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

Abstract

We present a novel language-driven ordering alignment method for ordinal classification. The labels in ordinal classification contain additional ordering relations, making them prone to overfitting when relying solely on training data. Recent developments in pre-trained vision-language models inspire us to leverage the rich ordinal priors in human language by converting the original task into a vision-language alignment task. Consequently, we propose L2RCLIP, which fully utilizes the language priors from two perspectives. First, we introduce a complementary prompt tuning technique called RankFormer, designed to enhance the ordering relation of original rank prompts. It employs token-level attention with residual-style prompt blending in the word embedding space. Second, to further incorporate language priors, we revisit the approximate bound optimization of vanilla cross-entropy loss and restructure it within the cross-modal embedding space. Consequently, we propose a cross-modal ordinal pairwise loss to refine the CLIP feature space, where texts and images maintain both semantic alignment and ordering alignment. Extensive experiments on three ordinal classification tasks, including facial age estimation, historical color image (HCI) classification, and aesthetic assessment demonstrate its promising performance. The code is available at https://github.com/raywang335/L2RCLIP.

1 Introduction

Ordinal classification aims to predict labels that are related in a natural or implied order, which can be considered as a special case of ordinal regression after label discretization, i.e. discretize the continuous labels and each bin is then treated as a class. Common examples of such tasks are facial age estimation (e.g., estimating the facial age from 1 to 100), historical color image classification (e.g., assigning a time period to color photographs, ranging from the 1930s to the 1970s), and aesthetics assessment(e.g., rating image quality on a scale from "unacceptable" to "exceptional").

Compared with common classification, ordinal property of labels need to be additionally considered in ordinal classification. Many algorithms [4, 35, 12] employ the ordinal classification framework, which trains a set of classifiers or integrates probabilistic priors to directly predict rank labels. However, these methods exhibit suboptimal performance as a result of insufficiently harnessing the ordering properties. Order learning algorithms [48, 39, 26, 12] demonstrate competitive performance by effectively capturing relative ordering relationships. These approaches determine the target order of a

⋆Corresponding author

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

novel instance by contrasting it with well-defined reference instances. Nonetheless, the algorithm's performance can be substantially compromised by inadequate indexing quality of the reference instances.

Furthermore, many metric learning techniques [3, 41, 13, 49] have been developed to construct an ordinal embedding space in which the distances between diverse features effectively represent the differences in their respective ranks. However, all these methods learn ranking concepts depending solely on training data, which renders them vulnerable to overfitting[29].

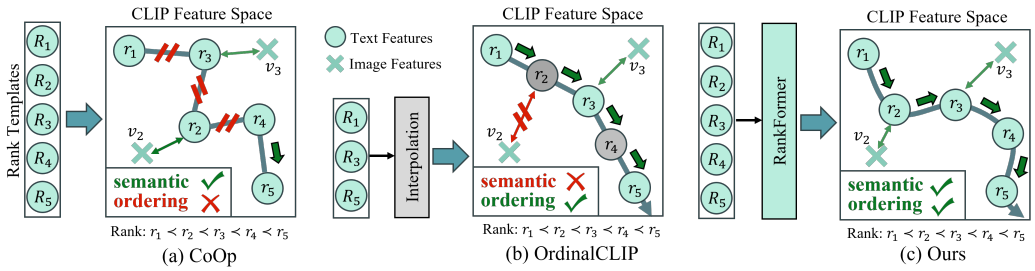

Figure 1: Comparison with CoOp(a) and OrdinalCLIP(b), where v represents the image features and $R; r$ represents the rank templates in word embedding space and CLIP feature space, respectively. (a) CoOp aligns each rank template with its corresponding images via contrastive loss but vanilla CLIP fails to ensure ordering alignment. (b) OrdinalCLIP considers additional interpolation to explicitly maintain ordering alignment. However, the interpolation term can not preserve semantic alignment in CLIP feature space. (c) Our method enhance the ordering relation of vanilla rank templates while ensuring the semantic alignment of CLIP space.

Fortunately, recent developments in large pre-trained vision-language models offer new insights for various visual tasks [32, 27, 29, 25, 56]. Compared with visual content, human language contains highly abstract concepts and rich semantic knowledge [29, 32]. Motivated by it, we attempt to borrow knowledge from language domain by two major observations. Firstly, rank templates inherently contains ordinal information, e.g. *"a sixty years old face" "a ten years old face"*, which is also demonstrated in Figure 3(a). Secondly, inspired by [25, 56], rank features encapsulate the average information of numerous image features within a well-aligned cross-modal feature space, which can be considered as a robust prior for cross-modal metric learning.

Hence, we propose L2RCLIP to boost learning-to-rank of CLIP-based models for ordinal classification. Specifically, we first introduce a complementary prompt tuning method, termed RankFormer, to enhance the ordering relation of original rank templates. Specifically, RankFormer employs a token-wise attention layer and performs residual-style prompt blending with the original templates for prompt tuning. Moreover, inspired by pairwise metric learning [2], we propose cross-modal ordinal pairwise loss to ensure both semantic and ordering alignment in the CLIP feature space. Concretely, we revisit the approximate bound optimization of conventional cross-entropy loss and reformulate it in the cross-modal embedding space with ordinal language priors. OrdinalCLIP [29] also incorporates language priors to model ordering alignment, demonstrating impressive performance. However, our method distinguishes itself from previous works, as depicted in Figure 1. CoOp achieves semantic alignment through contrastive loss but fails to maintain ordering alignment. OrdinalCLIP addresses ordering alignment at the cost of weakened semantic alignment. Conversely, by leveraging RankFormer and cross-modal ordinal pairwise loss, our approach simultaneously considers both semantic alignment and ordering alignment.

The contributions of this paper can be summarized as follows: (1) We incorporate learning-to-rank into vision-language pre-training model for ordinal classification, in which we present RankFormer to enhance the ordering relation of vanilla language prompts. (2) We explicitly utilize the language priors and further propose cross-modal ordinal pairwise loss to refine CLIP embedding space, in which image features and text features maintain both semantic and ordering alignment. (3) Extensive experiments demonstrate the competitive performance of L2RCLIP on age estimation, aesthetics assessment and historical image dating, as well as improvements in few-shot and distribution shift experiments.

2 Related Work

Ordinal Classi cation Ordinal classi cation attempts to solve classi cation problems in which all wrong classes are equally wrong. Early techniques [4, 45] adopted the classi cation framework and train a set of classi ers to directly estimate the rank labels. These methods got degraded performance due to ignoring the ordering relation. By incorporating probabilistic priors, Geng et al. [12] rstly proposed label distribution learning and assigned a Gaussian or Triangle distribution for an instance. The mean-variance loss was introduced [39] for learnable label distribution and penalizes the learned variance of estimated distribution to ensure a sharp distribution. Probabilistic embedding [28] was developed to model the data uncertainty for ordinal regression. Li et al. [33] proposed predicting the ordinal targets that fall within a certain interval with high con dence. These methods learn better rank concepts and signi cantly reduce the model's overcon dence toward incorrect predictions. In contrast to them, our L2RCLIP only focuses on enhancing vision-language alignment without complicated probabilistic distribution assumption.

Furthermore, many techniques [3, 41, 13, 49, 23, 24, 22] solve the ordinal classi cation task from the perspective of metric learning. These methods exploit the pairwise ordering relation in embedding space, where the distance between different features re ects ordinal information. For example, Ordinal log-loss (OLL) was presented in [3] with an additional distance-aware weighting term for ordinal classi cation. RankSim [13] proposed a regularization loss to ensure the ordinal consistency between label and feature. Suárez et al. [49] ranked the embedding distances between pairwise rank differences of features. Another way for ordinal classi cation is order learning [31, 43, 19], which learns ordering relation by comparison between instances. It usually show more promising results as learning relative ordering relation is much easier than learning absolute ordering relation. Lim et al. [31] rstly proposed this concept and determined the ordinal information of an unseen instance by compared to some known reference instances. Lee et al. [19] improved the quality of indexing to boost the performance. MWR [48] further proposed to learn a continuous regression score from reference instance. However, these methods often depend solely on training data to learn rank concepts, potentially leading to over tting. To mitigate these issues, we leverage rich priors in human language to learn language-driven rank concepts.

Vision-language Learning Vision-language pre-training (VLP) has signi cantly improved the performance on many downstream tasks by text and image matching, including segmentation [43, 8], object detection [14], image retrieval [36, 1], generation tasks [18, 40, 25, 5] and ordinal regression [29]. CLIP [42] and ALIGN [15] proposed to embed images and texts into the same representation space with two separate encoders in a contrastive-based approach. The experimental results show that the impressive "zero-shot" performance on downstream tasks, which show the power of language prior. Inspired by the recent advances in NLP, prompts and adapter-based tuning becomes prevalent in improving ability of CLIP. CLIP-Adapter [11] adds a light-weight module on top of image and text encoder. CoOp [55] proposes to learn context prompts for image classi cation. Due to lack of ordinal property, these methods lead to degraded performance in ordinal classi cation. Considering the great potential of language prior, we propose to incorporate learning-to-rank into CLIP for ordinal classi cation.

3 Proposed Approach

3.1 Problem Formulation

Ordinal classi cation is a unique case of image classi cation where labels possess ordinal properties. Mathematically, let $x_i \in \mathcal{X}$ denote the i-th input instance with $i = 1, 2, \ldots, N$, $y_i \in \{r_1, r_2, \ldots, r_M\}$ with ordered ranks $r_M \succ r_{M-1} \succ \cdots \succ r_1$ denote the ground truth value and $\hat{y}_i$ represent the predicted rank by the network, where N represents the total number of instances, M represents the number of ranks, and indicates the ordering between different ranks. Analogous to normal classi cation, ordinal classi cation aims to recover y_i by encoding the image to feature $z_i = \Phi(x_i)$ with encoder Φ and then using a classi er $f_i(\cdot)$ to compute probability p_i. The predicted label $\hat{y}_i$ is the result with the highest probability p_{y_i}. The classi cation probability can be calculated by:

Figure 2: An overview of the proposed L2RCLIP. We incorporate learning-to-rank into CLIP from two perspectives. First, RankFormer performs a token-level attention mechanism to enhance the ordering relation of vanilla rank prompts. Then, re ned rank-speci c prompts and randomly initialized context prompts are concatenated in the word embedding space and are sent to a text encoder to extract the corresponding text features. Moreover, we present two types of losses to re ne CLIP feature space by attraction and repulsion, respectively. Attraction refers to an asymmetrical contrastive loss and a tightness term to attract paired images and text features while repulsion refers to a reweighting diversity term to ensure the ordering alignment.

$$p_i = \frac{\exp(\omega_i^{\top} z_i)}{\sum_{j=1}^{M} \exp(\omega_j^{\top} z_i)}. \tag{1}$$

To exploit ordinal information in language, ordinal classi cation can be transformed into a vision-language alignment task. Speci cally, we use a pre-trained CLIP image feature extractor $\mathrm{Image}(\cdot)$ to extract features from input images $z_i := z_i = \mathrm{Image}(x_i)$. For text features, we rst construct hard rank templates $\mathcal{R} = \{R_1, R_2, \ldots, R_M\}$ for a given ordinal classi cation task. For each template, we convert it into xed-length tokens and then map them into 512-dimensional word embeddings. The language feature extractor $\mathrm{Text}(\cdot)$ encodes the embeddings as a classi er weight. The process can be formulated as $\omega_i = w_i = \mathrm{Text}(\mathrm{Tokenizer}(R_i))$. Finally, we can calculate the prediction probability p_i for rank i with Eq.(1).

3.2 Proposed Method

Our objective is to integrate learning-to-rank into CLIP and learn language-driven rank concepts for ordering alignment while preserving original semantic alignment. Inspired by [16, 27], we rstly learn rank concepts in the word embedding space and subsequently re ne the CLIP feature space to maintain both semantic and ordering alignment. Speci cally, we introduce RankFormer to enhance the ordering relation of the original language prompts. RankFormer employs a token-wise attention layer for rank prompt tuning. To better utilize language prior, we reformulate the approximate bound optimization within cross-modal embedding space. Furthermore, we propose a cross-modal ordinal pairwise loss to ensure ordering alignment. Moreover, randomly initialized global context prompts and an asymmetrical contrastive loss are adopted to ensure semantic alignment. The overall framework is illustrated in Figure 2.

Rank-speci c Prompts. As shown in Figure 3(a), rank templates in the vanilla CLIP contain some degree of ordinal information. However, a substantial proportion, nearly half of the pairs, lack clear ordinality. One intuitive approach to enhance this ordinal information is to further ne-tune these rank templates. However, implementing such a strategy introduces two primary challenges. First, the performance of different ranks for ordinal classi cation varies signi cantly driven by the imbalanced training data, and certain ranks cannot even be trained due to insuf cient training data in extreme cases. Second, the contrastive loss introduces dif culties in enhancing the ordering relation during the training [5, 13]. To address the rst challenge, OrdinalCLIP [29] trains only a small number of base rank templates and generates other rank templates through explicit interpolation, which con nes the ordinal information of vanilla rank templates. In this paper, we train a RankFormer with token-wise attention to enhance the ordinal information of the xed rank templates. Speci cally, we denote tokenized rank templates as $\mathcal{R} \in \mathbb{R}^{M \times n \times D}$, where M and D represent the number of templates and word embedding channels, respectively. Note that we only consider the ranking tokens with length of n. Subsequently, we perform token-wise attention for prompt tuning and apply residual-style prompt blending with the original prompts. The process can be formulated as: $\mathcal{R}^{0} = (1 - \alpha)\,\mathcal{R} + \alpha\, f_{FFN}(f_{MSA}(f_{LN}(\mathcal{R})))$, where α represents the residual ratio.

Cross-modal Ordinal Pairwise Loss. Since cross-entropy loss ignored the ordinal information during training [53], we turn to a solution that recovers the ordering relation while maintaining semantic alignment. Inspired by the pairwise metric learning, we first revisit the vanilla cross-entropy loss from the perspective of approximate bound optimization. Boudiaf et al. [2] proved that minimizing the cross-entropy loss accomplishes by approximating a lower bound pairwise cross-entropy loss L_{PCE}. L_{PCE} contains a tightness term and a diversity term, as follows:

$$L_{PCE} = \underbrace{\frac{1}{2N^2}\sum_{i=1}^{N}\sum_{j:y_i=y_j} z_i^{\top} z_j}_{\text{TIGHTNESS}} + \underbrace{\frac{1}{N}\sum_{i}^{N}\log\sum_{k=1}^{K}\exp\left(\frac{\sum_{j=1}^{N} p_{jk} z_i^{\top} z_j}{N}\right) - \frac{1}{2}\sum_{k=1}^{K}\|c_k\|}_{\text{DIVERSITY}}; \quad (2)$$

where z_i represents the image feature in the embedding space, p_{ik} represents the softmax probability of point z_i belonging to class k, $c_k = \sum_{i=1}^{N} p_{ik} z_i$ represents the soft mean of class k and $\frac{1}{2}\|c_k\|$ is to make sure L_{CE} is a convex function with respect to encoder w.

By incorporating the language priors, we turn Eq.(2) to a cross-modal pairwise cross-entropy loss. Specifically, as human language contains rich prior knowledge, text features can be considered as both hard mean $r_k = \frac{1}{N_k}\sum_{i=1}^{N_k} v_k$ and soft mean $r_k = \frac{1}{N}\sum_{i=1}^{N} p_{ik} v_i$ of image features at class k, where N_k represents the sample number of class k and N represents the total sample number. Then, we reformulate the original pairwise cross-entropy loss L_{PCE} as cross-modal pairwise cross-entropy loss L_{CPCE}:

$$L_{CPCE} = \underbrace{\frac{1}{2N}\sum_{i=1}^{N} v_i^{\top} r_{y_i}}_{\text{TIGHTNESS}} + \underbrace{\frac{1}{N}\sum_{i}^{N}\log\sum_{k=1}^{K}\exp\left(v_i^{\top} r_k\right) - \frac{1}{2}\sum_{k=1}^{K}\|r_k\|}_{\text{DIVERSITY}}; \quad (3)$$

Inspired by [53], we use meanNN entropy estimator [10] to estimate the diversity term in Eq.(3). The process is formulated as:

$$L_{CPCE}^{\text{diversity}} \propto \frac{D}{N(N-1)}\sum_{i=1}^{N}\sum_{j\neq i}^{N}\log(v_i^{\top} r_j + r_i^{\top} r_j); \quad (4)$$

$$L_{CPCE}^{\text{tightness}} \propto \frac{1}{N}\sum_{i=1}^{N} v_i^{\top} r_{y_i}; \quad (5)$$

where D represents the feature dimensions. To recover ordering relation, we propose an additional weighting term. Intuitively, each rank template will have a high similarity score with images of a close rank and a low similarity score with images of a distant rank. As such, we opt to weight the cross-modal features with w_{ij}, where w_{ij} are the distances in the label space. The final cross-modal ordinal pairwise loss is defined as follows:

$$L_{cop}(y_i) = \frac{1}{(B-1)}\sum_{j\neq i}^{B} w_{y_i j}\left[(v_{y_i} + \alpha r_{y_i})^{\top} r_j - v_{y_i}^{\top} r_{y_i}\right]; \quad (6)$$

where α controls the strength of the rank within rank templates and B represents the batchsize. To further refining the CLIP feature space, we also propose a simplified cross-modal ordinal pairwise loss L_{scop} with language-related parameters frozen (i.e. $\alpha = 0$).

Global Context Prompts. Given that global context prompts significantly surpass manually designed discrete prompts in vision tasks [55, 27, 54], we integrate them with our complementary rank-specific prompts in RankFormer to enhance semantic alignment. Specifically, we randomly initialize L global context prompts, denoted as $G = \{G_1, G_2, \ldots, G_L\}$, and concatenate them with rank-specific prompts in the word embedding space of CLIP.

Asymmetrical Contrastive Loss. In vanilla CLIP [42], models are optimized using the standard contrastive loss, including a text-image contrastive loss L_{t2i} and an image-text contrastive loss L_{i2t}. In this work, we replace the original contrastive loss with an asymmetrical contrastive loss due to

many-to-many image-text mappings within a batch. In other words, different images in a batch may have the same rank, and the same rank may have the same description. Therefore, we replace t_{y_i} in both text-image contrastive loss L_{t2i} and image-text contrastive loss L_{i2t}. Speci cally, our improved asymmetrical contrastive loss is de ned as follows:

$$L_{t2i}(y_i) = \frac{1}{|Z(y_i)|} \sum_{z \in Z(y_i)} \log \frac{\exp(v_z^{\top} r_{y_i} / \tau)}{\sum_{j=1}^{B} \exp(v_j^{\top} r_{y_i} / \tau)}; \tag{7}$$

where τ is the temperature parameter and $Z(y_i) = \{z \in \{1, \dots, B\} : y_z = y_i\}$.

3.3 Loss functions

We employ a two-stage training scheme to re ne our approach following [27]. In the rst stage, our focus is to enhance the ordinal information of the vanilla rank templates. We utilize cross-modal ordinal pairwise loss L_{cop} and the asymmetrical contrastive losses L_{t2i} and L_{i2t}. In the second stage, our objective is to simultaneously improve both semantic alignment and ordering alignment within the CLIP latent space. We use the simpli ed cross-modal ordinal pairwise loss L_{scop} and the cross-entropy loss L_{ce}. More detail can be found in Sec. 4.1 .

4 Experiments

4.1 Implementation details

We adopt the ViT-B/16 visual encoder and the text encoder from CLIP [42] as the backbone for our image and text feature extractor. All training data is resized to 224×224 and subjected to random horizontal ipping. For all experiments, we employ the AdamW [17] optimizer with default settings. The learning rates for the RankFormer and the visual backbone are 3.5×10^{-4} and 1×10^{-5}, respectively. The model is trained for 20 epochs in the rst stage and 40 epochs in the second stage, with the learning rate decayed by a factor of 0.1 in epoch 30. We set $K = 5$ in most of our experiments. For MORPH and CLAP2015, we set L_{t2i}, L_{i2t}, and L_{cop} with weights of 0.03, 0.03, and 3 for the rst stage, and L_{ce} and L_{scop} with weights of 1 and 1 for the second stage. For other datasets, we set L_{t2i}, L_{i2t}, and L_{cop} with weights of 0.1, 0.1, and 3 for the rst stage, and L_{ce} and L_{scop} with weights of 1 and 1 for the second stage. All experiments are conducted on a single NVIDIA 3090 GPU.

4.2 Age Estimation

Table 1: Results on MORPH II and CLAP2015.

Methods	Morph	CLAP2015
	MAE(#)	MAE(#)
AGEn [50]	2.52	2.94
BridgeNet [30]	2.38	2.87
AVDL [52]	2.37	-
POE [28]	2.35	-
DRC-ORID [19]	2.26	-
PML [6]	2.15	2.91
MWR [48]	2.13	2.77
Vanilla CLIP [42]	6.91	4.66
CoOp [55]	2.39	2.75
OridinalCLIP [29]	2.32	
L2RCLIP(Ours)	2.13	2.62

Table 2: Results on Adience dataset.

Methods	Adience	
	Accuracy(%;")	MAE(#)
OR-CNN [37]	56.7	0.54
CNNPOR [34]	57.4 ± 5.8	0.55 ± 0.08
GP-DNNOR [35]	57.4 ± 5.5	0.54 ± 0.07
SORD [7]	59.6 ± 3.6	0.49 ± 0.05
POE [28]	60.5 ± 4.4	0.47 ± 0.06
MWR [48]	62.6	0.45
GOL [20]	62.5	0.43
Vanilla CLIP [42]	43.3 ± 3.6	0.80 ± 0.02
CoOp [55]	60.6 ± 5.5	0.50 ± 0.08
OridinalCLIP [29]	61.2 ± 4.2	0.47 ± 0.06
L2RCLIP(Ours)	66.2 ± 4.4	0.36 ± 0.05

Datasets. Age estimation aims to predict the age of a given facial image. We train and evaluate our method on the widely-used MORPH II [44] dataset, CLAP2015 [9] dataset, and Adience [21] dataset. MORPH II [44] is one of the largest and most commonly used longitudinal face databases and we follow the widely adopted training and evaluation protocols [45, 28, 48, 47]. CLAP2015 [9] is used for apparent age estimation, where each image is rated by at least 10 annotators, and the mean rating

is set as the ground truth. Following [48], we split it into 2,476 for training, 1,136 for validation, and 1,079 for testing. Adience [21] contains discrete labels annotated with eight age groups. We adopt the same five-fold cross-validation protocol used in [21]. For evaluation metric, we use the mean average error (MAE) to measure the absolute differences between the ground truth labels and the predicted ones. Classification accuracy is additionally adopted for Adience. For detailed experimental settings, please refer to the supplementary material.

Comparison with State-of-the-art Methods. In Table 1, our L2RCLIP outperforms both conventional algorithms and language-powered algorithms in all tests. For results on the Morph II, compared with conventional algorithms, L2RCLIP achieves state-of-the-art performance on MAE at 2.13, verifying the significance of leveraging rich priors in human language. Furthermore, compared to methods utilizing language priors, our L2RCLIP exhibits substantial performance improvements. CoOp significantly enhances the vanilla CLIP's performance by learning global context prompts for semantic alignment. Subsequently, OrdinalCLIP further reduces MAE by 0.07 through fixed interpolation for ordering alignment. Nonetheless, a considerable margin remains compared to our methods, which validates the effectiveness of our approach.

Table 1 also compares the results on CLAP2015. Due to the great challenge, many previous methods adopt additional boosting schemes [30, 50]. However, without using such schemes, L2RCLIP outperforms all conventional algorithms. Compared with language-guided models, our L2RCLIP also shows much better results on MAE, specifically a significant MAE margin of 0.15 evaluated on the test set. Table 2 shows the comparison results on Adience [21] using the metrics of MAE and Accuracy. Compared with Morph II and CLAP2015, Adience is used for age group estimation. The underline indicates the mean value reported in the original paper. Our method outperforms the state-of-the-art algorithms by significant gaps of 5.7% in accuracy and 0.07 in MAE.

Figure 3: The similarity matrices of rank templates in different methods. The redder, the more similar the pair of rank templates. The percentages of templates pairs that obey the ordinality are: 55.36%, 59.92%, 65.94%, and 71.87%, respectively.

Ordinality of Learned Rank Templates. Following [29], we report the ordinality score by measuring the cosine similarity between rank templates. The ordinality score is the percentage of rank template pairs that obey the ordinal property. Figure 3 reports the ordinality scores in different methods. Vanilla CLIP in Figure 3(a) contains a certain degree of ordering relations, where more than half of the rank template pairs show correct ordering relation. Moreover, CoOp in Figure 3(b) introduces additional global context prompts and improves the ordinality score by 4.56%. OrdinalCLIP in Figure 3(c) adopts an explicit interpolation strategy and further improves the ordinality by 6.02%. However, there is a noticeable "red stripe" in the upper right corner, indicating that this section of rank template pairs significantly violates ordinality. We believe that explicit interpolation effectively ensures local ordinal properties, but may fail to preserve global ordinal properties. Our methods greatly alleviate this problem and achieve a higher ordinality score. The "red blob" in the lower right corner results from insufficient samples for old individuals. Therefore, to avoid the model's overconfidence towards incorrect predictions, this part of rank templates should exhibit higher similarity while ensuring ordering alignment. By comparison, our method outperforms previous methods.

Few-shot Learning. Table 4 demonstrates the generalization ability of L2RCLIP for few-shot learning. Following [29], the full dataset is split into 80% for training and 20% for testing. The entire test set is used for validation, and only 1/2/4/8/16/32/64 samples in the training set from each class of labels are chosen for training. We observe that introducing rank information in CLIP

i.e., $OS(\%) = \sum_{i=1}^{M} \sum_{j=i}^{M-1} \mathbb{1}(s_{i,j} > s_{i,j+1}) / ((M-1)M/2)$, where M is the number of templates and $s_{i,j}$ represents the cosine similarity of each pair of templates.

Table 3: We report the MAE results under few-shot settings on the MOPRH II dataset.

# Shots	1	2	4	8	16	32	64
CoOp [55]	5.09	4.50	3.81	3.57	3.23	2.87	2.61
OrdinalCLIP [29]	4.94	4.36	3.55	3.31	3.07	2.76	2.57
L2RCLIP(Ours)	4.54	3.92	3.40	3.28	2.81	2.55	2.38

actually bene ts few-shot learning tasks. Compared with CoOp, both OrdinalCLIP and our L2RCLIP achieve signi cant improvements across all settings, which veri es the effectiveness of ordering alignment. Moreover, our L2RCLIP further reduces the average MAE by 0.24, which demonstrates the effectiveness of the proposed learning-to-rank method in CLIP.

Table 4: The MAE results under the distribution shift setting on the MOPRH II. “re cls” denotes the number of reduced classes, and “re smp” means the percentage of reduced sampled in one class.

re cls - re smp	10-80	10-90	20-80	20-90	30-80	30-90	40-80	40-90
CoOp [55]	2.71	2.85	2.98	3.51	3.06	3.36	2.99	3.30
OrdinalCLIP [29]	2.61	2.67	2.77	3.06	2.86	3.21	2.84	3.12
L2RCLIP(Ours)	2.28	2.30	2.37	2.43	2.51	2.61	2.68	2.79

Distribution Shift. Following [29], we conduct data distribution shift experiments on the MORPH II dataset for generalization. We use the same setting as the general regression setting. For the training set, we randomly choose several rank labels, e.g., 10, 20, 30, and 40. Then, in those classes, we randomly discard some portion of training data, e.g., 80 and 90. We report our experiments in Table 4. For the most severe settings, CoOp, OrdinalCLIP, and our L2RCLIP methods incur performance losses of 42.24%, 34.49%, and 30.98%, respectively, which demonstrates that our approach exhibits superior robustness when faced with data distribution shift problems. For all shift settings, our method shows better performance than OrdinalCLIP, which highlights the effectiveness of the proposed learning-to-rank method in CLIP compared with interpolation.

Table 5: Ablation Study of L2RCLIP on the MORPH II and CLAP2015.

Ablation		Choices							
RankFormer			"				"	"	"
L_{cop}				"		"	"		"
L_{scop}					"	"		"	"
Morph II	MAE(#)	2.50	2.48	2.37	2.38	2.21	2.29	2.31	2.13
	OS(%,")	55.89	57.58	66.15	55.89	66.15	71.87	57.58	71.87
CLAP2015	MAE(#)	2.75	2.68	2.71	2.71	2.70	2.65	2.66	2.62
	OS(%,")	53.33	54.77	57.33	53.33	57.33	67.55	55.25	67.55

4.3 Analysis

Ablation Study. In this section, we conduct an ablation study to examine the respective roles of each component in L2RCLIP, and the results are reported in Table 5. Three key observations can be drawn from Table 5. First, each proposed component demonstrates improvements over the baseline model and complements one another. Second, although RankFormer marginally enhances the ordinal information of vanilla rank templates, its performance remains unsatisfactory compared to other methods when only contrastive loss is utilized for training. This result further highlights the importance of our cross-modal pairwise loss in re ning the CLIP feature space. Third, ne-tuning rank templates without RankFormer signi cantly impairs performance on CLAP2015, potentially due to imbalanced training for certain rank templates. This nding underscores the effectiveness of RankFormer in modeling the ordering relation derived from the complete set of rank templates.

Initialization Impact. As we aim to enhance the ordering relation vanilla rank templates, it is essential to examine the effects of initialization. We report our results in Table 6. Although our

Table 6: Initialization Impact Analysis.

Initialization	MAE(#)	OS(%; ")
A photo of {age} years old face.	2.13	71.87
Age estimation: a person at the age of {age}.	2.14	72.07
The age of the person in the portrait is {age}.	2.12	71.32
The age of the person is {age}.	2.16	71.52
The age of the face is {age}.	2.14	71.08
Mean/Std	2.14/0.01	71.57/0.40

methods x the original rank templates, different initializations lead to similar convergence and performance with a low standard deviation value of 0.014. This observation further substantiates the robustness of our methods against diverse initializations.

Table 7: Results on Morph, CLAP2015 and Adience datasets.

Methods	Morph (MAE, #)	CLAP2015 (MAE, #)	Adience (Accuracy,")		Adience (MAE, #)	
L2RCLIP-I	2.19	2.78	62.9	5.5	0.42	0.06
L2RCLIP (Ours)	2.13	2.62	68.2	7.2	0.36	0.05

Compared with interpolation-based method. To further prove the effectiveness of our proposed method, we compare our L2RCLIP with previous interpolation-based method (e.g. OrdinalCLIP[29]). We adopt the same setting except the process of ordinality learning and we term it as L2RCLIP-I. We report the results on the aging dataset. More experiments for few-shot learning and distribution shift and the implementation details setting for L2RCLIP-I can be found in appendix.

As illustrated in Table 7, our method outperforms interpolation-based methods with a signi cant margin in experiments involving a large number of rank categories. This outcome is attributable to the challenge posed by direct interpolation methods in modelling complex ordering relationships. Our approach continues to surpass interpolation-based methods even in experiments with a smaller number of rank categories. Collectively, these experiments corroborate the effectiveness of the methods proposed in this study.

4.4 Image Aesthetics Assessment

Datasets. CrowdBeauty [46] consists of 13,929 available Flickr photos across four categories: nature, animal, urban, and people. The aesthetic quality of each image is evaluated using ve absolute rating scales: "unacceptable", " awed", "ordinary", "professional", and "exceptional". Following previous methods [19, 20], we select 80% of the images for training and the rest for testing. Five-fold cross-validation is employed for fair comparisons. Both the mean MAE and accuracy are reported.

Table 8: Results on Image Aesthetics dataset.

Methods	Accuracy(%; ")					MAE(#)				
	Nature	Animal	Urban	People	Overall	Nature	Animal	Urban	People	Overall
CNNPOR [34]	71.86	69.32	69.09	69.94	70.05	0.294	0.322	0.325	0.321	0.316
SORD [7]	73.59	70.29	73.25	70.59	72.03	0.271	0.308	0.276	0.309	0.290
POE [28]	73.62	71.14	72.78	72.22	72.44	0.273	0.299	0.281	0.293	0.287
GOL [20]	73.8	72.4	74.2	69.6	72.7	0.27	0.28	0.26	0.31	0.28
Vanilla CLIP [42]	65.24	45.67	58.78	53.06	55.68	0.461	0.557	0.468	0.524	0.502
CoOp [55]	72.74	71.46	72.14	69.34	71.42	0.285	0.298	0.294	0.330	0.302
OridinalCLIP [29]	73.65	72.85	73.20	72.50	73.05	0.273	0.279	0.277	0.291	0.280
L2RCLIP(Ours)	73.51	75.26	77.76	78.69	76.07	0.267	0.253	0.216	0.246	0.245

Results. Table 8 presents our results on the CrowdBeauty dataset. Aesthetic score prediction is challenging due to the subjectivity and ambiguity of aesthetic criteria; however, our L2RCLIP demonstrates state-of-the-art performance on most experimental settings by fully exploring learning-to-rank with language prior. Compared with OrdinalCLIP, L2RCLIP improves accuracy by 3.02% and

reduces MAE by 0.35 overall, which further veri es the effectiveness of our proposed learning-to-rank method. When compared with the best approach without language prior, L2RCLIP improves overall MAE by 0.35 and overall accuracy by 3.37%. Consistent improvements are observed across all categories compared to previous methods, showcasing the effectiveness of our proposal in exploiting language ordinal information.

4.5 Historical Image Dating

Datasets. The historical image dating dataset [38] serves as a benchmark for automatically predicting the decade of historical colored images. The dataset comprises ve-decade categories, ranging from the 1930s to the 1970s. Following [34, 38], we adopt the same train-test split and ten-fold cross-validation. Both the mean and standard deviation for MAE and accuracy metrics are reported.

Results. Table 9 showcases improvements compared to other state-of-the-art models using the full dataset. Initially, zero-shot CLIP exhibits suboptimal performance resulting from inadequate semantic and ordinal alignment. CoOp enhances the results by incorporating global context prompts to facilitate semantic alignment. Furthermore, OrdinalCLIP exploits interpolation to improve the average MAE by 0.09. Our L2RCLIP further advances the average MAE by 0.25 with a lower standard deviation. In comparison with the previous conventional model, L2RCLIP achieves a new state-of-the-art performance with an MAE of 0.43 and an accuracy of 67.22%, thereby validating the effectiveness of our proposed method.

Table 9: Results on HCI dataset.

Methods	HCI			
	MAE(#)		Accuracy(%;")	
CNNPOR [34]	0.82	0.05	50.12	2.65
GP-DNNOR [35]	0.76	0.05	46.60	2.98
POE [28]	0.76	0.04	54.68	3.21
MWR [48]	0.58		57.8	
GOL [20]	0.55		56.2	
Vanilla CLIP [42]	1.01	0.03	30.41	3.32
CoOp [55]	0.76	0.06	51.94	2.60
OridinalCLIP [29]	0.67	0.03	56.44	1.66
L2RCLIP(Ours)	0.43	0.02	67.22	1.59

5 Discussions and Conclusions

In this paper, we propose L2RCLIP to boost learning-to-rank of CLIP for ordinal classi cation. Speci cally, we introduce a complementary prompt tuning method, termed RankFormer, to enhance the ordering relation of the original rank prompts. It performs token-level attention and residual-style prompt blending for prompt tuning. Additionally, we revisit the approximate bound optimization of cross-entropy and reformulate it in the cross-modal embedding space by incorporating the language knowledge. To additionally recover the ordinal information, we further introduce cross-modal ordinal pairwise loss to re ne the ordering alignment of CLIP feature space. Extensive experiments demonstrate the effectiveness of our approach on various ordinal classi cation tasks, including facial age estimation, historical image dating, and image aesthetics assessment. Furthermore, L2RCLIP outperforms in the challenging few-shot learning and data distribution shift learning scenarios. Lastly, we conduct a comprehensive ablation study to verify the effectiveness of each component of our proposed method.

Broader Impacts. As a versatile approach, L2RCLIP can be applied to any ordinal classi cation tasks such as image aesthetics assessment or other rank assessment. However, these tasks may pose a risk of unlawful surveillance or invasion of privacy if abused. Meanwhile, as L2RCLIP is based on a large-scale vision-language model, addressing demographic biases in pre-trained vision-language models is of signi cant importance. Therefore, we emphasize that L2RCLIP represents a research proof of language-driven learning and is not appropriate for real-world usage without strict technical controls.

Acknowledgement. This research is sponsored by National Natural Science Foundation of China (Grant No. 62306041, 62006228), Beijing Nova Program (Grant No. Z211100002121106, 20230484488, 20230484276), and Youth Innovation Promotion Association CAS (Grant No.2022132).

References

[1] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Effective conditioned and composed image retrieval combining clip-based features. In CVPR, 2022.

[2] Malik Boudiaf, Jérôme Rony, Imtiaz Masud Ziko, Eric Granger, Marco Pedersoli, Pablo Piantanida, and Ismail Ben Ayed. A unifying mutual information view of metric learning: cross-entropy vs. pairwise losses. In ECCV, 2020.

[3] François Castagnos, Martin Mihelich, and Charles Dognin. A simple log-based loss function for ordinal text classi cation. In COLING, 2022.

[4] Shixing Chen, Caojin Zhang, Ming Dong, Jialiang Le, and Mike Rao. Using ranking-cnn for age estimation. In CVPR, 2017.

[5] Xing Cui, Zekun Li, Peipei Li, Yibo Hu, Hailin Shi, and Zhaofeng He. I2edit: Towards multi-turn interactive image editing via dialogue. 2023.

[6] Zongyong Deng, Hao Liu, Yaoxing Wang, Chenyang Wang, Zekuan Yu, and Xuehong Sun. Pml: Progressive margin loss for long-tailed age classi cation. In IJCNN, 2021.

[7] Raul Diaz and Amit Marathe. Soft labels for ordinal regression. In CVPR, 2019.

[8] Xiaoyi Dong, Yinglin Zheng, Jianmin Bao, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, et al. Maskclip: Masked self-distillation advances contrastive language-image pretraining. arXiv preprint arXiv:2208.12262, 2022.

[9] Sergio Escalera, Junior Fabian, Pablo Pardo, Xavier Baró, Jordi Gonzalez, Hugo J Escalante, Dusan Misevic, Ulrich Steiner, and Isabelle Guyon. Chalearn looking at people 2015: Apparent age and cultural event recognition datasets and results. In ICCVW, 2015.

[10] Lev Faivishevsky and Jacob Goldberger. Ica based on a smooth estimation of the differential entropy. In NeurIPS, 2008.

[11] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. arXiv preprint arXiv:2110.04544, 2021.

[12] Xin Geng, Chao Yin, and Zhi-Hua Zhou. Facial age estimation by learning from label distributions. TPAMI, 2013.

[13] Yu Gong, Greg Mori, and Frederick Tung. Ranksim: Ranking similarity regularization for deep imbalanced regression. In ICML, 2022.

[14] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. arXiv preprint arXiv:2104.13921, 2021.

[15] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In ICML, 2021.

[16] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classi er for long-tailed recognition. arXiv preprint arXiv:1910.09217, 2020.

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

[18] Gihyun Kwon and Jong Chul Ye. Clipstyler: Image style transfer with a single text condition. CVPR, 2022.

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

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

[21] Gil Levi and Tal Hassncer. Age and gender classi cation using convolutional neural networks. CVPRW, 2015.

[22] Peipei Li, Yibo Hu, Ran He, and Zhenan Sun. Global and local consistent wavelet-domain age synthesis. TIFS, 14(11):2943–2957, 2019.

[23] Peipei Li, Yibo Hu, Xiang Wu, Ran He, and Zhenan Sun. Deep label re nement for age estimation. PR, 2020.

[24] Peipei Li, Huaibo Huang, Yibo Hu, Xiang Wu, Ran He, and Zhenan Sun. Hierarchical face aging through disentangled latent characteristics. ECCV, 2020.

[25] Peipei Li, Rui Wang, Huaibo Huang, Ran He, and Zhaofeng He. Pluralistic aging diffusion autoencoder. In ICCV, 2023.

[26] Qiang Li, Jingjing Wang, Zhaoliang Yao, Yachun Li, Pengju Yang, Jingwei Yan, Chunmao Wang, and Shiliang Pu. Unimodal-concentrated loss: Fully adaptive label distribution learning for ordinal regression. In CVPR, 2022.

[27] Siyuan Li, Li Sun, and Qingli Li. Clip-reid: Exploiting vision-language model for image re-identi cation without concrete text labels. arXiv preprint arXiv:2211.13977, 2022.

[28] Wanhua Li, Xiaoke Huang, Jiwen Lu, Jianjiang Feng, and Jie Zhou. Learning probabilistic ordinal embeddings for uncertainty-aware regression. In CVPR, 2021.

[29] 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.

[30] Wanhua Li, Jiwen Lu, Jianjiang Feng, Chunjing Xu, Jie Zhou, and Qi Tian. Bridgenet: A continuity-aware probabilistic network for age estimation. In CVPR, 2019.

[31] Kyungsun Lim, Nyeong-Ho Shin, Young-Yoon Lee, and Chang-Su Kim. Order learning and its application to age estimation. In *ICLR*, 2020.
[32] Zhiqiu Lin, Samuel Yu, Zhiyi Kuang, Deepak Pathak, and Deva Ramanan. Multimodality helps unimodality: Cross-modal few-shot learning with multimodal models. In *CVPR*, 2023.
[33] Liang Liu, Hao Lu, Haipeng Xiong, Ke Xian, Zhiguo Cao, and Chunhua Shen. Counting objects by blockwise classi cation. *TCSVT*, 2019.
[34] Yanzhu Liu, Adams Wai Kin Kong, and Chi Keong Goh. A constrained deep neural network for ordinal regression. In *CVPR*, 2018.
[35] Yanzhu Liu, Fan Wang, and Adams Wai Kin Kong. Probabilistic deep ordinal regression based on gaussian processes. In *ECCV*, 2019.
[36] Haoyu Ma, Handong Zhao, Zhe Lin, Ajinkya Kale, Zhangyang Wang, Tong Yu, Jiuxiang Gu, Sunav Choudhary, and Xiaohui Xie. Ei-clip: Entity-aware interventional contrastive learning for e-commerce cross-modal retrieval. In *CVPR*, 2022.
[37] Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Ordinal regression with multiple output cnn for age estimation. In *CVPR*, 2016.
[38] Frank Palermo, James Hays, and Alexei A Efros. Dating historical color images. In *ECCV*, 2012.
[39] Hongyu Pan, Hu Han, Shiguang Shan, and Xilin Chen. Mean-variance loss for deep age estimation from a face. In *CVPR*, 2018.
[40] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *ICCV*, 2021.
[41] Gorkem Polat, Ilkay Ergenc, Haluk Tarik Kani, Yesim Ozen Alahdab, Ozlen Atug, and Alptekin Temizel. Class distance weighted cross-entropy loss for ulcerative colitis severity estimation. *arXiv preprint arXiv:2202.05167*, 2022.
[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.
[43] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *CVPR*, 2022.
[44] Karl Ricanek and Tamirat Tesafaye. Morph: A longitudinal image database of normal adult age-progression. In *FGR*, 2006.
[45] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Dex: Deep expectation of apparent age from a single image. In *ICCVW*, 2015.
[46] Rossano Schifanella, Miriam Redi, and Luca Maria Aiello. An image is worth more than a thousand favorites: Surfacing the hidden beauty of ickr pictures. In *AAAI*, 2015.
[47] Wei Shen, Yilu Guo, Yan Wang, Kai Zhao, Bo Wang, and Alan L Yuille. Deep regression forests for age estimation. In *CVPR*, 2018.
[48] Nyeong-Ho Shin, Seon-Ho Lee, and Chang-Su Kim. Moving window regression: a novel approach to ordinal regression. In *CVPR*, 2022.
[49] Juan Luis Suárez, Salvador García, and Francisco Herrera. Ordinal regression with explainable distance metric learning based on ordered sequences: Extended abstract. In *IJCAI*, 2021.
[50] Zichang Tan, Jun Wan, Zhen Lei, Ruicong Zhi, Guodong Guo, and Stan Z Li. Ef cient group-n encoding and decoding for facial age estimation. *TPAMI*, 2017.
[51] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 2008.
[52] Xin Wen, Biying Li, Haiyun Guo, Zhiwei Liu, Guosheng Hu, Ming Tang, and Jinqiao Wang. Adaptive variance based label distribution learning for facial age estimation. In *ECCV*, 2020.
[53] Shihao Zhang, Linlin Yang, Michael Bi Mi, Xiaoxu Zheng, and Angela Yao. Improving deep regression with ordinal entropy. *arXiv preprint arXiv:2301.08915*, 2023.
[54] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, 2022.
[55] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 2022.
[56] Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. La te: Towards language-free training for text-to-image generation. In *CVPR*, 2022.