see arXiv version for details).

6.2 Semantic Co-speech Gesture Classification

The semantic gesture classifier is designed to answer: *Does a video clip contain a clear semantic gesture or not, given the surrounding motion context?* Note that the "motion-context" is important here to spot semantic gestures in a reliable way in unconstrained co-speech gestures.

Intuition. Semantic gestures are typically sparse, arbitrary, and highly variable across different words and are not necessarily unique to a person. In contrast, non-semantic beat gestures are frequent, repetitive, and serve as a person's default gesturing rhythm. By encoding a long temporal window, the model learns the speaker's baseline kinematics for non-semantic beat gestures and for their random gestures. This context makes it significantly easier for the model to contrast, isolate, and identify the sparse, unique semantic gestures, as such gestures physically deviate from the speaker's established baseline.

Task formulation. Given a *long* context video with T frames and a *query interval* $q = [s, e]$ with $Q = e - s + 1$, the goal is to predict a binary label

$y \in \{0, 1\}$, where $y = 1$ indicates that the query interval contains a *clear semantic gesture* and $y = 0$ indicates no semantic gesture.

Architecture. Visual features f are projected to d_{model} via a 2-layer MLP (LayerNorm, ReLU) and processed by a Transformer Encoder. To model relative motion timing across the long context (up to 10 seconds), we apply Rotary Position Embeddings (RoPE) [36], yielding a context-aware memory $M \in \mathbb{R}^{T \times d_{model}}$. Features spanning the target interval $[q_{start}, q_{end}]$ are extracted to form a query sequence $q \in \mathbb{R}^{Q \times d_{model}}$. This query cross-attends to the full memory M within a Transformer Decoder, effectively referencing the full motion context when interpreting the query interval. Finally, the decoder output is temporally mean-pooled and passed through an MLP with a sigmoid activation to yield the binary semantic probability p_{sem}.

Training. We train the semantic gesture classifier using the 'Semantic Gesture Classification' split of the GRW dataset in Table 2, consisting of 135,503 training examples, which exhibits a natural class imbalance (15,340 positive semantic gesture samples and 120,163 negative samples). To address this, we use class-balanced sampling to upsample positive class data in each batch, and additionally apply a class-weighting to the loss (assigns higher weight to positive samples). During training, we extract a long temporal context of $T = 250$ frames (10 seconds) and define a target query interval of $Q = 100$ frames (4 seconds). The model is optimized using a Binary Cross-Entropy (BCE) loss.

6.3 Co-speech Gesture Word Recognition and Localization

Given a *short* query clip containing a semantic gesture somewhere inside it, the goal is to predict: (i) the class of the word: $c \in \{1, \ldots, C\}$ corresponding to the co-speech gesture, and (ii) per-frame gesture presence: $\mathbf{a} = [a_1, \ldots, a_Q], \quad a_q \in \{0, 1\}$, where $a_q = 1$ indicates that frame q lies within the gesture interval and $a_q = 0$ otherwise. We use the SHuBERT features f_{multi}, which are then aggregated layer-wise to obtain $f \in \mathbb{R}^{Q \times d_{in}}$, same as in Section 6.1. These features are then temporally encoded using a Transformer encoder to get f_{enc}.

Gesture localization. To achieve precise temporal localization without requiring complex region-proposal pipelines, we formulate the alignment task as a dense, frame-level binary classification. An MLP + sigmoid operates directly on per-frame encoder features f_{enc} to get per-frame probabilities p_{loc}.

Word classification. We do a weighted average of the Transformer output features f_{enc} across Q frames to form a single, global video representation. The weights are determined by p_{loc}, which provides a per-frame probability of the relevant gesture being present. A classification head consisting of a two-layer MLP projects this global embedding to produce softmax probabilities p_{rec} over the C semantic word classes.

6.4 Training the Word Recognition and Localization Model

The training is divided into two steps: (i) the model is pre-trained on the pseudo-labeled GRW pre-training data (67k samples) – this is weak supervision because the pseudo-labeled data only has approximate gesture boundaries (only the word

Table 5: Semantic gesture classification performance. Our approach surpasses existing baselines, especially at high confidence thresholds (≥ 0.9). This validates its utility for extracting clean semantic gestures from unlabeled data.

Method	Accuracy (%)	Precision (%)	Recall (%)	Semantic class accuracy (%) for high-conf samples
Random	50.00	50.00	50.00	n/a
Frozen features				
Clip4Clip [29]	59.13	55.64	52.34	62.95
Sapiens [22]	60.10	60.92	57.80	68.01
SemMom_{Dinov3} [21]	61.77	60.91	58.85	67.31
Fine-tuned features				
GestSync [19]	61.58	59.90	62.17	64.91
JEGAL [20]	63.94	61.02	59.94	66.93
VLMs				
Intern-VL [9]	61.58	56.67	58.35	66.68
Qwen3-VL [6]	55.05	53.24	51.28	58.61
Gemini [17]	<u>67.30</u>	<u>65.78</u>	**71.10**	<u>68.21</u>
Ours	**75.83**	**79.91**	<u>69.00</u>	**93.20**

is known); (ii) the model is fine-tuned on the manually annotated GRW word recognition training dataset (15k samples) – this is strong supervision as the gesture boundaries are precisely annotated.

Weakly-supervised pre-training. We pre-train the word recognition and localization model on the 67,214 pseudo-labeled positive samples. Frames falling within the padded speech boundaries are treated as positives, and the rest are negative frames. The model is supervised using frame-level BCE loss for temporal localization, combined with a categorical cross-entropy loss for gestured word classification:

$$\mathcal{L} = -\sum_{c=1}^{C} y_c \log(p_{rec,c}) - \lambda_{loc} \frac{1}{Q} \sum_{q=1}^{Q} [a_q \log(p_{loc,q}) + (1 - a_q) \log(1 - p_{loc,q})]$$

where y_c is the one-hot word label, $p_{rec,c}$ is the predicted word class probability, a_q is the 'ground-truth' binary frame label, and $p_{loc,q}$ is the predicted frame-level gesture probability at frame q.

Fine-tuning on clean data. Finally, we fine-tune the recognition and localization model exclusively on the manually verified train split of 15,340 samples (Table 2). We use the same loss functions as in the pre-training stage.

7 Results

In this section, we evaluate on the unseen test set of GRW for semantic gesture classification (Section 7.1), and gestured word recognition & localization (Section 7.2). In addition to evaluating on the test sets of GRW, we also evaluate the gesture classifier on unseen words, and the localization ability of the gesture recognizer on the test set of [20]. The baseline models used for comparison and the task-specific evaluation metrics are described in the corresponding subsections.

Baselines. We evaluate on a diverse set of baselines, starting with strong pretrained video-language representations as *frozen* feature extractors, training only a lightweight classifier on top. We use Clip4Clip [29], Sapiens [22], and SemanticMoments (on Dinov3 features) [21], to measure how far generic pretrained representations go *without* end-to-end adaptation. Next, we compare with the recent state-of-the-art co-speech gesture models that learn dense frame-level representations. We finetune GestSync [19] and JEGAL [20] with additional task-specific prediction heads on the exact same data as our approach. Finally, we evaluate two state-of-the-art open-source VLMs, InternVL [9], Qwen3-VL [6] and a production-grade VLM, Gemini-3.5 in a prompted setting: given a video clip and a list of candidate word classes, the model is asked to: (i) classify if the video clip contains a semantic gesture, and (ii) recognize the gestured word and localize the temporal boundaries (start and end time) of the corresponding gesture instance. Detailed prompts are provided in the arXiv version.

7.1 Semantic Gesture Classification

Metrics: We evaluate performance using *Accuracy*, *Precision*, and *Recall* to account for inherent class imbalance. Additionally, to evaluate the classifier's reliability for generating pseudo-labels, we report its accuracy specifically on high-confidence predictions ($p_{sem} \geq 0.9$).
Results: Table 5 demonstrates that our model outperforms baseline methods across the board. Notably, it excels at high-confidence prediction, confirming that utilizing a ≥ 0.9 threshold effectively isolates semantic gestures in unlabeled data with minimal label noise.

Is the semantic classifier word-class agnostic? We evaluate this aspect by creating a new test set with 10 unseen words. Specifically, we manually annotate 500 new clips with binary labels indicating whether the clips contain a semantic gesture or not. We ensure that these clips do not contain any words that are present in the GRW dataset. We measure semantic class accuracy at high confidence thresholds for both our model and Gemini, which comes out to be 86.5% and 63.78% respectively. This indicates that our model generalizes across word classes and can be used for large-scale semantic gesture mining in future works.

7.2 Word Recognition and Localization

Metrics: We report top-k accuracy for recognition on the unseen test set of 100 word classes, for $k \in \{1, 5, 10\}$. For localization, we report mIoU to measure the temporal overlap between the ground-truth and predicted gesture segment.
Results: Table 6 shows the word recognition and localization results. We clearly outperform all the baselines on both tasks by a large margin. Gemini is particularly weaker on fine-grained gesture localization task. We also note that Acc@5 is a more indicative measure of performance due to the ambiguity between certain classes; for example, pairs such as *(bye, hello)* and *(little, small)* are often expressed using similar gestures.

Fig 7 provides qualitative examples demonstrating the model's accurate word recognition together with precise gesture localization results. Additionally, the

Table 6: Word-level gesture recognition & localization performance. Our model outperforms baselines on top-k recognition accuracy and temporal localization (mIoU).

Method	Recognition			Localization
	Acc @ 1	Acc @ 5	Acc @ 10	mIoU
Random	1.00	5.00	10.00	0.1975
Frozen features				
Clip4Clip [29]	8.70	19.25	27.60	n/a
Sapiens [22]	9.45	18.80	28.60	n/a
SemMom$_{Dinov3}$ [21]	10.05	21.15	29.65	n/a
Fine-tuned features				
GestSync [19]	9.12	20.37	30.18	0.5172
JEGAL [20]	10.43	22.71	31.88	0.5368
VLMs				
Intern-VL [9]	4.35	12.50	15.71	0.3932
Qwen3-VL [6]	5.92	13.19	19.27	0.2071
Gemini-3.5 [17]	13.95	30.15	41.10	0.4658
Ours	**18.35**	**37.30**	**51.90**	**0.6726**

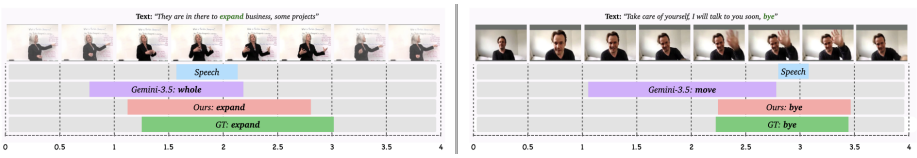

Fig. 7: We visualize sample predictions for joint word recognition and localization. In both instances, our model correctly assigns the target word (expand, bye) and closely matches the ground truth temporal gesture boundaries.

confusion matrices in Fig. 8 show that most misclassifications occur between semantically related words, highlighting the inherent ambiguity in gesture-based word recognition.

Generalization to other datasets. We evaluate on the AVS-Spot benchmark [20], which contains 500 samples with approximate temporal boundary annotations. The goal is to localize the semantic gesture in time, and the prediction is treated as correct if the predicted gesture segment has any overlap with the ground-truth segment. We obtain the gesture timestamps from our model's localization head, and compare the semantic gesture spotting accuracy directly with JEGAL [20]. We achieve an accuracy of **77.3%** compared to JEGAL's 63.6%, which is a significant improvement on an existing gesture benchmark.

7.3 Ablation Studies

Impact of context length. For the semantic gesture classification task, our approach leverages a substantially longer temporal context to determine the presence of a semantic gesture within a short target clip. As shown in Table 7, providing 10 seconds of temporal context for classifying a 4 second target segment improves accuracy by 5.9%.

Table 7: Providing a temporal context of 10s (250 frames) to make a prediction for a 4s (100 frames) target segment improves the accuracy by 5.9%.

	100 frames	150 frames	200 frames	250 frames	300 frames
High-conf Acc. (%)	87.27	89.82	91.56	**93.20**	92.95

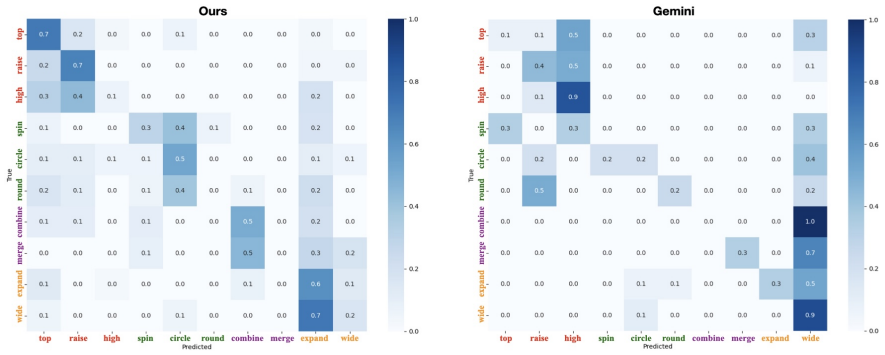

Fig. 8: Comparison of the confusion matrices for our model and Gemini. Most misclassifications occur between semantically related words.

Table 8: Pre-training before fine-tuning on the clean train set gives a clear performance improvement across all the metrics.

Method	Recognition			Localization
	Acc @ 1	Acc @ 5	Acc @ 10	mIoU
Train using Pre-train	11.50	28.15	38.70	0.5737
Train using Clean	16.35	35.60	49.40	0.6473
Pre-train + FT on clean	**18.35**	**37.30**	**51.90**	**0.6726**

Impact of pre-training. Table 8 shows that pre-training on noisy pre-training set improves both recognition and localization performance. Using manually verified labels provides a further boost. Interestingly, even without any manually labeled data, the model achieves competitive performance, suggesting that large-scale weakly supervised pre-training is a promising approach for these tasks.

8 Conclusion

We have introduced the GRW dataset, providing a large-scale benchmark for analysing, training, and evaluating models for in-the-wild co-speech gestures. We study the temporal offsets between speech and gestures, and also propose new models for three downstream gesture tasks using the dataset. We acknowledge, however, that our findings are conditioned on the nature of our source material – since GRW is derived from public-facing discourse such as lectures, talk shows, and interviews, the observed behaviors may reflect a specific register of formal communication. The gesture patterns may be slightly different in social settings or private conversations [7]. Despite this caveat, GRW represents a significant leap in scale and annotation granularity over existing benchmarks, enabling models to associate semantic gestures with spoken words.

Acknowledgements. We would like to thank Taein Kwon, Piyush Bagad, Jaesung Huh, and Paul Engstler for their valuable discussions. We are also grateful to Alyosha Efros, Jitendra Malik, and Justine Cassell for their valuable feedback and insightful suggestions; and to Abhishek Dutta, Prasanna Sridhar and David Pinto for setting up the data annotation tool, and Ashish Thandavan for his support with the infrastructure. This research is funded by EPSRC Programme Grant VisualAI EP/T028572/1 and a Royal Society Research Professorship RSRP\R\241003. Sindhu is supported by a Google PhD Fellowship.

References

1. Agrawal, V., Akinyemi, A., Alvero, K., Behrooz, M., Buffalini, J., Carlucci, F.M., Chen, J., Chen, J., Chen, Z., Cheng, S., et al.: Seamless interaction: Dyadic audiovisual motion modeling and large-scale dataset. arXiv preprint arXiv:2506.22554 (2025)
2. Ahuja, C., Lee, D.W., Ishii, R., Morency, L.P.: No gestures left behind: Learning relationships between spoken language and freeform gestures. In: Findings of the association for computational linguistics: EMNLP 2020. pp. 1884–1895 (2020)
3. Albanie, S., Varol, G., Momeni, L., Bull, H., Afouras, T., Chowdhury, H., Fox, N., Woll, B., Cooper, R., McParland, A., Zisserman, A.: BOBSL: BBC-Oxford British Sign Language Dataset. arXiv (2021)
4. Alexanderson, S., Henter, G.E., Kucherenko, T., Beskow, J.: Style-controllable speech-driven gesture synthesis using normalising flows **39**(2), 487–496 (2020)
5. Ao, T., Gao, Q., Lou, Y., Chen, B., Liu, L.: Rhythmic gesticulator: Rhythm-aware co-speech gesture synthesis with hierarchical neural embeddings. ACM Transactions on Graphics (TOG) **41**(6), 1–19 (2022)
6. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
7. Bavelas, J.B., Chovil, N., Coates, L., Roe, L.: Gestures specialized for dialogue. Personality and social psychology bulletin **21**(4), 394–405 (1995)
8. Cao, X., Virupaksha, P., Jia, W., Lai, B., Ryan, F., Lee, S., Rehg, J.M.: Socialgesture: Delving into multi-person gesture understanding. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 19509–19519 (2025)
9. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 24185–24198 (2024)
10. Duarte, A., Palaskar, S., Ventura, L., Ghadiyaram, D., DeHaan, K., Metze, F., Torres, J., Giro-i Nieto, X.: How2sign: a large-scale multimodal dataset for continuous american sign language. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2735–2744 (2021)
11. Dutta, A., Zisserman, A.: The VIA annotation software for images, audio and video. In: Proceedings of the 27th ACM International Conference on Multimedia. ACM (2019)
12. Ephrat, A., Mosseri, I., Lang, O., Dekel, T., Wilson, K., Hassidim, A., Freeman, W.T., Rubinstein, M.: Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation. ACM Transactions on Graphics (TOG) **37**(4), 1–11 (2018)
13. Ferstl, Y., McDonnell, R.: Investigating the use of recurrent motion modelling for speech gesture generation. In: Proceedings of the 18th international conference on intelligent virtual agents. pp. 93–98 (2018)
14. Ferstl, Y., Neff, M., McDonnell, R.: Expressgesture: Expressive gesture generation from speech through database matching. Computer Animation and Virtual Worlds **32**(3-4), e2016 (2021)
15. Ghaleb, E., Burenko, I., Rasenberg, M., Pouw, W., Uhrig, P., Holler, J., Toni, I., Özyürek, A., Fernández, R.: Co-speech gesture detection through multi-phase sequence labeling. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 4007–4015 (2024)

16. Ginosar, S., Bar, A., Kohavi, G., Chan, C., Owens, A., Malik, J.: Learning individual styles of conversational gesture. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3497–3506 (2019)
17. Google: Gemini 3.5: Frontier intelligence with action. https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-5/ (2026)
18. Gueuwou, S., Du, X., Shakhnarovich, G., Livescu, K., Liu, A.H.: Shubert: Self-supervised sign language representation learning via multi-stream cluster prediction. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (2025)
19. Hegde, S., Zisserman, A.: Gestsync: Determining who is speaking without a talking head. In: BMVC (2023)
20. Hegde, S.B., Prajwal, K., Kwon, T., Zisserman, A.: Understanding co-speech gestures in-the-wild. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9977–9987 (2025)
21. Huberman, S., Goldberg, K., Patashnik, O., Benaim, S., Mokady, R.: Semanticmoments: Training-free motion similarity via third moment features. arXiv preprint arXiv:2602.09146 (2026)
22. Khirodkar, R., Bagautdinov, T., Martinez, J., Zhaoen, S., James, A., Selednik, P., Anderson, S., Saito, S.: Sapiens: Foundation for human vision models. In: European Conference on Computer Vision. pp. 206–228. Springer (2024)
23. Köpüklü, O., Gunduz, A., Kose, N., Rigoll, G.: Real-time hand gesture detection and classification using convolutional neural networks. In: 2019 14th IEEE international conference on automatic face & gesture recognition (FG 2019). IEEE (2019)
24. Lee, G., Deng, Z., Ma, S., Shiratori, T., Srinivasa, S.S., Sheikh, Y.: Talking with hands 16.2 m: A large-scale dataset of synchronized body-finger motion and audio for conversational motion analysis and synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 763–772 (2019)
25. Li, D., Rodriguez, C., Yu, X., Li, H.: Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 1459–1469 (2020)
26. Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P., Yuan, L.: Video-llava: Learning united visual representation by alignment before projection. In: Proceedings of the 2024 conference on empirical methods in natural language processing. pp. 5971–5984 (2024)
27. Liu, H., Zhu, Z., Iwamoto, N., Peng, Y., Li, Z., Zhou, Y., Bozkurt, E., Zheng, B.: Beat: A large-scale semantic and emotional multi-modal dataset for conversational gestures synthesis. In: European conference on computer vision. pp. 612–630. Springer (2022)
28. Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.L., Yong, M.G., Lee, J., et al.: Mediapipe: A framework for building perception pipelines. arXiv preprint arXiv:1906.08172 (2019)
29. Luo, H., Ji, L., Zhong, M., Chen, Y., Lei, W., Duan, N., Li, T.: Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. Neurocomputing **508**, 293–304 (2022)
30. Materzynska, J., Berger, G., Bax, I., Memisevic, R.: The jester dataset: A large-scale video dataset of human gestures. In: Proceedings of the IEEE/CVF international conference on computer vision workshops (2019)
31. Molchanov, P., Yang, X., Gupta, S., Kim, K., Tyree, S., Kautz, J.: 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. pp. 4207–4215 (2016)
32. Mughal, M.H., Dabral, R., Habibie, I., Donatelli, L., Habermann, M., Theobalt, C.: Convofusion: Multi-modal conversational diffusion for co-speech gesture synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1388–1398 (2024)
33. Nyatsanga, S., Kucherenko, T., Ahuja, C., Henter, G.E., Neff, M.: A comprehensive review of data-driven co-speech gesture generation **42**(2), 569–596 (2023)
34. Prajwal, K., Hegde, S., Zisserman, A.: Scaling multilingual visual speech recognition. In: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2025)
35. Quiros, L.C., Tax, D.M.J., Hung, H.: Gestures in-the-wild: Detecting conversational hand gestures in crowded scenes using a multimodal fusion of bags of video trajectories and body worn acceleration. IEEE Trans. Multim. **22**(1), 138–147 (2020)
36. Su, J., Lu, Y., Pan, S., Wen, B., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding. CoRR **abs/2104.09864** (2021)
37. Ter Bekke, M., Drijvers, L., Holler, J.: Co-speech hand gestures are used to predict upcoming meaning. Psychological Science **36**(4), 237–248 (2025)
38. Uthus, D., Tanzer, G., Georg, M.: Youtube-asl: A large-scale, open-domain american sign language-english parallel corpus. Advances in Neural Information Processing Systems pp. 29029–29047 (2023)
39. Wan, J., Zhao, Y., Zhou, S., Guyon, I., Escalera, S., Li, S.Z.: Chalearn looking at people rgb-d isolated and continuous datasets for gesture recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 56–64 (2016)
40. Wang, Y., Li, K., Li, Y., He, Y., Huang, B., Zhao, Z., Yin, H., Chen, J., Jin, T., Wu, J., et al.: Internvideo2: Scaling video foundation models for multimodal video understanding. In: European Conference on Computer Vision. Springer (2024)
41. Zhang, Y., Cao, C., Cheng, J., Lu, H.: Egogesture: A new dataset and benchmark for egocentric hand gesture recognition. IEEE Transactions on Multimedia pp. 1038–1050 (2018)
42. Zhu, B., Lin, B., Ning, M., Yan, Y., Cui, J., Wang, H., Pang, Y., Jiang, W., Zhang, J., Li, Z., et al.: Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. arXiv preprint arXiv:2310.01852 (2023)