Scientific Data

https://doi.org/10.1038/s41597-026-06872-6

Article in Press

Sign4all: a Spanish Sign Language dataset

Received: 7 August 2025

Accepted: 9 February 2026

Cite this article as: Morillas-Espejo, F., Martinez-Martin, E. Sign4all: a Spanish Sign Language dataset. *Sci Data* (2026). https://doi.org/10.1038/s41597-026-06872-6

Francisco Morillas-Espejo & Ester Martinez-Martin

SCIENTIFIC DATA

Tracking no: SDATA-25-03784A

Sign4all: a Spanish Sign Language dataset

Authors: Francisco Morillas-Espejo (University of Alicante) and Ester Martinez-Martin (University of Alicante)

Abstract:
Sign Language Recognition is a critical component of human-machine interaction, enabling more inclusive technologies for the deaf and hard-of-hearing community. However, current datasets often suffer from data sparsity and a bias toward right-handed signs. In this regard, we present Sign4all, a dataset designed for Isolated Sign Language Recognition in Spanish Sign Language. The dataset consists of 7,756 high-resolution RGB video recordings and their corresponding skeletal keypoints, covering 24 signs related to a vocabulary centered in the catering field. Unlike sparse lexicons, Sign4all adopts a high-density approach, providing an average of 323 samples per sign to facilitate data-intensive deep learning models. Moreover, it provides a handedness balance, with equal representation of left- and right-handed signs to support handedness invariance. Each sample was manually segmented, temporally normalized and preprocessed through spatial normalization to guarantee consistency and compatibility with different deep learning pipelines. Technical validation using Transformer and skeletal models demonstrates the dataset's integrity and the need of providing pre-computed augmentation splits. Moreover, all data is formatted in widely supported file types.

Datasets:

Repository Name	Dataset Title	Accession Number or DOI	URL to data record	Private reviewer access URL/code
Science Data Bank	Sign4all: a Spanish Sign Language Dataset	10.57760/sciencedb.28304	https://www.scidb.cn/en/detail?dataSetId=12775cc0026841979bdaba60484ad067	https://www.scidb.cn/en/anonymous/Sk5WUmZx

Sign4all: a Spanish Sign Language dataset

Francisco Morillas-Espejo[1] and Ester Martinez-Martin[1]

[1]RoViT Lab, Department of Computer Science and Artificial Intelligence, University of Alicante, Carretera de San Vicente del Raspeig s/n, E-03690, Alicante, Spain
[*]corresponding author: Francisco Morillas-Espejo (francisco.morillas@ua.es)

Abstract

Sign Language Recognition (SLR) is a critical component of human-machine interaction, enabling more inclusive technologies for the deaf and hard-of-hearing community. However, current datasets often suffer from data sparsity and a bias toward right-handed signs. To support this effort, we present Sign4all, a dataset for Spanish Sign Language (LSE), specifically designed for Isolated Sign Language Recognition (ISLR). The dataset is composed of 7,756 high-resolution RGB video recordings and their corresponding skeletal keypoints, covering 24 signs related to daily activities, more specifically a vocabulary centered in the catering field. Unlike sparse lexicons, Sign4all adopts a high-density approach, providing an average of 323 samples per sign to facilitate data-intensive deep learning models. Moreover, the dataset provides a handedness balance, with equal representation of left- and right-handed signs for every sign to support handedness invariance. Each sample was manually segmented, temporally normalized and preprocessed through spatial normalization to guarantee consistency and compatibility with different deep learning pipelines. Technical validation using Transformer and skeletal models demonstrates the dataset's integrity and the need of providing pre-computed augmentation splits. All data is formatted in widely supported file types (AVI for video, HDF5 for keypoints), enabling direct use in machine learning frameworks such as TensorFlow or PyTorch.

Background & Summary

Automatic speech recognition has substantially enhanced human-machine interaction through applications such as text reading for visually impaired users, home appliance control via voice commands or translations between spoken languages. However, these advances do not directly benefit sign languages, where linguistic information is transmitted using the visual channel via positions, orientations and movements of the hands and body [1]. Consequently, research in Sign Language Recognition (SLR) relies on image and sequence processing. The use of high-quality datasets is crucial for advancing in SLR by providing the necessary data to develop and evaluate different deep learning models. Diverse datasets for different sign languages allow improvement in the generalization capabilities of the models, making the systems more robust to real-world scenarios. Moreover, sign languages show structural and linguistic differences depending on whether the aim is to analyze isolated signs or continuous conversation. According to this objective, it is necessary to use different recognition models and datasets. In this sense, there are two main SLR paradigms: Continuous Sign Language Recognition (CSLR) and Isolated Sign Language Recognition (ISLR) [2, 3].

CSLR focuses on recognizing signs within a continuous sequence, often involving large datasets of natural signing with sentence-level annotations. Some well-known CSLR datasets in the literature include RWTH-PHOENIX-Weather 2014 (and its extended version 2014T) for German Sign Language (DGS) [4, 5], which compiles broadcast weather reports over several

years, and How2Sign [6] for American Sign Language (ASL), featuring nearly 80 hours of multi-camera 2D and depth recordings from How2 dataset [7], a collection of instructional YouTube videos. Another example is CSL-Daily, which comprises around 20,000 high-resolution videos annotated at sentence and gloss levels for Chinese Sign Language (CSL) regarding people's daily lives (e.g., shopping, travel, medical care) [8], creating a vocabulary of 2,000 signs. However, while these datasets are fundamental for translation tasks, they introduce complexities related to co-articulation (where the beginning and end of a sign are modified by adjacent signs) which obscure the clear phonological boundaries required for the precise validation of ISLR architectures. [6, 9]

Consequently, ISLR focuses on recognizing signs performed one by one. However, the current landscape of ISLR datasets suffers from a significant trade-off between vocabulary size and sample density. While large-scale datasets offer extensive lexicons, they frequently suffer from data sparsity. For instance, WLASL [10] comprises 2,000 signs but averages only 10.5 samples per sign. Similarly, ASL-LEX [11, 12] includes 2,723 frequently used signs but typically provides one single exemplar, and MS-ASL [13] despite featuring a lexicon of 1,000 signs collected from social media averages only 25.5 samples per sign. Even the recently introduced NationalCSL-DP [14] which features an extensive vocabulary of 6,707 signs for Chinese Sign Language, provides only 10 samples per sign using the frontal view camera. This data sparsity is insufficient for training robust deep learning models that need to generalize across intra-class variations without overfitting.

Furthermore, datasets that achieve high sample counts, such as ASL Alphabet [15], MNIST ASL [16] or ASLAD-190k [17] use static images rather than video sequences. This limitation fails to capture the temporal dynamics and movement parameters essential for distinguishing complex signs. Additionally, most state-of-the-art datasets do not incorporate left-handed signers or they do in small portions like AUTSL [18] (2 left-handed out of 43 participants) or ChicagoFSWilds [19] (7.15% samples are left-handed), failing to account this variations in their datasets. Other datasets in languages such as Greek Sign Language (GSL) [3], Nepali Sign Language (NSL) [20], Indian Sign Language (ISL) [21] or Arabic Sign Language (ArSL) [22, 23] are constrained by challenging resolutions, low samples per sign or a lack of precise manual segmentation. Thus, a gap remains for high-resolution video-based ISLR datasets that provide sufficient sample depth and handedness diversity for fine-grained motion analysis.

Despite the variety and scale of available SLR datasets, Spanish Sign Language (LSE) remains underrepresented in publicly accessible corpora, a notable gap considering that it is the third most used sign language in the European Union [24, 25, 26, 27]. An example of a current LSE dataset is *LSE-Uvigo*, proposed by Docío-Fernández et al. [28]. They introduced a multi-source LSE dataset (RGB at 1920×1080 and depth at 512×424, both 30 fps) for 40 isolated signs and 40 continuous sentences but with only 43 repetitions per sign; however, this dataset is not publicly available. In addition, the authors do not specify the exact vocabulary included in the isolated signs nor in the continuous sentences.

Additionally, Docío-Fernández et al. proposed another LSE dataset, *LSE-eSaude_ UVIGO* [29], in the ECCV 2022 ChaLearn Sign Spotting Challenge. This dataset consists of Continuous Sign Language videos in the health domain, 100 signs annotated within full sentences with a variable amount of repetitions per sign, meaning that the repetitions are small for some instances and the dataset is imbalanced. The recordings were captured with an RGB camera with a 1080×720 resolution at 25 fps. Unlike *LSE-Uvigo*, this dataset is focused on real-world continuous signing scenarios, facilitating CSLR research. In this case, the authors specified the vocabulary, which is composed of the following words divided by grammatical categories:

- **Prepositions**: from (*a-partir-de*), against (*contra*), inside (*dentro*), until (*hasta*)
- **Determiners and adverbs**: some (*alguno*), no (*no*), other (*otro*)
- **Nouns**: artery (*arteria*), sugar (*azúcar*), case (*caso*), certificate (*certificado*), circulation

(*circulación*), body (*cuerpo*), day (*día*), exercise (*ejercicio*), epoch (*época*), etc. (*etc*), shape (*forma*), liver (*hígado*), reason (*motivo*), level (*nivel*), low level (*nivel-bajar*), level-b (*nivel b*), name (*nombre*), objective (*objetivo*), person (*persona* (three positional variations)), foot (*pie*), percentage (*porcentaje*), problem (*problema*), kidney (*riñón*), symptom (*síntoma*), therapy (*terapia*), type (*tipo*), vein (*vena*), truth (*verdad*), time (*vez*)

- **Verbs**: eat (*comer*), dinner (*crear*), discover (*descubrir*), not have (*haber-no*), increase (*incrementar*), want to say (*querer decir*), reduce (*reducir*), feel (*sentir*), suffer (*sufrir*), use (*usar*)

- **Adjectives**: complicated (*complicado*),ill (*enfermo*), easy (*fácil*), frequent (*frecuente*), strong (*fuerte*), lower (*inferior*), higher (*mayor*), nervous (*nervioso*), healthy (*sano*), superior (*superior*)

Another example is the proposal of Rodriguez-Moreno et al. [30] where they created a limited dataset of 5 signs with 175 samples each. This dataset is composed of the following words: well (*bien*), happy (*contento*), woman (*mujer*), man (*hombre*) and listener (*oyente*).

In an earlier study [31], we presented a Spanish Sign Language dataset focused on the 30 letters of the alphabet, recorded at a resolution of 640×480 pixels. While that work highlighted the importance of the spatial dimension for LSE alphabet and presented one of the first fingerspelling LSE datasets, it did not include additional day-to-day vocabulary or use high-resolution recordings. In this paper, we expand that dataset by introducing a new collection of 24 LSE signs related to daily activities, more specifically a vocabulary centered in the restoration area. Unlike sparse large vocabulary datasets, we adopted a “small class, high volume" architecture, providing an average of 323 ± 15 video samples per sign, a density significantly higher than major benchmarks like AUTSL or WLASL as it can be seen in Table 1.

Moreover, the vocabulary was not selected arbitrary but designed to challenge the fine-grained discrimination capabilities of computer vision models using phonological minimal pairs (e.g. signs sharing movement and location but differing in handshape), a concept further detailed in the Methods section. Each sign was manually segmented to ensure high-quality annotations suitable for machine learning applications. The dataset contains high-resolution RGB video recordings paired with extracted skeletal keypoints aimed to train deep learning models. Additionally, the dataset includes both right-handed and left-handed signs in order to mitigate possible biases introduced by right-hand-dominant datasets from the state-of-the-art.

Although the current dataset is focused on the ISLR task, the selected signs could be repurposed for CSLR in future work by collecting full-sentence recordings or applying synthetic data augmentation. By focusing on a real-world domain, this contribution complements the previously published alphabet corpus [32, 31], setting up the foundations for more robust and comprehensive LSE recognition.

Sign Language	Authors	Recording	Num. signs	Repetitions per sign	Participants	Dominant hand	Type	Availability
American Sign Language	*How2Sign* [6]	RGB-D (1280x720 @30)	79 h.	Not specified	11	Not specified	CSLR	Publicly available
American Sign Language	*How2Sign* [6]	Panoptic studio [33]	3 h.	Not specified	11	Not specified	CSLR	Publicly available
American Sign Language	*WLASL* [10]	RGB (256x256)	2,000 signs	10.5 avg.	119	Not specified	ISLR	Publicly available
American Sign Language	*ASL-LEX 2.0* [12]	RGB (640x480 @29.93)	2,700 signs	1	1	Not specified	ISLR	Partially available
American Sign Language	Athitsos et al. [34]	RGB (640x480 @60) + RGB (1600x1200 @30)	3,000 signs	1	3	Not specified	ISLR	Partially available
American Sign Language	*ASL alphabet* [15]	RGB (28x28)	29 signs	3,000 images	1	Right	ISLR	Publicly available
American Sign Language	*MNIST ASL* [16]	RGB	24 signs	1,700 images	1	Right	ISLR	Publicly available
American Sign Language	ChicagoFSWilds [19]	RGB	31 signs	Not specified	168	92.85% right 7.15% left 1.65% bimanual	ISLR	Publicly available
American Sign Language	Mavi and Dikle [35]	RGB (128x128)	26 signs	5	173	Right	ISLR	Publicly available
American Sign Language	*MS-ASL* [13]	RGB	1,000 signs	25.5	222	Not specified	ISLR	Publicly available
British Sign Language	BOBSL [36]	RGB (444x444 @25)	1,467 h.	198	39	Not specified	CSLR	Upon request
British Sign Language	*Dicta-Sign* [37]	RGB	>28 h.	Not specified	14	Not specified	CSRL	Not available
German Sign Language	*RWTH-PHOENIX-Weather 2024T [4]*	RGB (210x260 @25)	1,000 signs	71	9	Right	CSRL	Publicly available
German Sign Language	*Dicta-Sign* [37]	RGB	>28 h.	Not specified	14	Not specified	CSRL	Not available
German Sign Language	*Dicta-Sign* [37]	RGB	1,000 signs	10	2	Not specified	ISRL	Not available
Greek Sign Language	Adaloglou et al. [3]	RGB-D (840x840 @30)	331 sentences	Not specified	7	Not specified	CSRL	Publicly available
Greek Sign Language	Adaloglou et al. [3]	RGB-D (840x840 @30)	310 signs	Not specified	7	Not specified	ISLR	Publicly available
Greek Sign Language	*Dicta-Sign* [37]	RGB	>28 h.	Not specified	14	Not specified	CSRL	Not available
Greek Sign Language	*Dicta-Sign* [37]	RGB	1,000 signs	10	2	Not specified	ISLR	Not available
Indian Sign Language	Kumar et al. [21]	RGB-D	30 signs	90	10	Right	ISLR	Not available
Indian Sign Language	Bhatia and Wadhawan [38]	RGB	100 signs	350 images	Not specified	Not specified	ISLR	Not available
Chinese Sign Language	*CSL-Daily* [8]	RGB (1920x1080 @30)	2,000 signs	Not specified	10	Not specified	CSRL	Upon request
Chinese Sign Language	*DEVISIGN* [39]	RGB (640x480 @30)	2,000 signs	12	8	Not specified	ISLR	Upon request
Chinese Sign Language	*NationalCSL-DP* [14]	RGB (1920x1080 @50)	6,707 signs	10	10	Not specified	ISLR	Partially available
French Sign Language	*Dicta-Sign* [37]	RGB	>28 h.	Not specified	14	Not specified	CSLR	Not available
Arabic Sign Language	*ASLAD-190K* [22]	RGB (1280x720)	31 signs	6,100 images	Not specified	Not specified	ISLR	Publicly available
Arabic Sign Language	Al-Barham et al. [17]	RGB (different cameras)	31 signs	250 images	>200	Not specified	ISLR	Publicly available
Nepali Sign Language	*NSL23* [20]	RGB (1280x720)	49 signs	Variable	14	Right	ISLR	Publicly available
Serbian Sign Language	Radaković et al. [40]	RGB (1280x720)	30 signs	278	41	Not specified	ISLR	Publicly available
Russian Sign Language	*Slovo* [41]	RGB (different cameras)	1,000 signs	20	194	Not specified	ISLR	Publicly available
Turkish Sign Language	*AUTSL* [18]	RGB-D (512x512 @30)	226 signs	169.6	43	2 left-handed signers	ISLR	Publicly available
Spanish Sign Language	*LSE-Uvigo* [28]	RGB (1920x1080 @50) + Depth (512x424 @ 30)	40 sentences	Not specified	13	Not specified	CSLR	Not available
Spanish Sign Language	*LSE-Uvigo* [28]	RGB (1920x1080 @50) + Depth (512x424 @ 30)	40 signs	43	32	Not specified	ISLR	Not available
Spanish Sign Language	*LSE-eSaude_ UVIGO [29]*	RGB (1080x720 @25)	100 signs	Variable	10	Not specified	CSLR	Publicly available
Spanish Sign Language	Rodríguez-Moreno et al. [30]	RGB	5 signs	175	5	Right	ISLR	Not available
Spanish Sign Language	*Sign4all alphabet* [31]	RGB (640x480 @30)	30 signs	Not specified	12	Both	ISLR	Upon request
Spanish Sign Language	*Sign4all (ours)*	RGB (2560x1440 @30)	24 signs	323±15	8	Both	ISLR	Upon request

Table 1: Sign Language Recognition datasets. When the column for the dominant hand is designated as *Not specified*, it is most likely that the signers are right-handed.

Methods

This section describes the methodology used to produce the dataset. Specifically, it describes the selected vocabulary, the capture setup, the data composition, and the post-processing steps followed to clean the data.

Data selection

This study introduces Sign4all, a dataset designed for Isolated Sign Language Recognition (ISLR) in Spanish Sign Language (LSE), specifically targeting a vocabulary related to dining interactions. The construction of the dataset was driven by two primary methodological objectives: mitigating the handedness bias present in the existing literature and challenging the fine-grained discrimination capabilities of deep learning models through phonological similarity.

A critical limitation in current Sign Language Recognition (SLR) research is the over representation for right-handed signs since most state-of-the-art datasets contain only right-handed samples, ignoring the left hand, or left-handed signs comprise a small portion of the dataset, as depicted in Table 1. To address this, we present a dataset that takes a different perspective by introducing a balanced protocol where all signs are recorded using both the dominant and non-dominant hands for every participant. Unlike synthetic data augmentation techniques such as horizontal flipping, which may introduce artifacts or fail to capture the subtle kinematic differences of the non-dominant hand, our approach provides natural variations in sign execution. This allows us to build a dataset with greater variability to evaluate models invariant to lateral mirroring, a prerequisite for inclusive, real-world recognition systems.

The dataset consists of 24 selected signs, chosen based on their frequency of use, their ability to form meaningful short phrases and their phonological relationships. The included vocabulary is structured as follows:

- **Pronouns**: I (*yo*), you (*tú*)
- **Verbs**: eat (*comer*), dinner (*cenar*), breakfast (*desayunar*), like (*gustar*), dislike (*no-gustar*), want (*querer*)
- **Nouns**: meat (*carne*), spoon (*cuchara*), knife (*cuchillo*), hamburger (*hamburguesa*), eggs (*huevos*), fish (*pescado*), pizza (*pizza*), plate (*plato*), soup (*sopa*), fork (*tenedor*), omelet (*tortilla*), glass (*vaso*), vegetables (*verdura*)
- **Interrogative particles**: what (*qué*), when (*cuándo*), where (*dónde*)

The vocabulary was not selected arbitrarily but was curated to include specific phonological challenges, ranging from minimal pairs to compound signs. In Sign Language, signs are composed of specific hand parameters as stated by Stokoe [42]: hand configuration, hand location and hand movement. We deliberately selected signs that share these parameters while differing in one, as well as signs that require sequential temporal analysis. This selection challenges the models to distinguish between subtle inter-class variations rather than relying on gross motion features.

Specifically, the dataset presents varied phonological challenges. It features configuration contrasts where signs like [MEAT] and [FORK] share identical movement trajectories and locations (contact with the hand) but are different in hand configuration (see Figure 1); a model relying solely on wrist trajectory will fail to distinguish this pair. Movement contrasts are exemplified by the signs for [OMELET] and [SPOON] whose movement is different (up and down against rotational) but share the same hand configuration as displayed in Figure 2. We also introduced bimanual versus unimanual contrasts; for instance, the sign for [GLASS] and [BREAKFAST] share similar hand shapes and location near the mouth, yet [GLASS] is unimanual while [BREAKFAST] is a bimanual sign (Figure 3). Furthermore, the vocabulary includes sequential variations, such as the relationship between [LIKE] and [DISLIKE], the latter uses

the same initial movement as the former but appends the sign [NO] to indicate negation. This negation component shares similarities in location and movement with [VEGETABLES], adding another layer of complexity. This diversity ensures the dataset serves as a stress-test benchmark for skeletal pose estimation and temporal modeling architectures.

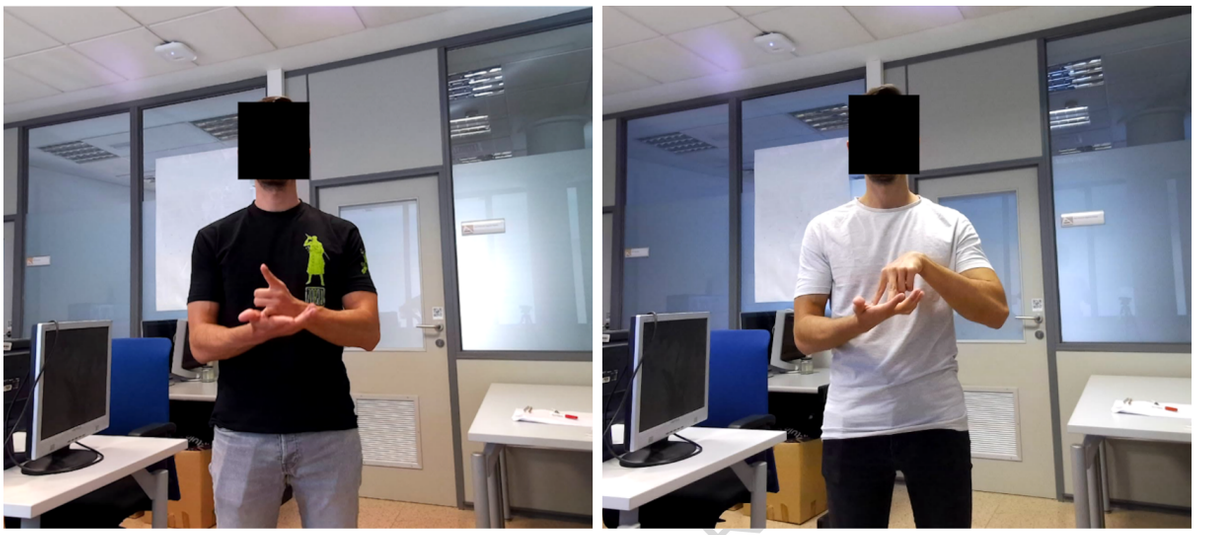


Figure 1: Comparison between signs [MEAT] (left) and [FORK] (right)

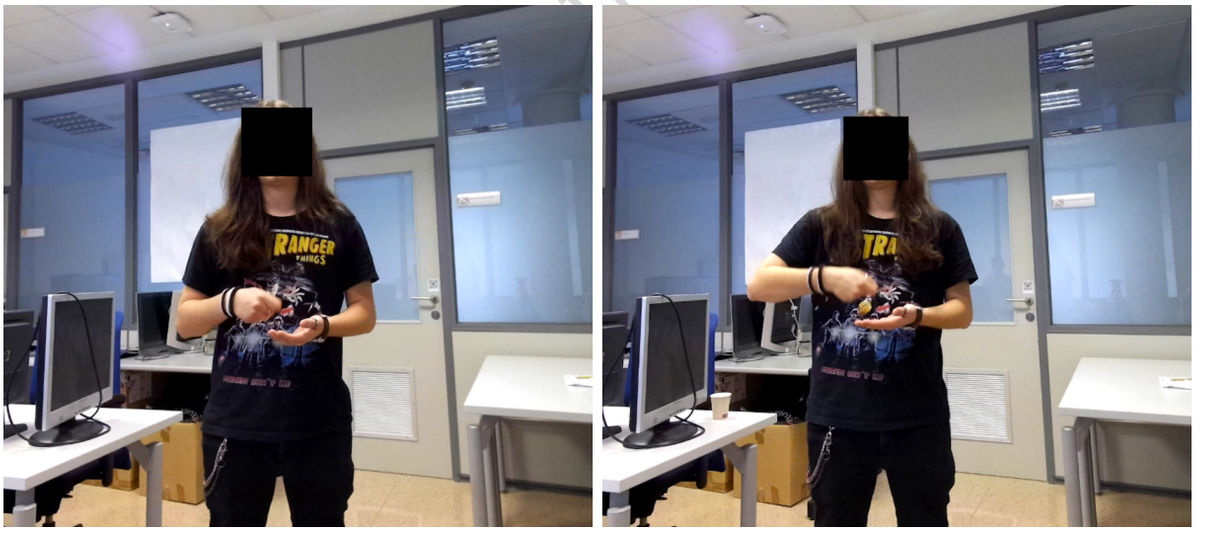


Figure 2: Comparison between signs [OMELET] (left) and [SPOON] (right)

Although this dataset is designed to be used in ISLR, the selected words allow forming short sign language sentences such as *I do not like meat* or *I have meat and eggs for breakfast*. This design allows training recognition systems capable of translating short sentences from their constituent signs by concatenating their predictions.

Data collection

In Sign Language recognition, there are two kinds of signs: static and in-motion signs. The key difference is that in the static signs, the user does not need to perform any hand movement to complete the sign; while for the in-motion or dynamic ones, the user needs to complete some kind of trajectory in order to have a meaningful sign. All signs comprising this dataset are

Figure 3: Comparison between monomanual sign [GLASS] (left) and bimanual sign [BREAKFAST] (right)

dynamic; thus, the recording procedure was structured to capture the full temporal evolution of every gesture.

Before detailing the acquisition strategy, it is necessary to categorize the signs based on manual articulation. The dataset includes unimanual signs, which involve the use of a single hand; and bimanual signs, which involve both hands. Special attention was given to bimanual signs that can be further categorized as symmetric (Figure 4) or asymmetric (Figure 5). On the one hand, asymmetric signs involve different trajectories or hand configurations for each hand, having a distinction between dominant and non-dominant execution. On the other hand, in symmetric bimanual signs, both hands perform identical trajectories and configurations, meaning that there is no visual distinction between left- or right-handed signs.

Figure 4: Symmetric bimanual sign: breakfast (*desayunar*)

Data collection involved 8 participants (4 male and 4 female). All participants signed an informed consent form agreeing to their participation in the study and the sharing of their potentially identifiable data as part of this. The Ethics Committee from the University of

Figure 5: Asymmetric bimanual signs: knife (*cuchillo*) on the left using same hand configuration but different trajectory and orientation; meat (*carne*) on the right using different hand configurations

Alicante approved this data collection under application number UA-2025-04-14_3.

The recording sessions were designed with a protocol to guarantee equal representation for every sign. Unlike crowd-sourced or wild-collected datasets where class distribution often exhibits significant skew, with some signs having abundant samples while others are scarce, we enforced a number of repetitions per class. This approach prevents the models from developing biases toward high-frequency classes, a common limitation in large-scale benchmarks. To achieve a handedness balanced dataset, all subjects were required to perform every unimanual and asymmetric bimanual signs using both their dominant and non-dominant hands, regardless of their natural handedness (one participant was left-handed). This resulted in two distinct sets of recordings per participant: one where the signs obey standard right-handed topology and another mirroring it for left-handed topology. By using this dual-execution protocol, we ensured that the dataset contains the same number of samples for both lateralities, removing the class imbalance present in other state-of-the-art dataset where left-handed samples are rare. Each participant was recorded an average of 20 repetitions per hand and per sign for these categories.

In contrast, for symmetric bimanual signs, instead of separating these into left- and right-handed, they were recorded with double the repetitions in a single batch. This guarantees that the total volume of data per class remains consistent for all types of signs.

To maintain consistency across recordings, participants followed a standardized protocol, keeping the signing hand centered in front of the torso before beginning to sign. After executing the sign according to its natural movement (see Figure 6), they returned to this position before continuing with the next iteration. This process simplified filtering the recordings by recognizing the start and end points of each sign. Moreover, when a sign only involves the dominant hand, each participant leaves the passive hand in a neutral position, which varies from one subject to another, depending on which position is most natural to them (see Figure 7).

Recording equipment

The dataset was recorded using an Azure Kinect DK camera [43], chosen for its ability to capture high-resolution RGB data. It uses an OV12A10 12MP CMOS sensor whose complete specifications are shown in Table 2. In this sense, a resolution of 2560×1440 pixels at 30 FPS was

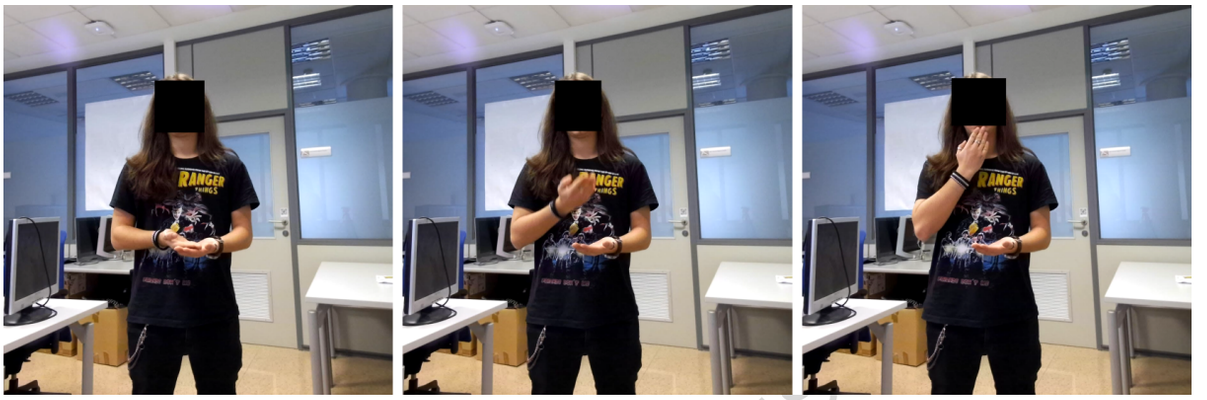


Figure 6: Different signing positions for the word soup (*sopa*)

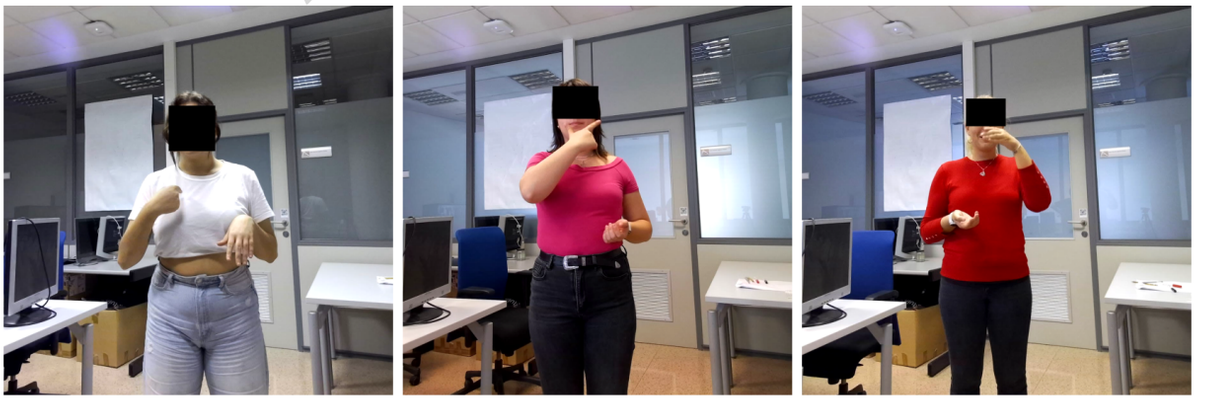


Figure 7: Passive hand position for different participants

used during the data recording process. This resolution was used as it provides an appropriate balance between data quality and size.

Resolution	Aspect Ratio	Format Options	FPS	Nominal FoV
3840x2160	16:9	MJPEG	0, 5, 15, 30	90°x59°
2560x1440	16:9	MJPEG	0, 5, 15, 30	90°x59°
1920x1080	16:9	MJPEG	0, 5, 15, 30	90°x59°
1280x720	16:9	MJPEG/YUY2/NV12	0, 5, 15, 30	90°x59°
4096x3072	4:3	MJPEG	0, 5, 15	90°x74.3°
2048x1536	4:3	MJPEG	0, 5, 15, 30	90°x74.3°

Table 2: Color camera supported operating modes [44]

Recording environment

To ensure consistency across all the recordings, the camera was mounted on a tripod at a fixed height of 117 cm, measured from the center of the lens to the floor. This height was determined after different experimental configurations as it captured all the participants from a frontal view. Since each participant has a different height, instead of changing the camera location, they were located at different distances from the camera, from 100 to 170 cm, ensuring optimal coverage of hand and body movements. This change in distance is depicted in Figure 8 where a subject 1.70 m tall is shown on the left and another subject 1.85 m tall on the right. As can be seen, in both cases the body is visualized in a similar way. Additionally, each sign was recorded with the participant centered in the frame.

Figure 8: Shorter person (left) compared to a taller person (right)

Recordings were conducted in a controlled laboratory environment to maintain consistent lighting conditions across all sessions, reducing potential variability in recognition performance. However, to ensure real-world applicability, the dataset was designed with natural variations in participant clothing and background settings (see Figure 9):

- No dress code restrictions: We deliberately abstained from imposing a uniform dress code to introduce high variance in visual appearance. Participants wore their regular clothing,

leading to natural variation in fabric textures, sleeve lengths and colors. Moreover, since the recording schedule force to record the same participants in multiple days they appear wearing different outfits across different signs. This introduces significant intra-subject variability, preventing the models from overfitting to specific signer-clothing combinations (e.g. associating a red shirt with a specific class). Instead, it forces the deep learning pipeline to focus on the skeletal motion rather than static appearance features. It also evaluates the systems' abilities to handle challenging scenarios such as low contrast between hands and clothing or partial wrist occlusions caused by loose sleeves.

- Fixed background: The recordings were captured in the same indoor location but without using a chroma-key environment (typically used in this kind of datasets). By retaining natural background elements (furniture, door frames) the dataset allows for the evaluation of segmentation algorithms against realistic background clutter rather than artificial silence.

- Illumination consistency: While no extreme lighting variations were introduced, the dataset reflects subtle shifts in ambient lighting due to sunlight coming through the windows, which varies depending on the day and time of the recordings.

Figure 9: Different lightning and cloth conditions for the same sign

Data composition and processing

To ensure the highest possible fidelity, all raw recordings were initially stored in the native Azure Kinect format (*k4record*). However, while this format retains maximum quality, it is not directly compatible with standard machine learning workflows, requiring conversion for efficient data handling.

To facilitate processing, annotation, and storage management, the dataset was converted to AVI JPEG format using the MJPEG codec. This conversion was selected due to its ability to significantly compress file size while maintaining high visual quality. The MJPEG codec provides frame-independent encoding, ensuring that each frame remains intact without predictive compression techniques that might degrade fine-grained motion details.

This format transformation had a direct effect on the dataset's size. Initially, the complete dataset composed of raw data occupied approximately 3 TB, with each sign contributing approximately 8 GB per participant per hand, considering all the repetitions of the sign. With the format conversion, this size was reduced to nearly 600 GB for the complete dataset and each sign occupies 1.5 GB per participant per hand.

Since each recording session included multiple sign repetitions, manual segmentation was performed to extract isolated instances of each sign. This process was done using Blender [45], an open-source 3D animation application that allows to store the resulting videos in the same format and codec as the original, preserving its maximum quality.

To perform segmentation in isolated signs, each recording session has been carefully analyzed to identify which frames correspond to each repetition. Since we used pauses at the beginning and end of each iteration, these are employed to easily analyze when a new sign starts, removing everything that is between these pauses. Figure 10 details this process, where in the left image it can be seen how the participant is not in the middle of a sign (for any reason) while in the right image he is at the beginning of the signing process. In this case, the first frame as well as the successive frames until reaching the signing start position are disregarded, generating isolated videos for each repetition.

Figure 10: Noise in the recording (left); initial signing position (right)

After this process, a detailed analysis is made on the resulting frames where these initial and final positions are removed, storing only the frames that constitute a sign. This is depicted in Figure 11 in which, despite both frames being part of the signing process, the left image does not represent the sign; instead, it is the movement required to reach the signing position, while the right image actually depicts the sign. For this reason, those frames where a participant is reaching or leaving the signing position are discarded.

In this way, a corpus of isolated signs is generated in which only the signing process appears. In addition, those signs that are misinterpreted or ambiguous are also rejected, ensuring that only quality samples are kept.

This processing resulted in a final dataset comprising $7,756$ high-resolution RGB videos, whose distribution among classes is shown in Table 3. A defining characteristic of this dataset is its high-density sample volume; unlike dictionary-style datasets that prioritize vocabulary breadth at the cost of depth, our dataset provides a robust mean of 323 ± 15 samples per sign. This density is particularly valuable for data-intensive deep learning architectures, which require substantial training volumes to converge effectively and generalize across intra-class variations.

Furthermore, the dataset achieves a good balance in handedness representation, addressing a gap in the current state-of-the-art. As illustrated in Figure 12, the distribution of samples is evenly split between right-handed and left-handed execution across all signs. Specifically, the dataset provides a mean of 160 ± 7 samples per sign for the right hand and 162 ± 9 samples per sign for the left hand. This balance ensures that models trained on this data are not biased toward right-handed topology, enabling the development of truly inclusive, hand-agnostic recognition systems. By providing statistically equivalent volumes of data for both handedness, this dataset serves as a reliable benchmark for evaluating hand invariant skeletal pose estimation and feature

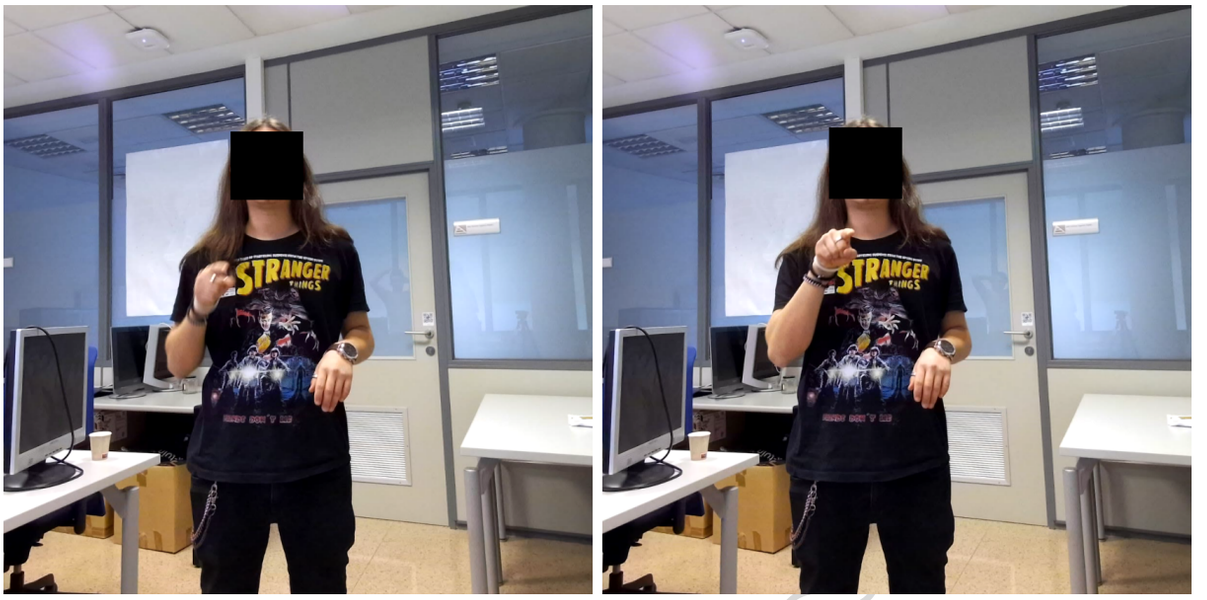


Figure 11: Reaching signing position (left) versus start signing (right)

extraction architectures.

Temporal normalization

One challenge in Sign Language datasets is the variable duration of sign execution. This variation arises from two main factors: the phonological complexity of the sign (e.g. the sign [EAT] requires a short and direct trajectory, whereas [DISLIKE] involves a longer and compound movement) and the inter-subject variability in articulation speed (i.e. each person signs at different speeds). In raw data, these differences create inconsistent sequence lengths that difficult the formation of batches during model training. To address this we applied a fixed-length temporal normalization, standardizing all video sequences to 48 frames.

The selection of this specific parameter was riven by a methodological trade-off between sign execution fidelity and hardware optimization. From a kinematic perspective, our visual analysis of the raw recordings determined that the average sign execution time at 30 fps is approximately 1.6 seconds. A 48-frame window covers this duration, ensuring that the full motion trajectory is preserved without excessive truncation or unnecessary frame duplication. Simultaneously, this value is optimized for high-performance computing pipelines; deep learning libraries using NVIDIA Tensor Cores achieve maximum matrix multiplication efficiency when dimensions are multiples of 8 as stated by Sarge et al. [46] By using this temporal dimension of 48 frames we ensure that the dataset is optimized for modern GPU architectures, reducing training latency.

To perform this temporal normalization, two techniques were applied:

- **Downsampling**: If a video exceeds 48 frames, a subset of frames is evenly sampled across the sequence to reduce its length while preserving the full motion trajectory of the sign. In this sense, the frame selection is performed using:

$$I_k = \lfloor k \cdot \frac{N-1}{48-1} \rfloor, k = 0, 1, ..., 47 \quad (1)$$

  Where N is the original number of frames of the video and I_k represents the indices of the selected frames. This preserves critical movement details.

- **Upsampling**: If a video contains fewer that 48 frames, additional frames are generated through duplication. Instead of randomly inserting extra frames or adding them at the end

Sign	Right-hand videos	Left-hand videos	Total videos
Meat (*Carne*)	162	165	327
Hamburger (*Hamburguesa*)	168	175	343
Eggs (*Huevos*)	179	178	357
Fish (*Pescado*)	165	182	347
Pizza (*Pizza*)	160	167	327
Soup (*Sopa*)	167	173	340
Omelet (*Tortilla*)	165	163	328
Vegetables (*Verdura*)	163	164	327
When (*Cuándo*)	152	162	314
Where (*Dónde*)	152	155	307
What (*Qué*)	165	155	320
You (*Tú*)	155	152	307
Me (*Yo*)	153	152	305
Spoon (*Cuchara*)	151	158	309
Knife (*Cuchillo*)	165	153	318
Plate (*Plato*)	157	163	320
Fork (*Tenedor*)	150	154	304
Glass (*Vaso*)	164	149	313
Like (*Gustar*)	159	166	325
Dislike (*No-gustar*)	171	175	346
Want (*Querer*)	155	154	309
Dinner (*Cenar*)	163	171	334
Eat (*Comer*)	157	161	318
Breakfast (*Desayunar*)	155	156	311
Total (average)	3,853 (160±7)	3,903 (162±9)	7,756 (323±15)

Table 3: Dataset distribution after manual filtering

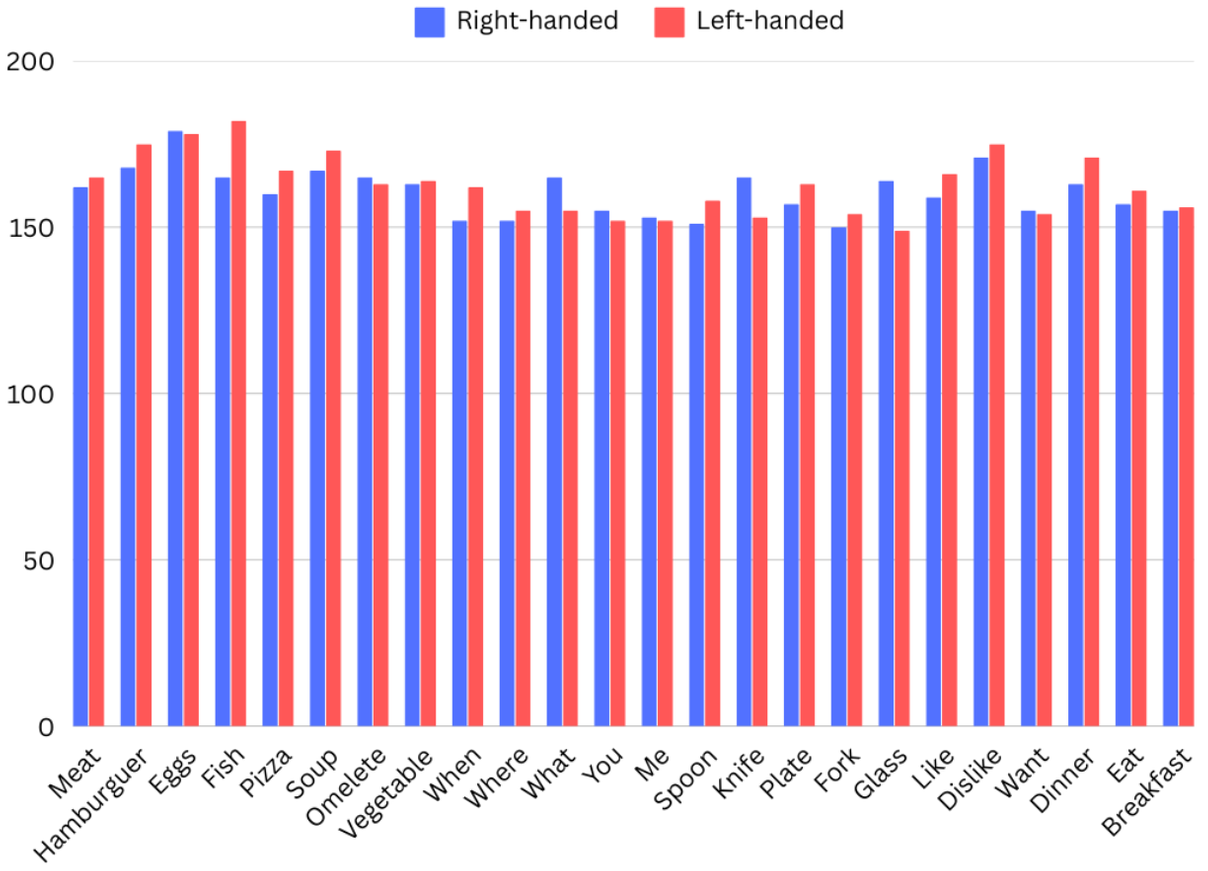


Figure 12: Graphical distribution of right and left samples

of the videos, the frames are duplicated at regular intervals throughout the video using the same method as in the downsampling case. This ensures an even distribution of additional frames while minimizing sudden jumps in motion.

Methodologically, this normalization serves a dual purpose beyond storage convenience. First, it ensures direct compatibility with standard deep learning architectures, facilitating batched training without the need for zero-padding or complex masking strategies. Second, and more importantly, it prevents the model from exploiting sequence duration as a classification shortcut. By enforcing a uniform temporal footprint, the trained models cannot classify signs based on their length (e.g. assuming all short videos correspond to the sign [GLASS]) and they are instead forced to rely on spatial configuration and motion dynamics to discriminate classes.

Background reduction

Raw video recordings often contain significant dead space such as empty walls, floor or ceiling, that contributes no linguistic information but consumes computational resources. To guarantee dataset uniformity, a spatial normalization process was performed. This step involves extracting the signer's Region of Interest (ROI) and cropping the frame to a square resolution of 1224×1224 pixels. This specific resolution was derived from the sensor's native height (1440 pixels), keeping approximately 85% of the vertical field of view. This resolution removes only the non-informative floor and ceiling margins while maintaining high pixel density, which is critical for ISLR as it preserves the high-frequency details of the hands and fingers.

The decision of providing precalculated square crops rather than leaving spatial preprocessing to downstream tasks is grounded in the necessity for geometric fidelity and benchmarking consistency. Most state-of-the-art deep learning architectures (e.g. ResNet [47], EfficientNet [48], Video Vision Transformers [49]) require square input tensors. If raw rectangular videos are resized directly to a square during training, the image suffers from anamorphic distortion, squashing the signer horizontally and altering the skeletal topology essential for recognition. Alternatively, zero-padding wastes computational cycles on non-informative pixels. By standardizing the ROI at the dataset level we guarantee that the signer's aspect ratio is preserved naturally. Moreover, this standardization ensures that all future benchmarks using this dataset are performed on the exact same visual input, removing the possibility of using different cropping strategies among researchers.

To achieve this, EfficientDetD1 [50], an object detection model which contains a *person* class, was used to identify the signer in different videos and extract their bounding box. For each participant, a bounding box was determined for every recorded sign, storing the most extreme corner values. Specifically, the process involved tracking the maximum (x_{max}, y_{max}) and minimum (x_{min}, y_{min}) positions across all signs, ensuring that the bounding box captured the signer in every instance. After manually reviewing the bounding box to ensure it fully captured the participant and adjusting its width and height to reach the desired resolution, the image was cropped by removing excess pixels on the left, right, top, and bottom, with the signer remaining centered. An example of this cropping process can be seen in Figure 13.

Skeletal keypoint representation

To augment the dataset with structured motion information, skeletal keypoints were extracted from every video frame using MediaPipe [51], a widely recognized tool for real-time human pose estimation. MediaPipe works using a two-stage pipeline: a detector first locates the ROI within the frame and a landmark model then predicts the precise keypoint coordinate within that ROI. This architecture allows for efficient tracking by re-running the detector only when the tracking confidence drops. This skeletal framework offers a succinct yet highly informative abstraction

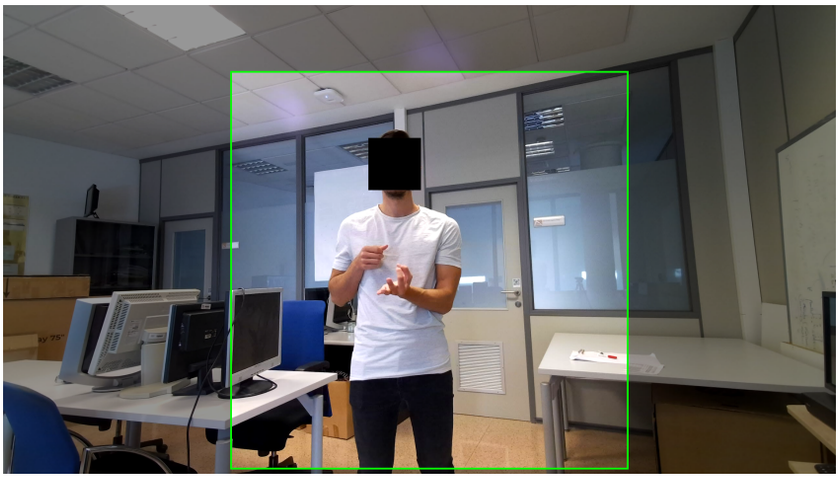


Figure 13: Colored image inside green square corresponds to the result after background removal

of the signer's motions, identifying critical anatomical landmarks crucial for Sign Language Recognition. In contrast to raw video data, that includes additional visual information, skeletal data focuses only on the spatial and temporal dynamics of signing, making it particularly suitable for deep learning models that prioritize movement over visual appearance.

MediaPipe's human pose estimation model detects 33 keypoints for the complete body, including legs, arms and torso; while the hand representation is composed of 21 keypoints per hand. The skeletal model used in this dataset focuses on landmarks of the hands, wrists, elbows, shoulders and torso as depicted in Figure 14, as these regions convey the most critical linguistic information for sign language. More precisely, the extracted subset consists of 50 landmarks: 21 per hand (fingers and wrist) and 8 from the upper body pose (shoulders, elbows, wrists and hips).

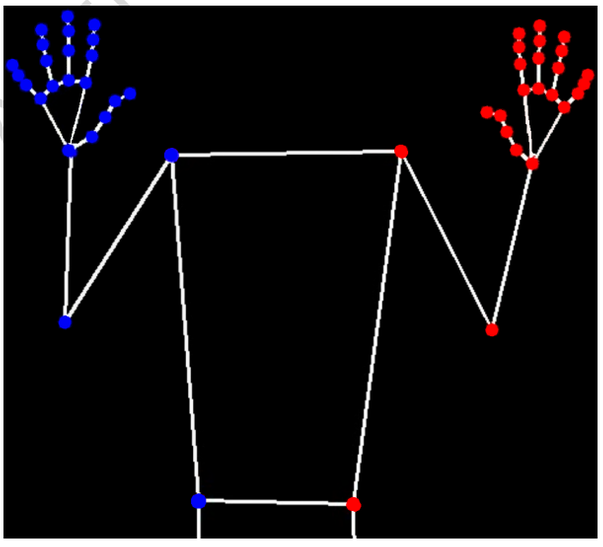


Figure 14: Keypoints used during recognition, blue color for right part of the body and red color for left one

As a normalization process, all the extracted keypoints are transformed into a relative coordinate system, centered at the midpoint between the shoulders (see Figure 15). This process helps to reduce variations due to differences in body proportions, participant positioning, or camera distance. The transformation is achieved by identifying the left and right shoulder keypoints, computing their midpoint and shifting all extracted landmarks relative to this new reference

system. This approach ensures that all skeletal representations are aligned, regardless of the signer's initial position.

Additionally, if any of the selected keypoints are not visible, they are assigned a value of $3,000$. This determination is based on MediaPipe's internal visibility and confidence score, landmarks with a value below 0.5 are treated as missing. This high value was chosen because it represents an unreachable position, regardless of the sign that is being performed. This approach enables deep learning models to selectively ignore missing values during training, ensuring that in real-time execution, non-visible points do not introduce noise into the system.

Despite an initial manual validation phase where most samples with significant execution errors were removed, some samples within the dataset may still exhibit specific skeletal detection artifacts inherent to MediaPipe. Rapid ballistic movements such as those in the sign [DISLIKE] (see Figure 16a), can occasionally cause motion blur, resulting in the tracker producing jittery or inconsistent coordinates across consecutive frames. Moreover, bimanual signs involving crossing hands, e.g. sign [HAMBURGER] (Figure 16b), may lead to depth ambiguity, where the model struggles to resolve finger configurations during momentary self-occlusion. Rather than removing all such instances, these were retained to ensure the dataset reflects the realistic noise conditions that recognition models must overcome in real-world deployment.

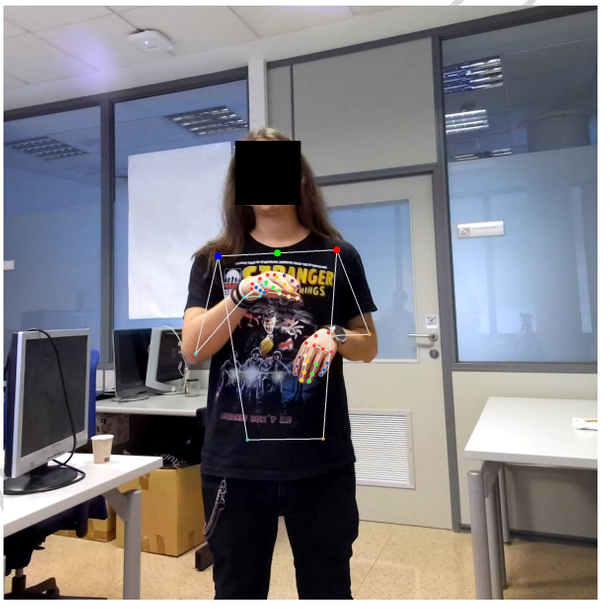
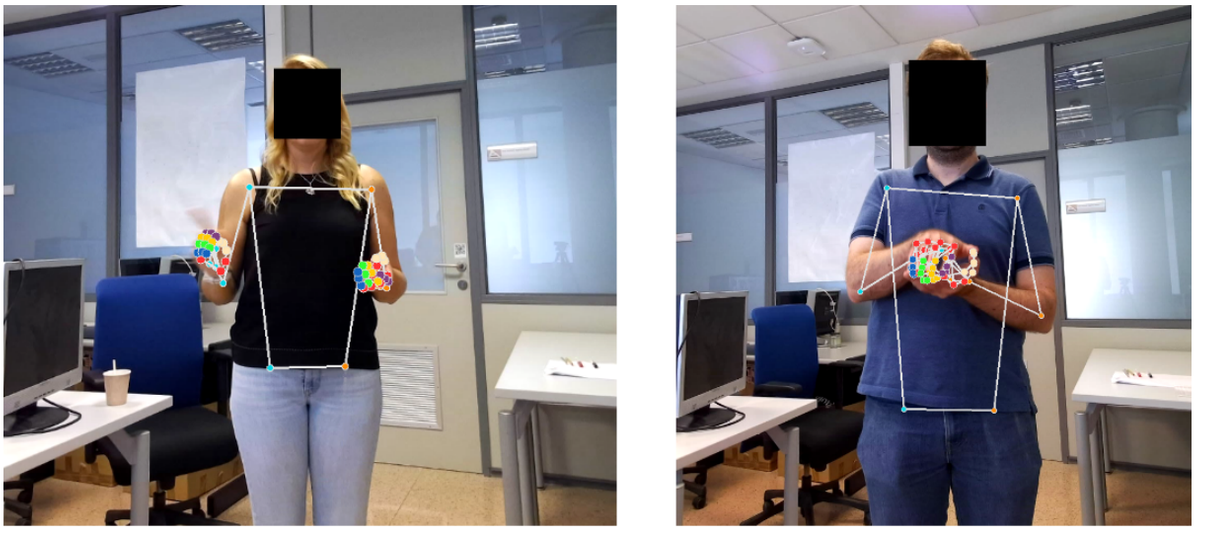


Figure 15: MediaPipe image of a sign with the used keypoints and the points used for normalization: Left shoulder (red dot), right shoulder (blue dot), middle point (green dot)

For the construction of this part of the dataset, a matrix was generated for each repetition of every sign. This matrix has a size of 48×100, where each column represents a keypoint, and each row captures its evolution over time, as shown in Equation 2. The fixed-length RGB videos generated in the previous steps were used to compute these matrices. It is important to note that each of these matrices is stored in *h5* format, resulting in a dataset with a significantly reduced storage size.

$$M = \begin{bmatrix} (q_0^x, q_0^y) & (q_1^x, q_1^y) & (q_2^x, q_2^y) & \cdots & (q_n^x, q_n^y) \\ (q_0^x, q_0^y) & (q_1^x, q_1^y) & (q_2^x, q_2^y) & \cdots & (q_n^x, q_n^y) \\ (q_0^x, q_0^y) & (q_1^x, q_1^y) & (q_2^x, q_2^y) & \cdots & (q_n^x, q_n^y) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ (q_0^x, q_0^y) & (q_1^x, q_1^y) & (q_2^x, q_2^y) & \cdots & (q_n^x, q_n^y) \end{bmatrix} \tag{2}$$

Where:



(a) [DISLIKE] sign (b) [HAMBURGER] sign

Figure 16: MediaPipe detection errors

- q_i^x, q_i^y represent the normalized position in coordinates of the point q_i.
- Each row corresponds to a frame in time, capturing the temporal evolution of skeletal movement.

Data augmentation

With the aim of increasing the dataset's size, data augmentation techniques were applied. Given that the dataset includes both right-handed and left-handed signs, an essential augmentation technique consisted of mirroring the existing videos. By flipping each sequence horizontally, additional left-handed samples were generated from right-handed signs and vice versa. This strategy serves a dual purpose: first, it effectively doubles the number of available samples (as shown in Table 4); second, and more importantly, it systematically simulates the kinematic structure of opposite-handed signing. Including these mirrored samples in the dataset rather than relying on random flipping during training we ensure that the model is deterministically exposed to both lateral topologies for every single instance. This guarantees that the trained models learn to be invariant to handedness based on a balanced distribution avoiding the uneven exposure that occurs with stochastic augmentation.

While generic data augmentation is conventionally implemented dynamically as part of the downstream training pipeline (on-the-fly), we opted to provide a pre-computed augmented split as a core component of the dataset release. This methodological choice serves two critical functions aimed at standardization and accessibility. First, regarding benchmarking consistency, dynamic augmentation pipelines often vary significantly across deep learning frameworks (e.g. PyTorch or TensorFlow) and specific library versions. Using a pre-calculated and standardized augmented samples enable the establishment of a fixed hard evaluation set. This ensures that future comparisons between models are attributable to architectural differences instead of discrepancies in augmentation aggressiveness or random seed generation. Second, this approach improves computational accessibility; video-based augmentation is computationally expensive and can become a bottleneck during training. Thus, offloading this process to a preprocessing stage lowers the hardware barrier for researchers, enabling those with limited CPU/GPU bandwidth to train on a robust, high-variability dataset without incurring runtime overhead.

Consequently, various transformations were applied to the dataset using the AugLy [52] library, a multimodal data augmentation tool for video, image, audio, and text. The selected

transformations were carefully designed to preserve the integrity of the signs, ensuring that no excessive rotations or translations were applied that could semantically alter the gesture or move the hands out of the central crop. More precisely, geometric rotation was capped at $\pm 15°$ to simulate natural camera tilt without invalidating the pose.

As a result of these increases, the total size of the dataset increased to 61,409 samples, distributed as shown in Table 5. Particularly, the applied transformations along with their application probabilities are as follows:

- Contrast = 1.5; Probability = 0.35
- Brightness = 0.2; Probability = 0.35
- Saturation = 2.8; Probability = 0.35
- Blur = 2.5; Probability = 0.30
- Noise = 50; Probability = 0.30
- Rotation between -15° and -5°; Probability = 0.40
- Rotation between 5° and 15°; Probability = 0.40

Among the different augmentation techniques employed, the mirroring transformation was applied to both the video sequences and the skeletal keypoints. In contrast, transformations affecting the image appearance, such as color and brightness adjustments, were only applied to the video data. This distinction was necessary because changes in image tone do not affect the extracted skeletal joint points, which remain unaffected by variations in lighting or color.

Furthermore, since the dataset is built from high-resolution RGB recordings, all augmentations were carefully selected to avoid distortions that could degrade the clarity of the sign. For instance, geometric transformations were applied conservatively to ensure that hand movements and key signing regions remained intact.

Data Records

The complete dataset is available at Science Data Bank repository under the name of *Sign4all: a Spanish Sign Language Dataset* [53]. This dataset consists of six different versions, each tailored for specific applications in Sign Language Recognition (SLR). This version includes raw RGB recordings, temporally normalized and background-cropped RGB videos, their augmented variants, and skeletal keypoints formatted for SLR. The dataset is structured to facilitate research in both vision-based and pose-based sign recognition models.

Dataset versions

The dataset is distributed in six different formats to facilitate multiple experimental settings:

1. **RGB Original Dataset**: The raw high-resolution RGB videos, cropped per sing repetition but without background removal or temporal normalization.
2. **RGB Normalized Dataset**: The RGB dataset after temporal normalization and cropped to 1224×1224 pixels to remove background elements.
3. **RGB Normalized Dataset with Mirroring**: A version of the normalized dataset that applies horizontal flipping, effectively doubling the dataset size.
4. **RGB Normalized Dataset with Augmentation**: A version of previous dataset with additional transformation, such as brightness and contrast adjustments, to enhance model generalization.

Sign	Num. Videos
Meat (*Carne*)	654
Hamburger (*Hamburguesa*)	686
Eggs (*Huevos*)	714
Fish (*Pescado*)	694
Pizza (*Pizza*)	654
Soup (*Sopa*)	680
Omelet (*Tortilla*)	656
Vegetables (*Verdura*)	634
When (*Cuándo*)	628
Where (*Dónde*)	614
What (*Qué*)	640
You (*Tú*)	614
Me (*Yo*)	610
Spoon (*Cuchara*)	618
Knife (*Cuchillo*)	636
Plate (*Plato*)	640
Fork (*Tenedor*)	608
Glass (*Vaso*)	626
Like (*Gustar*)	650
Dislike (*No*)-gustar	692
Want (*Querer*)	618
Dinner (*Cenar*)	668
Eat (*Comer*)	636
Breakfast (*Desayunar*)	622
Total	15,512

Table 4: Dataset distribution after applying mirror technique

Sign	Num. Videos
Meat (*Carne*)	2,312
Hamburger (*Hamburguesa*)	2,760
Eggs (*Huevos*)	2,856
Fish (*Pescado*)	2,808
Pizza (*Pizza*)	2,616
Soup (*Sopa*)	2,720
Omelet (*Tortilla*)	2,632
Vegetables (*Verdura*)	2,624
When (*Cuándo*)	2,520
Where (*Dónde*)	2,464
What (*Qué*)	2,544
You (*Tú*)	2,472
Me (*Yo*)	2,448
Spoon (*Cuchara*)	2,472
Knife (*Cuchillo*)	2,392
Plate (*Plato*)	2,568
Fork (*Tenedor*)	2,425
Glass (*Vaso*)	2,520
Like (*Gustar*)	2,256
Dislike (*No-gustar*)	2,776
Want (*Querer*)	2,480
Dinner (*Cenar*)	2,704
Eat (*Comer*)	2,552
Breakfast (*Desayunar*)	2,488
Total	61,409

Table 5: Dataset distribution after data augmentation

5. **Skeletal Keypoint Dataset**: Extracted keypoints aligned with the RGB Normalized Dataset, with 48 frames per sample, keypoints manually selected for SLR and normalized to the shoulder midpoint for consistency.

6. **Skeletal Keypoint Dataset with Mirroring**: The mirrored version of the skeletal keypoint dataset, aligned with the mirrored RGB dataset.

All RGB videos are provided in AVI format using the MJPEG codec to ensure high visual quality and frame-by-frame integrity. Skeletal keypoints are stored in HDF5 (*.hd*) format, preserving spatial relationships while reducing storage size.

File organization

The dataset is structured in a hierarchical folder system as follows:

```
Dataset_Version/
    |-- personX_hand
    |   |-- signClass/
    |   |   |-- label/
    |   |   |   |-- repetition1.avi (or repetition1.h5)
    |   |   |   |-- repetition2.avi (or repetition2.h5)
```

Each file follows this convention:

- **personX_right** / **personX_left**: Since each participant signs with both hands, each individual is represented twice, effectively creating 16 unique identities instead of 8.
- **signClass**: Signs are categorized into six linguistic groups according to their meaning, as shown in Table 6.
- **label**: The specific sign being performed.
- **repetitionN**: Each iteration of a given sign.

Category	Signs included (numerical label)
Food items	carne (0), hamburguesa (1), huevos (2), pescado (3), pizza (4), sopa (5), tortilla (6), verdura (7)
Interrogative particles	cuándo (8), dónde(9), qué (10)
Pronouns	tú (11), yo (12)
Kitchen utensils	cuchara (13), cuchillo (14), plato (15), tenedor (16), vaso (17)
Verbs	gustar (18), no-gustar (19), querer (20)
Eating-related verbs	cenar (21), comer (22), desayunar (23)

Table 6: Sign categories in the dataset. The number between each sign in parenthesis corresponds to its numerical label

Skeletal keypoints data structure

The skeletal information is distributed in HDF5 files that store a $48 \times 50 \times 2$ matrix. The second dimension of the matrix corresponds to a specific skeletal keypoint for MediaPipe, where the last dimension refers to X and Y axis, and the first dimensions represents an RGB frame. So, the rows represent the temporal evolution of the keypoints.

Key landmarks for Sign Language Recognition, including the hand, wrist, elbow, shoulder, and torso, are chosen. Each column is linked to a corresponding keypoint (with X, Y location) and is structured in the following order: thumbR, indexR, middleR, ringR, pinkyR, bodyR, bodyL, pinkyL, ringL, middleL, indexL, thumbL; where:

- R denotes points on the right side of the body, while L denotes the left side.
- Each point is represented by its spatial (X, Y) coordinates.
- The detailed composition of each category is as follows:
  - Thumb = wrist, thumb1, thumb2, thumb3, thumb4
  - Index = index1, index2, index3, index4
  - Middle = middle1, middle2, middle3, middle4
  - Ring = ring1, ring2, ring3, ring4
  - Pinky = pinky1, pinky2, pinky3, pinky4
  - Body = wrist, elbow, shoulder, hip

Dataset splits: training, validation and testing

To ensure robust generalization and prevent overfitting to specific individuals, a train, validation and test split is proposed for every dataset. In this sense, both validation and test sets contain unseen participants for better generalization and signer-independent evaluation. Also, the splits are structured to maintain class balance across all categories.

For these splits, each dataset folder contains three csv files (`train.csv`, `val.csv`, `test.csv`) with the necessary information to load the data. The files follow the format shown in Table 7 as an example. In it, the headers of the table correspond to the headers of the csv files, where:

- **sequence**: Path to video (AVI) or keypoints (HDF5) file, depending on the dataset.
- **person**: Numerical identifier (0-7) corresponding to each participant. Left-handed and right-handed versions of the same signer share the same number, allowing for different person-based training strategies.
- **label**: Numerical class identifier (0-23) assigned to each sign.
- **name**: Text label of the sign, written in Spanish.

sequence	person	label	name
person0_right/alimentacion/carne/repetition1.avi	0	0	carne
person1_left/alimentacion/sopa/repetition2.avi	1	5	sopa
person5_left/utensilios/tenedor/repetition15.avi	5	11	tenedor
person7_right/verbos/querer/repetition6.avi	7	21	querer

Table 7: Example of a csv file organization

Technical Validation

This section presents the procedures used to verify data integrity, confirm preprocessing effectiveness, evaluate the impact of data augmentation, and evaluate the usability of skeletal keypoints for sign recognition. Finally, we outline potential future improvements.

Visual inspection and data consistency

In order to achieve a high-quality and precise dataset, each video was manually revised to segment individual sign repetitions accurately. Misperformed or ambiguous samples were removed, guaranteeing that only high-quality data was kept. This verification process was essential for maintaining clean and well-structured annotations across all samples.

Additionally, background removal was applied using EfficientDetD1 together with manual annotation, to effectively isolate the signer while maintaining a fixed resolution of 1224×1224 pixels. Figure 13 illustrates the effectiveness of this method, where the cropped images show a clear focus on the signing region. The successful training of the models described below empirically validated this spatial normalization strategy, confirming that the square aspect ratio removes the geometric distortions common in resizing operations and provides a direct compatibility with standard deep learning input tensors.

To standardize input length, all videos were normalized to 48 frames, a value determined through empirical evaluation to ensure complete gesture representation while keeping computational efficiency.

Impact of data augmentation

Vision transformer models require large-scale datasets to generalize effectively, as stated by Dosovitskiy et al. [54], particularly the Video Vision Transformer (ViViT) [49] architecture. A model focused on video recognition which achieves state-of-the-art values when a great amount of data is used for training. To empirically validate the necessity of including the pre-computed augmentation split within the dataset release, we conducted an ablation study comparing the performance of identical architectures on raw versus augmented data.

Both models used the train-test-validation split described in *Dataset splits* subsection to ensure a fair comparison. Without augmentation, the model achieved less than 10% accuracy on the validation set. In contrast, the model trained on the pre-computed augmented split achieved 39% accuracy. This four-times performance increase validates the methodological decision to offload augmentation to the dataset construction phase, proving that the proposed transformations successfully bridge the gap between the raw sample size and the high data requirements of Transformer architecture. Although the model did not achieve outstanding results, these data-intensive architectures were used to evaluate the impact of the proposed augmentation techniques compared to training on the non-augmented dataset.

Figure 17 illustrates examples of augmented samples, showing how the applied transformations maintain the semantic integrity of the signs while enhancing dataset diversity. These results confirm that data augmentation plays a crucial role in improving recognition accuracy, particularly for data-hungry architectures like ViViT.

Validation of skeletal keypoints

To evaluate the effectiveness of the extracted skeletal keypoints in Sign Language Recognition, a Fully Connected Network (FCN) was trained exclusively on skeletal data. The model was trained with the same train-test-validation split as described in *Dataset splits* subsection and in the ViViT architectures and achieved an 81.64% of accuracy on the test set, confirming that skeletal keypoints have enough information for distinguishing between different signs. An analysis of the confusion matrix (Figure 18) reveals a consistent performance across the different classes, particularly 15 out of 24 signs achieved high recognition rates (accuracy higher than 90%). However, the confusion matrix also highlights specific confusions due to phonological similarity. For instance, the sign [VEGETABLES] is mostly confused with sign [DISLIKE] since both of them have the same movement and location but change the number of fingers involved in the signing. In a similar way, the sign [FORK] is confused with [HAMBURGER] because

Figure 17: Augmented samples

they share the same contact point in the hand, which may cause occlusions; another example is the sign [WANT] which is confused with sign [FISH], in this case they share the same location and hand shape but the movement is in opposite directions. Even though there exists this small errors, the results demonstrates that the dataset provides a robust and structured representation of motion, which is essential for ISLR applications.

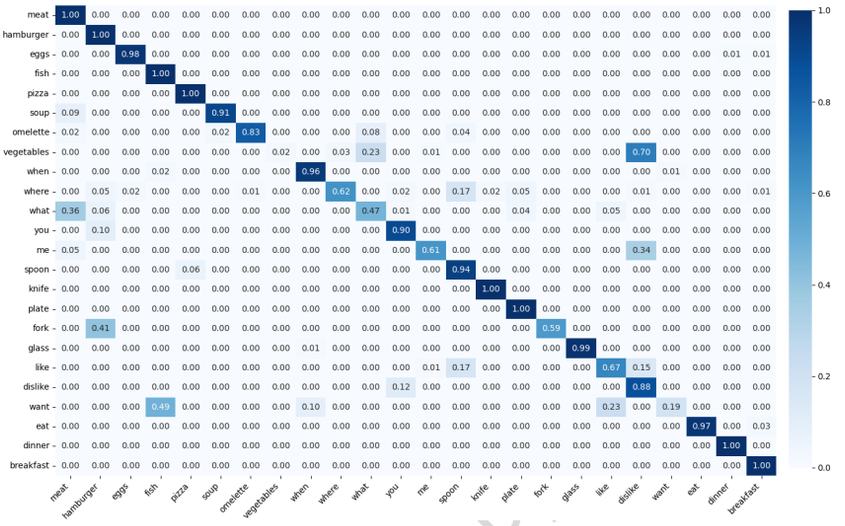


Figure 18: Confusion matrix for skeletal keypoints model

Furthermore, this performance achieves similar results to some state-of-the-art methods and even surpasses them [10, 21], emphasizing the high precision and consistency of the extracted keypoints. Additionally, the normalization process used in each keypoint ensures generalization of the model, making it invariant to the signers location. For instance, Figure 19 depicts two signs performed in different locations from the camera and both of them were properly recognized.

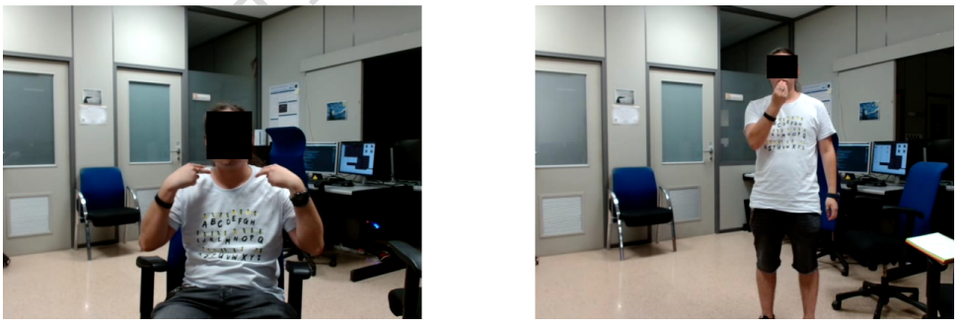


Figure 19: Different location for a sign

Usage Notes

This dataset is designed for Isolated Sign Language Recognition and is already formatted to be used in deep learning models. In this sense, the RGB videos are provided in AVI format (MJPEG codec) and the skeletal keypoints are stored in HDF5 files, both of them are widely supported formats in deep learning frameworks like TensorFlow or PyTorch.

Given that the dataset is distributed with temporal normalization, background removal, square cropping, and keypoints extraction, there is no need to add any further preprocessing step before training most deep learning architectures; this allows researchers to directly load and train models without any custom processing. However, if for any reason researchers require raw

video with the complete background and the full length of the signs, an unaltered version of the RGB recordings is also provided. Additionally, we provide an augmented version of the dataset, which adds custom transformations aimed to improve model robustness in real-world scenarios.

Data availability

The complete Sign4all dataset is available at Science Data Bank [53]. Because the dataset contains identifiable facial and body features –including characteristics from which gender may be inferred– participants are exposed to a potential risk of re-identification. For this reason, access to the dataset is restricted and subject to manual request. To obtain the data, researchers must agree to a Data Usage Agreement (DUA) and provide contact information such as name, email address and affiliation details. Once the request is verified, access will be granted via a secure download link sent by email.

Code availability

None of the six variations of the proposed dataset require any custom code for access or processing since all the data is provided in widely supported formats, as mentioned before.

The dataset was processed with Blender 4.0 as video editing software, MediaPipe 0.10.1 for keypoint extraction, and TensorFlow 2.12.0 for model training and testing; all of them under an Arch Linux operative system with Python 3.8. For the dataset recording, Azure Kinect SDK 1.3 [55] with PyKinect Azure [56] as Python wrapper was used under Ubuntu 18.04 LTS.

Acknowledgments

This work has been partially funded by a PhD grant under the reference UAFPU21-78 from the University of Alicante (Spain).

Author contributions statement

F.M.E. and E.M.M. defined the vocabulary; F.M.E. recorded, filtered and processed the data, also performed the technical validation; E.M.M. supervised the experiments. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

References

[1] Duda, R. O., Hart, P. E. & Stork, D. G. *Pattern Classification (2nd Edition)* (Wiley-Interscience, USA, 2000).

[2] Tao, T., Zhao, Y., Liu, T. & Zhu, J. Sign language recognition: A comprehensive review of traditional and deep learning approaches, datasets, and challenges. *IEEE Access* **PP**, 1–1, 10.1109/ACCESS.2024.3398806 (2024).

[3] Adaloglou, N. *et al.* A comprehensive study on deep learning-based methods for sign language recognition. *IEEE Transactions on Multimedia* **24**, 1750–1762, 10.1109/tmm.2021.3070438 (2022).

[4] Koller, O., Forster, J. & Ney, H. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding* **141**, 108–125, https://doi.org/10.1016/j.cviu.2015.09.013 (2015). Pose and Gesture.

[5] Camgoz, N. C., Hadfield, S., Koller, O., Ney, H. & Bowden, R. Neural sign language translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2018).

[6] Duarte, A. *et al.* How2sign: A large-scale multimodal dataset for continuous american sign language (2021). 2008.08143.

[7] Sanabria, R. *et al.* How2: A large-scale dataset for multimodal language understanding (2018). 1811.00347.

[8] Zhou, H., Zhou, W., Qi, W., Pu, J. & Li, H. Improving sign language translation with monolingual data by sign back-translation (2021). 2105.12397.

[9] Armstrong, D. F., Stokoe, W. C. & Wilcox, S. E. Gesture and the nature of language. *Cambridge University Press* (1995).

[10] Li, D., Opazo, C. R., Yu, X. & Li, H. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison (2020). 1910.11006.

[11] Caselli, N., Sehyr, Z., Cohen-Goldberg, A. & Emmorey, K. Asl-lex: A lexical database of american sign language. *Behavior Research Methods* **49**, 10.3758/s13428-016-0742-0 (2016).

[12] Sehyr, Z. S., Caselli, N., Cohen-Goldberg, A. M. & Emmorey, K. The asl-lex 2.0 project: A database of lexical and phonological properties for 2,723 signs in american sign language. *The Journal of Deaf Studies and Deaf Education* **26**, 263–277, 10.1093/deafed/enaa038 (2021). https://academic.oup.com/jdsde/article-pdf/26/2/263/36643382/enaa038.pdf.

[13] Joze, H. R. V. & Koller, O. Ms-asl: A large-scale data set and benchmark for understanding american sign language. *ArXiv* **abs/1812.01053** (2018).

[14] Jin, P. *et al.* A large dataset covering the chinese national sign language for dual-view isolated sign language recognition. *Scientific Data* **12** (2025).

[15] Asl alphabet. https://www.kaggle.com/dsv/29550, 10.34740/KAGGLE/DSV/29550.

[16] MNIST. Sign language mnist. https://www.kaggle.com/datasets/datamunge/sign-language-mnist (2018). Accessed: January 2025.

[17] Al-Barham, M. *et al.* Rgb arabic alphabets sign language dataset, 10.48550/arXiv.2301.11932 (2023).

[18] Sincan, O. M. & Keles, H. Y. Autsl: A large scale multi-modal turkish sign language dataset and baseline methods. *IEEE Access* **8**, 181340–181355, 10.1109/access.2020.3028072 (2020).

[19] Kumwilaisak, W., Pannattee, P., Hansakunbuntheung, C. & Thatphithakkul, N. American sign language fingerspelling recognition in the wild with iterative language model construction. *APSIPA Transactions on Signal and Information Processing* **11**, 10.1561/116.00000003 (2022).

[20] Sunuwar, J., Borah, S. & Kharga, A. Nsl23 dataset for alphabets of nepali sign language. *Data in Brief* **53**, 110080, 10.1016/j.dib.2024.110080 (2024).

[21] Kumar, P., Saini, R., Roy, P. P. & Dogra, D. P. A position and rotation invariant framework for sign language recognition (slr) using kinect. *Multimedia Tools Appl.* **77**, 8823–8846, 10.1007/s11042-017-4776-9 (2018).

[22] Boulesnane, A., Bellil, L. & Ghiri, M. Aslad-190k: Arabic sign language alphabet dataset, 10.31219/osf.io/n236q (2024).

[23] El Kharoua, R. & Jiang, X. Deep learning recognition for arabic alphabet sign language rgb dataset. *Journal of Computer and Communications* **12**, 32–51, 10.4236/jcc.2024.123003 (2024).

[24] Wikipedia. List of sign languages. https://en.wikipedia.org/wiki/List_of_sign_languages (2025). Accessed: March 2025.

[25] Wikipedia. List of sign languages by number of native signers. https://en.wikipedia.org/wiki/List_of_sign_languages_by_number_of_native_signers (2025). Accessed: March 2025.

[26] Ethnologue. Ethnologue. https://www.ethnologue.com (2025). Accessed: March 2025.

[27] Ethnologue. Spanish Sign Language by Ethnologue. https://www.ethnologue.com/language/ssp/ (2025). Accessed: March 2025.

[28] Docío-Fernández, L. *et al.* LSE_UVIGO: A multi-source database for Spanish Sign Language recognition. In Efthimiou, E. *et al.* (eds.) *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives*, 45–52 (European Language Resources Association (ELRA), Marseille, France, 2020).

[29] Vázquez Enríquez, M., Castro, J. L. A., Fernandez, L. D., Jacques Junior, J. C. S. & Escalera, S. Eccv 2022 sign spotting challenge: Dataset, design and results. In Karlinsky, L., Michaeli, T. & Nishino, K. (eds.) *Computer Vision – ECCV 2022 Workshops*, 225–242 (Springer Nature Switzerland, Cham, 2023).

[30] Rodríguez-Moreno, I., Martinez-Otzeta, J. M. & Sierra, B. *A Hierarchical Approach for Spanish Sign Language Recognition: From Weak Classification to Robust Recognition System*, 37–53 (2022).

[31] Morillas-Espejo, F. & Martinez-Martin, E. A real-time platform for spanish sign language interpretation. *Neural Computing and Applications* (2024).

[32] Martinez-Martin, E. & Morillas-Espejo, F. Deep learning techniques for spanish sign language interpretation. *Computational Intelligence and Neuroscience* 10.115/2021/553280 (2021).

[33] Joo, H. *et al.* Panoptic studio: A massively multiview system for social interaction capture (2016). 1612.03153.

[34] Athitsos, V. *et al.* The american sign language lexicon video dataset. *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops* **0**, 1–8, 10.1109/CVPRW.2008.4563181 (2008).

[35] Mavi, A. & Dikle, Z. A new 27 class sign language dataset collected from 173 individuals (2022). 2203.03859.

[36] Albanie, S. *et al.* Bbc-oxford british sign language dataset (2021). 2111.03635.

[37] Efthimiou, E. *et al.* Sign language recognition, generation, and modelling: A research effort with applications in deaf communication. In Stephanidis, C. (ed.) *Universal Access in Human-Computer Interaction. Addressing Diversity*, 21–30 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2009).

[38] Bhatia, P. & Wadhawan, A. Deep learning-based sign language recognition system for static signs. *Neural Computing and Applications* 10.1007/s00521-019-04691-y((2021).

[39] Chai, X., Wang, H. & Chen, X. The devisign large vocabulary of chinese sign language database and baseline evaluations (2014).

[40] Radakovic, M. *et al.* The serbian sign language alphabet: A unique authentic dataset of letter sign gestures. *Mathematics* **12**, 525, 10.3390/math12040525 (2024).

[41] Kapitanov, A., Karina, K., Nagaev, A. & Elizaveta, P. *Slovo: Russian Sign Language Dataset*, 63–73 (Springer Nature Switzerland, 2023).

[42] Stokoe, W.C. *Sign Language Structure: An Outline of the Visual Communication Systems of the American Deaf.* Studies in Linguistics. Occasional Papers (University of Buffalo, 1960).

[43] Azure Kinect DK oficial page. https://azure.microsoft.com/es-es/products/kinect-dk/ (2021). Accessed: February 2025.

[44] Azure Kinect DK hardware specifications. https://learn.microsoft.com/en-us/previous-versions/azure/kinect-dk/hardware-specification#depth-camera-supported-operating-modes (2021). Accessed: February 2025.

[45] Blender's page. https://www.blender.org/ (2025). Accessed: May 2025.

[46] Sarge, V., Andersch, M., Fabel, L., Micikevicius, P. & Tran, J. Tips for Optimization GPU Performance using Tensor Cores. https://developer.nvidia.com/blog/optimizing-gpu-performance-tensor-cores/ (2019). Accessed: February 2025.

[47] He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition (2015). 1512.03385.

[48] Tan, M. & Le, Q. V. Efficientnet: Rethinking model scaling for convolutional neural networks (2020). 1905.11946.

[49] Arnab, A. *et al.* Vivit: A video vision transformer (2021). 2103.15691.

[50] Tan, M., Pang, R. & Le, Q. V. Efficientdet: Scalable and efficient object detection (2020). 1911.09070.

[51] Lugaresi, C. *et al.* Mediapipe: A framework for building perception pipelines (2019). 1906.08172.

[52] Papakipos, Z. & Bitton, J. Augly: Data augmentations for robustness (2022). 2201.06494.

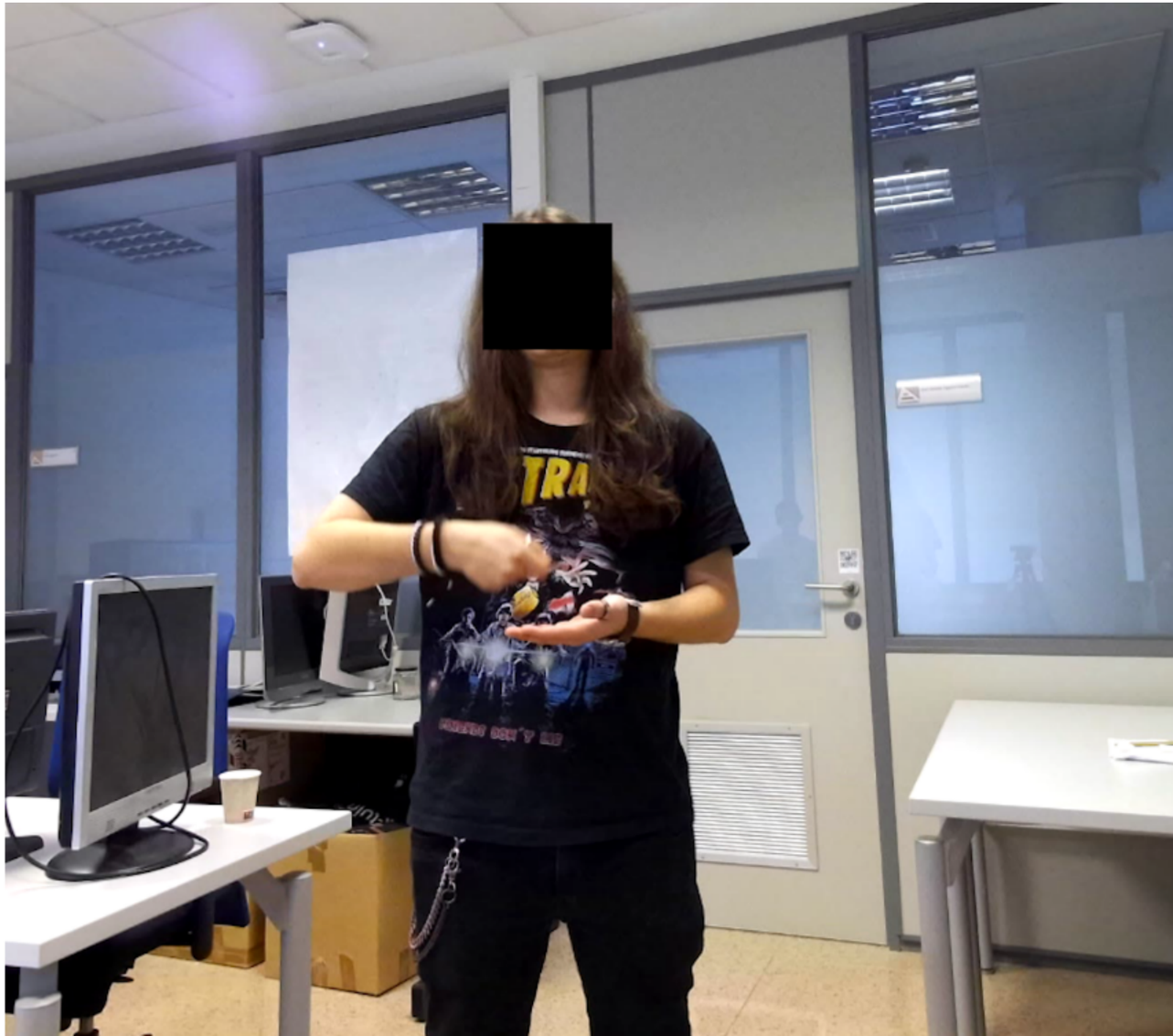
[53] Morillas-Espejo, F. & Martinez-Martin, E. Sign4all: a spanish sign language dataset, 10.57760/sciencedb.28304 (2025).

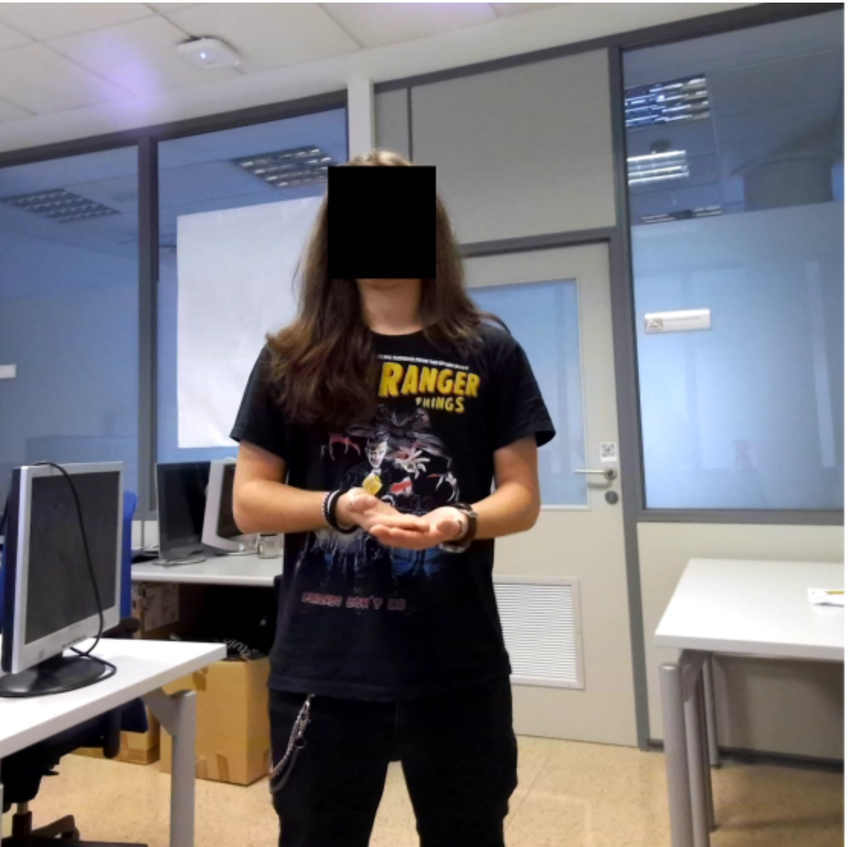
[54] Dosovitskiy, A. *et al.* An image is worth 16x16 words: Transformers for image recognition at scale (2021). 2010.11929.

[55] Azure Kinect Sensor SDK oficial documentation. https://microsoft.github.io/Azure-Kinect-Sensor-SDK/master/index.html. Accessed: February 2025.

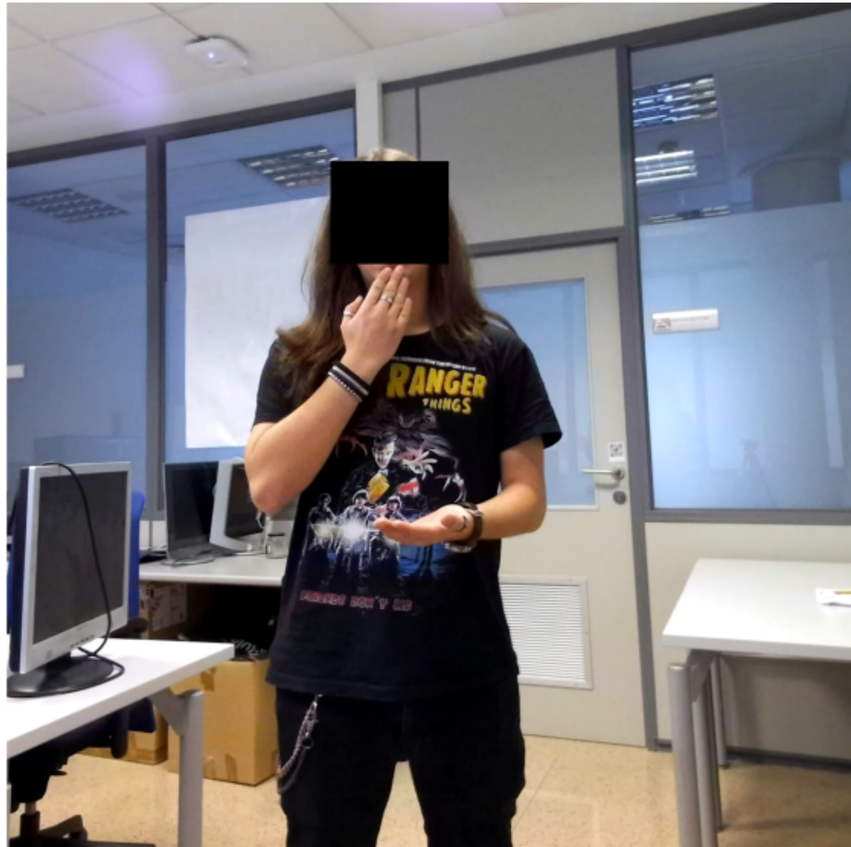
[56] Gorordo, I. PyKinectAzure GitHub page. https://github.com/ibaiGorordo/pyKinectAzure?tab=readme-ov-file (2020). Accessed: February 2025.

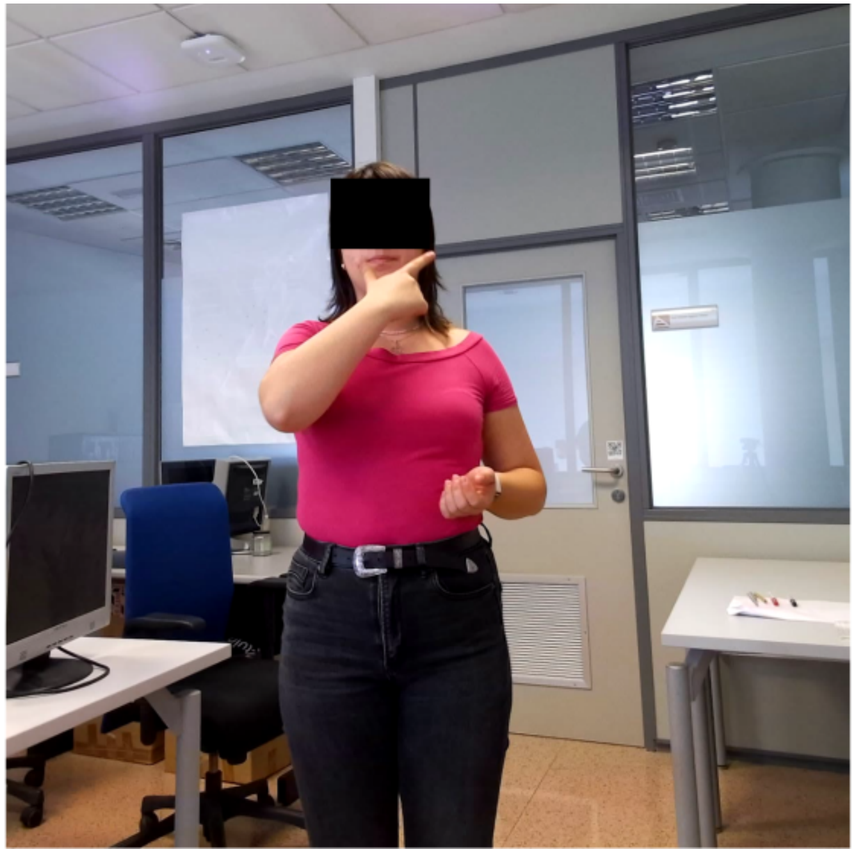

RANGER

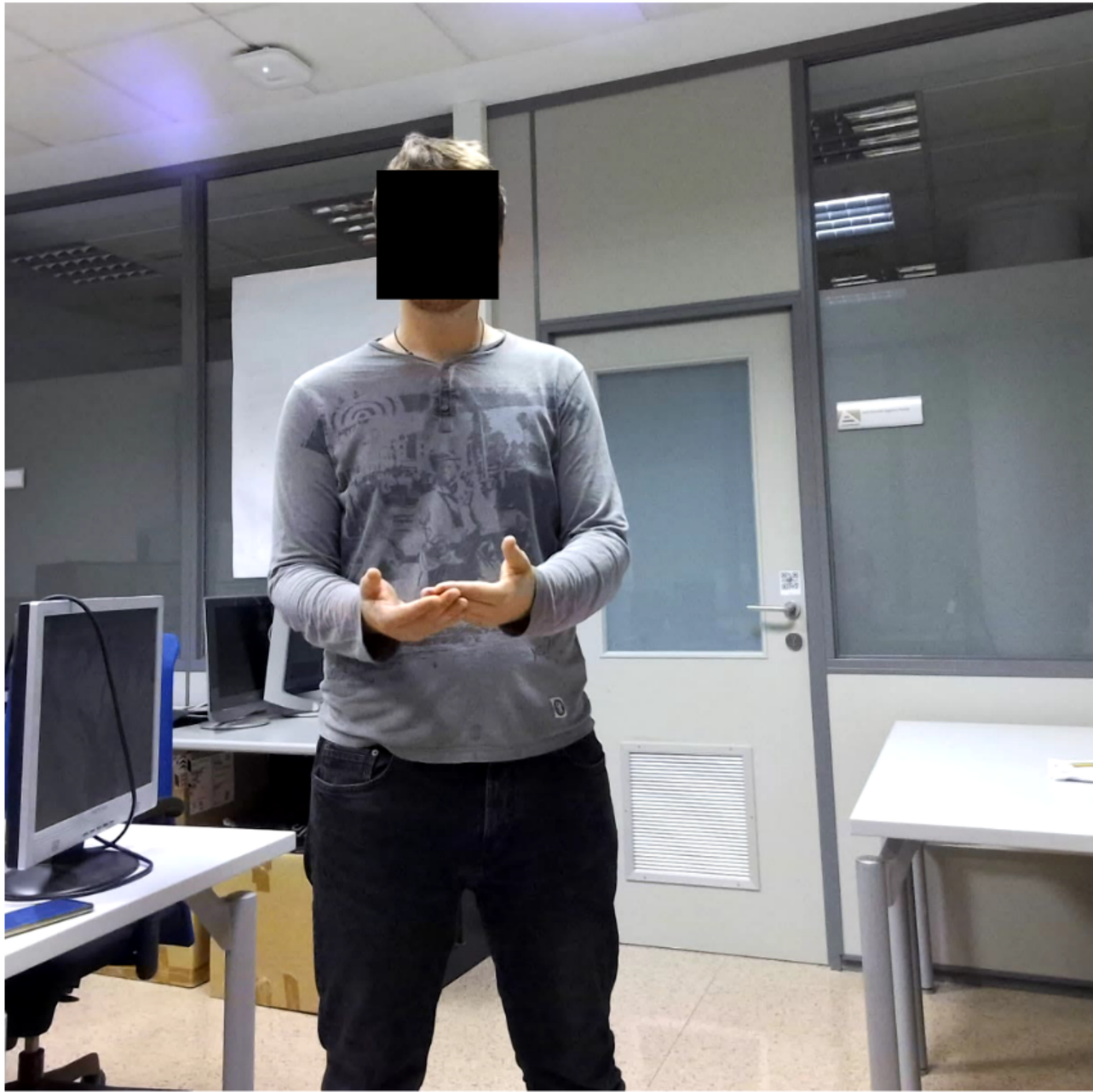


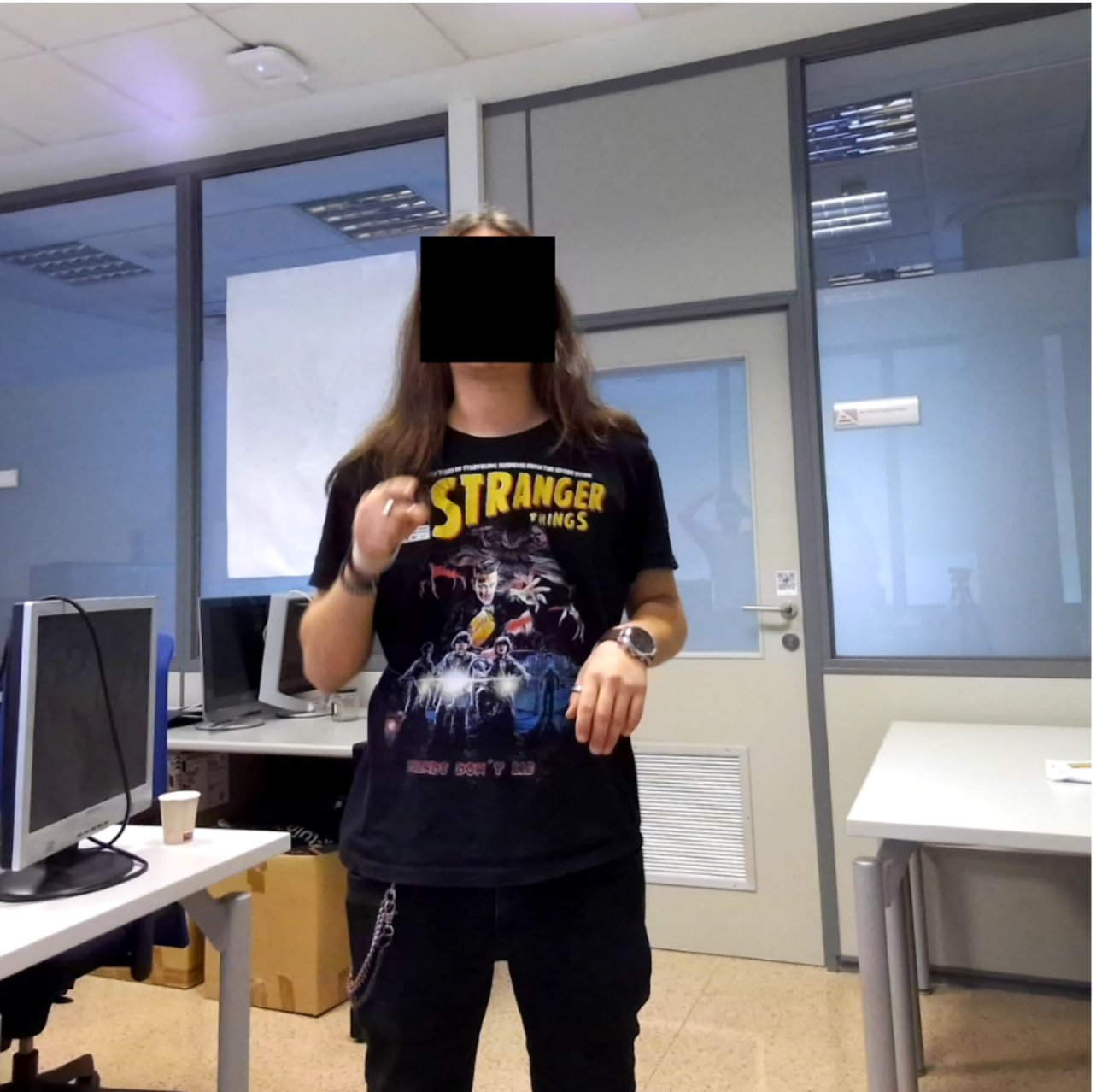

RANGER
THINGS

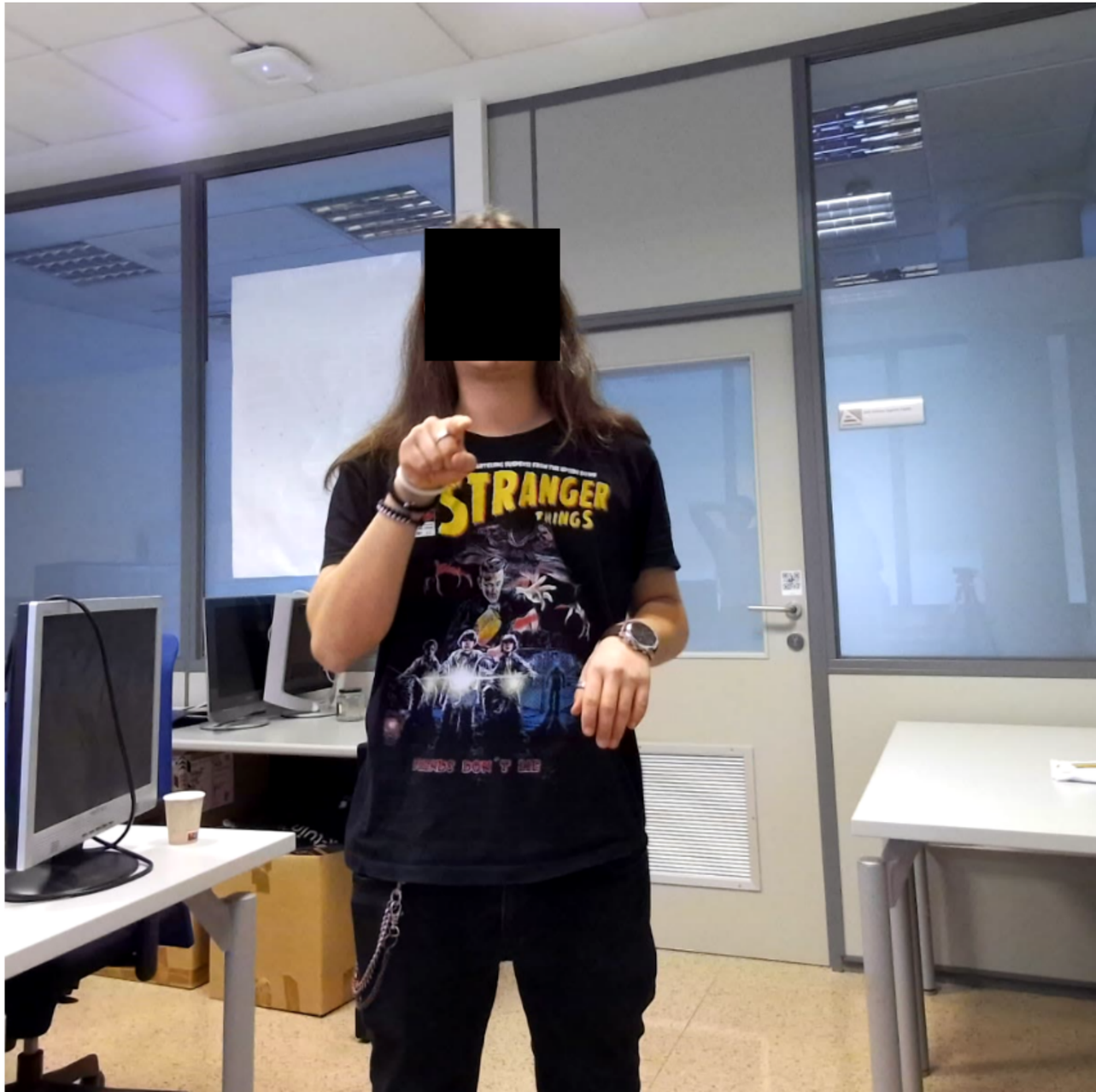

RANGER
THINGS

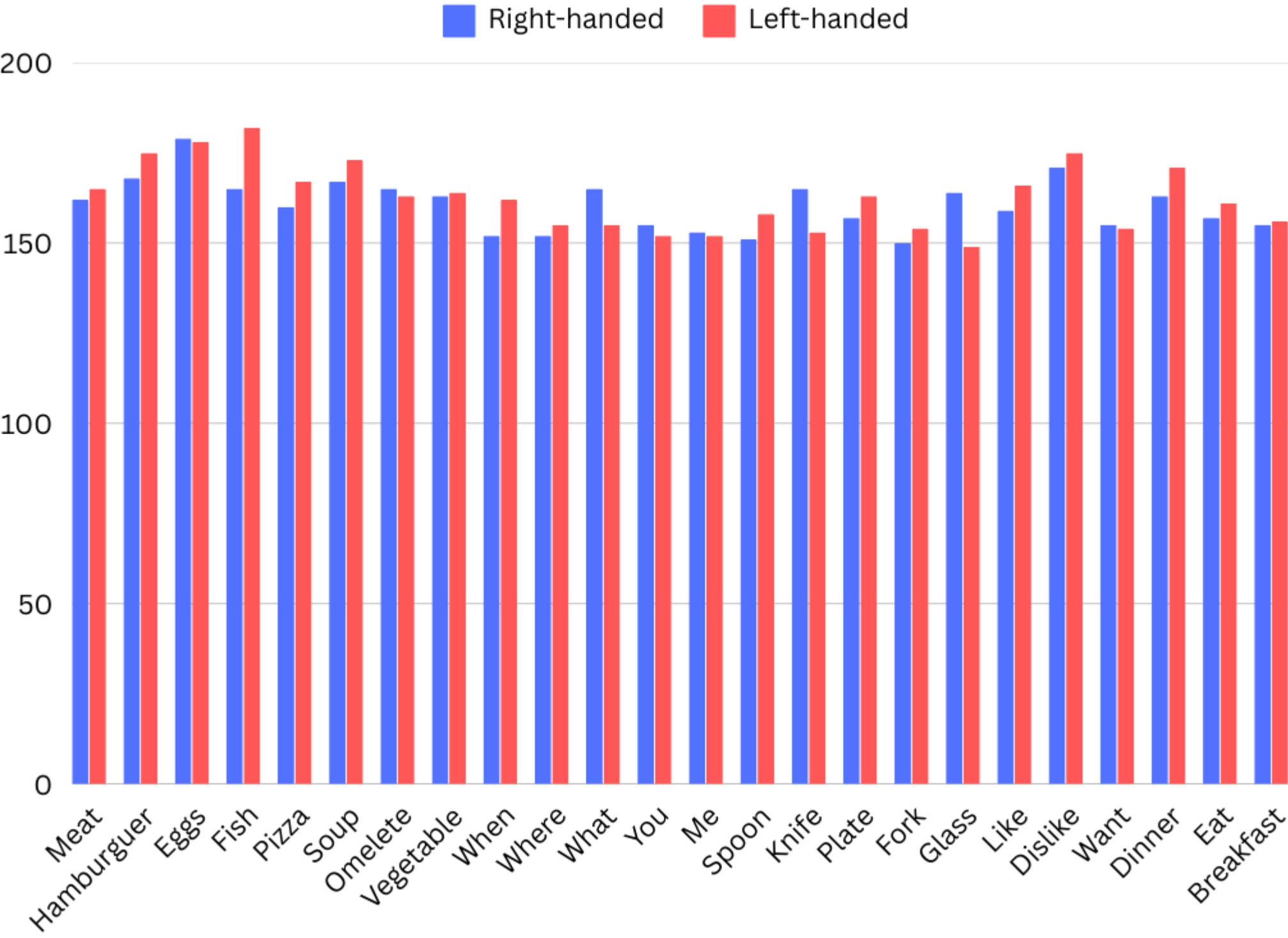

RANGER
THINGS

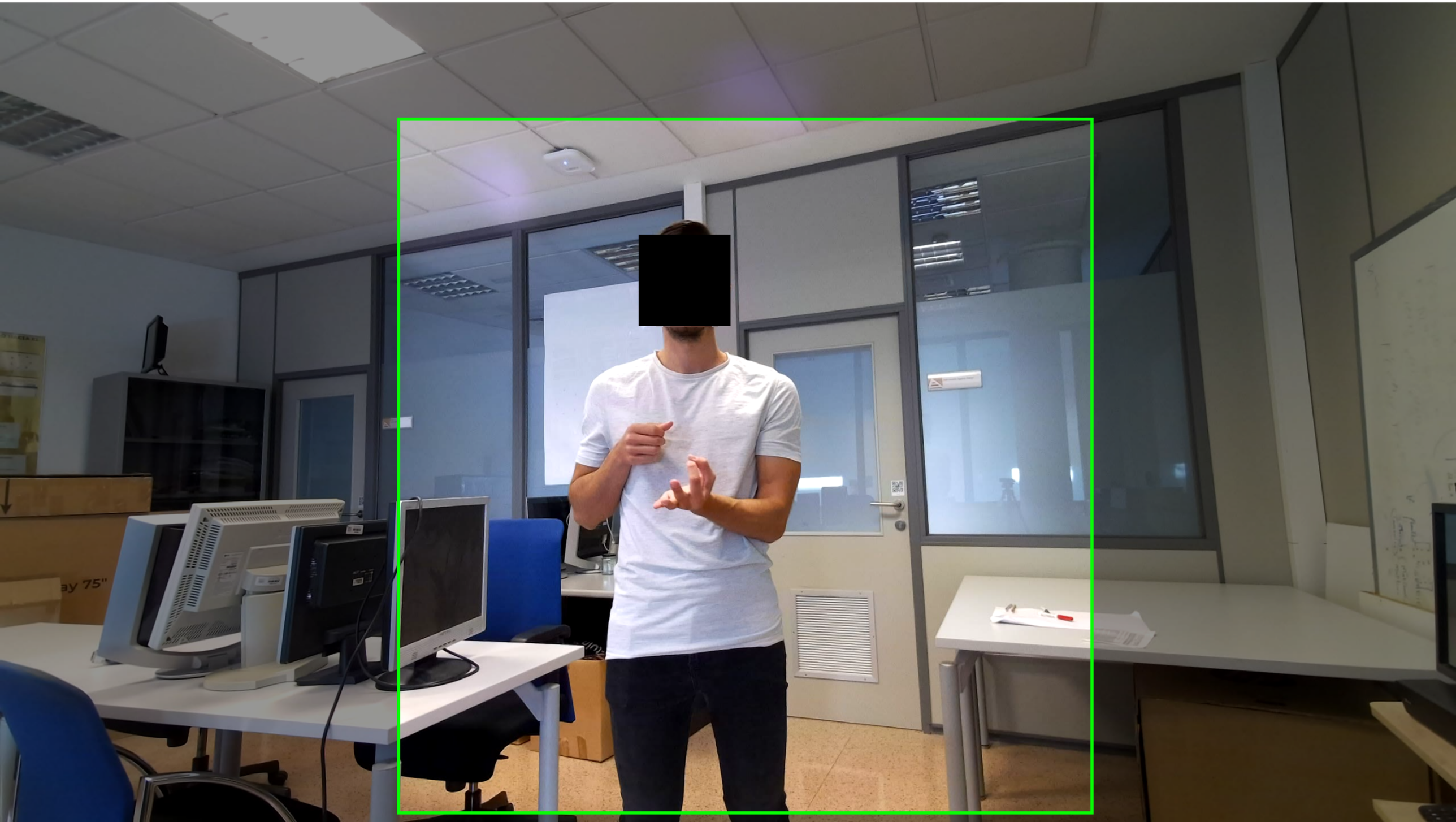


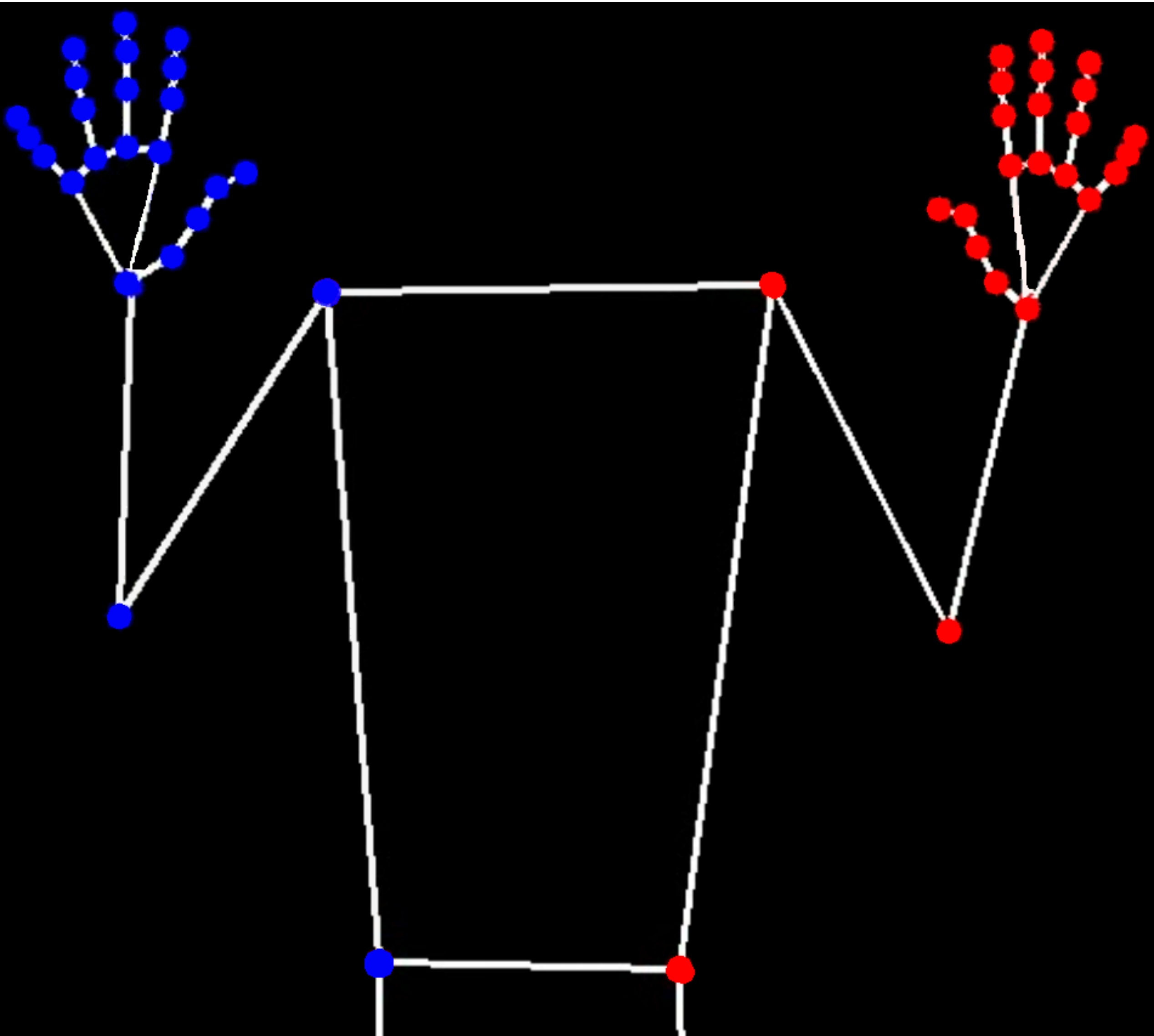


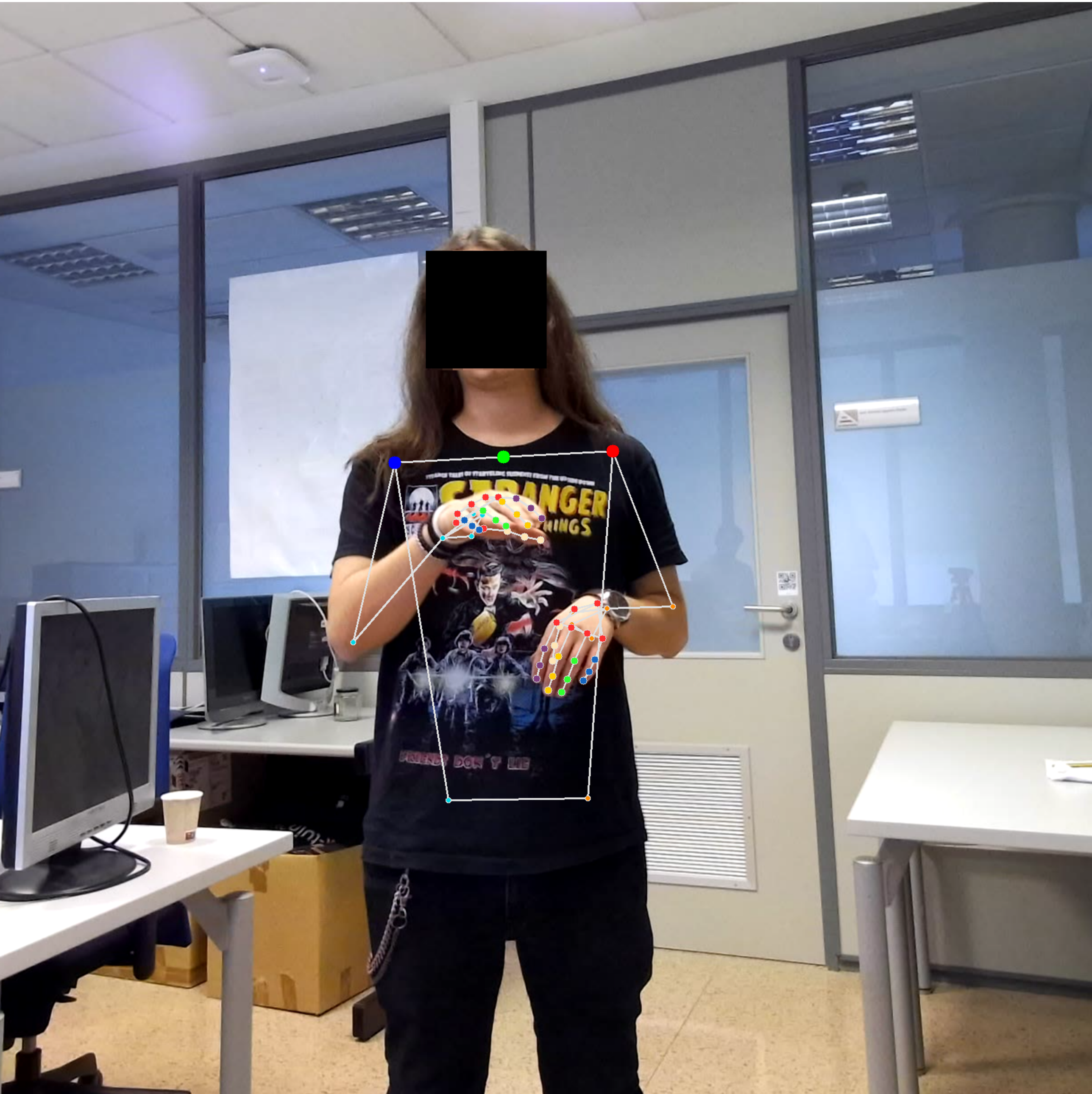

STRANGER
THINGS

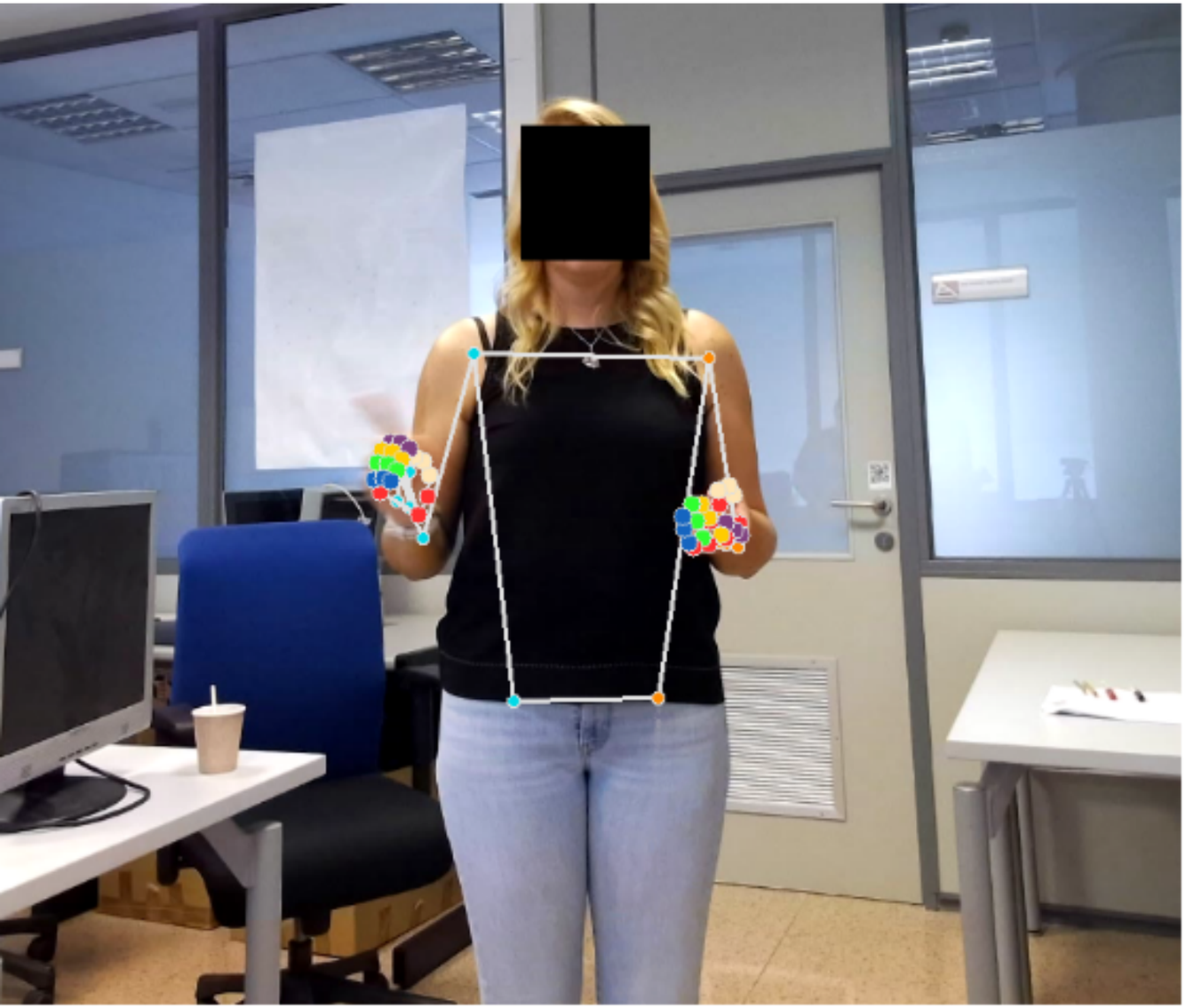

STRANGER
THINGS


Right-handed
Left-handed
200
150
100
50
0
Meat
Hamburguer
Eggs
Fish
Pizza
Soup
Omelete
Vegetable
When
Where
What
You
Me
Spoon
Knife
Plate
Fork
Glass
Like
Dislike
Want
Dinner
Eat
Breakfast



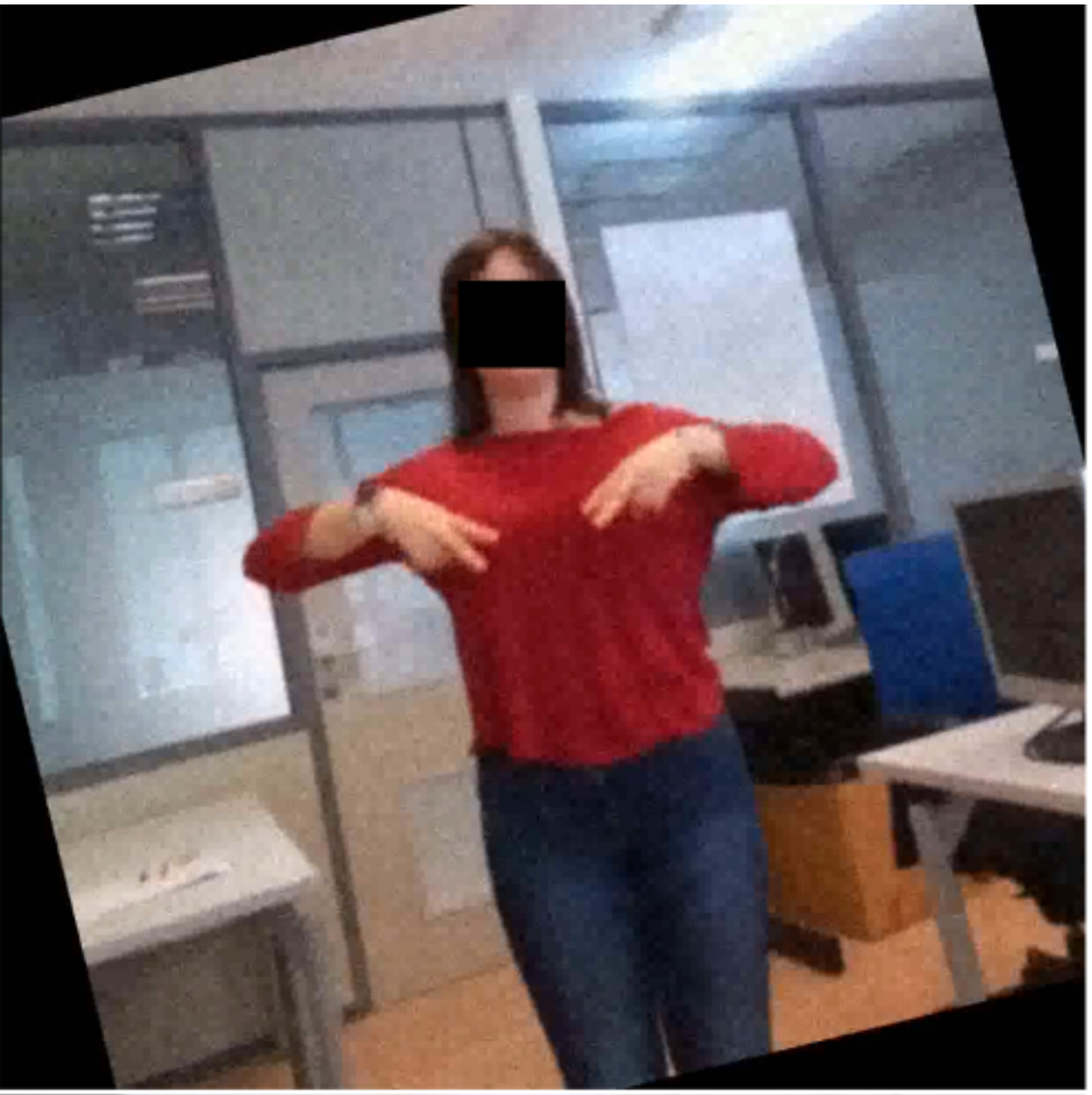


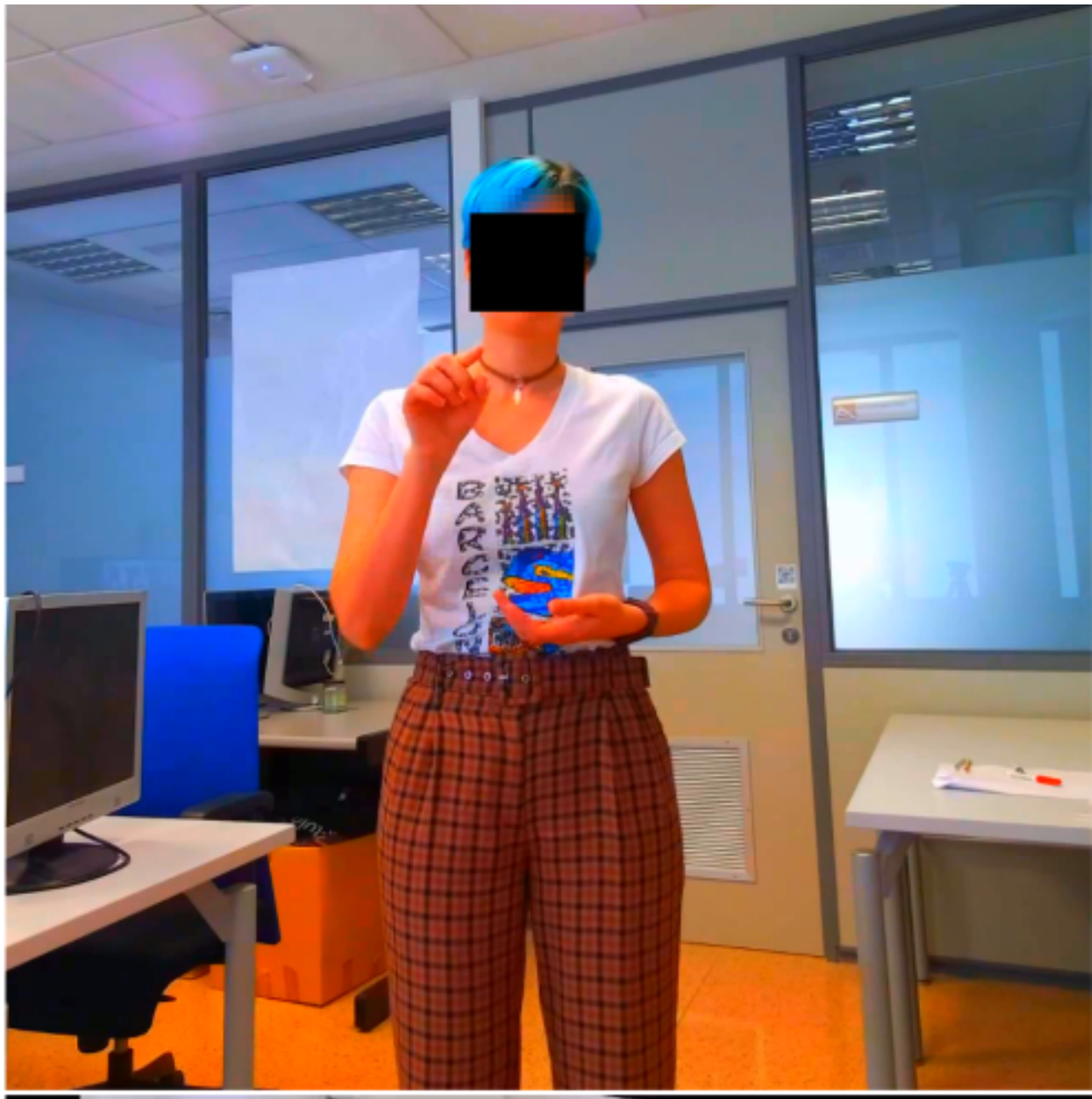

STRANGER
HINGS
FRIENDS DON'T LIE

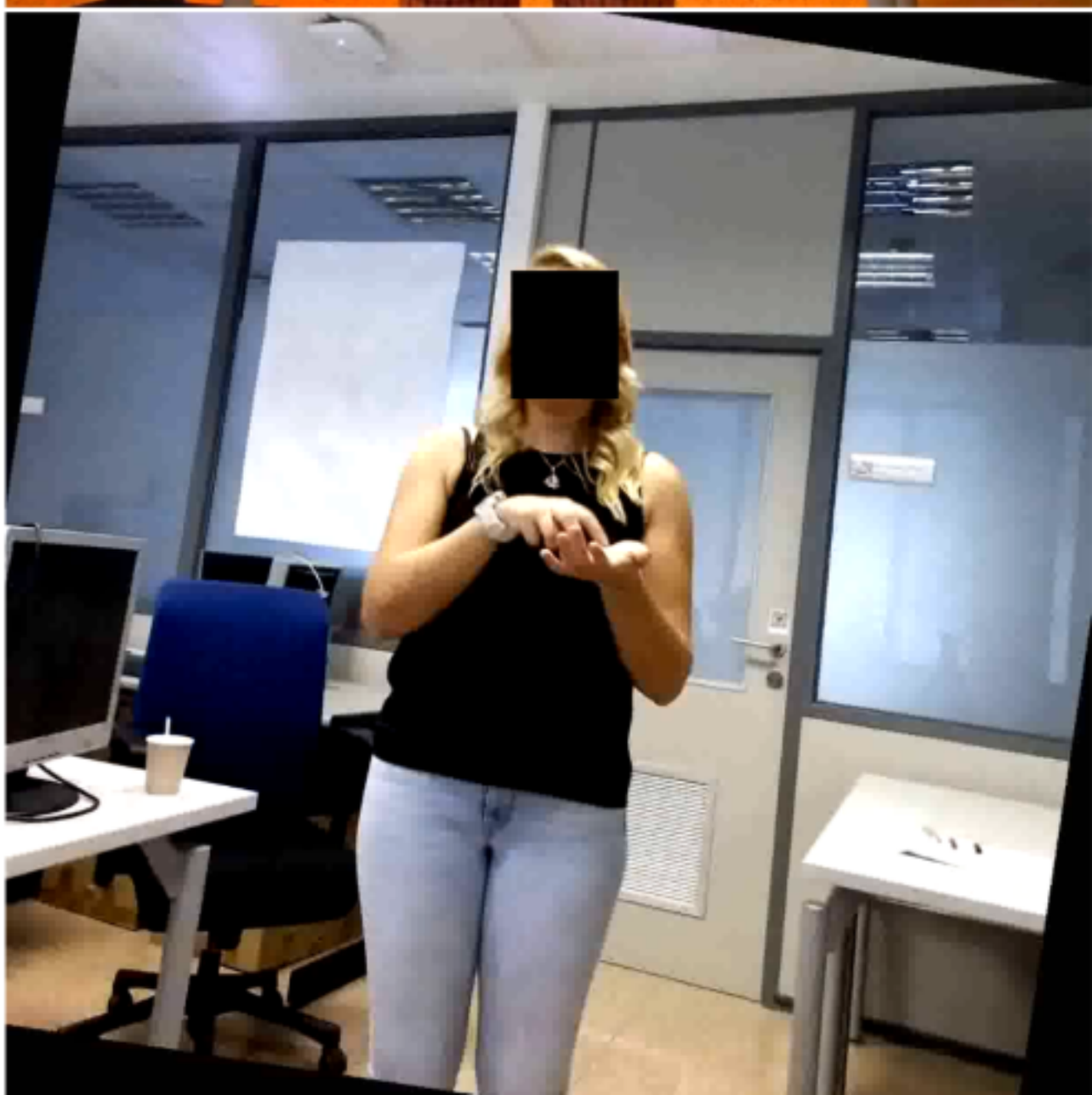


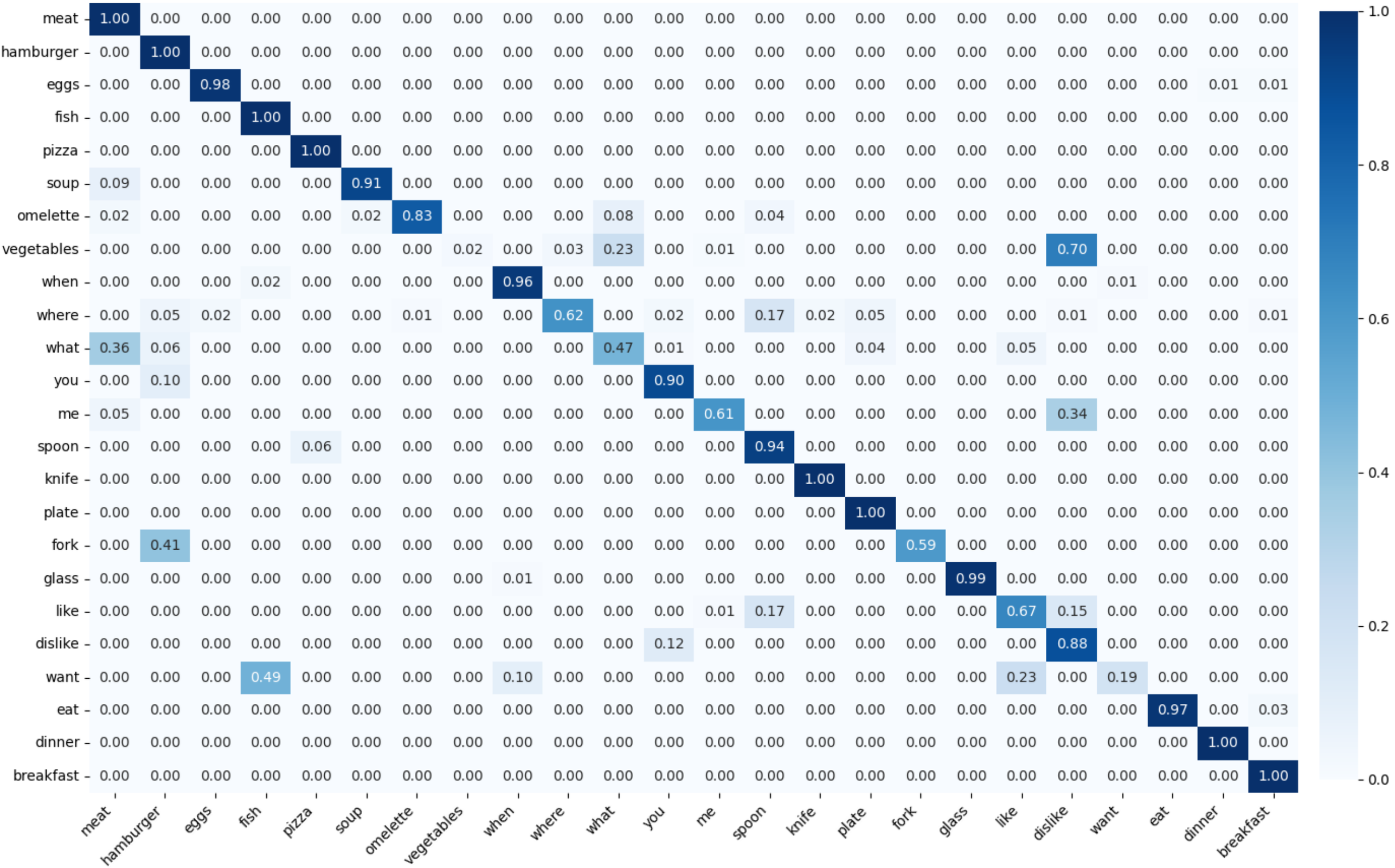
(a) [DISLIKE] sign

(b) [HAMBURGER] sign








meat 1.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
hamburger 0.00 1.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
eggs 0.00 0.00 0.98 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.01 0.01
fish 0.00 0.00 0.00 1.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
pizza 0.00 0.00 0.00 0.00 1.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
soup 0.09 0.00 0.00 0.00 0.00 0.91 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
omelette 0.02 0.00 0.00 0.00 0.00 0.02 0.83 0.00 0.00 0.00 0.08 0.00 0.00 0.04 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
vegetables 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.02 0.00 0.03 0.23 0.00 0.01 0.00 0.00 0.00 0.00 0.00 0.00 0.70 0.00 0.00 0.00 0.00
when 0.00 0.00 0.00 0.02 0.00 0.00 0.00 0.00 0.96 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.01 0.00 0.00 0.00
where 0.00 0.05 0.02 0.00 0.00 0.00 0.01 0.00 0.00 0.62 0.00 0.02 0.00 0.17 0.02 0.05 0.00 0.00 0.00 0.01 0.00 0.00 0.00 0.01
what 0.36 0.06 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.47 0.01 0.00 0.00 0.00 0.04 0.00 0.00 0.05 0.00 0.00 0.00 0.00 0.00
you 0.00 0.10 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.90 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
me 0.05 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.61 0.00 0.00 0.00 0.00 0.00 0.00 0.34 0.00 0.00 0.00 0.00
spoon 0.00 0.00 0.00 0.00 0.06 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.94 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
knife 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 1.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
plate 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 1.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00
fork 0.00 0.41 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.59 0.00 0.00 0.00 0.00 0.00 0.00 0.00
glass 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.01 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.99 0.00 0.00 0.00 0.00 0.00 0.00
like 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.01 0.17 0.00 0.00 0.00 0.00 0.67 0.15 0.00 0.00 0.00 0.00
dislike 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.12 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.88 0.00 0.00 0.00 0.00
want 0.00 0.00 0.00 0.49 0.00 0.00 0.00 0.00 0.10 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.23 0.00 0.19 0.00 0.00 0.00
eat 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.97 0.00 0.03
dinner 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 1.00 0.00
breakfast 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00 1.00
meat
hamburger
eggs
fish
pizza
soup
omelette
vegetables
when
where
what
you
me
spoon
knife
plate
fork
glass
like
dislike
want
eat
dinner
breakfast
1.0
0.8
0.6
0.4
0.2
0.0


A B C D E F G H
I J K L M N O P Q
R S T U V W X Y Z