Tune-An-Ellipse: CLIP Has Potential to Find What You Want

Supplementary Material

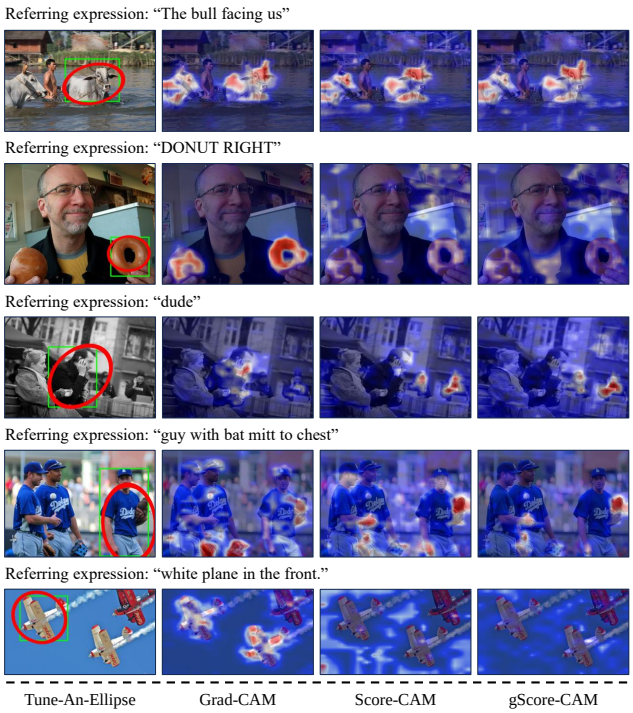

Figure 8. Visual comparisons among various CAM techniques.

Visualization Results. We present visual comparisons among Grad-CAM [22], Score-CAM [29], and gScore-CAM [4] in Fig. 8. It is evident that among these CAM techniques, Grad-CAM generally excels in attending to the most relevant regions. However, Grad-CAM tends to focus on some irrelevant regions. Our Tune-An-Ellipse, on the other hand, effectively eliminates such interferences and achieves accurate localization.

We show some failure selection/initialization of CLIP and our tuned results in Fig. 9. For instance, as shown in the first row, both RedCircle and our initial selection incorrectly identify the object when given "babychair on the left". However, after the visual prompt tuning, our method successfully locates the correct target regions.

Visualizations of the localization results achieved by Tune-An-Ellipse are presented in Fig. 10. It is clearly observed that Tune-An-Ellipse exhibits significant potential in accurately, comprehensively, and compactly localizing the target regions described in the referring expressions, even in complicated scenes. As illustrated in the right section of the fourth column, when provided with referring expressions "cat looking at us" and "cat reflection", Tune-An-Ellipse successfully distinguishes the two target regions.

Localization results of Tune-An-Ellipse at various tuning steps are visualized in Figs.11, 12, and 13. It is evident that the initial red ellipses typically cover only a small part of the ground-truth regions. As Tune-An-Ellipse progresses through the tuning steps, the initial red ellipse gradually refines its shape and position, achieving accurate and complete localization of the target regions. For example, in the top of Fig.11, given a referring expression "guy in black shorts", the red ellipse is initially located on the regions of black shorts. From the first step to the final step, the initial red ellipse is gradually inflated and covers the target object.

Loss	Grad-CAM	RefCOCO		
		val	*testA*	*testB*
RN50×16	ViT-B/16, ViT-L/14@336px	25.95	30.40	20.57
ViT-B/16		26.25	30.99	20.01
ViT-B/32		24.18	27.42	19.35
ViT-L/14		26.34	31.35	20.73
ViT-L/14@336px		26.47	31.39	20.31
RN50×16, ViT-B/16		26.56	31.11	20.51
RN50×16, ViT-B/32		26.21	29.91	20.14
RN50×16, ViT-L/14		26.64	31.12	**20.84**
RN50×16, ViT-L/14@336px		26.02	31.43	20.57
ViT-B/16, ViT-B/32		25.95	30.55	20.19
ViT-B/16, ViT-L/14		26.33	**32.31**	19.84
ViT-B/16, ViT-L/14@336px		**26.74**	31.85	20.11
ViT-B/32, ViT-L/14		25.88	30.65	20.09
ViT-B/32, ViT-L/14@336px		26.39	30.92	20.01
ViT-L/14, ViT-L/14@336px		26.53	30.72	20.43

Table 5. Results of ensembles of various CLIP's image encoders.

Ablation Studies. Evaluation results obtained using different ensembles of CLIP's image encoders are presented in Table 5. It is evident that an ensemble comprising two distinct CLIP image encoders generally yields improved performance on RefCOCO *val*, *testA*, and *testB* sets.

Tuning Steps	Accuracy (%)	Inference Cost (s)
0	3.13	2.19 / 9.17
50	31.4	+ 0.99
100	32.0	+ 1.29
150	31.9	+ 1.13
200	**32.3**	+ 1.09
250	31.8	+ 1.10
300	31.9	+ 1.12

Table 6. Impact of various tuning steps. The preparation of Grad-CAM and initial ellipse resulted in inference times of 2.19s and 9.17s, respectively, with and without parallelization.

The evaluation results for Tune-An-Ellipse at various tuning steps are presented in Table 6. Without any tuning, the initial localization accuracy is notably low at 3.13%. As the tuning steps increase, there is a corresponding improvement in localization accuracy. After surpassing 100 tuning steps, the performance of Tune-An-Ellipse stabilizes, indicating robust and consistent localization results.

Figure 9. Inaccurate selection by CLIP and our tuned results.

Figure 10. Localization results of the proposed Tune-An-Ellipse on RefCOCO, RefCOCO+ and RefCOCOg.

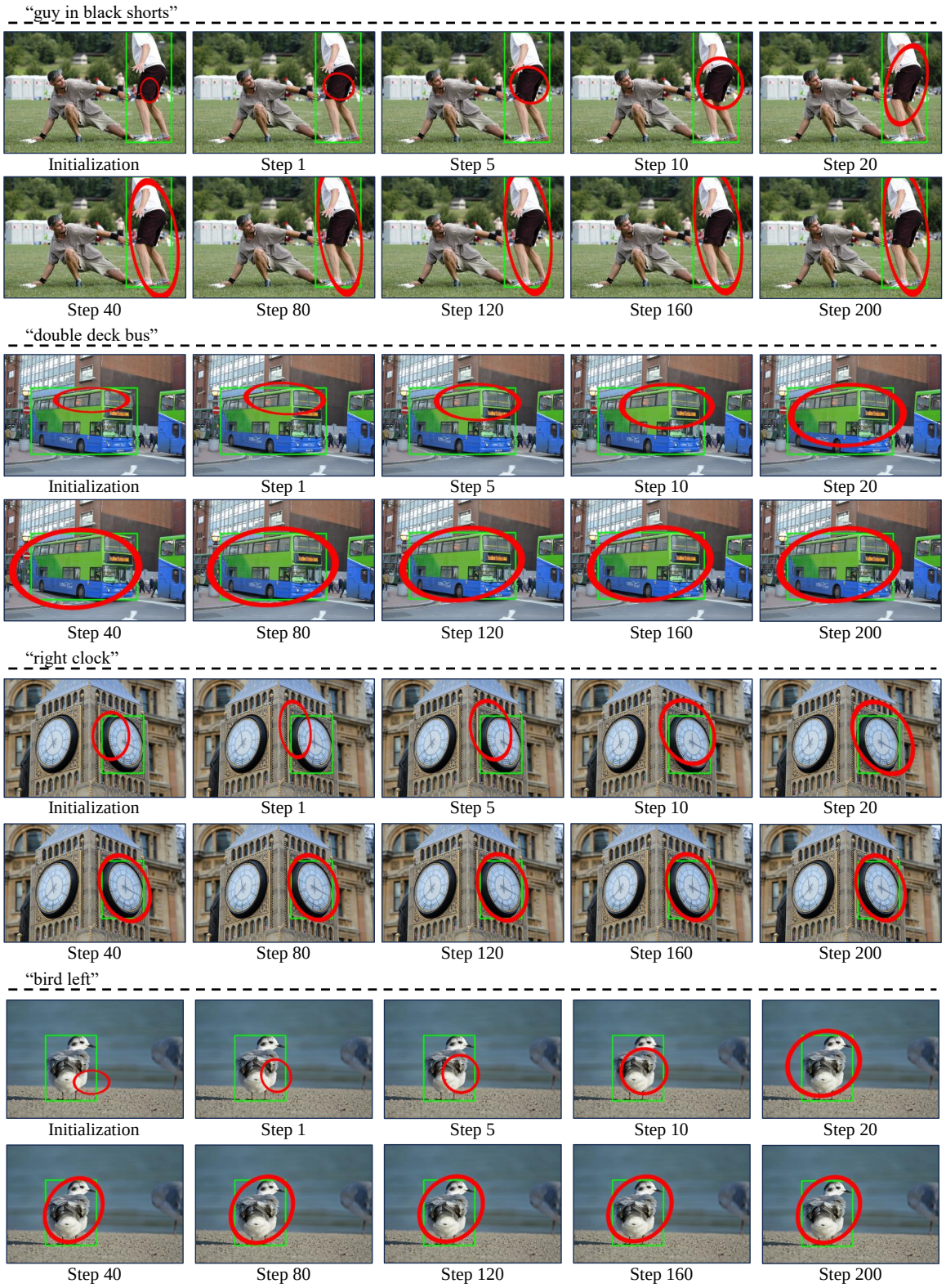

Figure 11. Visualization of the localization results at different tuning steps. Samples are drawn from RefCOCO.

Figure 12. Visualization of the localization results at different tuning steps. Samples are drawn from RefCOCO+.

Figure 13. Visualization of the localization results at different tuning steps. Samples are drawn from RefCOCOg.