

# Supplementary Material: Selective Vision-Language Subspace Projection for Few-shot CLIP

Anonymous Author(s)

## 1 MORE ANALYSIS OF MODALITY GAPS .

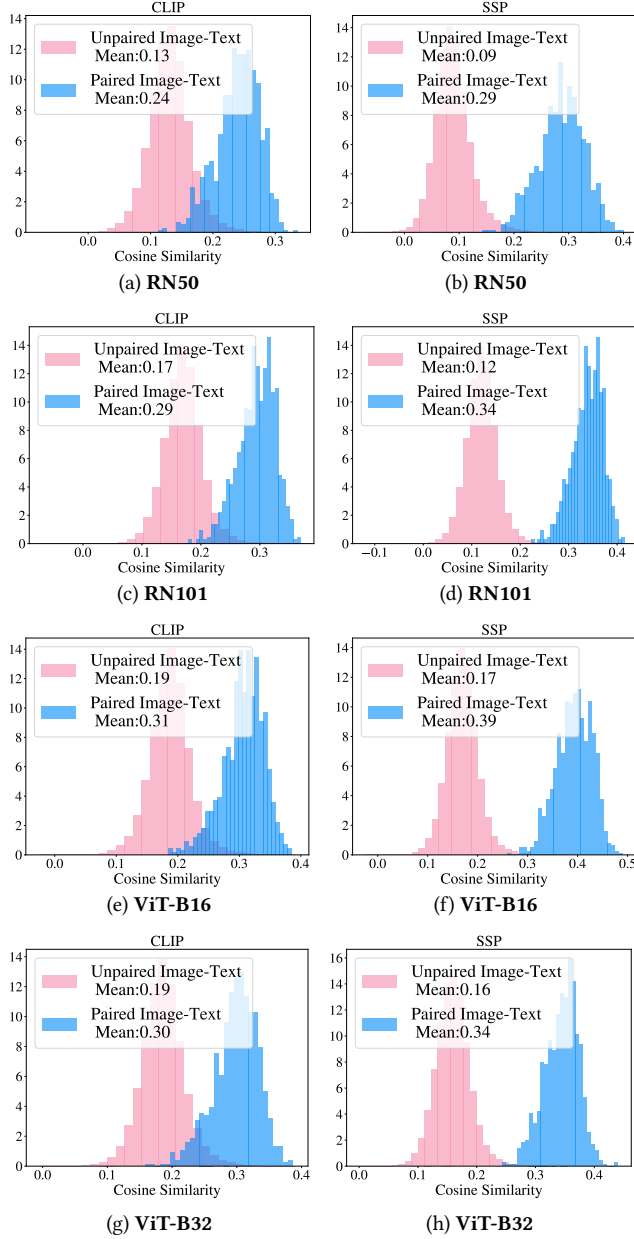


Figure A: Comparison of similarity distributions between text-image features of CLIP and SSP using different encoders on the ImageNet dataset.

To further analyze the effectiveness of our SSP working in narrowing modality gaps, we visualize the similarity distributions for both paired and unpaired image and text features. The comparisons are illustrated in Figure A. It is evident that in the original CLIP feature space, the similarities of paired image-text features only slightly exceed those of unpaired (the first column in Figure A). However, with the application of our SSP (the second column of Figure A), there is a noticeable improvement in reducing the similarities between unpaired image-text features (0.09 vs. 0.13 under RN50) and simultaneously enhancing the similarities between paired features (0.39 VS. 0.31 under ViT-B/16), which has effectively reduce the modality gaps.

## 2 DETAILS OF VON MISES-FISHER (VMF) DISTRIBUTION FORMULATION

As discussed in the previous works [4, 5], text and image features encoded from CLIP are located on a unit hypersphere. Consequently, these features follow a von Mises–Fisher (vMF) distribution [1]. The probability density function of the vMF distribution for the random  $d$ -dimensional unit vector  $f$  is expressed as:

$$p_d(f; \mu, \kappa) = \frac{1}{C_d(\kappa)} \exp(\kappa \mu^T f), \quad (1)$$

$$C_d(\kappa) = \frac{(2\pi)^{d/2} I_{(d/2-1)}(\kappa)}{\kappa^{d/2-1}},$$

where  $\kappa \geq 0$ ,  $\|\mu\|_2 = 1$  and  $I_{(d/2-1)}$  denotes the modified Bessel function of the first kind at order  $I_{(d/2-1)}$  as discussed in [1]. We utilize maximum likelihood estimation to calculate parameters  $\mu$  and  $\kappa$ , where  $\mu$  is the average of all features and  $\kappa$  is approximated as follows [3]:

$$\hat{\kappa} = \frac{\bar{R}(d - \bar{R}^2)}{1 - \bar{R}^2}. \quad (2)$$

Here  $\bar{R}$  is the length of sample mean, i.e.,  $\bar{R} = \|\bar{z}\|_2$ ,  $\bar{z} = \frac{1}{N} \sum_{i=1}^N f_i$ , and we calculate the  $I_{(d/2-1)}(\kappa)$  based on the work [2]. Based on the above estimation, we can calculate metrics of distributions between the image and text features, as discussed in Figure 1 of the manuscript.

## 3 MORE PERFORMANCE COMPARISONS

Our SSP is a training-free method that can also be applied to training-required methods, such as Tip-F [4] and APE-T [5] mentioned in the manuscript. We evaluate our method against Tip-F and APE-T to demonstrate its efficacy. The summarized results of these applications are presented in Table A. We consistently observed that our SSP achieves the highest average accuracy across all benchmarks. However, in specific datasets like Aircraft and Cars under the 1-shot setting, it may not always outperform APE-T. When compared to Tip-F, our SSP demonstrates an average accuracy improvement of 0.68% in the 16-shot setting. Similarly, when

Table A: The classification accuracy (%) comparison on few-shot learning, *i.e.*, 1-/2-/4-/6-/8-/16-shot, across 11 datasets. The results for LFA, Tip, and APE from our implementation by open public project, and the datasets include F102.(Flowers102), Euro(EuroSAT), F101.(Food101), SUN.(SUN397), C101.(Caltech101), UCF.(UFC101), and ImgN.(ImageNet).

Method	Pets	F102.	FGVC	DTD	Euro.	Cars	F101.	SUN.	C101.	UCF.	ImgN.	Avg.
CLIP	85.77	66.14	17.28	42.32	37.56	55.61	77.31	58.52	86.29	61.46	58.18	58.77
<i>16-shot</i>												
Tip-F (ECCV22)	89.70	94.40	35.28	67.71	84.95	76.41	79.33	71.28	92.84	76.74	65.48	75.83
Tip-F + SSP	<b>91.17</b>	<b>95.21</b>	<b>35.58</b>	<b>68.62</b>	<b>85.62</b>	<b>76.70</b>	<b>79.50</b>	<b>71.89</b>	<b>93.27</b>	<b>78.46</b>	<b>65.63</b>	<b>76.51</b> $\uparrow$ 0.68
APE-T (ICCV23)	89.59	96.26	37.38	69.56	86.35	76.15	79.35	72.45	93.06	79.38.	66.12	76.63
APE-T + SSP	<b>90.11</b>	<b>96.43</b>	<b>37.65</b>	<b>70.45</b>	<b>86.75</b>	<b>76.40</b>	<b>79.49</b>	<b>72.65</b>	<b>93.57</b>	<b>79.57</b>	66.12	<b>77.20</b> $\uparrow$ 0.57
<i>8-shot</i>												
Tip-F (ECCV22)	88.20	91.88	30.30	63.12	77.72	69.72	78.59	69.02	92.06	74.62	63.93	72.65
Tip-F + SSP	<b>90.35</b>	<b>93.71</b>	<b>30.57</b>	<b>63.56</b>	<b>79.23</b>	<b>71.68</b>	<b>78.92</b>	<b>69.80</b>	<b>92.48</b>	<b>74.89</b>	<b>64.14</b>	<b>73.58</b> $\uparrow$ 1.04
APE-T (ICCV23)	88.25	94.52	32.25	67.08	81.04	71.72	78.73	70.94	92.13	77.24	64.80	74.43
APE-T + SSP	<b>88.88</b>	<b>94.84</b>	<b>32.52</b>	<b>67.49</b>	<b>81.16</b>	71.19	<b>78.64</b>	<b>71.08</b>	<b>92.45</b>	<b>77.77</b>	<b>64.83</b>	<b>74.62</b> $\uparrow$ 0.20
<i>4-shot</i>												
Tip-F (ECCV22)	87.50	89.40	25.86	57.80	73.43	64.85	78.06	66.32	91.81	71.87	62.53	69.95
Tip-F + SSP	<b>87.77</b>	<b>89.77</b>	<b>26.13</b>	<b>58.45</b>	<b>74.01</b>	<b>66.18</b>	<b>78.36</b>	<b>67.20</b>	<b>92.18</b>	<b>73.13</b>	<b>63.09</b>	<b>70.57</b> $\uparrow$ 0.62
APE-T (ICCV23)	87.98	91.68	26.85	65.66	73.48	67.62	78.27	68.81	91.68	73.78	63.61	71.77
APE-T + SSP	<b>88.23</b>	<b>91.92</b>	<b>27.12</b>	<b>65.78</b>	<b>74.35</b>	<b>68.28</b>	<b>78.30</b>	<b>68.88</b>	<b>91.68</b>	<b>74.76</b>	<b>63.65</b>	<b>72.09</b> $\uparrow$ 0.32
<i>2-shot</i>												
Tip-F (ECCV22)	87.23	83.19	22.92	54.31	66.26	63.00	77.71	63.92	90.30	68.28	61.76	67.17
Tip-F + SSP	<b>87.51</b>	<b>84.41</b>	<b>24.78</b>	<b>54.91</b>	<b>67.64</b>	<b>64.53</b>	<b>77.76</b>	<b>65.94</b>	<b>90.72</b>	<b>70.07</b>	<b>62.53</b>	<b>68.25</b> $\uparrow$ 1.08
APE-T (ICCV23)	87.03	87.41	24.41	59.28	71.89	63.96	77.58	67.14	90.18	69.92	63.27	69.28
APE-T + SSP	<b>87.57</b>	<b>88.75</b>	<b>24.93</b>	59.04	<b>72.37</b>	63.89	<b>77.64</b>	67.00	<b>90.87</b>	<b>69.97</b>	<b>63.28</b>	<b>69.57</b> $\uparrow$ 0.29
<i>1-shot</i>												
Tip-F (ECCV22)	86.41	80.02	20.82	48.52	58.95	58.65	77.50	62.56	90.30	65.01	61.17	64.54
Tip-F + SSP	<b>86.67</b>	<b>81.93</b>	<b>21.15</b>	<b>49.29</b>	<b>61.19</b>	<b>59.91</b>	<b>77.62</b>	<b>62.85</b>	<b>90.56</b>	<b>65.44</b>	<b>61.95</b>	<b>65.32</b> $\uparrow$ 0.79
APE-T (ICCV23)	86.54	83.23	22.08	55.02	67.59	60.65	77.22	65.66	89.82	65.69	62.56	66.91
APE-T + SSP	<b>86.73</b>	<b>84.21</b>	21.63	<b>55.61</b>	<b>68.63</b>	60.55	<b>77.53</b>	<b>65.71</b>	<b>90.43</b>	<b>66.56</b>	<b>62.58</b>	<b>67.29</b> $\uparrow$ 0.37

compared to APE-T, our SSP shows a 0.57% improvement. These consistent enhancements underscore the versatility and effectiveness of our SSP approach.

## REFERENCES

- [1] Chaoqun Du, Yulin Wang, Shiji Song, and Gao Huang. 2024. Probabilistic Contrastive Learning for Long-Tailed Visual Recognition. *CoRR* abs/2403.06726 (2024).
- [2] Wittawat Jitkrittum, Aditya Krishna Menon, Ankit Singh Rawat, and Sanjiv Kumar. 2022. ELM: Embedding and Logit Margins for Long-Tail Learning. *CoRR* abs/2204.13208 (2022).
- [3] Suvrit Sra. 2012. A short note on parameter approximation for von Mises-Fisher distributions: and a fast implementation of  $I_s(x)$ . *Computational Statistics* 27 (2012), 177–190.
- [4] Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. 2022. Tip-Adapter: Training-Free Adaption of CLIP for Few-Shot Classification. In *ECCV* (35), Vol. 13695. 493–510.
- [5] Xiangyang Zhu, Renrui Zhang, Bowei He, Aojun Zhou, Dong Wang, Bin Zhao, and Peng Gao. 2023. Not All Features Matter: Enhancing Few-shot CLIP with Adaptive Prior Refinement. *CoRR* abs/2304.01195 (2023).