VILA: Learning Image Aesthetics from User Comments with Vision-Language Pretraining

CVPR 2023

Introduction

- Humanlabeled rating scores simplifies the process of aesthetic perception.

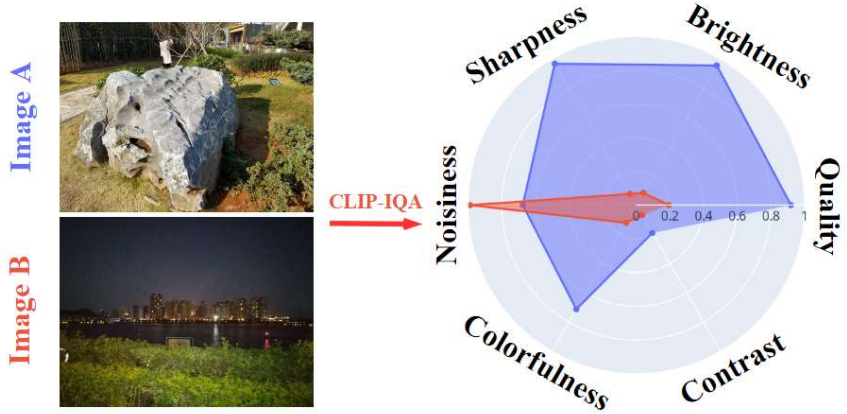
  User comments provide more comprehensive information and human are better at expressing aesthetic preferences through natural language.

we propose learning image aesthetics from user comments, and exploring vision-language pretraining methods to learn multimodal aesthetic representations

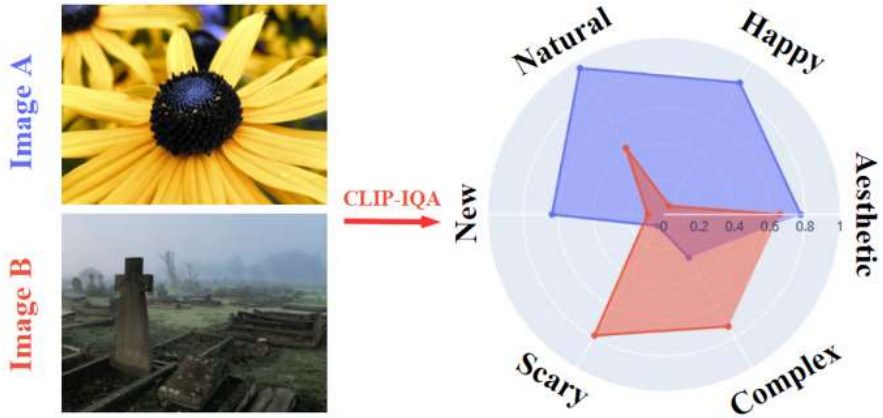
CLIP-IQA

Look Brightness、Colorfulness、Noise quality perception

Feel Emotion、 Aesthetic abstract perception



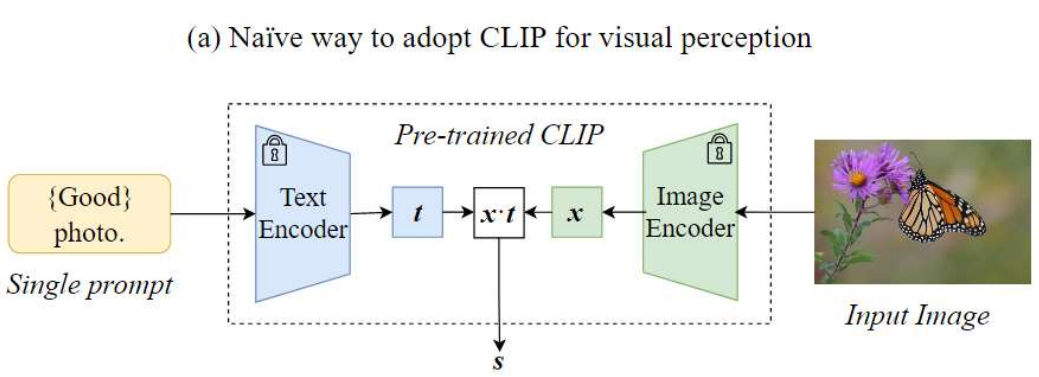
(a) CLIP for quality perception

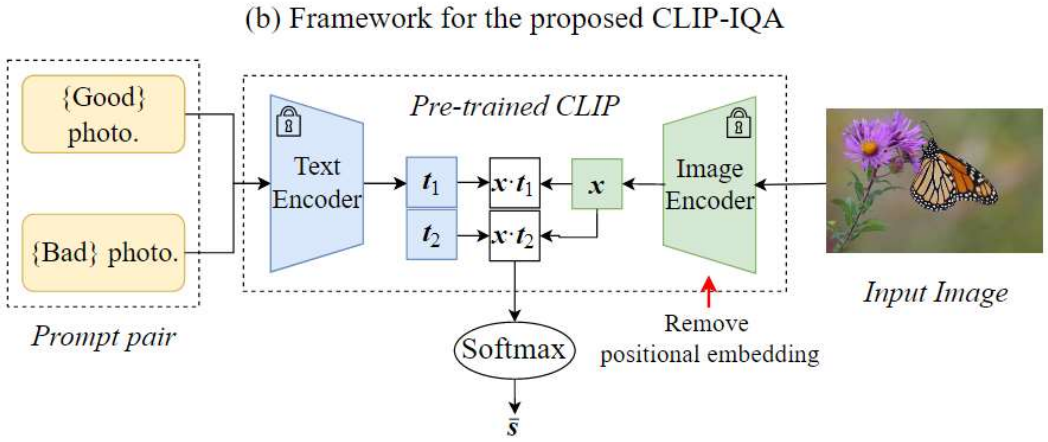


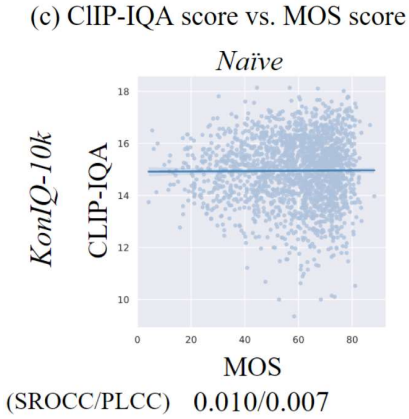
(b) CLIP for abstract perception

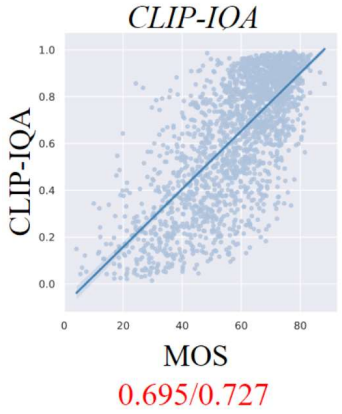
Wang J, Chan K C K, Loy C C. Exploring clip for assessing the look and feel of images[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2023, 37(2): 2555-2563.

CLIP-IQA









➢ **Antonym prompt pairing**

$$s = \frac{\boldsymbol{x} \odot \boldsymbol{t}}{||\boldsymbol{x}|| \cdot ||\boldsymbol{t}||},$$

$$s_i = \frac{\boldsymbol{x} \odot \boldsymbol{t}_i}{||\boldsymbol{x}|| \cdot ||\boldsymbol{t}_i||}, \quad i \in \{1, 2\},$$

$$\bar{s} = \frac{e^{s_1}}{e^{s_1} + e^{s_2}}.$$

Wang J, Chan K C K, Loy C C. Exploring clip for assessing the look and feel of images[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2023, 37(2): 2555-2563.

VILA

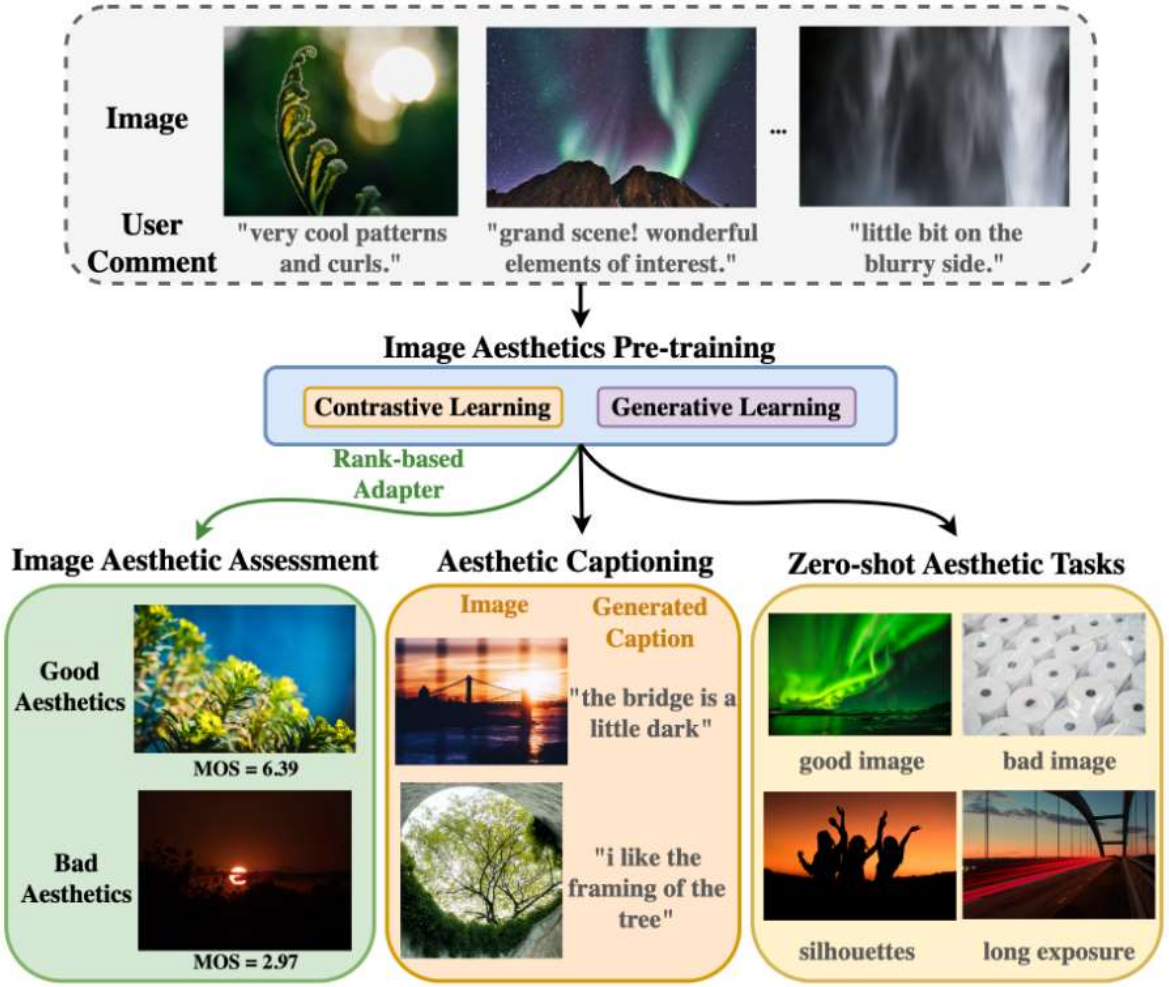
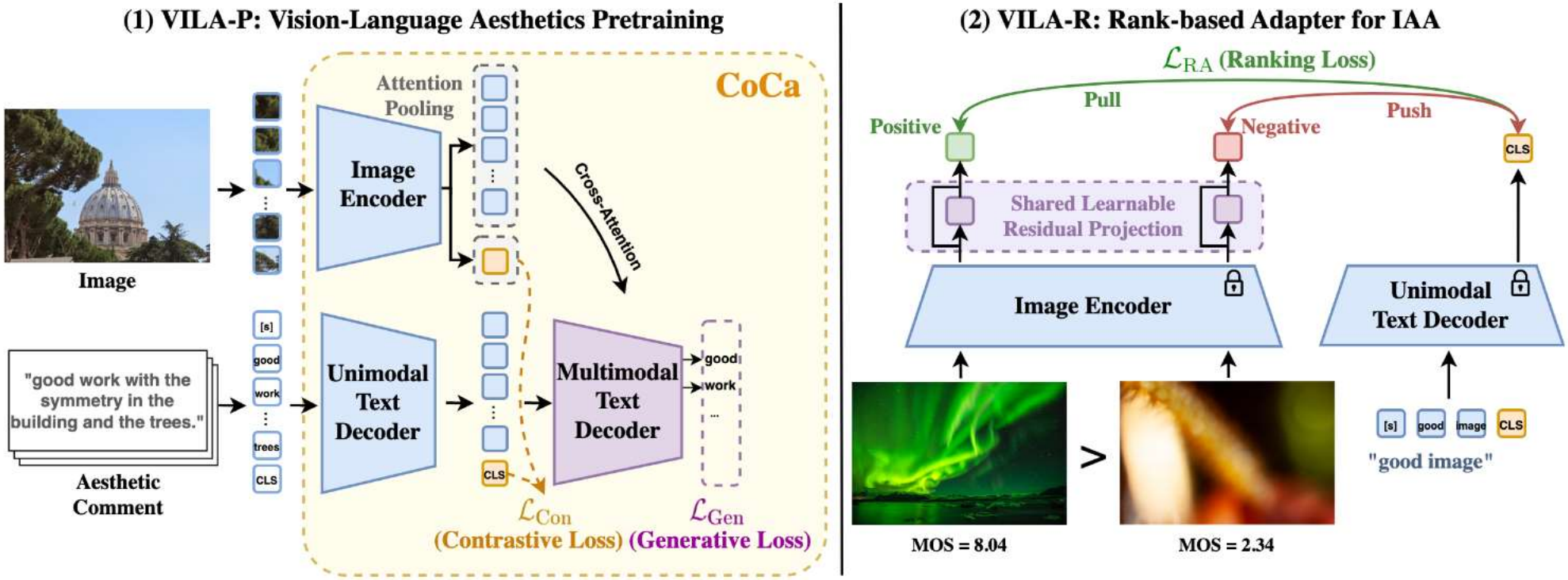
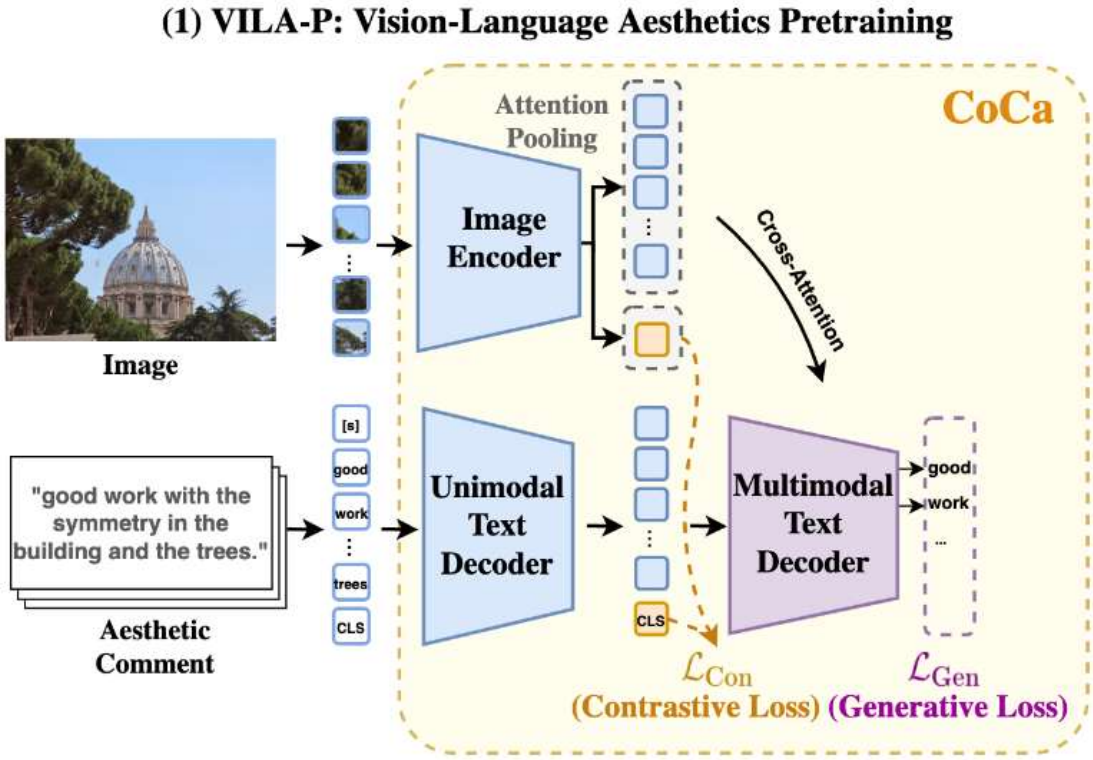


Image
User Comment
"very cool patterns and curls."
"grand scene! wonderful elements of interest."
...
"little bit on the blurry side."
Image Aesthetics Pre-training
Contrastive Learning
Generative Learning
Rank-based Adapter
Image Aesthetic Assessment
Good Aesthetics
MOS = 6.39
Bad Aesthetics
MOS = 2.97
Aesthetic Captioning
Image
Generated Caption
"the bridge is a little dark"
"i like the framing of the tree"
Zero-shot Aesthetic Tasks
good image
bad image
silhouettes
long exposure

Method


(1) VILA-P: Vision-Language Aesthetics Pretraining
CoCa
Attention Pooling
Image Encoder
Cross-Attention
Image
"good work with the symmetry in the building and the trees."
Aesthetic Comment
[s]
good
work
trees
CLS
Unimodal Text Decoder
Multimodal Text Decoder
good
work
$\mathcal{L}_{\text{Con}}$
(Contrastive Loss)
$\mathcal{L}_{\text{Gen}}$
(Generative Loss)
(2) VILA-R: Rank-based Adapter for IAA
$\mathcal{L}_{\text{RA}}$ (Ranking Loss)
Pull
Push
Positive
Negative
CLS
Shared Learnable Residual Projection
Image Encoder
Unimodal Text Decoder
>
[s]
good
image
CLS
"good image"
MOS = 8.04
MOS = 2.34

VILA-P



Preliminary of CoCa

Contrastive Learning Objective

$$\mathcal{L}_{\text{Con}}^{i2t} = -\frac{1}{N}\Big(\sum_{i}^{N} \log \frac{\exp(\boldsymbol{x}_i^\top \boldsymbol{y}_i)/\tau}{\sum_{j=1}^{N} \exp(\boldsymbol{x}_i^\top \boldsymbol{y}_j/\tau)}\Big)$$

$$\mathcal{L}_{\text{Con}}^{t2i} = -\frac{1}{N}\Big(\sum_{i}^{N} \log \frac{\exp(\boldsymbol{y}_i^\top \boldsymbol{x}_i)/\tau}{\sum_{j=1}^{N} \exp(\boldsymbol{y}_i^\top \boldsymbol{x}_j/\tau)}\Big)$$

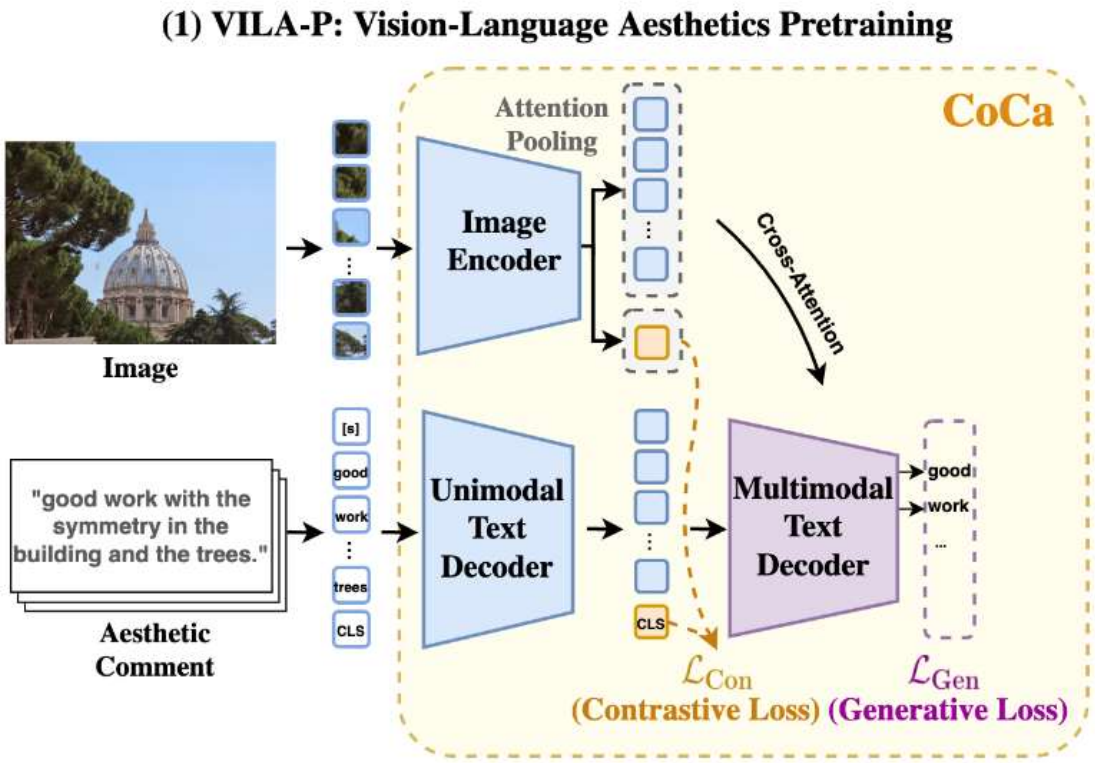
$$\mathcal{L}_{\text{Con}} = \mathcal{L}_{\text{Con}}^{i2t} + \mathcal{L}_{\text{Con}}^{t2i}$$

Generative Learning Objective

$$\mathcal{L}_{\text{Gen}} = -\sum_{t=1}^{L} \log P(\boldsymbol{w}_t | \boldsymbol{w}_{<t}, \boldsymbol{V}).$$

$$\mathcal{L} = \alpha \mathcal{L}_{\text{Con}} + \beta \mathcal{L}_{\text{Gen}}.$$

VILA-P



Two-stage pretraining approach

- General Pretraining

  LAION-5B-English dataset

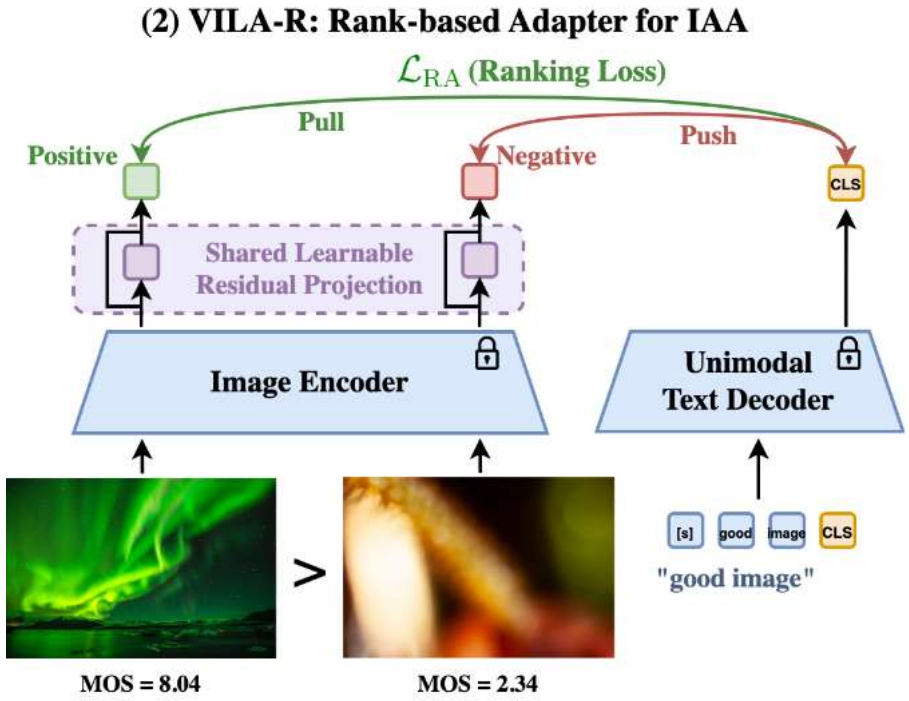
  650M images – text pairs

- Aesthetic Pretraining

  AVA-Captions dataset

  230k images + 1.5M captions

VILA-R



IAA Formulation

$$\boldsymbol{v} = E(\boldsymbol{I}, \boldsymbol{\theta}_{frozen}),$$
$$r = F(\boldsymbol{v}, \boldsymbol{\gamma}),$$

Rank-based Adapter

$$\tilde{\boldsymbol{v}} = \mathrm{normalize}(\boldsymbol{v}^\top \boldsymbol{H} + \boldsymbol{v}),$$
$$r = \tilde{\boldsymbol{v}}^\top \boldsymbol{w}_p$$

triplet ranking loss:

$$\mathcal{L}_{\mathrm{RA}} = \frac{1}{P} \sum_{i,j,i\neq j,l_i>l_j} \max\left(0, m - \tilde{\boldsymbol{v}}_i^\top \boldsymbol{w}_p + \tilde{\boldsymbol{v}}_j^\top \boldsymbol{w}_p\right)$$

Experiments

Method	SRCC	PLCC
Kong *et al.* [24]	0.558	-
NIMA (Inception-v2) [42]	0.612	0.636
AFDC + SPP [2]	0.649	0.671
MaxViT [45]	0.708	0.745
AMP [31]	0.709	-
Zeng *et al.* (resnet101) [55]	0.719	0.720
MUSIQ [19]	0.726	0.738
Hentschel *et al.* [13]	0.731	0.741
Niu *et al.* [33]	0.734	0.740
MLSP (Pool-3FC) [15]	0.756	0.757
TANet [12]	0.758	**0.765**
GAT$_{\times 3}$-GATP [11]	**0.762**	0.764
Zero-shot Learning		
VILA-P (single prompt)	0.605	0.617
VILA-P (ensemble prompts)	0.657	0.663
VILA-R	**0.774**	**0.774**

Table 1. Results on AVA dataset. **Blue** and **black** numbers in bold represent the best and second best respectively. First group shows baselines, second group shows ZSL results using our model from Sec. 3, final line shows our result combining Sec. 3 and Sec. 4.

	Prompts	
	$\boldsymbol{p}_g$	$\boldsymbol{p}_b$
Single Prompt	*"good image"*	*"bad image"*
Ensemble of Prompts	*"good image"*	*"bad image"*
	"good lighting"	*"bad lighting"*
	"good content"	*"bad content"*
	"good background"	*"bad background"*
	"good foreground"	*"bad foreground"*
	"good composition"	*"bad composition"*

Table 11. Text prompts used in ZSL for IAA.

Experiments

- Effects of image-text pretraining rank-based adapter

	ZSL Ens. Prompts			w/ Our Adapter		
General Pretraining	✓		✓	✓		✓
Aesthetic Pretraining		✓	✓		✓	✓
SRCC	0.228	0.265	**0.657**	0.746	0.566	**0.774**
PLCC	0.228	0.276	**0.663**	0.750	0.575	**0.774**

Table 2. Effects of image-text pretraining on AVA. Different pretraining schema are employed for each column and two settings are reported: 1) ZSL using an ensemble of prompts; 2) further finetuned using our proposed rank-based adapter.

Method	SRCC	PLCC
VILA-P w/ L2 Loss	0.757	0.756
VILA-P w/ EMD Loss [42]	0.759	0.759
VILA-R w/o Text Anchor	0.763	0.764
VILA-R w/o Residual	0.766	0.766
VILA-R (Ours)	**0.774**	**0.774**
VILA-R Finetune Image Encoder	0.780	0.780

Table 3. Ablation for the proposed rank-based adapter (Sec. 4) on AVA. First two groups use frozen pretrained image encoder.

Experiments

Zero-shot Aesthetic Style Classification

Method	mAP (%)
Murray *et al.* [32]	53.9
Karayev *et al.* [18]	58.1
Lu *et al.* [29]	64.1
MNet [41]	65.5
Sal-RGB [9]	71.8
Zero-shot Learning	
General Pretraining (single prompt)	29.3
General Pretraining (ensemble prompts)	32.6
VILA-P (single prompt)	62.3
VILA-P (ensemble prompts)	**69.0**

AVA Comments Generation

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	CIDEr
CWS [10]	**0.535**	0.282	0.150	0.074	0.254	0.059
Yeo *et al.* [52]	0.464	0.238	0.122	0.063	**0.262**	0.051
VILA	0.503	**0.288**	**0.170**	**0.113**	**0.262**	**0.076**

Table 5. Results on AVA-Captions dataset.

Figure 5. Aesthetic comments generated by VILA.

Thanks!