

Beyond Image Classification: A Video Benchmark and Dual-Branch Hybrid Discrimination Framework for Compositional Zero-Shot Learning

Supplementary Material

This appendix is organized as follows:

- C-EgoExo dataset detailed data screening, labeling, splitting process in Sec. 1
- Analysis of datasets including UT-Zappos and CGQA in Sec. 2;
- Details for training and inference setting in Sec. 3;
- Evaluating cross different domain in Sec. 4;
- Potential societal impacts and limitations in Sec. 5.

1. The Compositional EgoExo-4D Benchmark

Ego-Exo4D[8] is a large-scale, multimodal, multiview video dataset created to facilitate research on skilled human activities from both egocentric (first-person) and exocentric (third-person) perspectives. Developed collaboratively by 12 research institutions over two years, the dataset encompasses 1,286 hours of video, featuring 740 participants performing diverse activities across 123 natural scenes in 13 cities worldwide.

We selected it as the foundation for our compositional zero-shot learning benchmark for the following reasons: 1) Diverse Activity Domains: The dataset spans eight domains, covering both physical tasks (e.g., soccer, dance, bouldering) and procedural tasks (e.g., cooking, bike repair). 2) Expertise Spectrum: Activities range from novice to professional levels, enabling fine-grained analysis of skill and expertise. 3) Rich Annotations: It includes detailed annotations such as 3D body and hand poses, key steps, procedural dependencies, proficiency ratings, and video-language pairs. 4) Unique Linguistic Resources: The dataset provides first-person narrations, third-person action descriptions, and expert commentary focusing on execution quality, offering valuable linguistic diversity for compositional tasks.

We first analyzed the characteristics of the EgoExo-4D dataset. As illustrated in Fig. A1, we observed that for compositional zero-shot learning (CZSL) tasks, the ego perspective (i.e., the data collector’s first-person view of the action) often fails to adequately represent the task. This limitation arises because actions involve an agent affecting an object, and the agent’s perspective may not capture the object’s changes comprehensively or objectively. Instead, the agent might focus on specific local details or other objects influencing their actions, introducing ambiguity.

Meanwhile, although each action is recorded from four exo perspectives, these views are frequently occluded, limiting their ability to describe the action in sufficient detail. Consequently, for each action annotation in EgoExo-4D, we

Domain	#Num.	#Ego	#Exo	#Objects	#Verbs	#Actions
Cooking	160270	69074	91196	614	357	6033
Health	45704	19381	26323	214	211	1849
Bike Repair	33596	16448	17148	254	205	1922
Music	6358	1771	4587	102	93	322
Basketball	50009	14275	35734	81	84	288
Rock Climbing	34486	13751	20735	108	132	476
Soccer	36189	16780	19409	78	91	267
Dance	22983	314	22669	170	165	935

Table A1. Dataset distribution.

#Original annotation	#Mistake	#Correction
deops the nozzle	deop	drop
plavces a container	plavce	place
cpovers a container	cpover	cover
adjusst the wheel	adjusst	adjust
passese the covid test manual	passese	pass
bouces the ball	bouce	bounce
stes the cucumber	ste	set

Table A2. Part of data set Egoexo-4D is incorrectly labeled.

selected video clips from only a subset of ego samples and the exo sample with the highest observation quality.

This selection process resulted in a dataset with a distribution distinct from the original EgoExo-4D dataset. The final C-EgoExo dataset comprises 387,000 video clips. To highlight the differences between C-EgoExo and the original dataset, we present the distribution of selected samples across different real-world domains, as shown in Tab. A1.

In Tab. A1, Ego represents the number of samples captured from the first-person (egocentric) perspective, while Exo denotes those observed from third-person (exocentric) perspectives. Objects, Verbs, and Actions indicate the number of distinct objects, verbs, and action types, respectively.

The EgoExo-4D dataset does not provide separate annotations for verbs and objects. To address this, we employed Stanza[1] and SpaCy[10] to perform independent annotation tasks. We retained only the samples with consistent results between the two tools, resulting in 353,490 samples. The remaining 33,510 samples were manually annotated.

Additionally, the original annotations in EgoExo-4D contained numerous spelling errors and inconsistencies. To correct these, we utilized LLaMA and ChatGPT as aids for error correction. These tools helped refine 101 object labels and 46 verb labels. Furthermore, 249 text annotations were manually revised. Examples of the corrected errors are shown in Tab. A2.

After determining the total number of samples, we pro-

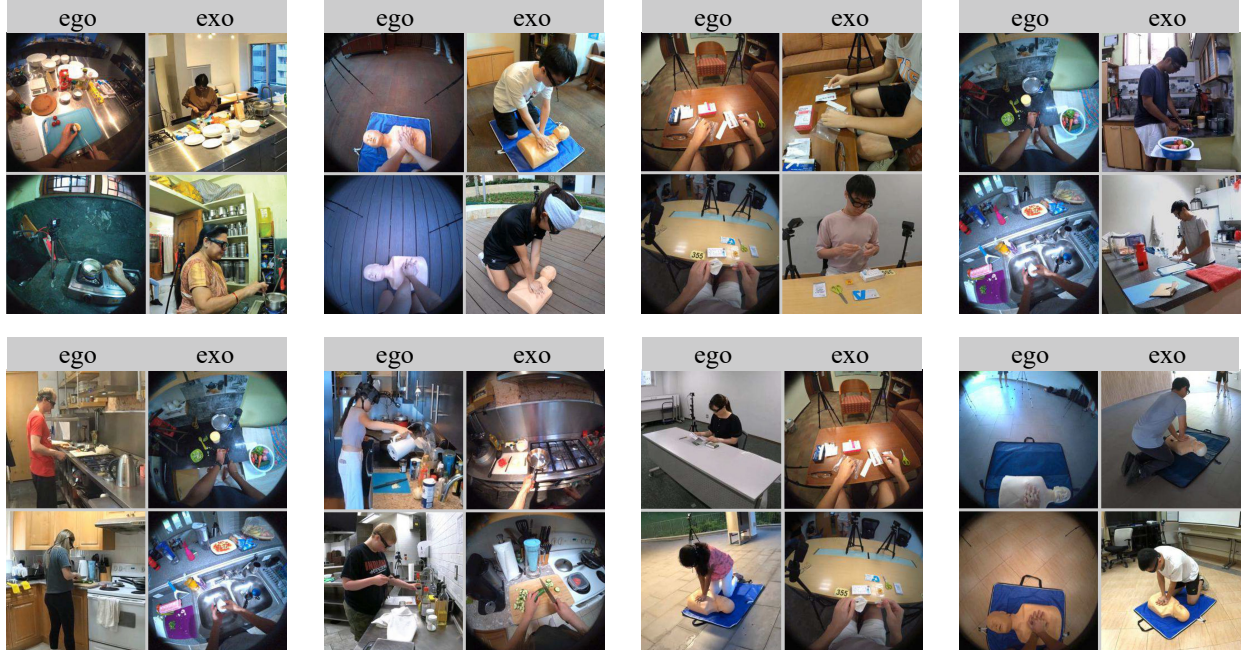


Figure A1. C-Egoexo dataset data example.

	#Objects	#Verbs	#Verb-Object Actions	#Sample
Train	1042	575	6709	186970
Val	591	303	1401 Seen + 1450 Unseen	95584
Test	641	331	1905 Seen + 1934 Unseen	104446
All	1042	575	10112	387000

Table A3. Dataset statistics of C-EgoExo.

ceeded to split the dataset. Our initial goal was to create a training set of approximately 190,000 samples. First, for each object and verb primitive, we randomly selected one action containing it and included a corresponding sample in the training set, resulting in 1,617 samples. Next, we continued sampling from all possible combinations until we reached 6,516 unique compositions in the training set. We then extracted 185,891 samples from the selected categories, adding any remaining categories with only a single sample to the training set. This yielded a final training set of 186,970 samples. The remaining data contained 3,384 unseen categories. These were split into a validation set and a test set in a 3:4 ratio. Additionally, the remaining *seen* category samples were randomly distributed in approximately the same proportion. The final dataset split is summarized in Tab. A3.

2. Other Datasets

The UT-Zappos dataset¹ is provided by Yu *et al.* [38], with permission granted for non-commercial research purposes.

¹<https://vision.cs.utexas.edu/projects/finegrained/utzap50k/>

Hyperparameter	UT-Zappos	CGQA	C-EgoExo
Learning rate	2.5×10^{-4}	10^{-4}	1.25×10^{-4}
Batch size	64	64	256
Number of epoch	50	50	50
Warm up (linear)	3	5	3
Feature dimension	1024	1024	1024
Attention head	16	16	16
Dropout	0.5	0.5	0.5
\alpha	1	1	1
\beta	1	1	1
\sigma	1	1	1
\delta	0.5	0.5	0.5

Table A4. Hyperparameter for different datasets.

The CGQA[24] dataset is an extension of the Stanford GQA dataset², and both are licensed for non-commercial research applications. The data distribution is shown in Fig. A2.

3. Hyperparameters

As shown in Tab. A4, we summarize the key hyperparameters of the DHD framework. The parameters α , β , σ , and δ are used to balance different components of the loss function (Eq. (A1)). Ablation studies on these parameters can be found in Sec. 5.

$$\mathcal{L} = (\mathcal{L}_{con} + \alpha\mathcal{L}_{ocd}) + \beta(\mathcal{L}_{obj} + \sigma\mathcal{L}_{ccd}) + \delta\mathcal{L}_{ort}, \quad (A1)$$

²<https://cs.stanford.edu/people/dorarad/gqa/index.html>

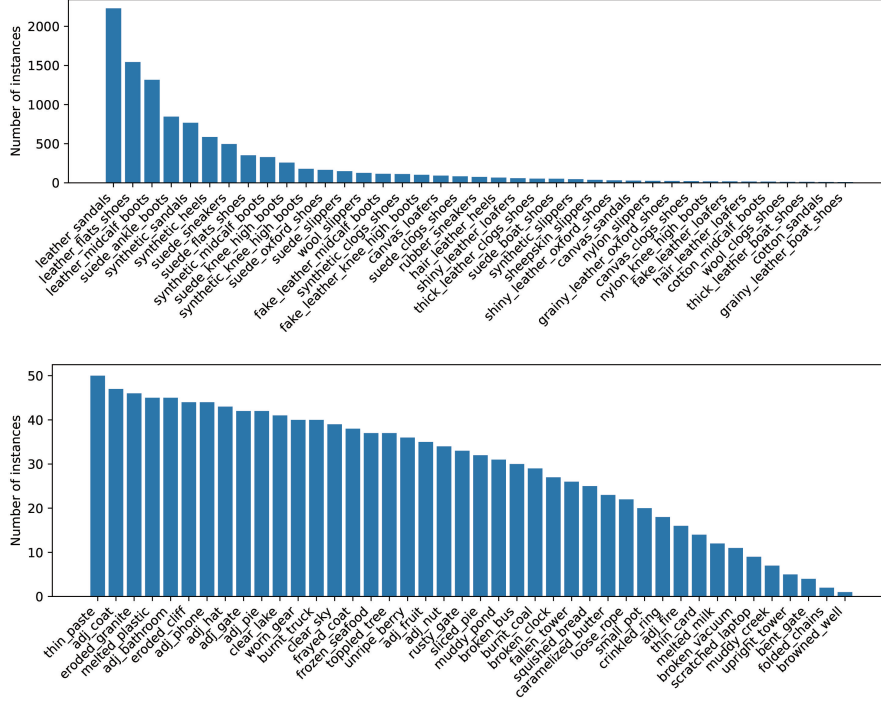


Figure A2. Dataset distribution in the UT-Zappos and C-GQA.

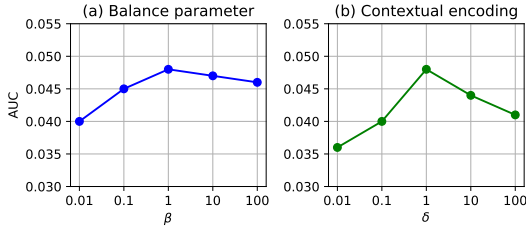


Figure A3. Ablation study on the parameter in loss function.

The hyperparameters α and σ control the balance between the losses generated during the independent decoding of the first step and the conditional decoding of the second step in each branch. Since both modules utilize cross-entropy loss with similar magnitudes, we set these values to 1, which provides a general-purpose configuration. For tasks that require stronger contextual dependence, these values can be increased to better suit the specific task requirements.

β balances the decoding weights between the two branches. Based on our hypothesis, the two branches should have equal importance in the absence of prior information, so we set $\beta = 1$. For ablation experiments on β , please refer to Sec. 5 and Fig. A3.

Finally, δ adjusts the Copula-based orthogonal decoding loss between the two branches. We recommend setting

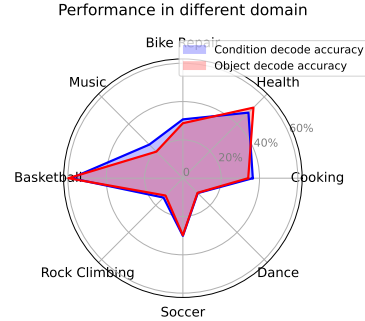


Figure A4. Accuracy performance on different domain.

$\delta = \frac{1}{1+\beta}$, as this ensures a balance with \mathcal{L}_{con} and \mathcal{L}_{obj} , which mainly regulate the first-step decoding process. If β is large, an equally large δ would overly emphasize the first-step decoding results, potentially compromising the accuracy of the second-step decoding.

4. Domain Performance in C-EgoExo

We also evaluated the performance of our method across different domains, as shown in Fig. A4. The results indicate that the model still struggles to balance performance in cross-domain experiments. This challenge is largely at-

tributable to the characteristics of annotations in different domains. For instance, in domains like *music* and *dance*, actions are often fine-grained and composite, such as simultaneously waving arms while lifting a leg or swaying. These complex actions are difficult to encapsulate with a single label. In contrast, domains with coarser-grained annotations, such as *basketball*, feature broader actions like "kicking the ball," which are relatively easier for the model to recognize.

5. Potential Societal Impacts and Limitations

Although our work focuses on classification, it may carry unintended social implications. This approach relies on pre-trained text and image encoders, which, if trained on biased or flawed datasets, could result in misunderstandings, biases, or inaccuracies. Additionally, the dual-branch hybrid decoding design introduces increased computational complexity, which may impact its real-time applicability.