Supplementary Materials of Learning-to-Rank Meets Language: Boosting Language-Driven Ordering Alignment for Ordinal Classification

Rui Wang[1], **Peipei Li**[1⋆], **Huaibo Huang**[2], **Chunshui Cao**[3], **Ran He**[2], **Zhaofeng He**[1]
[1]Beijing University of Posts and Telecommunications
[2]CRIPAC&MAIS, Institute of Automation, Chinese Academy of Sciences
[3]WATRIX.AI
{wr_bupt, lipeipei, zhaofenghe}@bupt.edu.cn
huaibo.huang@cripac.ia.ac.cn, chunshui.cao@watrix.ai, rhe@nlpr.ia.ac.cn

1 Appendix

This supplementary material begins with a comprehensive visualization of the datasets central to our study. The specifics of our experimental settings are subsequently outlined in Section 1.2. Section 1.1 features an expanded analysis, including results from ablation studies. A key highlight of this section is the visual interpretation of the CLIP image features facilitated by t-SNE [6]. Concurrently, a comparative analysis is conducted, comparing the efficacy of interpolation-based strategies with our learning-based methods(i.e. L2RCLIP).

1.1 More Analysis of L2RCLIP

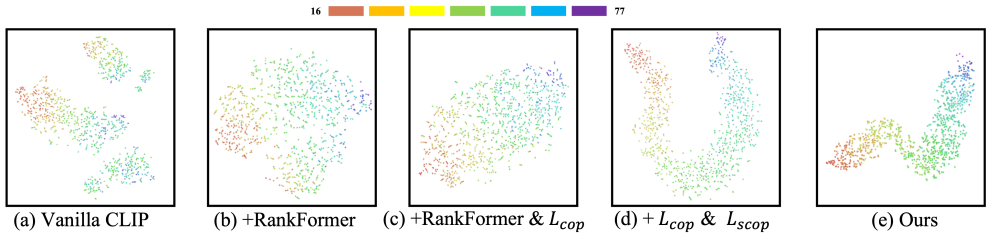

Figure 1: Visualizing the ablation effects on the MORPH II dataset: t-SNE visualizations of 512D spaces in CLIP latent space.

Additional Ablation Study. Figure 1 presents the embedding spaces corresponding to various ablation settings, mapped from the original 512D embedding spaces via t-SNE [6]. In Panel (a), the vanilla CLIP implementation reveals a sub-optimal ordering relationship among images of distinct ranks. Feature demarcations across different ranks are ambiguous, displaying considerable overlap. The incorporation of RankFormer with global context prompts [7], illustrated in Panel (b), aids in improving and consolidating the order alignment of these image features. As shown in Panel (c), the implementation of our proposed loss function, L_{cop}, further enhances this order alignment. Conversely, Panel (d) indicates that without the RankFormer module, the order alignment between the lowest and highest ranks falls short of the desired outcome. These findings substantiate the efficacy of the modules proposed in this study, with each component playing a significant role in the overall model performance. This underscores the criticality of their synergistic implementation.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Table 1: Batchsize Analysis.

Batchsize	16	32	64	96	128
MAE(↓)	2.15	2.13	2.13	2.17	2.17
OS(%, ↑)	67.21	68.55	71.87	71.71	75.42

We also explore the role of different batchsize settings. We report the result in Table 1. According to the result, we choose batchsize with 64 to conduct other experiments.

Table 2: The MAE results under the distribution shift setting and few shot setting on the MOPRH II.

re cls - re smp	10-90	20-80	20-90	30-80	30-90	40-80	40-90
L2RCLIP-I	2.39	2.45	2.50	2.57	2.70	2.73	2.93
L2RCLIP(Ours)	**2.30**	**2.37**	**2.43**	**2.51**	**2.61**	**2.68**	**2.79**
#Shots	#1	#2	#4	#8	#16	#32	#64
L2RCLIP-I	**4.31**	4.02	3.63	3.48	3.13	2.80	2.62
L2RCLIP(ours)	4.54	**3.92**	**3.40**	**3.28**	**2.81**	**2.55**	**2.38**

Compared with Interpolation-based method. In this study, we also propose an interpolation technique for our basic rank templates, which we term L2RCLIP-I. Two distinguishing characteristics set L2RCLIP-I apart from OrdinalCLIP [3]. Firstly, we utilize the ViT-B/16 visual backbone of CLIP for image feature extraction, whereas OrdinalCLIP employs a pre-trained VGG-16 network supplemented by a linear projection layer. Secondly, our method relies on a two-stage training strategy, in contrast to the one-stage approach adopted by OrdinalCLIP. Importantly, both training strategies require a comparable time commitment. The results are reported in Tables **??** and 2.

Local ordinality score of L2RCLIP. To further prove our L2RCLIP have learned better ordering relationship, we follow OrdinalCLIP[3] and use the local ordinality score. The formula is defined as: $LOS(\%) = \sum_{i=1}^{K} \sum_{j=i}^{K} \mathbb{I}\{s_{i,j} > s_{i,j+1}\}/(K \times (K-1)/2)$, where K is the size of local window and $s_{i,j}$ represents the cosine similarity of each pair of templates. We propose that the locally linear manifold can be preserved within a fixed small window size. Therefore, we calculate the local ordinality score using window sizes of 2, 4, 8, 16, and 32. The results of the local ordinality score are shown in Table 3.

Table 3: The local ordinality score results on the MORPH II dataset.

# window size	2	4	8	16	32
Vanilla CLIP	100.00	83.33	78.57	70.83	60.08
OrdinalCLIP[3]	100.00	100.00	100.00	96.19	-
L2RCLIP(Ours)	**100.00**	**100.00**	**100.00**	**100.00**	**97.78**

More ablation study of L2RCLIP To avoid the effect of token mixing and the type of architecture of RankFormer, we conduct more detailed ablation study. The results are shown in Table4. Our proposed method is complementary to previous prompt tuning methods and our L2RCLIP can achieve comparable performance with OrdinalCLIP even without context prompt. We have also compared L2RCLIP performance with an MLP-based architecture to avoid effects driven by extra computation. Note that both RankFormer and MLP have similar training parameters. The results show that our token-wise RankFormer can enhance the ordinality between input rank templates.

More results on Morph II datasets We have conducted experiments on the other three settings of Morph II. The results are presented in Table 5. The details for each settings are as follows:

- Setting A: A total of 5,492 images of Caucasians are sampled and then randomly divided into training and test sets with a ratio of $8:2$.
- Setting B: Approximately 21,000 images of Caucasians and Africans are randomly selected, ensuring a balanced ratio of $1:1$ between Caucasians and Africans, as well as a ratio of

Table 4: Ablation study of global context prompts and architecture.

Method	Morph(MAE)	Morph(OS%)	CLAP2015(MAE)	CLAP2015(OS%)
Vanilla CLIP	6.91	55.36	4.66	52.51
w/o context prompt	2.23	65.46	2.76	67.17
RankFormer→ MLP	2.27	67.48	-	-
L2RCLIP(Ours)	**2.13**	**71.87**	**2.62**	**67.55**

Table 5: Additional results on Morph II.

Methods	Setting A	Setting B	Setting C	Setting D
DRC-ORID[1]	2.26	2.51	2.58	2.16
OL[4]	2.41	2.75	2.68	2.22
MWR-G[5]	2.24	2.55	2.61	2.16
GOL[2]	2.17	2.60	**2.51**	2.09
L2RCLIP(Ours)	**2.13**	**2.53**	2.56	**1.95**

$1:3$ between females and males. The dataset is then divided into three subsets (S1, S2, S3). The training and testing process is repeated twice: 1) training on S1 and testing on S2+S3, and 2) training on S2 and testing on S1+S3.

- Setting C: The entire dataset is randomly partitioned into five folds, with the constraint that images of the same person belong to only one fold. The 5-fold cross-validation is then performed.
- Setting D: The entire dataset is randomly divided into five folds without any restrictions. The 5-fold cross-validation is then performed.

1.2 Experiment settings

Dataset Details. In the scope of this study, we only utilize publicly available data. To provide a comprehensive understanding of the tasks at hand, we illustrate a selection of random samples from the image aesthetics assessment dataset (Figure 2) and the historical image dating dataset (Figure 3). To further enhance our exposition, Figure 4 depicts both the original and adjusted distributions of the MORPH II dataset.

(a) Exceptional. (b) Professional. (c) Ordinary. (d) Flawed. (e) Unacceptable.

Figure 2: Samples from the urban collections of the aesthetics dataset.

(a) 1930s. (a) 1940s. (a) 1950s. (a) 1960s. (a) 1970s.

Figure 3: Samples from the historical image dating dataset.

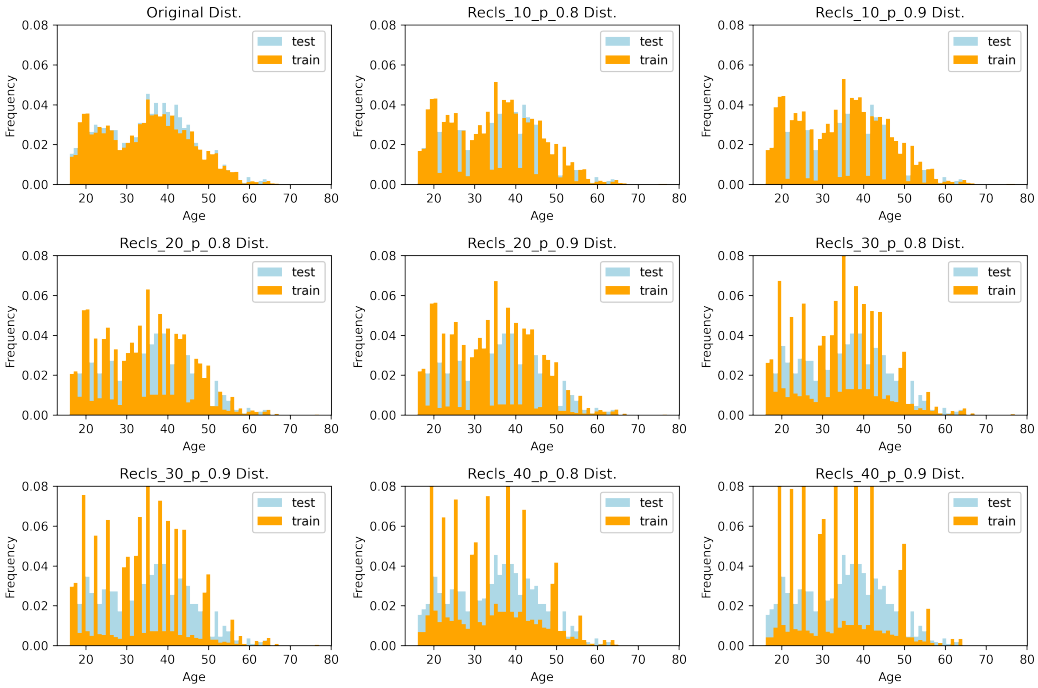

Figure 4: Original and shifted distributions of the MORPH II dataset.

References

[1] Seon-Ho Lee and Chang-Su Kim. Deep repulsive clustering of ordered data based on order-identity decomposition. In *ICLR*, 2021.

[2] Seon-Ho Lee, Nyeong Ho Shin, and Chang-Su Kim. Geometric order learning for rank estimation. In *NeurIPS*, 2022.

[3] Wanhua Li, Xiaoke Huang, Zheng Zhu, Yansong Tang, Xiu Li, Jie Zhou, and Jiwen Lu. Ordinalclip: Learning rank prompts for language-guided ordinal regression. In *NeurIPS*, 2022.

[4] Kyungsun Lim, Nyeong-Ho Shin, Young-Yoon Lee, and Chang-Su Kim. Order learning and its application to age estimation. In *ICLR*, 2020.

[5] Nyeong-Ho Shin, Seon-Ho Lee, and Chang-Su Kim. Moving window regression: a novel approach to ordinal regression. In *CVPR*, 2022.

[6] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 2008.

[7] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 2022.