SIGN-LANGUAGE

Spring 2026

Meeting No.2

Datasets and Models

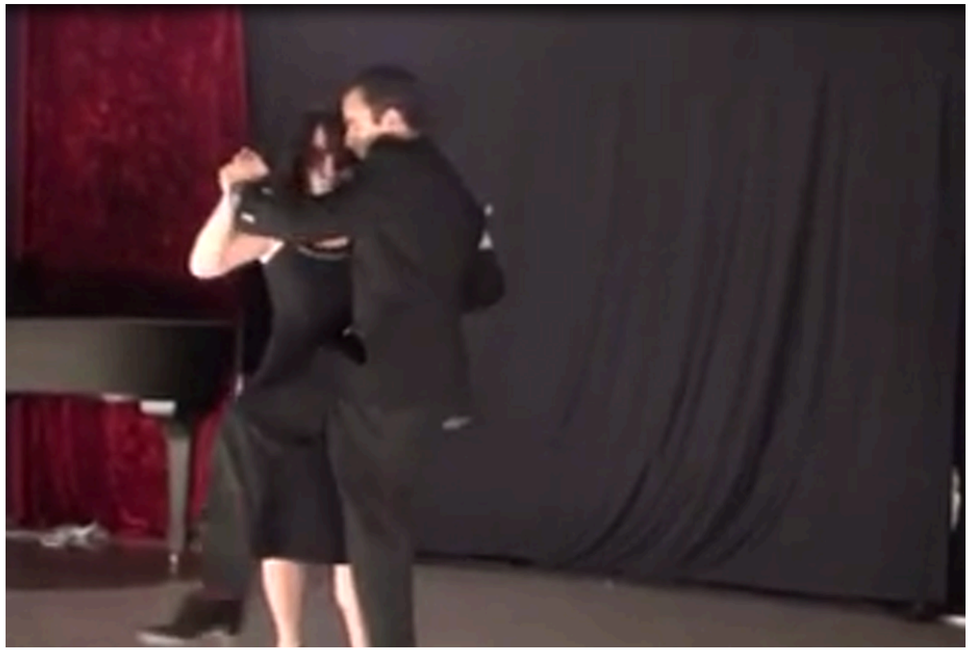
Kinetics dataset

The Kinetics dataset is dataset for human action recognition in videos.
The dataset consists of around 500,000 video clips covering 400/600 human action classes with at least 400/600 video clips for each action class.
Each video clip lasts around 10 seconds and is labeled with a single action class.

	Classes	Training videos	Validating videos
Kinetics-400	400	246,245	60 000
Kinetics-600	600	392,622	102 925
Kinetics-400 (5%)	400	9600	2400
Kinetics-600 (5%)	600	15680	3920

Kinetics dataset example

Brushing teeth



Tango dancing

VIVIT TRIALS

Kinetics-400: **top1:** 0.37 **top5:** 0.71 **avg_loss:** 3.65

Kinetics- 600: **top1:** 0.26 **top5:** 0.61 **avg_loss:** 4.18

SlowFast Training (ICCV 2019)

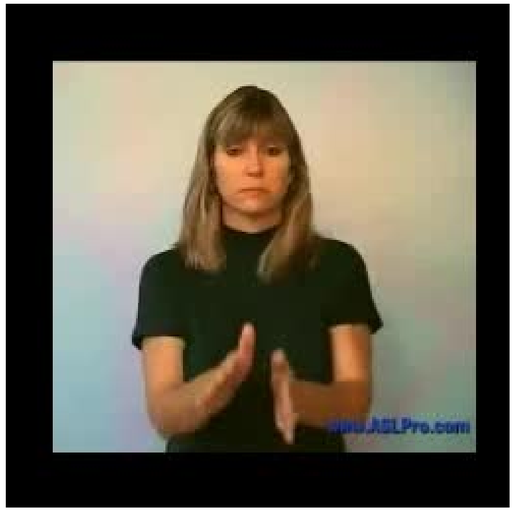
	Kinetics-400 (5%)	Kinetics-600 (5%)
Validation accuracy (Epoch 1)	42.3%	44.7%
Validation accuracy (Epoch 10)	67.4%	71.1%
Theoretical data (Top-1 accuracy)	**75.6%**	**78.8%**

WLASL-100

Classes: 100
Videos: 2038
Examples:

Apple

Book



RESULTS FOR WLASL-100

Models:	ViViT	SlowFast-R50	Pose-TGCN (Hands)	Pose-TGCN (hands, face)	Pose-TGCN (hands, face, body)
Accuracy:	62%	71%	50%	58%	53%

Khan, Mahwish Maqsood (2025)

WLASL-300

Classes: 300
Videos: 5118
Examples:

Animal

Bad

RESULTS FOR WLASL-300

Models:	ViViT	SlowFast-R50	Pose-TGCN (Hands)	Pose-TGCN (hands, face)	Pose-TGCN (hands, face, body)
Accuracy:	35%	43%	37%	39%	38%

Khan, Mahwish Maqsood (2025)

References

- Feichtenhofer, C., Fan, H., Malik, J., & He, K. (2019). SlowFast networks for video recognition. https://openaccess.thecvf.com/content_ICCV_2019/html/Feichtenhofer_SlowFast_Networks_for_Video_Recognition_ICCV_2019_paper.html
- A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lucic, and C. Schmid, “ViViT: A Video Vision Transformer,” Proc. IEEE/CVF International Conference on Computer Vision (ICCV), 2021.

THANK YOU