Open-Vocabulary Temporal Action Localization using Multimodal Guidance (Supplementary Material)

Akshita Gupta[1,3]
agupta22@uoguelph.ca
Aditya Arora[2,3]
adityac8@yorku.ca
Sanath Narayan[4]
sanath.narayan@tii.ae
Salman Khan[5]
salman.khan@mbzuai.ac.ae
Fahad Shahbaz Khan[5]
fahad.khan@mbzuai.ac.ae
Graham W. Taylor[1,3]
gwtaylor@uoguelph.ca

[1] University of Guelph, Guelph, Ontario
[2] York University, Toronto, Ontario
[3] Vector Institute, Toronto, Ontario
[4] Technology Innovation Institute, Abu Dhabi, UAE
[5] Mohamed Bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE

In this supplementary material, we provide additional quantitative and qualitative analysis of our proposed Open-Vocabulary Temporal Action Localization (OVTAL) framework, OVFormer. Additional implementation details and quantitative results are discussed in § S1, § S2, followed by qualitative analysis in § S3. Finally, we provide details for the LLM-generated text descriptions for THUMOS14 (§ S4) and ActivityNet-1.3 (§ S5) used in the main manuscript.

S1 Additional Implementation details

Datasets: We evaluate OVFormer on two datasets: THUMOS14 [6] and ActivityNet-1.3 [5]. THUMOS14 consists of 20 classes and contains 413 untrimmed videos, while Activi-Net-1.3 is a large-scale dataset with 200 classes and 14,950 videos. Following [7], we divide the datasets into training and testing sets. Furthermore, we consider two settings: (A) training on 75% of the action categories and testing on the remaining 25%, and (B) training on 50% of the categories and testing on the other 50%. For THUMOS14, setting (A) involves 15 categories for training and 5 for testing, whereas setting (B) uses 10 categories for both training and testing. For ActivityNet-1.3, setting (A) assigns 150 categories for training and 50 for testing, while setting (B) uses 100 categories for both training and testing. In each setting, we randomly sample the categories 10 times to create training and testing splits, and we report the average performance across these splits. For pretraining, we utilize the HACS

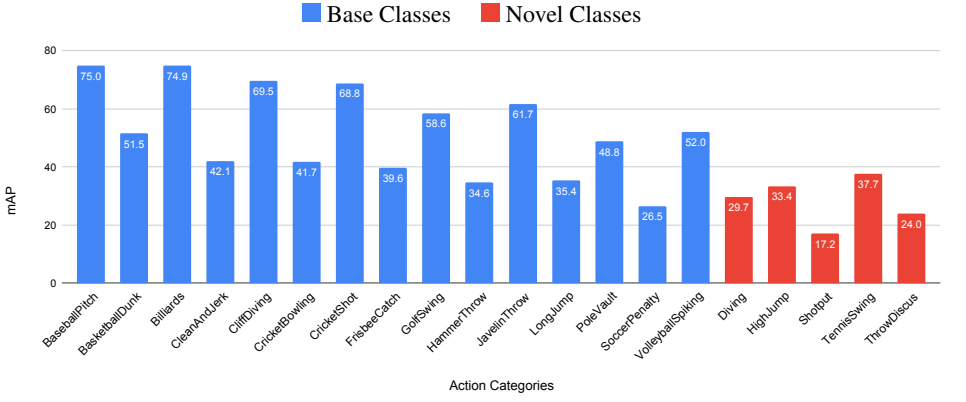

Figure S1: Class-wise average *mAP* for THUMOS14 for 75-25 train-test split.

dataset [15], a large-scale dataset with dense annotations. Importantly, the HACS OV split, consisting of 24,407 videos, does not overlap with the testing splits of THUMOS14 and ActivityNet-1.3, ensuring a fair evaluation of OVFormer generalization capabilities.

Evaluation Metrics: Following other image-based open-vocabulary approaches [9, 12, 16] and TAL methods [3, 8, 13, 14], we report mean average precision over base (mAP_{base}), novel (mAP_{novel}), and all (mAP_{all}) categories. The mAP_{all} is used to show the model's performance across all action classes when both base and novel categories are present during inference. The mAP_{all} is the most important metric: achieving a balance between mAP_{base} and mAP_{novel} is important, and while improving mAP_{novel}, a model should not improve mAP_{novel} at the cost of degrading mAP_{base}. For ZSTAL [7, 9], we report mAP averaged over novel action categories.

Implementation Details: Our architecture is based on ActionFormer [14]. Frame-level features and snippet-level features are extracted using DINOv2 [10] and a two-stream I3D video encoder [2] for HACS, THUMOS14 and ActivityNet-1.3 datasets. For pretraining using the HACS dataset, we use a temporal length of 512, a learning rate of $1e-3$, 40 epochs, and an NMS threshold of 0.75. Furthermore, for finetuning with THUMOS14, we use a temporal length of 2304, a learning rate of $1e-4$, 13 epochs, and an NMS threshold of 0.5. Similarly, for finetuning with ActivityNet-v1.3, we use a temporal length of 192, a learning rate of $1e-3$, 15 epochs, and an NMS threshold of 0.7. To generate text descriptions, we use the `gpt-3.5-turbo-instruct` model available from OpenAI and compute the text embedding using the CLIP ViT-B/32 text encoder model [11]. All experiments are performed using a single NVIDIA A100 GPU.

S2 Additional Quantitative Results

S2.1 Class-wise Average mAP

In Figure S1 and Figure S2, we report class-wise results of OVFormer on THUMOS14 for one of the 10 random splits [7] on 75-25 and 50-50 train-test splits, respectively. Both

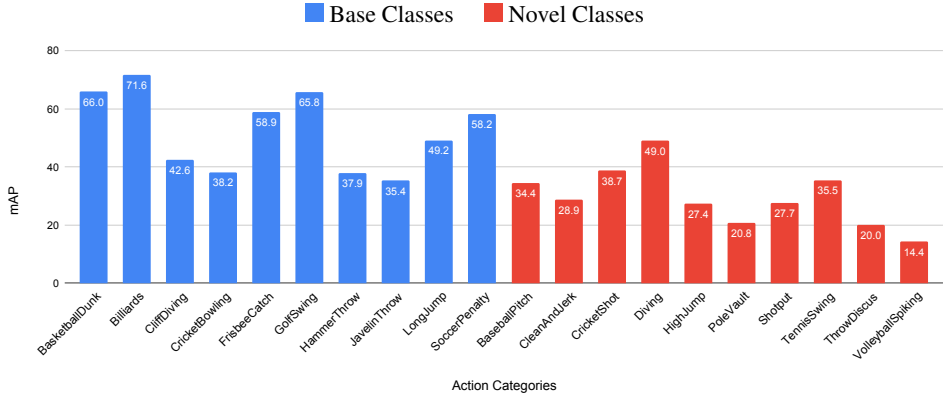

Figure S2: Class-wise average *mAP* for THUMOS14 for 50-50 train-test split.

plots show a high variance in average mAP among the classes, specifically for actions with very similar visual cues. For example, `HammerThrow` and `JavelinThrow` have mAP values of 37.9% and 35.4%, respectively, for the 50-50 split, while `FrisbeeCatch` and `CricketBowling` have mAP values of 39.6% and 41.7%, respectively, for the 75-25 split. We attribute this variance in mAP to the similarity in visual cues and body movements between these actions. For instance, a person in a throwing motion is a common visual cue shared by both `HammerThrow` and `JavelinThrow`. The similarity between these actions motivated us to incorporate rich class-specific language descriptions and integrate the learning of these descriptions alongside the snippet-level features in the form of multimodal guided features. Also, incorporating Stage I training aids in mitigating the issue of overfitting on the base dataset $\mathcal{V}_{base}$. As a result, our approach learns to distinguish these close similarities between fine-grained actions better and enhances the detection of novel action categories without overfitting on the base action categories.

Our OVFormer achieves higher mAP values for the base action categories (shown in blue) compared to the novel ones (shown in red). This is expected, as the model has been trained on the base categories and can better recognize them during inference. However, OVFormer is able to maintain a reasonable performance on the novel action categories. The effectiveness of this method can be observed in the performance on novel action categories. For instance, in the 75-25 split, the novel action categories such as `Diving`, `HighJump`, `Shotput`, `TennisSwing`, and `ThrowDiscus` have mAP values ranging from 17.2% to 37.7%. Similarly, in the 50-50 split, the novel action categories have mAP values ranging from 14.4% to 49.0%. These results demonstrate that OVFormer can effectively generalize to unseen action categories by incorporating rich class-specific language descriptions and the multimodal guided features. OVFormer is able to better distinguish between visually similar actions and improve performance on novel action categories that were not seen during training.

S3 Additional Qualitative Results

In this section, we show additional qualitative results comparing the performance of OVFormer to the baseline method P-ActionFormer on the THUMOS14 and ActivityNet-1.3 datasets. We show results for both novel action categories (Figure S4 and Figure S6) and

base and novel action categories (Figure S3 and Figure S5). In each figure, the top row displays the ground truth action boundaries, the middle row shows the predictions from P-ActionFormer, and the bottom row presents the predictions from OVFormer. We observe that OVFormer improves localization performance for novel action categories compared to P-ActionFormer. Specifically, in Figure S3(a), which shows results on base and novel action categories from THUMOS14, P-ActionFormer confuses `Throw Discus` (novel class) and `Basketball Dunk` (base class) actions when the body movements hold a very strong similarity. However, OVFormer can correctly separate these action categories, showing the significance of the multimodal guided features that capture rich scene information and semantic context related to the actions. Furthermore, in Figure S3(b), also on THUMOS14, P-ActionFormer confuses `Javelin Throw` (base class) and `Volleyball Spiking` (novel class) actions, while OVFormer can correctly distinguish between them. In Figure S4, which shows results on novel action categories from THUMOS14, P-ActionFormer misses the action boundaries for the ground-truth classes `Diving` (Figure S4(a)) and `Volleyball Spiking` (Figure S4(b)), whereas OVFormer is able to correctly localize the action boundaries.

On the ActivityNet-1.3 dataset, Figure S5 shows the localization comparison between OVFormer and P-ActionFormer on base and novel action categories. In Figure S5(a), P-ActionFormer gets confused between visually similar action categories, such as `Ice Fishing` (base class) and `Removing Ice from Car` (novel class), leading to inaccurate localization of the action boundaries when the action category holds visual similarity with other action categories. Similarly, in Figure S5(b), P-ActionFormer confuses `Tennis Throw` (novel class) and `Playing Badminton` (base class), while OVFormer can correctly distinguish between them. In Figure S6, which shows results on novel action categories from ActivityNet-1.3, P-ActionFormer misses the action boundaries for the ground-truth classes `Platform Diving` (Figure S6(a)) and `Discus Throw` (Figure S6(b)), whereas OVFormer is able to correctly localize the action boundaries. All these qualitative examples demonstrate OVFormer's strong open-vocabulary capability, as it leverages multimodal representations to effectively recognize and localize novel action categories that were unseen during training. This is in contrast to P-ActionFormer, which struggles to distinguish between visually similar actions, especially for novel categories.

In Figure S7, we perform a false positive (FP) analysis at tIOU=0.5 for THUMOS14 for 50-50 split on base and novel action categories. For clarity, we choose to show the results on one of the splits from the 10 random splits. We compare the baseline method P-ActionFormer (Figure S7(a)) and OVFormer (Figure S7(b)). We can see a significant improvement in true positive prediction which clearly shows the significance of Stage I training on a larger vocabulary dataset and multimodal guided features for OVTAL. For more detailed explanations regarding the FP analysis chart and error categorization, we refer the readers to the work [1], which introduced this diagnostic tool for evaluating temporal action localization models.

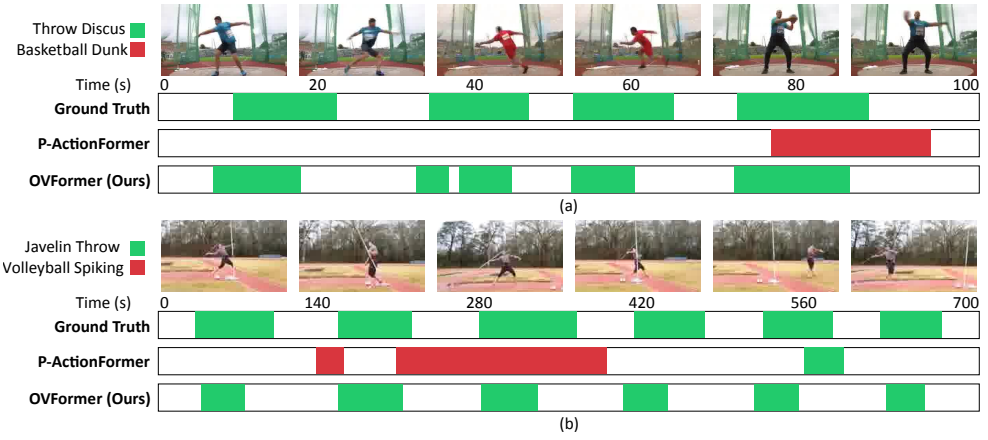

Figure S3: **OVTAL comparison between OVFormer and P-ActionFormer on the test set for THUMOS14 with a 50-50 split on base and novel action categories.** The top row shows the ground truth action boundaries, the middle row shows the baseline method P-ActionFormer's performance, and the bottom row shows the performance of our proposed method OVFormer. In (a), P-ActionFormer struggles to differentiate between the novel action category `Throw Discus` and the base action category `Basketball Dunk`. Similarly, in (b), P-ActionFormer confuses the novel action category `Javelin Throw` with the base action category `Volleyball Spiking`. These errors occur due to the visual similarities between the action categories. In contrast, our proposed method is able to correctly localize the action boundaries. See § S4 for more details.

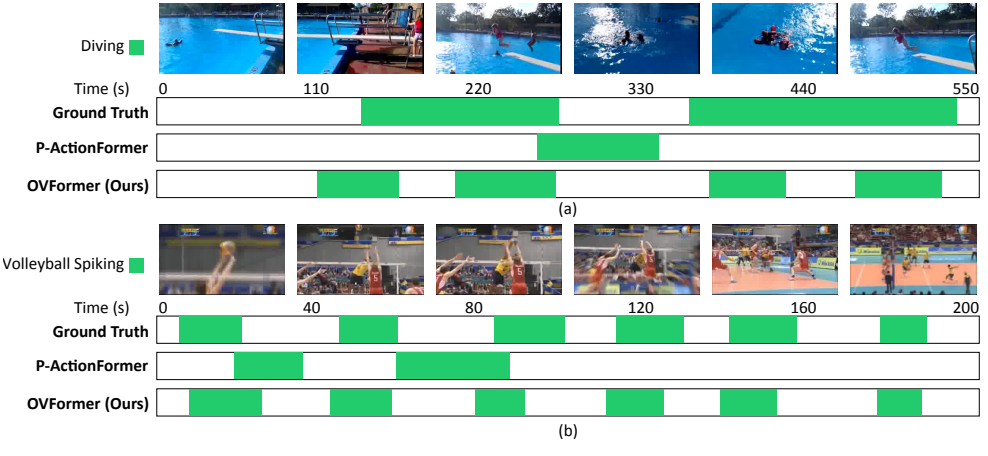

Figure S4: **OVTAL comparison between OVFormer and P-ActionFormer on the test set for THUMOS14 with a 50-50 split on novel action categories.** The top row shows the ground truth action boundaries, the middle row shows the baseline method P-ActionFormer performance, and the bottom row shows the performance for our proposed method OVFormer. We can see that P-ActionFormer misses the action boundaries for the ground-truth classes `Diving` in (a) and `Volleyball Spiking` in (b) whereas our proposed method is able to localize the action boundaries correctly. See § S4 for more details.

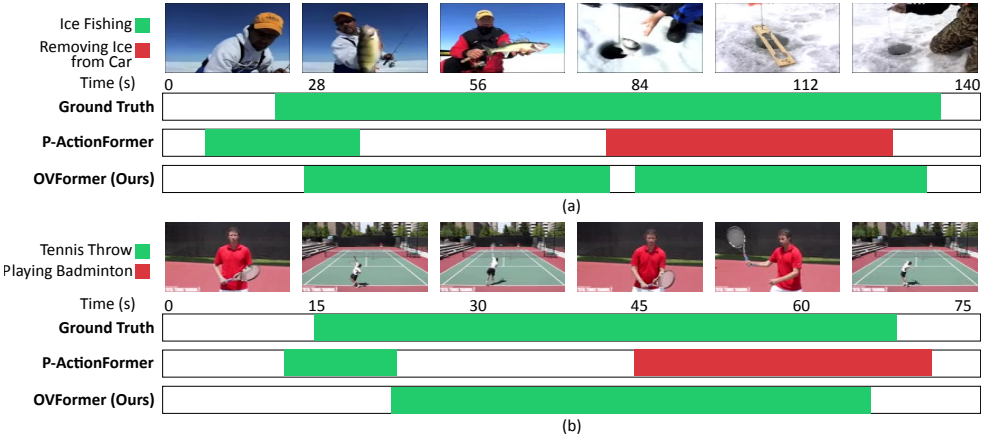

Figure S5: **OVTAL comparison between OVFormer and P-ActionFormer on the test set for ActivityNet-1.3 with a 50-50 split on base and novel action categories.** The top row shows the ground truth action boundaries, the middle row shows the baseline method P-ActionFormer performance, and the bottom row shows the performance for our proposed method OVFormer. In (a), P-ActionFormer struggles to differentiate between the novel action category `Removing Ice from Car` and the base action category `Ice Fishing`. Similarly, in (b), P-ActionFormer confuses the novel action category `Tennis Throw` with the base action category `Playing Badminton`. These errors occur due to the visual similarities between the action categories. Our proposed method is able to localize the action boundaries correctly. See § S5 for more details.

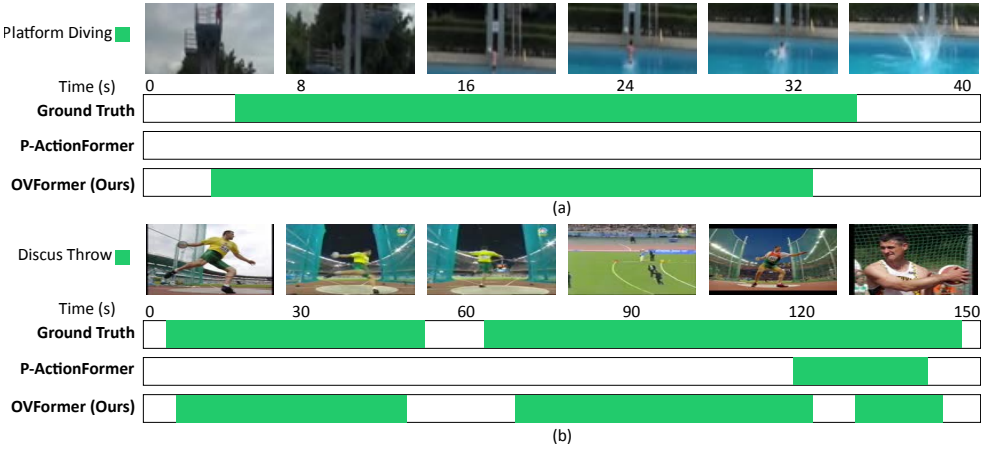

Figure S6: **OVTAL comparison between OVFormer and P-ActionFormer on the test set for ActivityNet-1.3 with a 50-50 split on novel action categories.** The top row shows the ground truth action boundaries, the middle row shows the baseline method P-ActionFormer performance, and the bottom row shows the performance for our proposed method OVFormer. We can see that P-ActionFormer misses the action boundaries for the ground-truth classes `Platform Diving` in (a) and `Discus Throw` in (b) whereas our proposed method is able to localize the action boundaries correctly. See § S5 for more details.

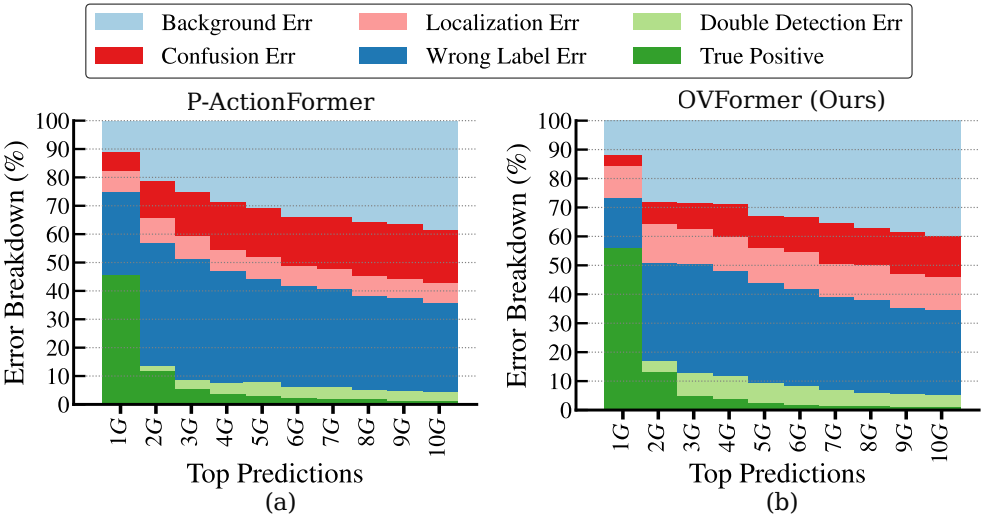

Figure S7: False positive (FP) profiling on THUMOS14 on 50-50 split using the approach from [1]. The figure shows the FP error breakdown for the top 10 ground truth predictions per action category. On the left (a), we have the baseline method P-ActionFormer, and on the right (b), we present our proposed method OVFormer. We observe a significant improvement in true positives for our proposed method and a substantial decrease in confusion errors compared to the baseline method.

S4 Generated Class Description Examples: THUMOS14

In this section, we show 10 rich text descriptions generated using the `gpt-3.5-turbo-instruct` model from OpenAI for five classes from THUMOS14. All text descriptions will be released publicly along with the code.

S4.1 Generated Description for 'BaseballPitch':

1. You can recognize a video of a person performing the BaseballPitch action by looking for certain key actions such as a raised arm, a windup involving a back and forth motion of the arms and a follow-through, a powerful transfer of weight.
2. A video of a person performing a BaseballPitch action can be recognized by the player throwing the baseball with their arm, with their body facing forward and their arm in a slightly bent position, and then releasing the ball with a snapping motion of.
3. The most reliable way to recognize a video of a person performing a BaseballPitch action is by looking for certain visual cues.
4. These cues include the pitcher raising their leg in a kicking motion, a forward-leaning torso, arms bent at a 90.
5. A video of a person performing a BaseballPitch action can be recognized by looking for the following clues: the person holding the ball in an overhand grip, bringing the arm back with the elbow raised, cocking the wrist, and then.
6. A video of a person performing a Baseball Pitch action can be recognized by looking for certain movements in the video.
7. Key features of a Baseball Pitch include the pitcher winding up by swinging backwards with their arm, bringing their body straight, and then bringing.
8. A video of someone performing a Baseball Pitch action can be identified by looking for a sequence of distinct motions.
9. These motions should include the windup, transitioning to the leg kick, driving their arm towards the plate, and releasing the ball.
10. You can recognize a video of a person performing a Baseball Pitch action by looking for features such as arm movement in the windup position, a smooth overhand delivery, and the followthrough of the pitch.

S4.2 Generated Description for 'CliffDiving':

1. One way to recognize a video of a person performing CliffDiving action is by looking for the following visual cues: a high elevation from the ground, a person diving from the cliff, and either a pool, lake, or ocean nicely situated below.
2. Cliff diving can be easily identified by looking for a person performing high jumps and dives off a high cliff into the water below.
3. The cliff diving locations will generally have a steep drop off which is why it is considered a high-risk sport.
4. A video of a person performing a CliffDiving action can be recognized by looking for key traits of cliff diving, such as jumping off a cliff, performing a flip or spin, and entering the water feet first.
5. Cliff diving is an extreme sport that involves diving off a cliff or other high structure into water.
6. To recognize a video of a person performing a cliff diving action, look for visuals of a person leaping off a high structure into water and flashing.
7. You can recognize a video of someone performing a cliff diving action by looking for clues such as a high cliff or outcropping of rock, a person in swimming gear or a wet suit, and the person leaping into the water from the cliff.
8. A video of someone performing a CliffDiving action would typically involve a person diving off of a tall cliff or precipice into the water below.
9. In the video, you may see the person taking a running start, executing a somersault.
10. A video of a person performing a CliffDiving action can be recognized by looking for visuals of an individual jumping and/or diving off a high cliff into a body of water.

S4.3 Generated Description for 'FrisbeeCatch':

1. You can recognize a video of a person performing FrisbeeCatch action by looking for the motions of throwing and catching a Frisbee in the video.
2. You should also look for visual cues such as the Frisbee itself and any.
3. To recognize a video of a person performing the FrisbeeCatch action, look for the following visual cues: the individual throwing the Frisbee, the Frisbee in the air, the person catching the Frisbee, and.
4. A person performing the FrisbeeCatch action can be recognized by their stance – a low athletic position ready to catch the fly-by disc, and by the way they're moving – arms outstretched and eyes tracking the fr.
5. A video of someone performing the Frisbee Catch action can be recognized by seeing them throw a frisbee in the air, and then quickly running to catch it before it hits the ground.
6. There should also be an obvious throwing and catching.
7. You can recognize a video of someone performing FrisbeeCatch by looking for one or more persons throwing and catching a Frisbee.
8. It should be clear that the persons are attempting to catch the Frisbee while it is in the.
9. You can recognize a video of someone performing the FrisbeeCatch action by looking for images of someone throwing a Frisbee and watching to see if they catch it in their hands.
10. Additionally, the video should include the person running, jumping and stretching to catch the Frisbee.

S4.4 Generated Description for 'JavelinThrow':

1. In a video of someone performing the JavelinThrow action, you should look for a person throwing a javelin with good technique and form as well as the javelin leaving their hands and flying through the air.
2. A video of a person performing a JavelinThrow action can be recognized by observing the person's technique as they grip the javelin tightly in their hand, run towards the throwing line, and hurl the javelin.
3. You can recognize a video of a person performing the JavelinThrow action by looking for visual clues.
4. The presence of a javelin in the video
5. A person gripping the javelin, winding.
6. You can recognize a video of a person performing a JavelinThrow action if you observe the person holding a javelin in their dominant hand and throwing it with their arm up in an arching motion.
7. You may also see them run.
8. You can recognize a video of person performing JavelinThrow action by looking for certain key elements.
9. These elements include a person gripping the javelin, running down the field/track, throwing the javelin and watching it soar.
10. The person will have a javelin in their hand.

S4.5 Generated Description for 'Billiards':

1. You can recognize a video of a person performing billiards action by looking for the visual cues of a billiards table, the holding and playing of the billiards cue by the person, and the striking of the billiard balls.
2. To recognize a video of a person performing billiards action, you can look for cues such as a pool table or billiards equipment, a person holding a billiards cue, shots of the ball impacting other balls or the cushion.
3. You can recognize a video of a person performing billiards action by looking for a number of visuals.
4. These visuals could include a person gripping a pool cue, a pool table, pool balls, and objects being struck by the cue ball.

5. A video of someone performing a billiards action can be identified by seeing them take shots with a standard billiards cue stick, striking balls on the billiards table and positioning them with the use of their cue stick, as well.
6. To recognize a video of someone performing billiards action, you should look for the following elements: a billiards table with two pockets at either end, cues or pool sticks and a complete set of billiard balls.
7. You can recognize a video of a person performing billiards action by looking for cues such as the cue ball, pool stick, and billiard table, as well as watching for specific movements associated with billiards.
8. To recognize a video of a person performing a billiards action, you should look for recognizable cues such as the person grabbing a pool cue, the sound of a ball being hit, and the movement of balls on the table.
9. You can recognize a video of a person performing Billiards action by looking for cues such as the person holding a pool cue, a pool table with the balls arranged in a rack, and the sound of the balls being struck together during the.
10. To recognize a video of person performing billiards action, look for cues such as the billiard table, billiard balls, cues, and the various motions of the person playing the game.

S5 Generated Class Description Examples: ActivityNet-1.3

In this section, we show 10 rich text descriptions generated using the `GPT-3.5-turbo-instruct` model from OpenAI for five classes from ActivityNet-1.3. All text descriptions will be released publicly along with the code.

S5.1 Generated Description for 'Applying sunscreen':

1. You can recognize a video of a person performing the action of applying sunscreen by watching them slather the sunscreen on their skin, rubbing it in until their skin is covered, and seeing them put the sunscreen away when they are finished.
2. You can recognize a video of person performing the action of applying sunscreen by looking for the typical signs of the action.
3. You can recognize a video of a person performing the action of applying sunscreen if the person is seen taking out a topical sunscreen product from its container, then applying the product to their skin, ensuring that all exposed skin areas are covered.
4. If you are looking for a video of someone performing the action of applying sunscreen, you may search for terms such as "applying sunscreen video", "sunscreen application", or "sunscreen application tutorial".
5. You can recognize a video of a person performing the action of applying sunscreen by looking for visual cues such as the person applying a white creamy sunscreen product to their face, ears, arms, legs, etc.
6. Visual cues you may look out for in a video of a person applying sunscreen may include seeing someone's hands applying lotion or cream onto their exposed skin, rubbing the lotion into the skin, and/or seeing the person use a sun.
7. You can recognize a video of a person applying sunscreen action by looking for someone taking out a bottle of sunscreen from a bag and then applying it to exposed skin.
8. You can recognize a video of a person performing the action of applying sunscreen by looking for certain items used when applying sunscreen.
9. The video could show the person taking sunscreen in the palm of their hand and applying it on their skin.
10. You can recognize a video of a person performing the action of applying sunscreen by looking for visual cues such as a person of any age, gender, or ethnicity.

S5.2 Generated Description for 'Braiding hair':

1. You can recognize a video of someone performing Braiding hair by looking for someone with a comb in their hand who is separating the hair into sections, twisting the sections of hair around each other and securing each section with a hair tie or clip.
2. You can recognize a video of someone performing the Braiding Hair action by looking for distinct movements such as: sectioning the hair into 3 or more sections, crossing the outer sections over the inner section, looping the strands around each other,.
3. You can recognize a video of person performing braiding hair action by looking for someone holding several strands of hair, parting it into sections, and weaving them into a tight plait or braid.
4. You can recognize a video of someone performing a braiding hair action by looking for visual cues, such as images of someone with their hands braiding another person's hair and/or visible motion of someone's hands doing a braid.
5. Look for a video that shows a person with their hands weaving together strands of hair.
6. You can recognize a video of a person performing the Braiding Hair action by looking for someone who is using their hands to weave and braid hair strands together and forming patterns.
7. You can recognize a video of a person performing Braiding hair action by looking for a person with their hands moving back and forth as if they are weaving together sections of hair.
8. You can recognize a video of a person performing the braiding hair action by looking for someone separation sections of the hair with their hands and weaving them together over and over to create a woven pattern.
9. You can recognize a video of someone performing braiding hair by looking for visual indications of the person or people in the video performing the action of braiding hair.
10. You can recognize a video of a person performing a Braiding hair action by looking for specific visuals such as a person with their hair parted in the middle, with three strands of hair taken and twisted together in a specific pattern.

S5.3 Generated Description for 'Drinking coffee':

1. The person will typically be seen stirring or mixing their coffee, picking up the mug and bringing it to their mouth, and drinking from the mug.
2. You can recognize a video of someone drinking coffee by looking for visual cues such as someone picking up a cup, pushing a lid off of a cup, pouring a liquid into a cup, or putting a spoonful of sugar into a cup.
3. You can recognize a video of someone drinking coffee by looking for certain visuals and sounds.
4. You can look for video footage of the person holding a coffee cup, drinking from the cup, or stirring the coffee with a spoon.
5. You can recognize a video of someone performing the action of drinking coffee by looking for familiar motions, like lifting a cup to their lips, and the characteristic sound of a person savoring a sip of hot drink.
6. A video of a person performing the Drinking Coffee action can be recognized by visual cues, such as the person picking up a mug, bringing the mug to their lips, and then taking a sip of coffee.
7. You can recognize a video of a person performing the drinking coffee action by looking for visual cues such as the person holding a mug, steam rising from a cup, and/or the person taking a sip of the coffee.
8. You could look for video footage of someone taking a sip of coffee, preparing coffee, or pouring coffee into a cup.
9. You can recognize a video of a person performing the Drinking Coffee action by looking for the action of a person picking up a cup of coffee and putting it to their mouth.
10. You can recognize a video of a person performing the Drinking coffee action by looking for visuals such as a person holding a mug of coffee, making the drinking motion with their hand, or looking into a cup of coffee.

S5.4 Generated Description for 'Skiing':

1. A video of someone performing a skiing action can be recognized by observing how the person moves their body and skis down a slope.
2. You can recognize a video of a person performing a skiing action by looking for recognizable ski clothing, skis, ski poles and other ski equipment, and by watching for the person to make recognizable skiing motions, such as gliding down a hill.
3. You can recognize a video of someone performing a skiing action by looking for the following elements: the person wearing ski apparel, the skiing equipment and the environment (snow-covered slopes, ski-lifts, and other skiers).
4. One way to recognize a video of someone performing the skiing action is to look for telltale signs such as the person wearing alpine skiing equipment, such as ski boots, skis, poles, and a helmet.
5. Look for someone skiing down a hill with skis, poles, and ski boots.
6. You may recognize a video of someone skiing by looking for recognizable skiing positions and movements, such as edging, carving, and making turns.
7. One way to recognize a video of a person performing the skiing action is to look for clues such as snow, skis, ski poles, and the crouched position that a skier assumes when skiing.
8. You can recognize a video of person performing skiing action by looking for visual elements that include a person skiing down a slope or off a jump and make turns, wearing ski equipment like boots, bindings, and skis.
9. You can recognize a video of someone performing skiing action by looking for specific visual cues.
10. You can recognize a video of someone performing skiing by looking for recognizable skiing movements such as a two-footed gliding motion, making turns in the snow, or controlling speed by using pole plants.

S5.5 Generated Description for 'Making a sandwich':

1. You can recognize a video of person performing the action of Making a sandwich by looking for visual clues such as seeing a person assembling bread, meat, cheese, and other ingredients; slicing these ingredients; and arranging them on a plate.
2. You can recognize a video of a person performing the action of making a sandwich by observing the physical movement of the person putting ingredients between two slices of bread, such as meat, cheese, and condiments, and then finishing off the process by.
3. You can recognize a video of someone making a sandwich by looking for footage of them putting bread, meat, and vegetables onto a plate and combining them into a sandwich.
4. You can recognize a video of a person performing the action of making a sandwich by observing the person going through the steps of constructing the sandwich, such as spreading the condiments, arranging the ingredients, and slicing the sandwich in half.
5. You can recognize a video of someone performing the action of making a sandwich by looking for visual cues such as a person cutting, spreading, and arranging various ingredients on bread or an alternative base.
6. You can recognize a video of a person making a sandwich by looking for several key components.
7. You can recognize a video of a person making a sandwich by observing the visual of the person assembling the sandwich, such as spreading butter, putting slices of meat and cheese, adding condiments and vegetables, then cutting it in half.
8. You can recognize a video of someone making a sandwich action by looking for someone with bread, fillings, and any other necessary items such as knives, cutting boards, etc.
9. You can recognize a video of a person performing the action of making a sandwich by looking for visuals such as: someone assembling two pieces of bread, adding condiments such as meat, cheese and/or vegetables, and putting condiments like mayo.
10. To recognize a video of a person performing the action of making a sandwich, you can look for visuals of the person gathering the ingredients for a sandwich, assembling the sandwich together, and then cutting the sandwich into slices.

References

[1] H. Alwassel, F. C. Heilbron, V. Escorcia, and B. Ghanem. Diagnosing error in temporal action detectors. In *Proceedings of the European conference on computer vision (ECCV)*, pages 256–272, 2018.

[2] J. Carreira and A. Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *CVPR*, 2017.

[3] F. Cheng and G. Bertasius. TALLFormer: Temporal action localization with a long-memory transformer. In *ECCV*, 2022.

[4] X. Gu, T.-Y. Lin, W. Kuo, and Y. Cui. Zero-shot detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021.

[5] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles. ActivityNet: A large-scale video benchmark for human activity understanding. In *CVPR*, 2015.

[6] H. Idrees, A. R. Zamir, Y.-G. Jiang, A. Gorban, I. Laptev, R. Sukthankar, and M. Shah. The THUMOS challenge on action recognition for videos "in the wild". *CVIU*, 2017.

[7] C. Ju, T. Han, K. Zheng, Y. Zhang, and W. Xie. Prompting visual-language models for efficient video understanding. In *ECCV*, 2022.

[8] T. Lin, X. Liu, X. Li, E. Ding, and S. Wen. BMN: Boundary-matching network for temporal action proposal generation. In *ICCV*, 2019.

[9] S. Nag, X. Zhu, Y.-Z. Song, and T. Xiang. Zero-shot temporal action detection via vision-language prompting. In *ECCV*, 2022.

[10] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[11] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[12] A. Zareian, K. D. Rosa, D. H. Hu, and S.-F. Chang. Open-vocabulary object detection using captions. In *CVPR*, 2021.

[13] R. Zeng, W. Huang, M. Tan, Y. Rong, P. Zhao, J. Huang, and C. Gan. Graph convolutional networks for temporal action localization. In *ICCV Workshops*, 2019.

[14] C.-L. Zhang, J. Wu, and Y. Li. ActionFormer: Localizing moments of actions with transformers. In *ECCV*, 2022.

[15] H. Zhao, A. Torralba, L. Torresani, and Z. Yan. Hacs: Human action clips and segments dataset for recognition and temporal localization. In *ICCV*, 2019.

[16] Y. Zhong, J. Yang, P. Zhang, C. Li, N. Codella, et al. RegionCLIP: Region-based language-image pretraining. In *CVPR*, 2022.