Viewpoint Rosetta Stone: Unlocking Unpaired Ego-Exo Videos for View-invariant Representation Learning

Supplementary Material

6. ViewpointRosetta Details

For the contrastive losses $\mathcal{L}_{\text{v-v}}$, $\mathcal{L}_{\text{v-t}}$, and $\mathcal{L}_{\text{v-s}}$ described in Section 3.4, we present the formulation for only one direction of the contrastive loss—specifically, using x_1 as the anchor for the positive pair (x_1, x_2). However, in our experiments, we compute the contrastive loss in both directions (using x_1 as anchor and x_2 as anchor) and use their average as the final loss.

The video encoder in Figure 2 is pretrained using video-text contrastive learning [49] on video-narration pairs from Ego-Exo4D [17]. This training ensures that the ego and exo video features processed by our Rosetta Stone Translator (RST) effectively capture meaningful information about the input videos, facilitating accurate and context-aware translation between perspectives.

7. Downstream Task Details

For cross-view action recognition, the dataset includes a total of 188 action classes. To provide an overview of the distribution of action labels, we present a word cloud visualization in Figure 8, where larger words indicate higher frequencies of their occurrence in the labels.

For cross-view skill assessment, we showcase a few examples of the ”making omelet” task in Figure 9. Common instances of poor execution typically involve incorrect hand-holding gestures or improper procedural order.

8. Implementation Details

Training details For the video-text dual encoder, the model is trained using the Adam optimizer with hyperparameters $(\beta_1, \beta_2) = (0.9, 0.999)$ and a weight decay of 0.01, over 5 epochs. A fixed learning rate of 3×10^{-5} is employed.

For the two downstream tasks—cross-view action recognition and cross-view skill assessment, which involve fine-tuning the pre-trained video encoder on task-specific data—the model is trained for 200 epochs. Training is conducted using the SGD optimizer with a weight decay of 4×10^{-5} and a Cosine Annealing learning rate schedule. An initial learning rate of 3×10^{-3} is used to optimize performance. We leverage PyTorch's native FP16 mixed precision training and gradient checkpointing to enable efficient use of computational resources. Training is conducted with a per-GPU batch size of 32 across 32 GPUs, resulting in a total batch size of 1,024.

The temperature τ for contrastive learning is set to 0.07, while λ_1 and λ_2 are set to 0.5 and 0.2, respectively. The projection head following the dual encoders consists of a linear layer with an output dimension of 256.

Dataset Details Following [49], for Ego4D's input pre-processing, each video is divided into 5-minute segments, and the shorter side is scaled to 288 pixels. For the Ego-Exo4D dataset, we use the downscaled version with a resolution of 796 × 448 to improve data loading efficiency. Given the large dataset size, for Ego4D, we utilize only the videos from cooking scenarios for pretraining, while for HowTo100M, we select videos from the "Food and Entertaining" category.

9. Additional Results of Cross-view Retrieval

We present full quantitative results for both exo-to-ego and ego-to-exo cross-view retrieval tasks in Table 2, along with additional qualitative examples illustrated in Figure 10. The conclusion drawn from the quantitative is consistent with that in Section 4, further validating the effectiveness of our approach. Looking at Figure 10, Compared to the VI Encoder [17], our ViewpointRosetta leverages unpaired data to capture richer semantic information, enabling the retrieval of samples that are not only visually aligned with the input view but also semantically connected.

Category	Method	Cross-view Retrieval (ego → exo)				Cross-view Retrieval (exo → ego)			
		HR@1	HR@5	HR@10	HR@20	HR@1	HR@5	HR@10	HR@20
Egocentric Video Representation Learning	TimeSformer [6]	1.33	6.68	12.87	22.07	1.46	6.95	12.87	22.82
	EgoVLP [24]	13.28	29.33	40.40	52.83	4.65	13.47	19.74	29.21
	LaViLa [49]	17.71	34.91	47.10	58.48	4.46	12.02	17.78	26.99
	LaViLa* [49]	19.18	43.63	56.91	70.10	7.87	21.93	32.31	44.81
View-invariant Representation Learning	Random Align *	6.51	23.33	35.66	51.26	5.94	20.08	31.39	46.87
	ActorObserverNet † [36]	9.64	29.50	43.05	59.74	7.63	24.85	37.70	54.80
	VI Encoder † [17]	8.65	29.53	44.31	60.63	7.05	24.40	37.72	53.92
	SUM-L * [41]	20.37	47.14	62.19	76.09	13.15	32.77	44.35	59.04
	VIEWPOINTROSETTA (Ours)	**25.34**	**58.14**	**71.70**	**83.41**	**19.28**	**47.21**	**61.78**	**74.17**

Table 2. Full results for cross-view retrieval when choosing different k values for the hit rate (HR). * means having access to all the same paired and unpaired data as ours. † means only training with paired data.

stove heat
skillet
pot
stove
Adjust
Turn
garlic cloves
Put away
trash
egg whites
spoon
yolks
Check
paper
recipe
Cut tomato
Transfer food
Add water
Add salt
well integrated
Add oil
Wash hands
empty space
Whisk
Crack eggs
mixture
spatula
plate
Throw
Stir noodles
cup
rotate
frying
spring onions
waste
Check heat
egg mixture
pan
tea
butter flow
Tilt
allow oil
Pour egg
mixing bowl
Wipe hands
Add tomato
green chilies
Cut onions
black pepper
bell peppers
side
Stir
lid
milk
noodles
Place
knife

Figure 8. Word cloud visualization of cross-view action recognition task's action label distribution. Larger words indicate higher frequencies of their occurrence in the labels.

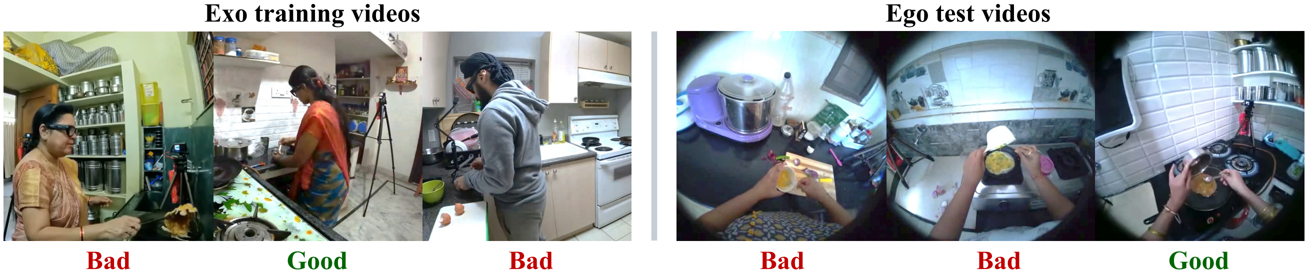

Figure 9. Visualization of cross-view skill assessment task's training exo samples and test ego videos. Common instances of bad execution typically involve incorrect hand-holding gestures or improper procedural order.

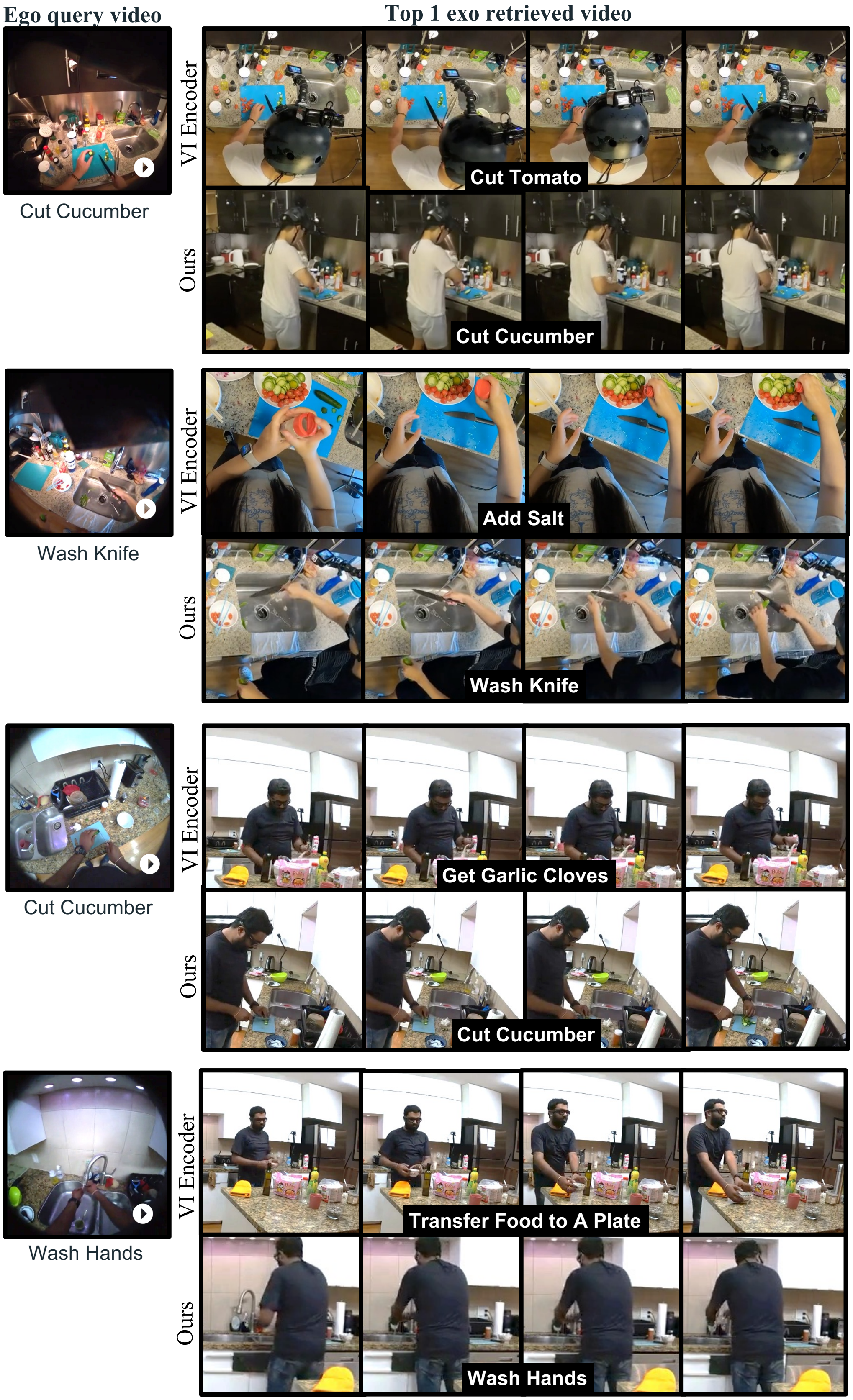

Figure 10. Extras qualitative results of ego-to-exo cross-view retrieval. Compared to the VI Encoder [17], our ViewpointRosetta unlocks unpaired data to capture rich semantic information, enabling retrieval of samples that are not only visually similar to the input view but also semantically related.