Zero-Shot Egocentric Video Action Recognition

Team Name- Random 1

Team Members- Abhilash Rajendra Sarwade, Shifali Agrahari, Gautam Singha

Introduction

- Egocentric video action recognition is important where understanding actions from a first-person perspective is crucial.
- Zero-shot learning (ZSL) offers a promising solution, but it faces limitations in handling complex, dynamic, and egocentric scenarios.
- Applications - Virtual reality and augmented reality, Security and surveillance, Sports analytics, Human-computer interaction and so on.

Related Works

Works related to Zero shot ego-centric action recognition

- Use of external knowledge
- Integrating human gaze into attention
- GCN for zero shot learning

Methodology

Dataset Preparation

Dataset

EGTEA GAZE +

- 28 hours (de-identified) of seven meal-preparation activities from 86 unique sessions performed by 32 subjects.

- Activities: Continental Breakfast, Pizza, Bacon and Eggs, Greek Salad, Pasta Salad, Turkey Sandwich and Cheese Burger.

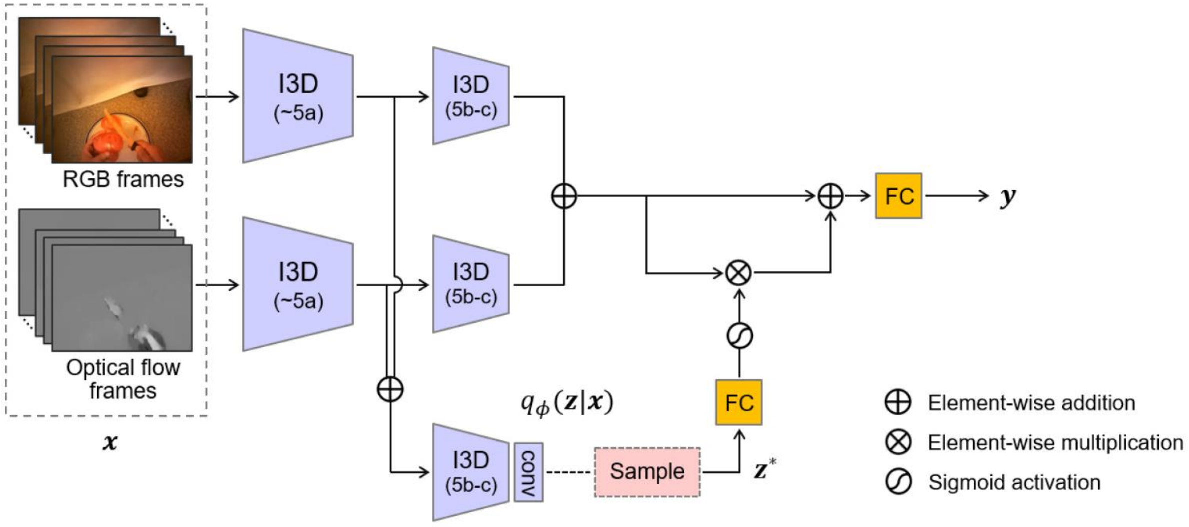
R-Split Dataset

- Splitting Dataset- R-split (Recipe Split). 6121 training video clips and 1464 test video clips.
- 65 seen (eg.- cut tomato, open cabinet) and 16 unseen (eg.- cut bell_pepper, put bread) classes.
- Data preparation: RGB frames and optical flow extracted from each video.

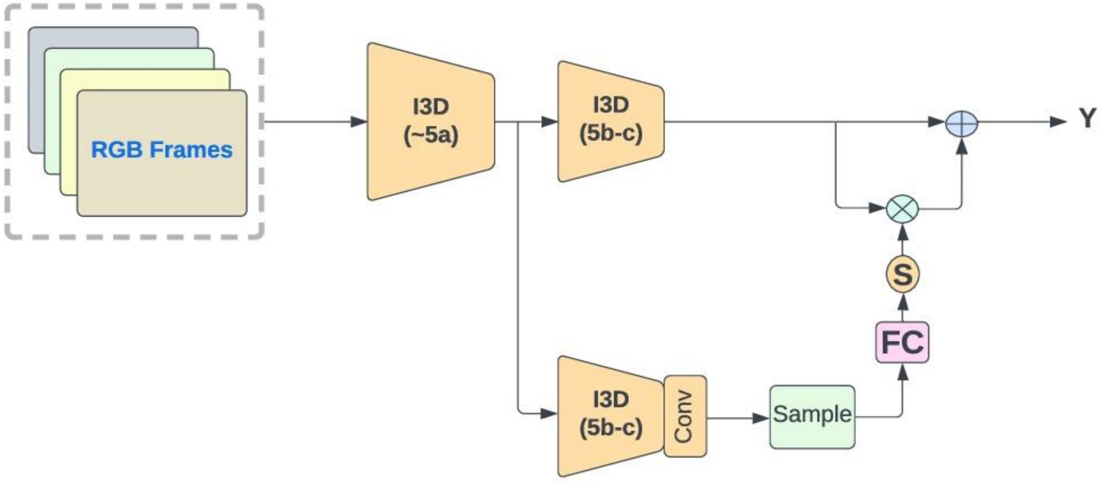
RGB Frames and Optical Flow

Feature Extraction

Min et al 2021



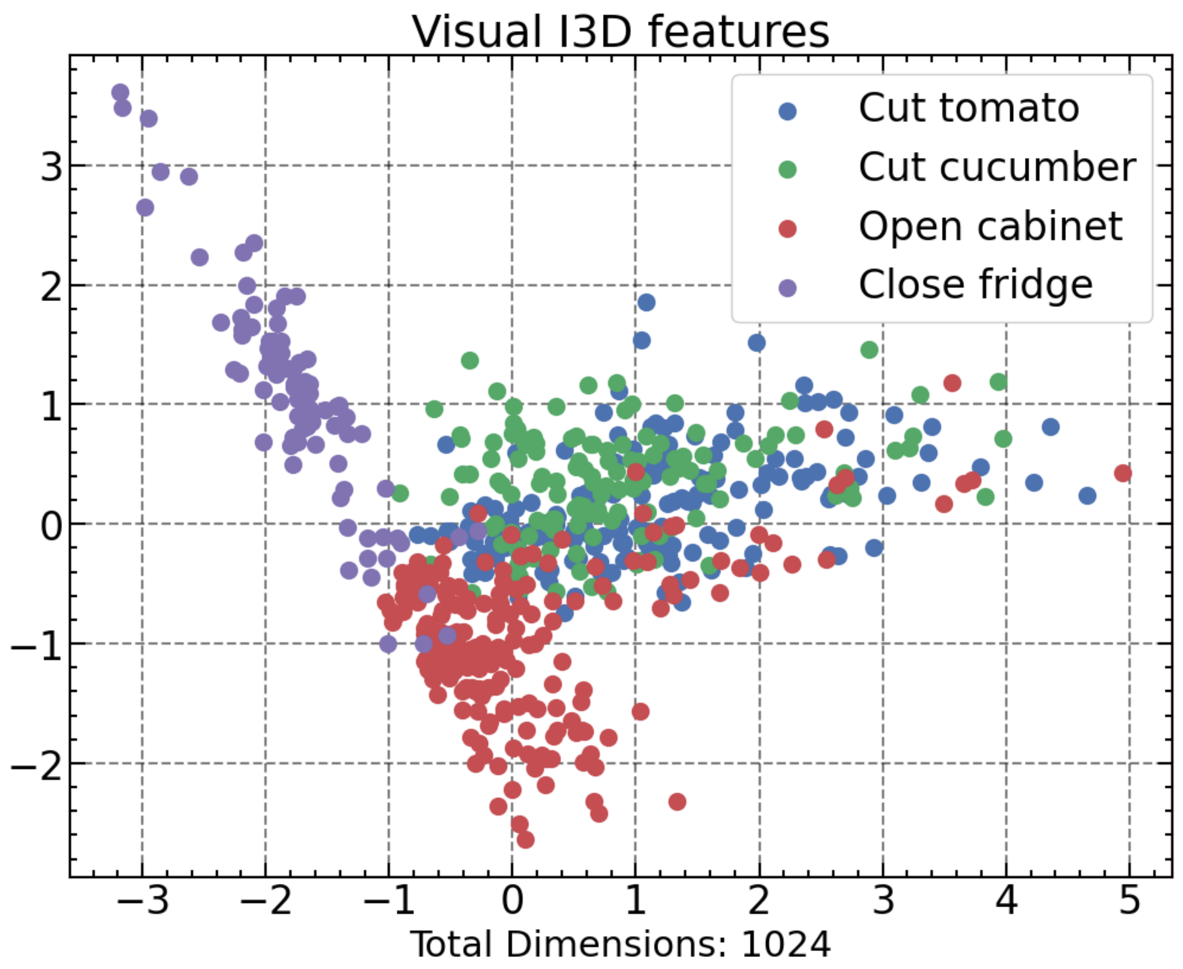
- Pre-trained convolutional model- I3D network having gaze attention for image feature extraction.
- Network structure modified to ignore optical flow due to high computational time.
- Last fully connected layer removed to get 1024 features.



Gaze Estimation

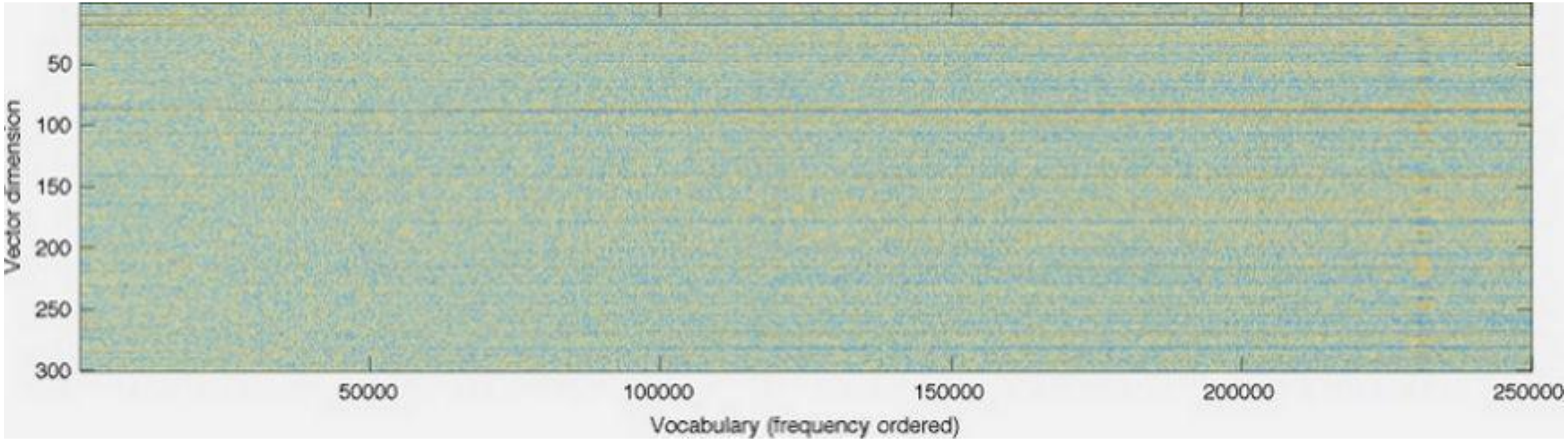
"crack eggs"

I3D Feature Visualization



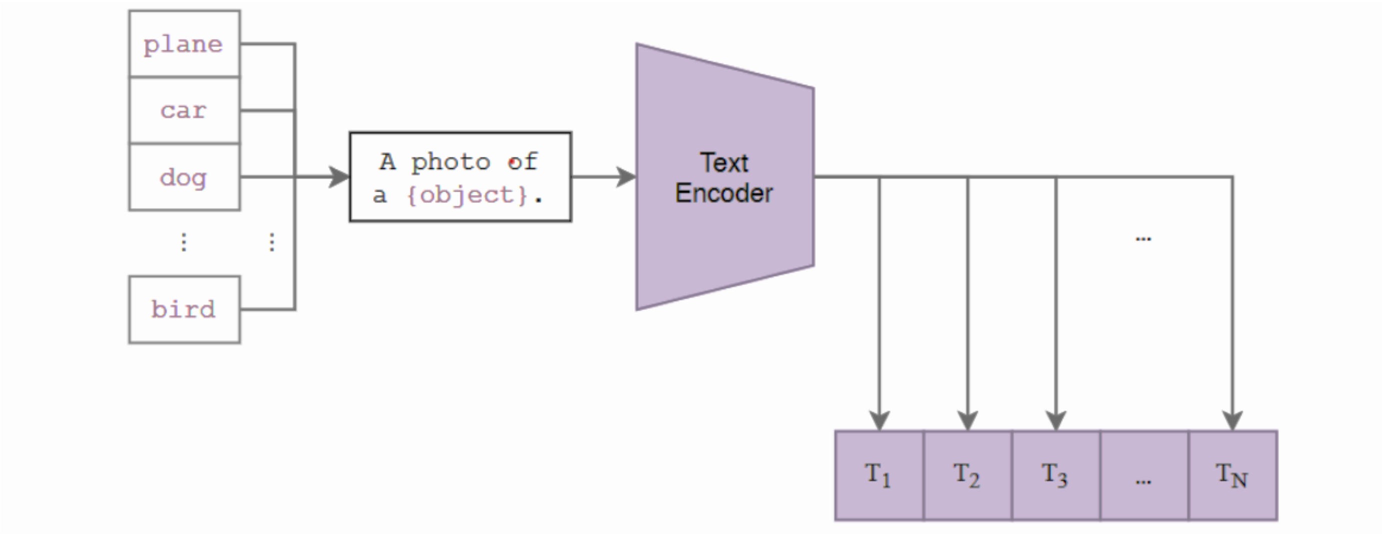
Semantic Embedding of Class Labels

- GloVe semantics



Semantic Embedding of Class Labels

- CLIP semantics
- Clip backbone model = ViT- B/32
- Prompt
- "a video of an object is {action label}"



Experiments

1st Model Classifier Neural Network

Set Up

- Training Data- 65 label
- Zeroshot Datase(test Data) - 16 label
- PyTorch – 2.1.0 Version
- Label Embedding

  Clip (512 size vector) Vit-B/32

  GloVe (300 size vector)
- Gaze Attention I3D feature extraction (1024 size vector)

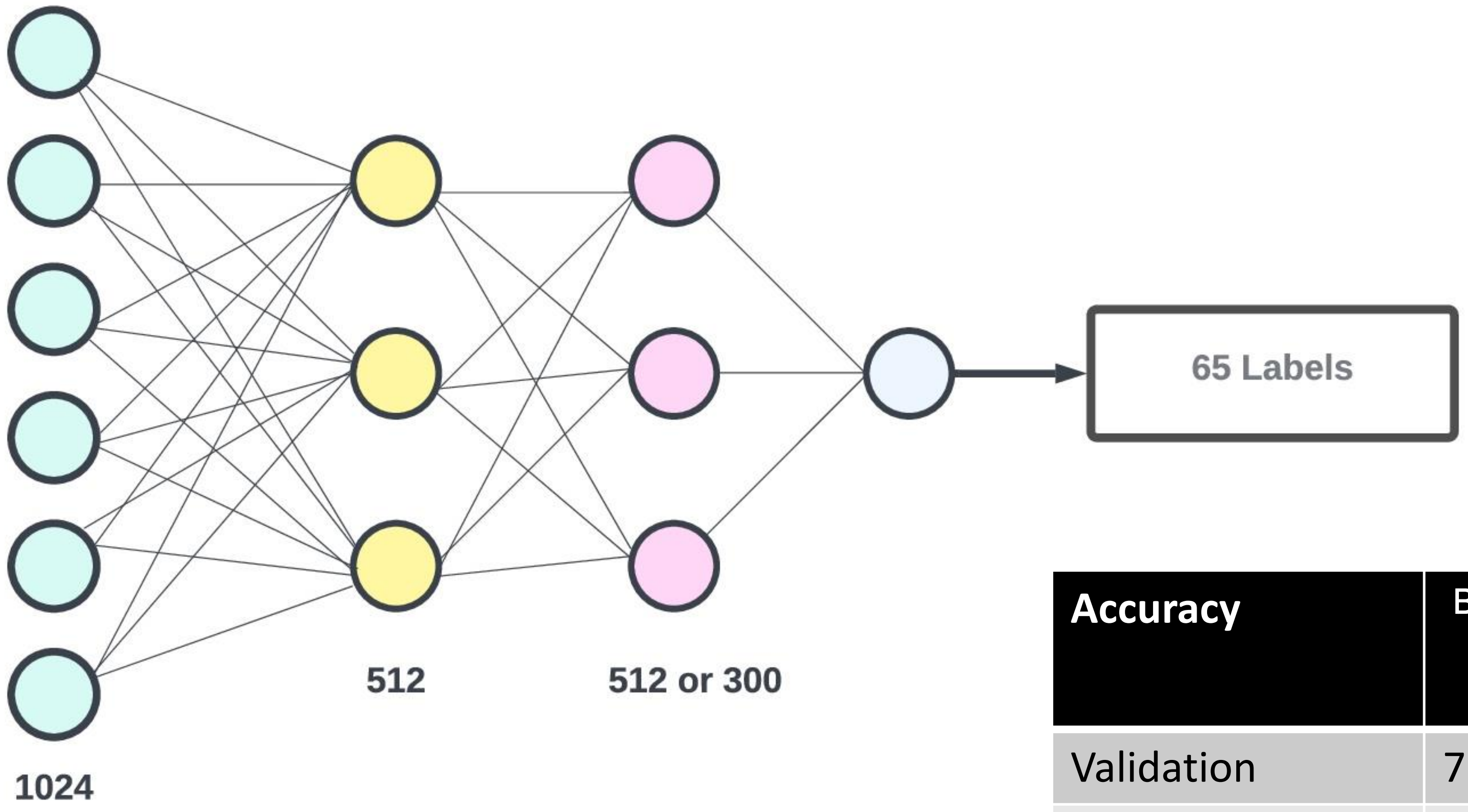
Classifier Based Method

- For training Dataset- 3416 clip
- For validation Dataset – 1230 clip
- Epoch - 125
- Batch size - 128
- Loss function
- BCEWithLogitsLoss, CrossEntropyLoss()
- Optimizer – Adam
- Learning rate 0.00005

Training based True Label

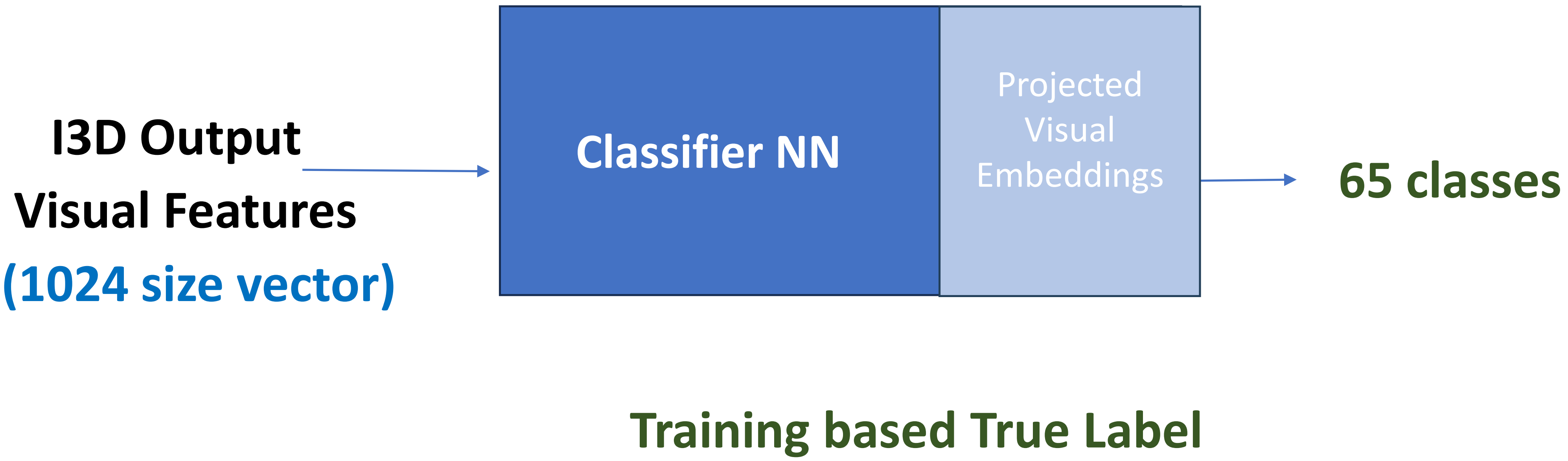
Classifier Architecture

Execution Time – 11 min

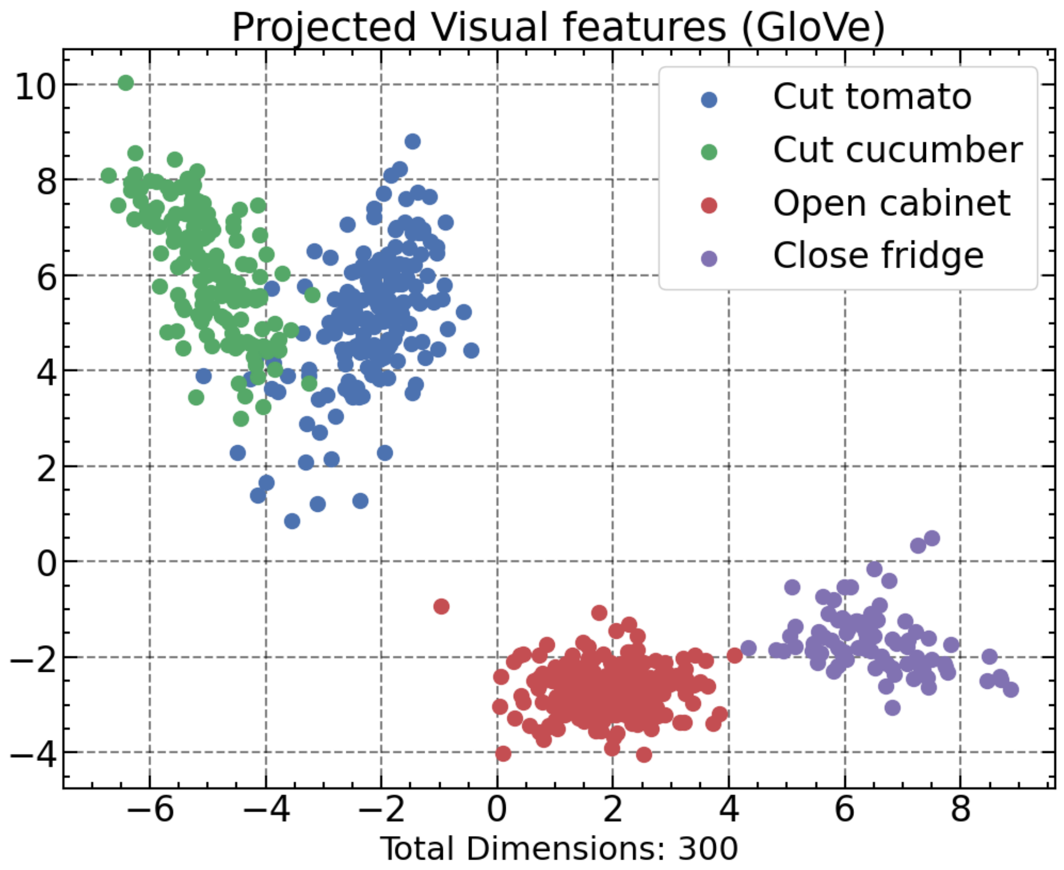


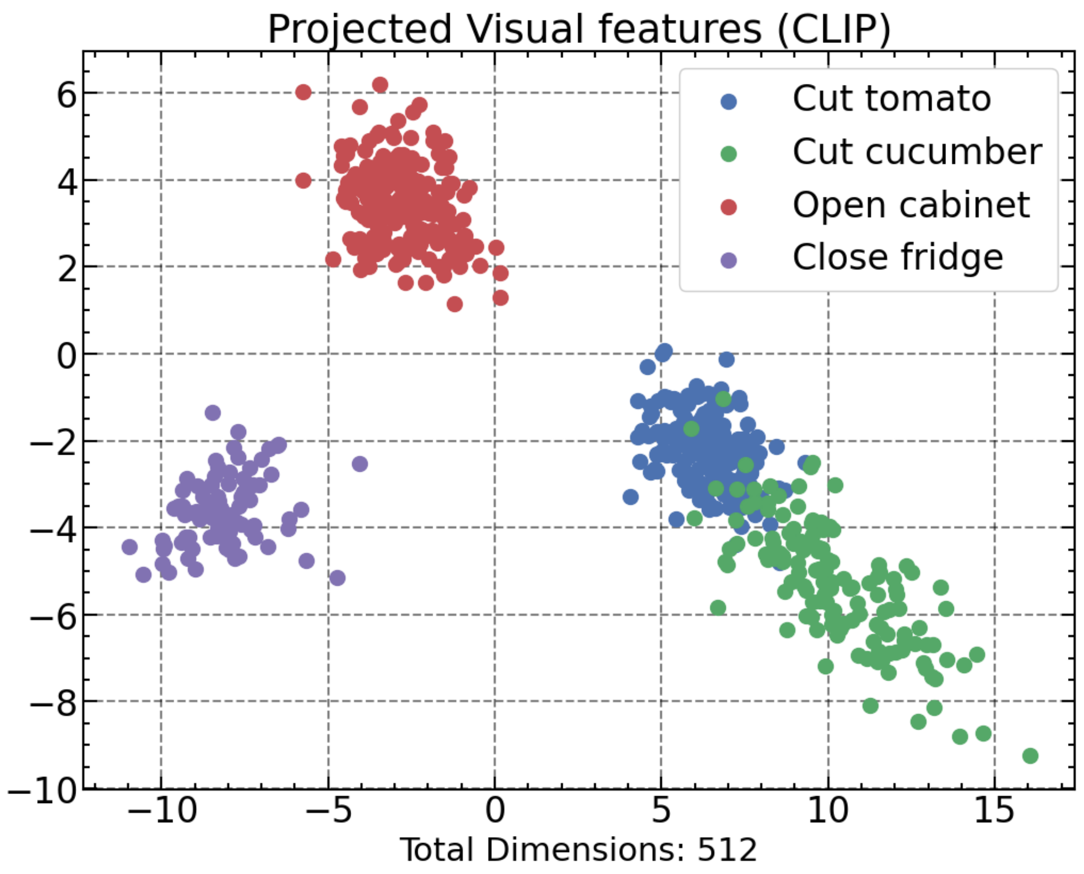
Accuracy	BCEWithLogitsLoss	CrossEntropyLoss()
Validation	75.67%	56.3%
Testing	49.8%	46.2%

Classifier Based Method



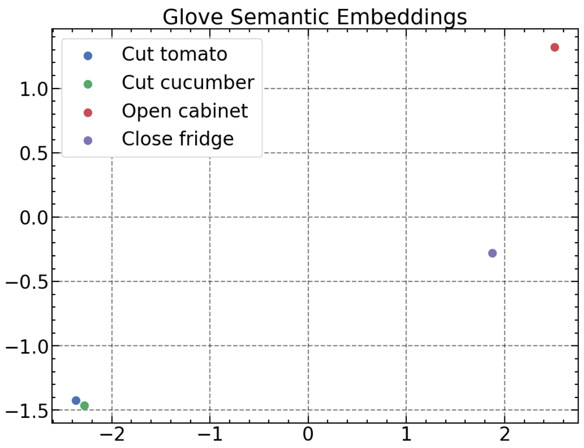
Projected Visual Features



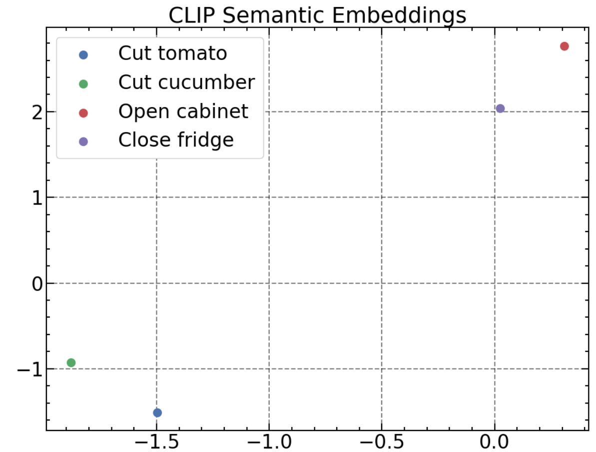


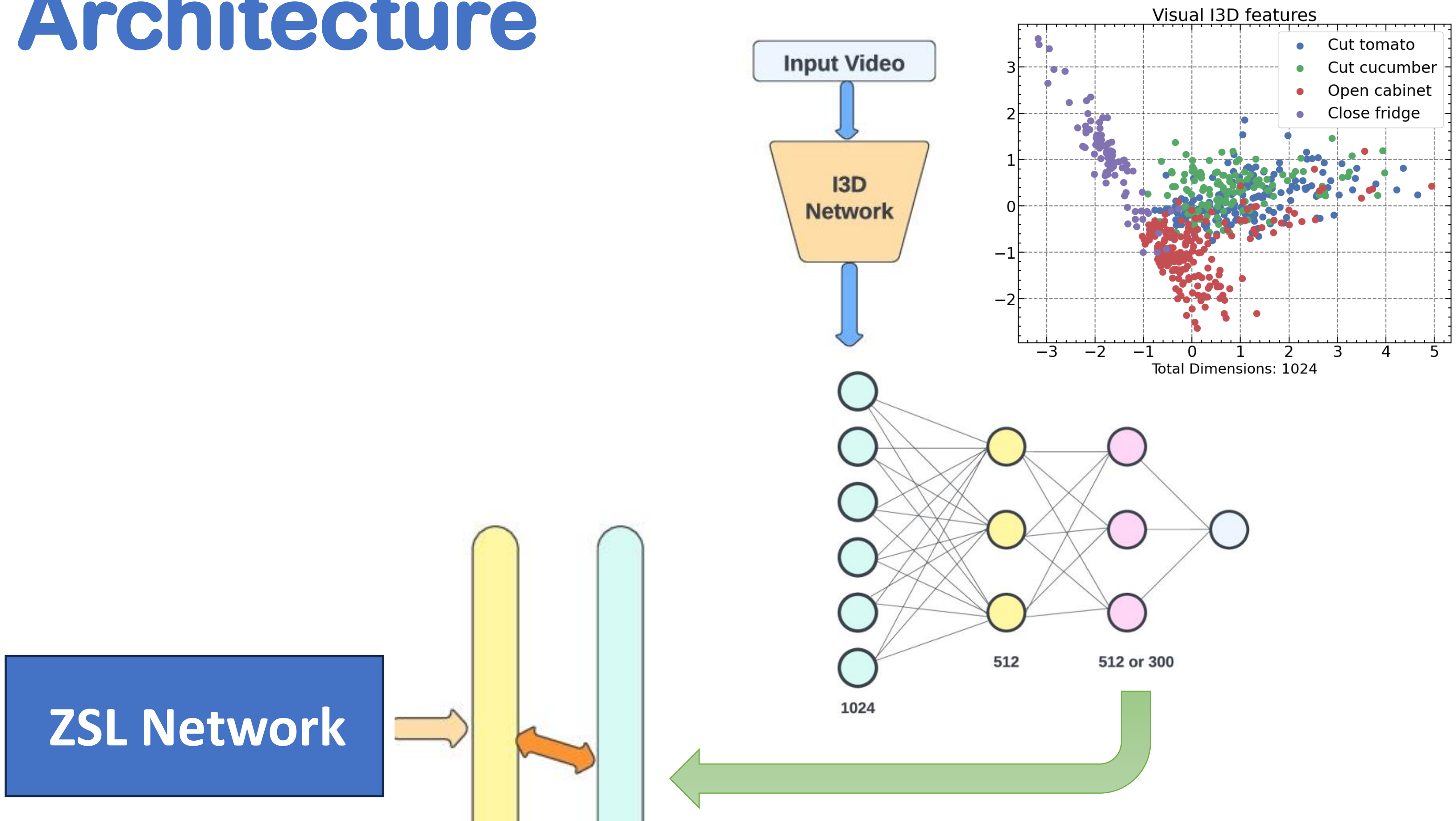
Zero-Shot Classification Network

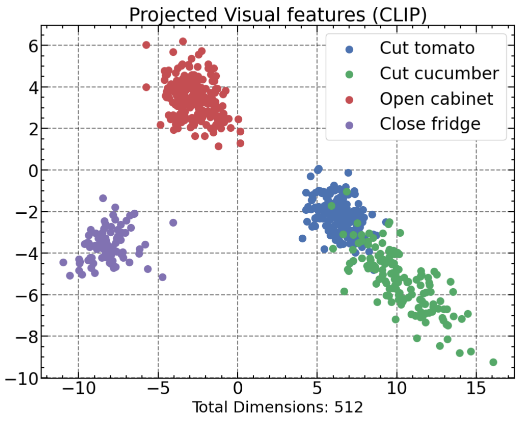
Zero Shot Architecture

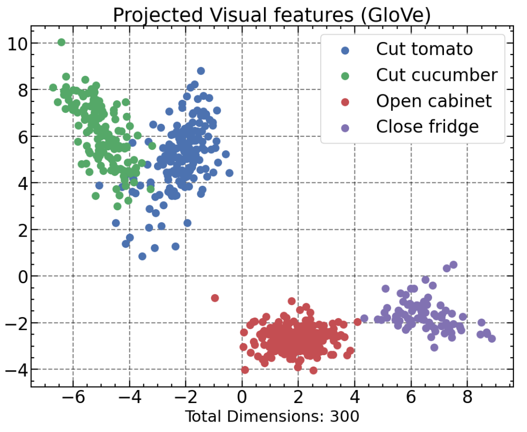


Semantic Embeddings



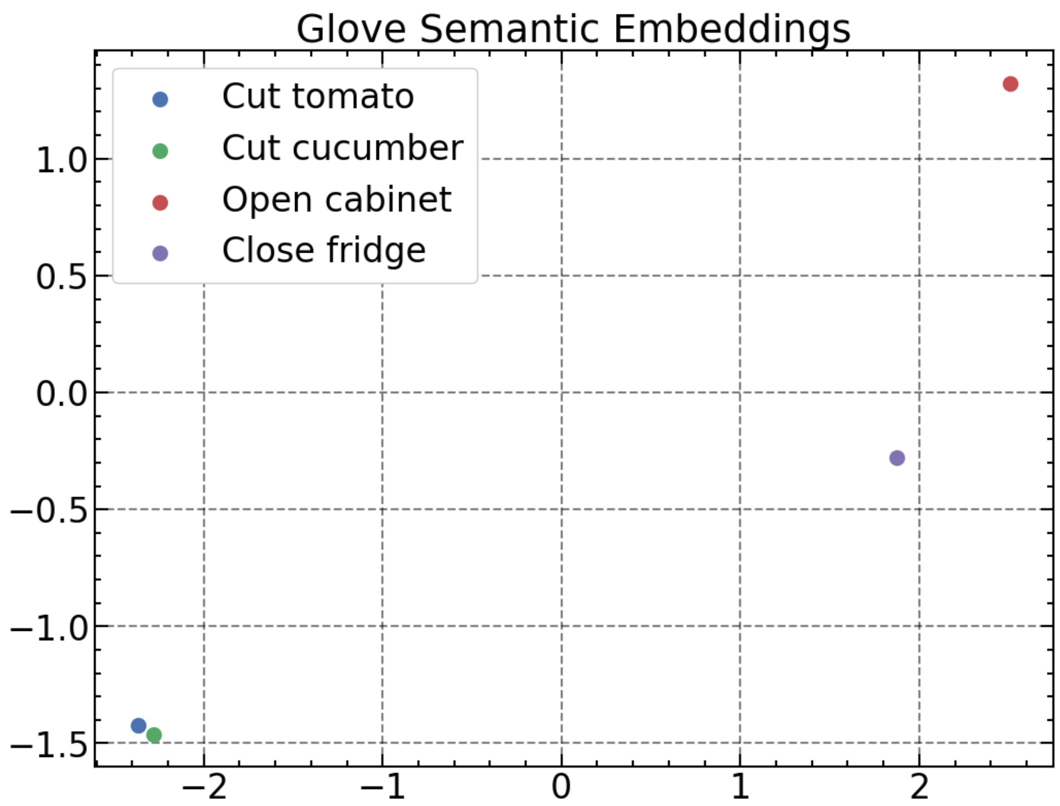




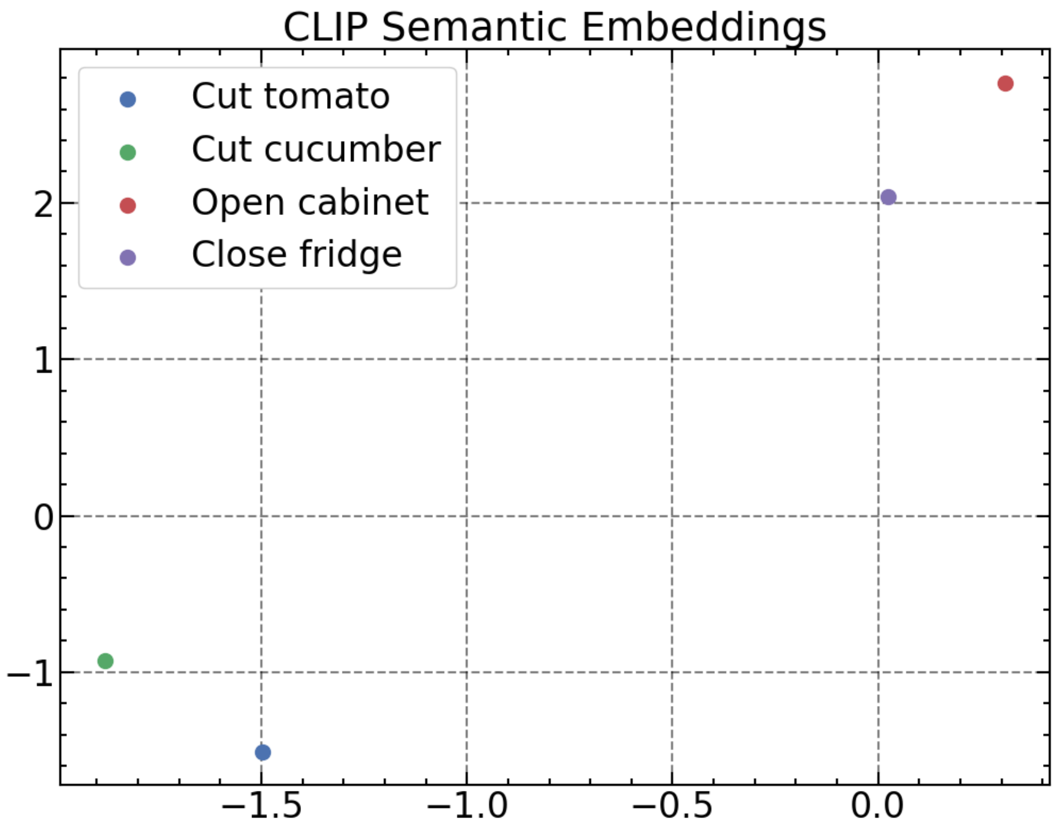


Semantic Embedding of Class Labels

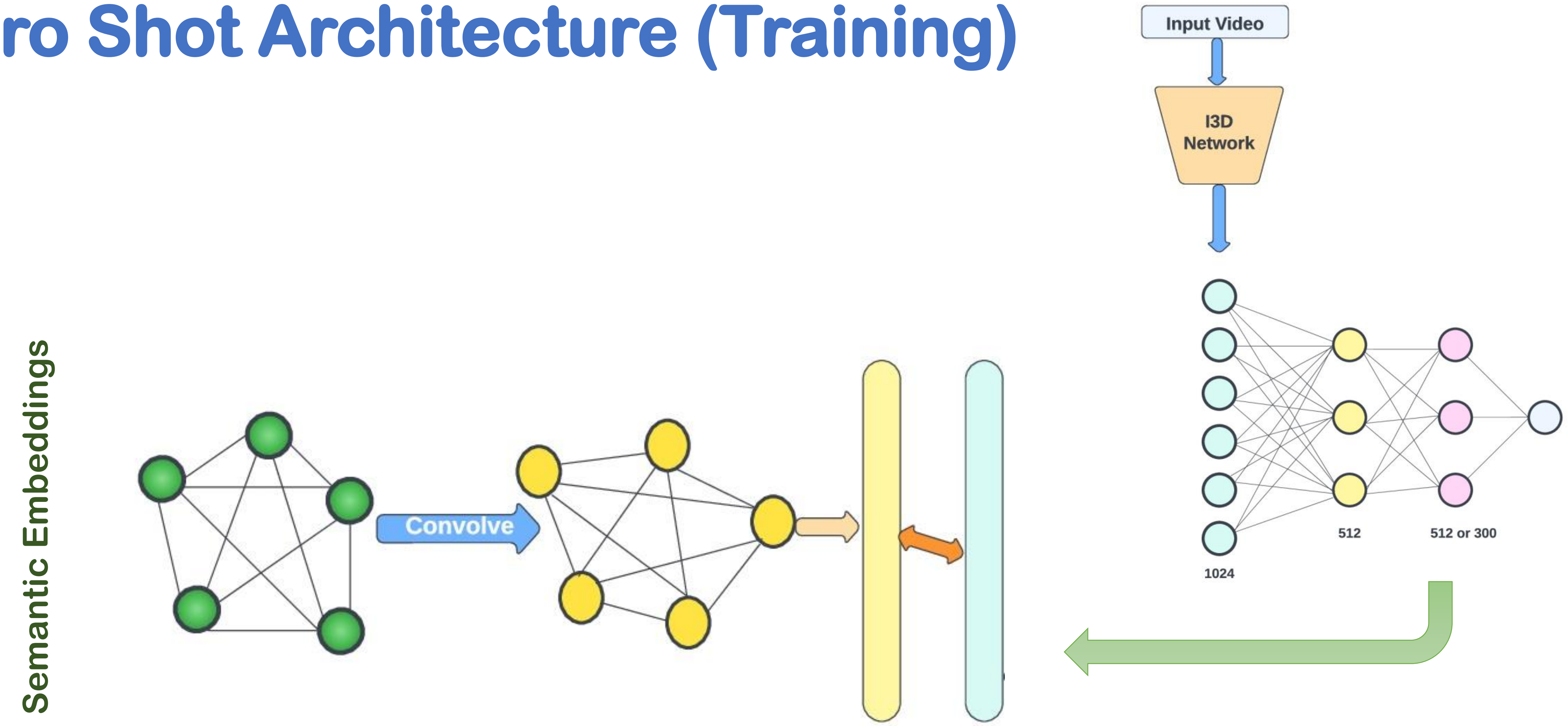
- GloVe semantics



- CLIP semantics

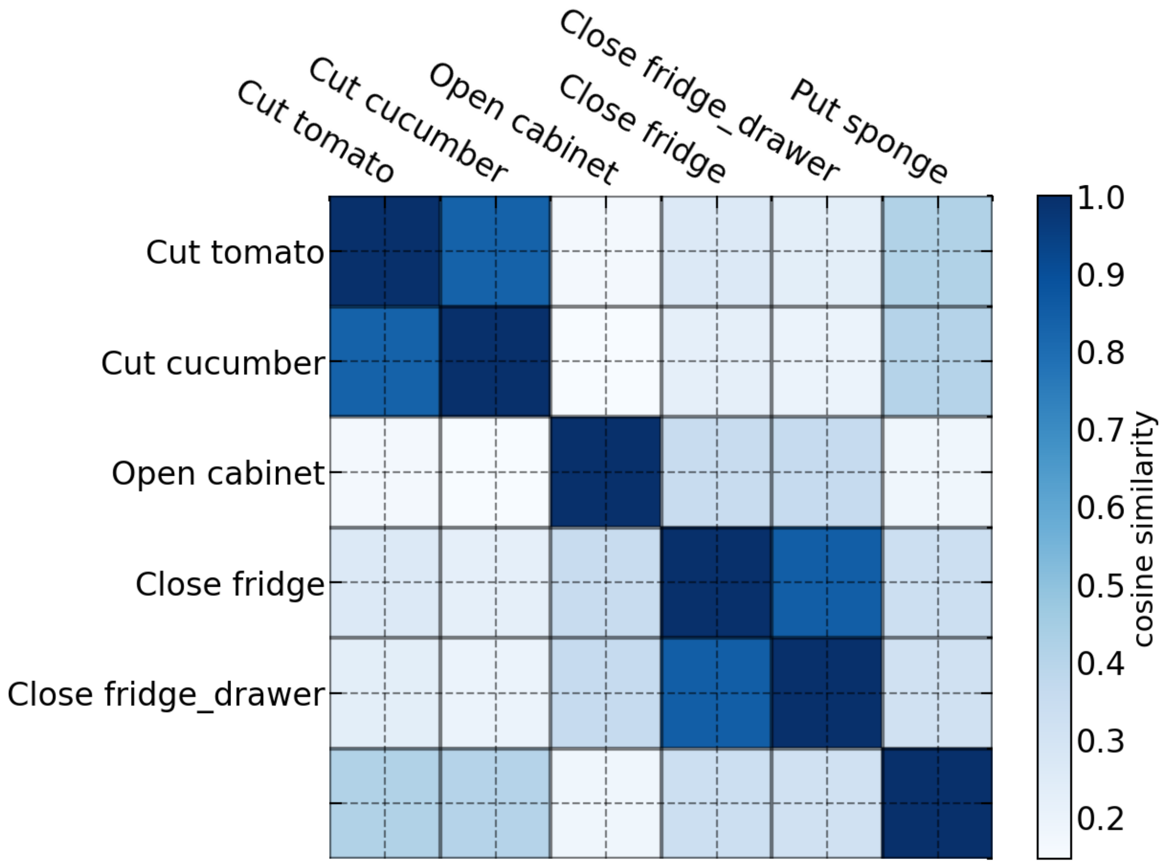


Zero Shot Architecture (Training)

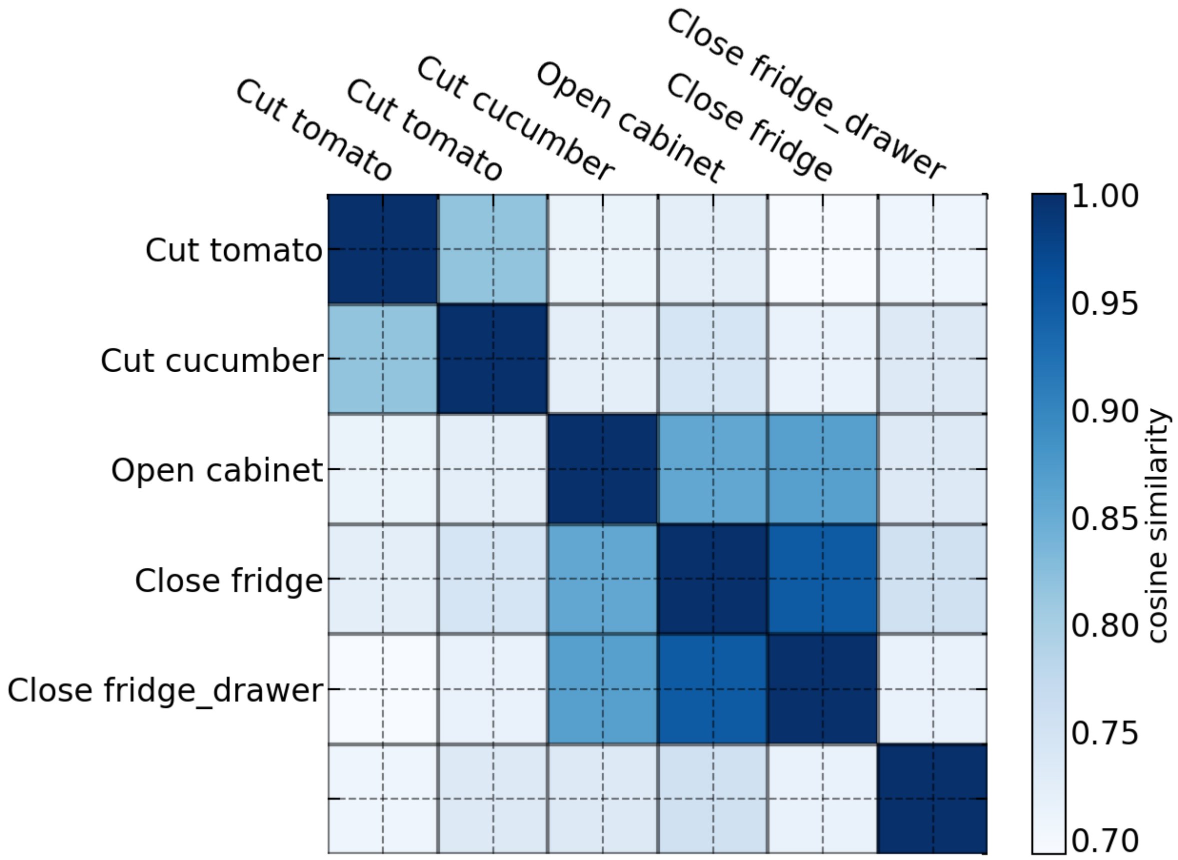


Semantic Adjacency Matrix

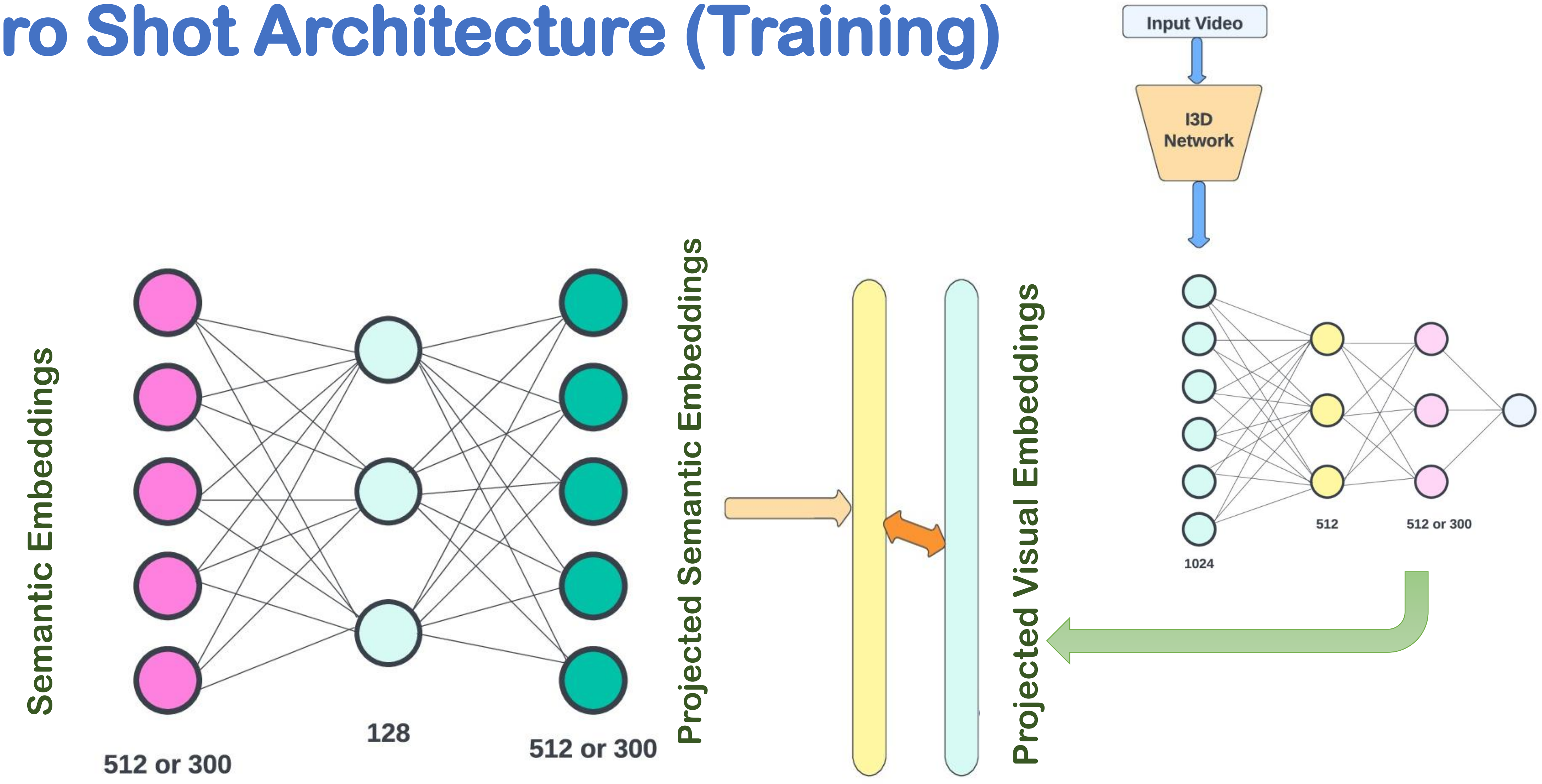
GloVe semantics



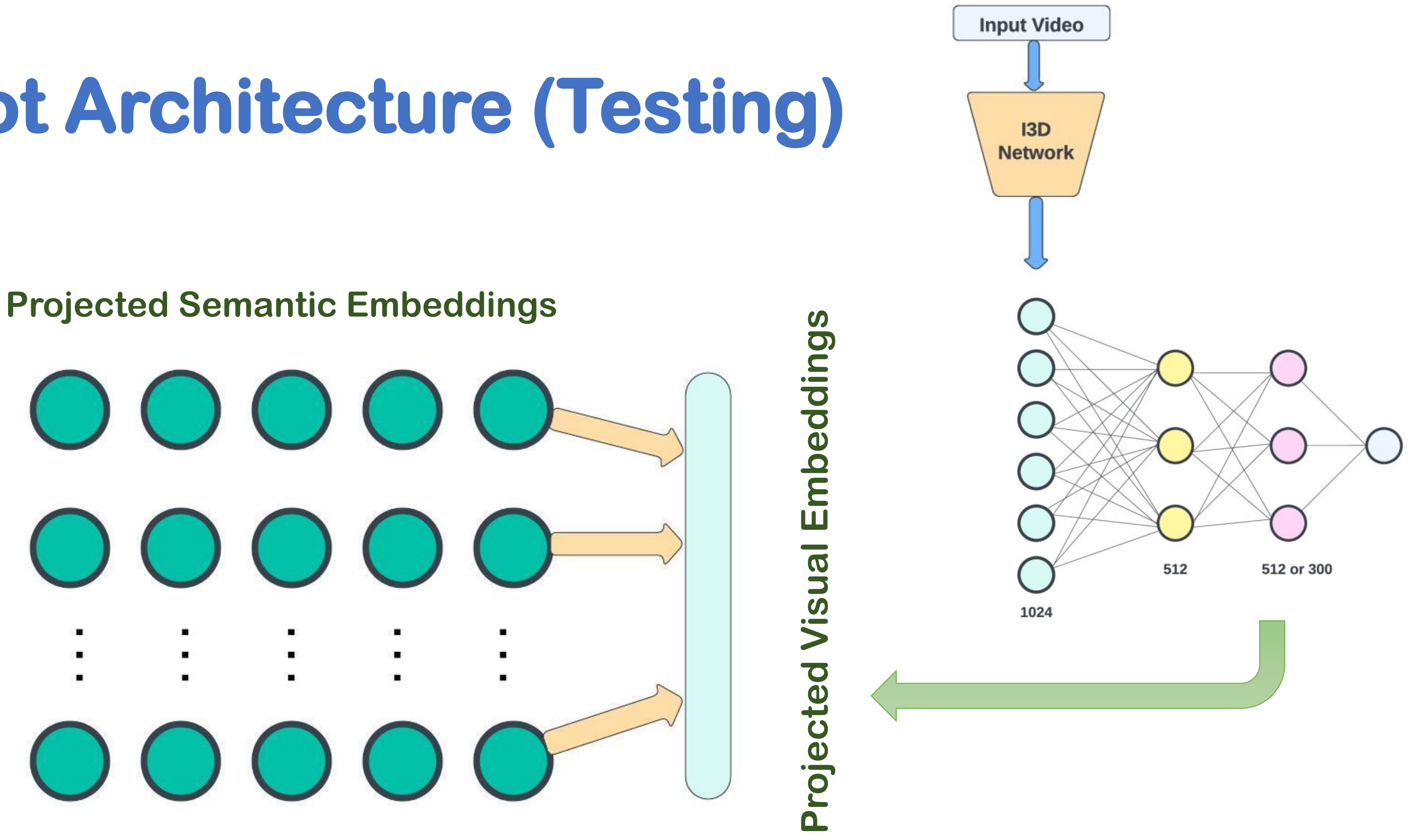
CLIP semantics



Zero Shot Architecture (Training)

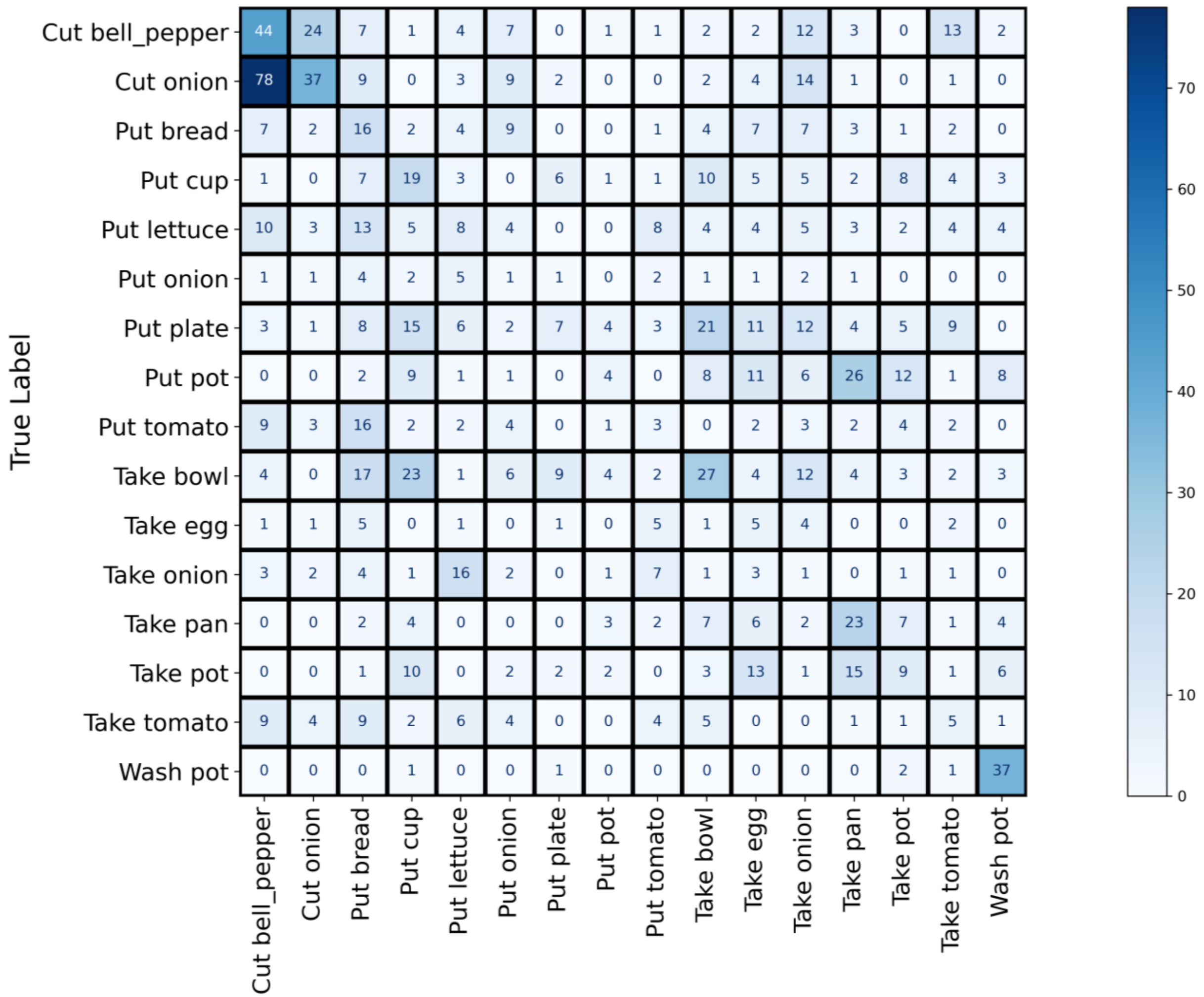


Zero Shot Architecture (Testing)

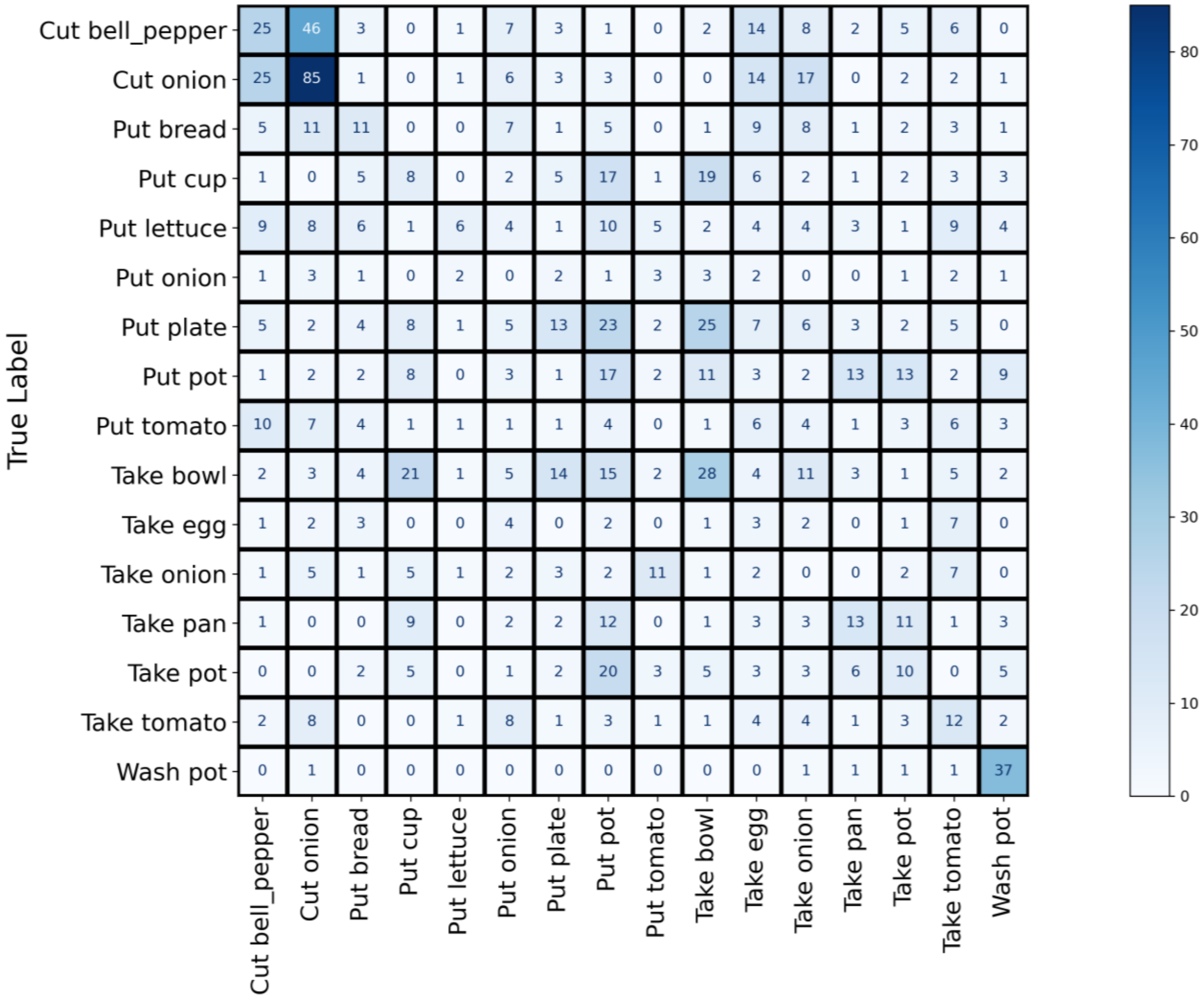


Results (Confusion Matrix)

- **GloVe semantics (20.87%)**



- **CLIP semantics (22.63%)**



Accuracy Comparison with Baseline paper for unseen data

Class (R split)	Cookbook prior	Ours (clip)	Ours(Glove)
cut bell pepper	23.28%	20.01%	36.32%
cut onion	8.05%	53.13%	23.05%
put bread	25.89%	17.43%	25.82%
put cup	14.17%	11.02%	25.01%
put lettuce	41.75%	7.87%	10.23%
put onion	10.26%	0.00%	4.51%
put plate	29.41%	12.34%	6.30%
put pot	28.05%	19.67%	4.53%
put tomato	3.17%	0.00%	5.73%
take bowl	18.00%	23.81	22.61%
take egg	0.98%	12.14%	19.05%
take onion	17.22%	0.00%	2.30%
take pan	17.98%	21.23%	33.04%
take pot	7.26%	15.09%	14.82%
take tomato	3.07%	24.29	9.08%
wash pot	13.95%	88.78%	88.23%

Results comparison with Baseline

Paper	Zero shot	Dataset	Accuracy
Using external knowledge	Yes	EGTEA	18.08%
Integrating Human Gaze into Attention	No	EGTEA	62.84%
GCN	Yes	Epic- Kitchens	
Classifier(Our)	No	EGTEA Gaze+	58.9%
MLP (with Glove)	Yes	EGTEA Gaze+	20.87%
MLP (with CLIP)	Yes	EGTEA Gaze+	22.63%

Conclusion/Limitation

- Multiple action label instances not considered (NR- Split)
- Pre-trained I3D model used (Gaze attention)
- Optical flows not used as an input to gaze attention I3D
- Could not employ GCN (Simple Neural Network used instead)
- Classifier network and zero-shot network not trained simultaneously.

Future Work

- Generation of Optical flows
- Employing GCN
- Simultaneous training of Classifier network and zero-shot network.

THANK YOU