A APPENDIX

A.1 DATASETS

The details of the 11 downstream datasets are shown in Table 5. The accuracy metric of each dataset follows CLIP.

Dataset	Classes	Train size	Test size
ImageNet	1000	1,281,167	50,000
Caltech-101	102	3,060	6,085
Oxford Pets	37	3,680	3,669
Stanford Cars	196	8,144	8,041
Oxford Flowers 102	102	2,040	6,149
Food-101	102	75,750	25,250
FGVC Aircraft	100	6,667	3,333
SUN397	397	19,850	19,850
Describable Textures	47	3,760	1,880
EuroSAT	10	10,000	5,000
UCF101	101	9,537	1,794

Table 5: 11 Datasets statistics.

A.2 IMPLEMENTATION DETAILS

We use a few-shot training strategy in all experiments at 16 shots which are randomly sampled for each class. We apply APPLe on a pre-trained ViT-B/16 CLIP model where the textual feature dimension is 512, the visual feature dimension is 768 and then transformed to 512. We generate 50 prototypes per category for all 11 datasets with the GPT-3 model (Brown et al., 2020). The textual features of prototypes are fine-tuned for 100 epochs. The batch size is set to 64, of which the samples are strictly from base classes only. We use an SGD optimizer with a learning rate of 0.002. All experiments are conducted on an NVIDIA A6000 GPU. We report base and new class accuracies and harmonic mean (HM) averaged over 3 runs.

A.3 DISCUSSION AND COMPARISON TO CONTEXT OPTIMIZATION METHODS

APPLe significantly differs from existing context optimization methods. With the above experiments, APPLe provides a different perspective to understand CLIP:

1. Is the generalization ability claimed in context optimization methods validated in practice?
   - **No.** APPLe* outperforms context optimization methods on new classes, and new datasets by a large margin. We confirm that the original CLIP already possesses strong generalization ability on new datasets, but a singular prompt cannot exploit such ability.
2. When disregarding the generalization to new classes, is optimizing the context information more effective than fine-tuning the textual features?
   - **No.** The results of our comprehensive experiments indicate that tuning the textual features stands out as more beneficial, enhancing performance predominantly on base classes.
3. Does fine-tuning the prompt textual features cause the overfitting issue?
   - **No.** As evidenced in Fig. 4, utilizing a minimal number of prototypes per class (1 to 3) does lead to overfitting issues. However, increasing the prototype count to more than 10 significantly mitigates this, fostering improved generalization to new classes.

A.4 ADAPTIVE ATTENTION VISUALIZATION

To better understand the impact of adaptive attention, we visualize the learned attention matrix on a heatmap as shown in Fig. 7. The weights of different prompts in each class vary significantly, ranging from around 0.68 to 0.93. As in Equation 2, the adaptive attention matrix $\boldsymbol{W}$ is applied to

the computed similarity matrix, resulting in weighted confidence on the prompt prototypes. Thus, the attention matrix plays an important role in controlling the contribution of each prototype.

We also show the two prototypes, which show the least and the maximum contribution to the corresponding categories. The prototype of least attention is "`The image is of a beige, cream, and brown rug.`" We can identify that this prompt involves the wrong category name `Rug`, rather than the ground-truth name `Woven`. Therefore, our proposed adaptive attention mechanism has *the ability to handle noisy prototypes*. Thus, a As for the prompt of maximum attention "`The image is of a close-up of a knitted cream-colored sweater.`" Other prompts in the same category involve specific colors that cover a small portion of the visual samples, but `cream-colored` is a more general keywords that may match more visual samples. Thus, a higher attention value is assigned.

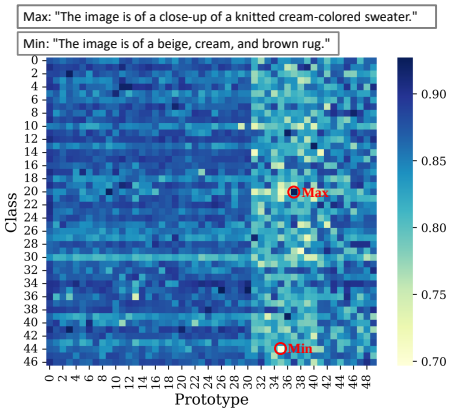


Figure 7: Heatmap visualization of the learned attention matrix on DTD dataset.

A.5 Cross-Dataset Transfer

APPLe can be easily extended to perform cross-dataset transfer by involving the target class prompt prototypes during training. Specifically, we incorporate all prompt prototypes (50 prototypes in each class) from 11 datasets. When fine-tuning on base classes from the source dataset ImageNet, we include prototypes for both base and new classes. This inclusion allows the model to learn and optimize features that are not only specific to the base classes but also relevant to the new classes in the target dataset. By training on the visual samples from the ImageNet dataset, these prompt prototypes can enhance the overall generalization ability. In addition, we note that the training-free version of APPLe, denoted as APPLe*, can already achieve much better cross-dataset transferability. We confirm that ***the original CLIP already possesses strong generalization ability on new datasets, but a singular prompt cannot exploit such ability.***

Table 6: Comparison of baseline methods and APPLe in the cross-dataset transfer setting.

	Source	*Target*										
	ImageNet	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	Aircraft	SUN397	DTD	EuroSAT	UCF101	Average
CoOp	71.51	93.70	89.14	64.51	68.71	85.30	18.47	64.15	41.92	46.39	66.55	63.88
Co-CoOp	71.02	**94.43**	90.14	65.32	71.88	86.06	22.94	67.36	45.73	45.37	68.21	65.74
MaPLe	70.72	93.53	90.49	66.29	72.23	86.20	24.74	67.01	46.49	**48.06**	68.69	66.30
APPLe*	69.88	93.35	91.11	66.30	**73.37**	**86.37**	28.17	67.72	54.79	46.16	70.16	67.75
APPLe	**71.90**	93.51	**91.22**	**66.32**	73.16	86.31	**28.26**	**67.98**	**55.26**	46.15	**70.26**	**67.84**

A.6 Impact of Training Epochs

We show the impact of the training epochs on CoOp in Table 7. The default training epoch of CoOp is 200, which makes the model overfit to the base classes. However, when training on fewer epochs (e.g., 10), we see a large performance improvement.

A.7 More Retrieval Results

We visualize the retrieved samples on the DTD dataset. As shown in Fig. 8, when retrieving the top 5 image samples for cobwebbed, CLIP retrieves all false positive samples. In contrast, our method can accurately find the 5 cobwebbed images. Our correct retrieval results are based on the averaged confidence of prototypes.

We present 4 cases in Fig. 9. The ambiguous keywords in the prompts are highlighted in red.

Table 7: Impact of Training Epochs on CoOp

Datasets	Sets	CLIP	CoOp200	CoOp150	CoOp100	CoOp50	CoOp20	CoOp15	CoOp10	CoOp5	CoOp2	Co-CoOp	APPLe
DTD	Base	53.24	79.44	79.25	78.82	79.98	80.90	78.01	76.85	73.15	61.46	77.01	82.41
	New	59.90	41.18	36.15	35.39	36.67	41.43	50.97	49.28	47.10	48.55	56.00	69.57
	H	56.37	54.24	49.65	48.85	50.28	54.78	61.66	60.05	57.30	54.25	64.85	75.45
Flowers	Base	72.08	97.60	97.53	97.60	97.47	96.58	96.68	95.35	92.02	79.68	94.87	96.58
	New	77.80	59.67	58.59	60.17	61.72	61.35	68.44	70.35	71.63	73.33	71.75	78.58
	H	74.83	74.06	73.20	74.44	75.58	81.33	80.15	80.96	80.55	76.37	81.71	86.66
UCF	Base	72.08	84.69	84.38	84.07	84.80	82.57	83.14	82.99	79.89	76.32	82.33	86.56
	New	77.80	56.05	50.82	55.22	58.36	64.47	66.79	67.17	67.71	69.50	73.45	81.99
	H	74.83	67.46	63.43	66.66	69.14	72.41	74.07	74.25	73.30	72.75	77.64	84.21

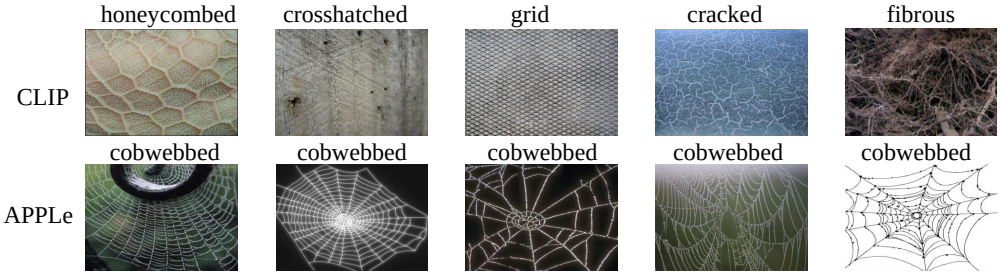


Figure 8: Retrieved image samples for category *cobwebbed* on DTD dataset.

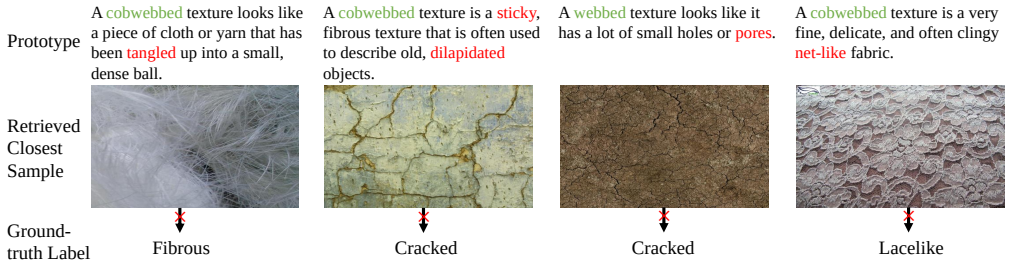


Figure 9: Failure cases of retrieved closest samples with single prompt prototypes on DTD dataset.

A.8 Results on few-shot learning

We have provided experimental results for the few-shot learning setting. Our experiments suggest that our method (APPLE) shows promising results compared to both PLOT (Chen et al., 2022) and ProDA (Lu et al., 2022).

A.9 Comparison to the generic 80 prompts in the CLIP model

CLIP's performance on ImageNet benefits from 80 generic prompts. However, our work focuses on the limitations of these generic prompts, particularly for fine-grained datasets. These prompts, as limited in the official CLIP repository [1], are less effective for such datasets due to their lack of specificity. For example, prompts like 'a plastic ' or 'a in a video game' do not capture fine-grained distinctions well. It is worth noting that the CLIP performance reported in our paper uses the customized prompts as indicated in the CLIP paper, e.g., "A photo of a label, a type of pet." for OxfortPets.

These context prompts can indeed improve performance. We have evaluated the base-to-new setting on ImageNet in Table 9 (CLIP-80-emb). As for the embedding ensembling method mentioned in (Radford et al., 2021), we have tested the performance between embedding ensembling and logits ensembling methods. It can be seen that without fine-tuning, the performance results between the

[1] https://github.com/openai/CLIP/blob/main/notebooks/Prompt$_{E}ngineering_{f}or_{I}mageNet.ipynb$

#Shot	Methods	ImageNet	UCF	Food101	DTD	Stanford Cars	Caltech101	OxfordPets	Flowers102	Food101	FGVC AirCraft	SUN397
1	CoOp	66.32	71.68	81.26	47.04	64.39	90.87	86.10	82.14	81.30	25.95	61.64
	ProDA	68.73	69.92	85.39	49.23	68.88	93.47	89.15	75.40	84.91	28.95	68.55
	PLOT	66.45	74.31	86.16	54.57	68.81	93.14	91.89	80.48	86.16	28.60	66.77
	APPLe	**70.60**	**75.39**	**86.36**	**56.56**	**69.80**	**94.55**	**91.91**	**83.39**	**86.23**	**33.15**	**69.90**
2	CoOp	67.87	74.33	82.32	49.88	69.28	93.51	86.26	89.77	82.87	31.29	66.86
	ProDA	69.44	71.98	85.63	54.61	69.48	94.36	91.55	79.94	85.98	30.15	70.50
	PLOT	68.28	76.76	86.33	56.72	**73.17**	94.69	**92.29**	**89.81**	86.33	31.14	68.06
	APPLe	**70.65**	**77.72**	**86.41**	**59.93**	72.34	**95.21**	92.10	87.29	**86.47**	**35.22**	**71.21**
4	CoOp	69.76	78.30	84.65	57.45	72.70	94.20	89.42	92.85	83.15	33.27	69.89
	ProDA	70.96	76.05	86.38	61.23	69.99	95.01	92.23	86.24	86.33	32.67	72.38
	PLOT	70.40	79.76	86.46	62.43	**76.25**	95.13	92.55	**92.93**	86.46	35.29	71.73
	APPLe	**71.77**	**80.07**	**86.85**	**65.25**	75.64	**95.33**	**92.72**	90.42	**86.59**	**36.81**	**73.44**
8	CoOp	70.81	79.12	85.14	64.42	76.42	95.25	92.01	**95.93**	84.59	30.15	72.33
	ProDA	71.89	79.75	87.07	67.08	71.88	95.05	93.16	92.33	87.01	34.32	73.99
	PLOT	71.31	82.80	86.58	66.49	**81.26**	95.51	93.02	95.44	86.58	41.42	73.93
	APPLe	**72.80**	**84.75**	**87.29**	**67.73**	79.65	**96.06**	**93.54**	93.91	**87.37**	**41.94**	**74.85**
16	CoOp	71.92	81.95	86.32	69.42	78.70	95.42	92.80	95.94	85.62	35.89	74.28
	ProDA	72.93	82.69	87.24	70.09	75.09	95.09	93.54	96.02	87.41	37.17	75.64
	PLOT	72.60	**85.34**	87.11	71.43	**84.55**	96.04	93.59	**97.56**	87.11	**46.74**	76.03
	APPLe	**73.73**	85.11	**87.61**	**72.52**	82.99	**96.63**	**94.22**	95.90	**87.65**	46.53	**76.43**

Table 8: Few-shot Learning results

two methods are relatively similar. However, if we want to further fine-tune the textual features, embedding ensembling method tends to overfit to base classes. This phenomenon has been discussed in Figure 4.

As for the test time speed, logits ensembling indeed causes more computation overhead, but only for dot product between the visual and textual features. Because the textual features of the 80 generic prompts also need to be inferred through the language transformer 80 times. The dot product between visual and textual features needs relatively less computational resources.

Methods	CLIP	CLIP-80-emb	CLIP-80-logits	APPLe*-50-emb	APPLe*-50-logits	APPLe-50-emb	APPLe-50-logits
Base	72.43	73.56	73.54	74.62	74.62	74.16	78.17
New	68.14	69.97	69.99	71.79	71.94	71.93	72.12
HM	70.22	71.72	71.72	73.17	73.26	73.02	75.02

Table 9: Comparison to the generic 80 prompts in the CLIP model

A.10 Generalizability across different CLIP variants and other VLMs

We conducted extensive experiments with various CLIP models and other VLMs, as detailed in Table 11. Our method's consistent performance improvement across different architectures, including ViT-B/16, ViT-B/32, ViT-L/16, ViT-L/16@336, LAION's CLIP replication (Schuhmann et al., 2022), and the BLIP model (Li et al., 2022), attests to its robustness. These results underscore our method's adaptability to different model architectures and training datasets. This is particularly significant given that each of these variants and models has unique characteristics and was trained on diverse datasets.

A.11 Prompt Quality Impact

The quality of prompts is a critical factor in our experiments, as they directly influence the model's ability to accurately interpret and classify images based on textual descriptions.

Models	Set	CLIP	CLIP-80	APPLe*	APPLe
ViT-B/16	Base	72.43	73.56	74.62	78.17
	New	68.14	69.97	71.94	72.12
	HM	70.22	71.72	73.26	75.02
ViT-B/32	Base	67.43	67.52	69.40	72.66
	New	64.04	65.84	67.78	67.83
	HM	65.69	66.67	68.58	70.16
ViT-L/14	Base	79.20	79.96	81.03	83.51
	New	74.02	76.43	78.06	78.13
	HM	76.52	78.16	79.52	80.73
ViT-L/14@336	Base	80.25	81.04	82.00	84.26
	New	75.50	77.60	79.08	79.15
	HM	76.52	78.16	80.51	81.63
LAION ViT-B-32	Base	70.01	70.07	70.53	73.45
	New	69.06	69.66	70.13	70.22
	HM	69.53	69.86	70.33	71.80
BLIP ViT-B	Base	43.63	50.10	54.40	67.56
	New	48.42	59.16	63.38	66.28
	HM	45.90	54.25	58.55	66.91

Table 10: Generalizability across different CLIP variants and other VLMs

In our experiments, we observed that the performance with prompts generated by GPT-3 is relatively consistent with those generated by GPT-4. This could be attributed to the fact that generating category descriptions for our tasks may not require advanced reasoning capabilities, a domain where GPT-4 has more significant improvements over GPT-3. Hence, GPT-3's capacity appears to be sufficient for this specific task.

Regarding the column labeled 'mixed1' in our table, it represents a mix of prompts generated by both GPT-3 and GPT-4. We included this to examine the impact of using a heterogeneous set of prompts on model performance. The results indicate that there is not a significant deviation in performance when using mixed-quality prompts compared to those generated solely by GPT-3 or GPT-4.

We acknowledge the limitations in our current testing due to the time constraints of the rebuttal period. In the next version of our paper, we plan to conduct more comprehensive testing across all 11 datasets to demonstrate the consistency of performance regardless of the prompt generation source. This will provide a more complete picture of how different qualities of prompts impact the overall effectiveness of our model.

Datasets	Set	GPT-3	GPT-4	mixed1
ViT-B/16	Base	82.41	83.10	81.95
	New	69.57	70.41	70.72
	HM	75.45	76.23	75.92
ViT-B/16	Base	95.64	95.43	95.69
	New	98.04	97.93	98.21
	HM	96.83	96.66	96.93
ViT-B/16	Base	44.66	45.14	45.20
	New	43.13	39.41	41.81
	HM	43.88	42.08	43.44

Table 11: Prompt Quality Impact

A.12 Prototype Calibration Strategies

We have explored alternative balancing methods, including the Boltzmann operator and logsumexp. We conducted additional experiments on ImageNet using these methods and present the results in the

Table 13. We test different Boltzmann temperatures (T=20, 10, 5) to understand their impact on performance. Our findings show that the mean/max balancing method outperforms both the Boltzmann operator and logsumexp. This superiority may be attributed to the specific way mean/max balancing integrates information from all prototypes while emphasizing the most relevant ones, which seems particularly effective for our model and dataset.

Regarding the Boltzmann operator, we observed that varying the temperature has a noticeable impact on performance. Lower temperatures (T=5) led to results closer to our method, suggesting that a tighter focus on the most relevant prototypes can be beneficial. However, none of the temperatures tested could match the performance achieved by our mean/max balancing. The performance with logsumexp, especially in the Base class, was notably lower. This could be due to the mathematical properties of logsumexp, which might lead to a less effective balance between prototype relevance and diversity in our specific application.

These experiments reinforce the robustness and effectiveness of our chosen mean/max balancing method. It demonstrates that our approach is well-suited for handling the classification challenges in our model, outperforming other common balancing strategies.

Methods	Mean/Max Balancing	Boltzmann T=20	T = 10	T = 5	Logsumexp
Base	78.17	76.96	77.35	77.30	70.48
New	72.12	71.96	71.96	72.07	71.64
HM	75.02	74.27	74.56	74.59	73.73

Table 12: Prototype Calibration Strategies

A.13 Best prototype selection

We experiment with training attention weights only and select the best prototype with the highest weights. We conduct this experiment on ImageNet and compare the performance with CLIP's hand-designed prompts and our APPLe* model.

As shown in Table 13, the performance using the best prototype selected by the model is inferior to that of the hand-designed prompts used in CLIP. This outcome suggests that relying on a single prototype, even if it's the 'best' as determined by the model, may not effectively capture the diverse and complex nature of objects in images. Our results indicate that a single prototype is often biased towards a specific representation of an object, which limits its generalizability.

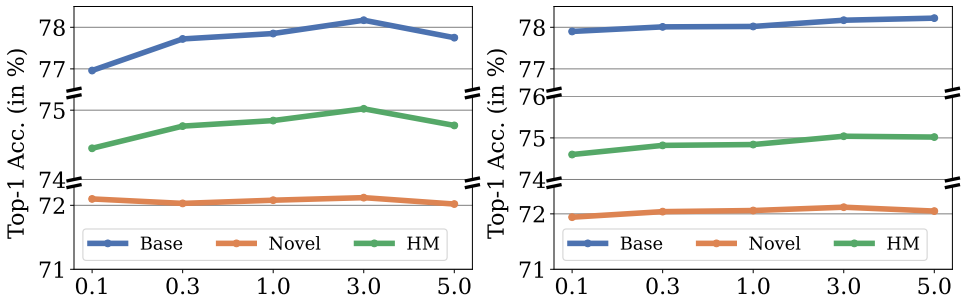
For example, as highlighted in Figure 1 of our paper, each prompt for an apple pie depicts a particular state or aspect of the pie, such as a round pie or a slice on a plate. While each prompt is accurate in its description, none can encompass all the variations of apple pies alone. This specificity is where the limitation lies. It becomes challenging for a single, even well-crafted prototype to represent the breadth of variations that an object can have in the real world.

Methods	CLIP	CLIP-closest prompt	APPLe*
Base	72.43	71.84	74.62
New	68.14	67.90	71.94
HM	70.22	69.81	73.26

Table 13: Performance Comparison with the Best Prototypes.

A.14 Hyper-parameters Analysis

We report the effects of varying the loss coefficients λ_1 and λ_2 for ℓ_{max} and ℓ_{dec} on ImageNet. As shown in Fig. 10, both hyper-parameters achieve the best performance at 3.



(a) Impact of ℓ_{max} (b) Impact of ℓ_{dec}

Figure 10: Hyper-parameter sensitivity

A.15 PROMPTS

We randomly select some prompt samples from 11 datasets to present in Table 14. We mainly use the following five prompts for GPT-3 to generate the prototypes. c is the category name and length is the expected length of the generated prompts.

1. t1 = f"Describe a photo of c) in one short sentence, no more than length words."
2. t2 = f"How does a c look like? Answer in no more than length words."
3. t3 = f"Summarize visual features of c in no more than length words."
4. t4 = f"Tell me what c looks like in a short sentence, less than length words."
5. t5 = f"Use less than length words to outline the look of c."

Table 14: Prompt Samples of 11 Datasets

Dataset	Prompt Samples
ImageNet	There are many different types of military uniforms, but they all share some common features. The image is of a red and white mitten with a green background. A graduation cap typically has a square or pyramid shape and is made of stiff paper or fabric.
Caltech-101	A motorbike with two wheels and a seat. The image shows an airplane flying through the sky. You can identify a ant by its long, segmented body and its long, bent antennae.
Oxford Pets	The Russian Blue has a sleek and elegant coat of bluish-gray fur with bright green eyes. The Scottish Terrier is a small, sturdy breed with a distinctive wiry coat and a distinctive beard. A hairless cat breed with wrinkled skin, large ears, and a slim muscular body.
Stanford Cars	The 2007 Dodge Dakota Club Cab is a four-door truck with aRaised Crew Cab and Extended Cab. A 1993 Geo Metro Convertible would look like a small, boxy car with a convertible top. The image is of a silver 2012 Dodge Charger Sedan.
Oxford Flowers	Cyclamen have heart-shaped leaves with vibrant flowers that come in various shades of pink, red, white, and purple. Frangipani is a stunning tropical flower with vibrant, star-shaped petals and a captivatingly sweet fragrance. The sword lily features tall, slender stalks topped with vibrant, sword-shaped blooms in various hues.
Food101	A baklava is a layered pastry made with nuts, honey, and phyllo dough. The image is of a beet salad with goat cheese, arugula, and pistachios. A bread pudding generally has a bread base with eggs, milk, and sugar added.
FGVC Aircraft	A 737-300 aircraft operated by SouthWest Airlines takes off from the Ronald Reagan Washington National Airport. The image is of an aircraft 737-400 with the engines running. The plane is on the runway ready for take off. Bustling with activity, this Gulfstream IV is parked on the tarmac, with ground crew attending to it.
SUN397	This image is of the Abbey of Saint-Denis, a large abbey located in the northern suburbs of Paris, France. An image of an amusement arcade shows a large room with brightly lit machines and people playing games. A ticket booth and information desk are visible in this shot of the indoor seating area of a movie theater.
Describable Textures	A bubbly texture can look like small mountains with peaks that are round and smooth. A chequered texture is a texture that has a repeating pattern of light and dark squares. The crystalline texture is characterized by having a distinct, three-dimensional crystal structure.
EuroSAT	A centered satellite image of a River displays areas of human settlement protected by levees or embankments. A centered satellite image of Residential Buildings displays clusters of chimneys or rooftop vents. A centered satellite image of Forest displays clear linear patterns, indicating logging roads or forest trails.
UCF101	A person is applying lipstick in the image. A person doing Baby Crawling looks like a human infant crawling on all fours. The person is doing a handstand on the balance beam.