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

Junjie Ke, Keren Ye, Jiahui Yu, Yonghui Wu, Peyman Milanfar, Feng Yang
Google Research
{junjiek, yek, jiahuiyu, yonghui, milanfar, fengyang}@google.com

Abstract

Assessing the aesthetics of an image is challenging, as it is influenced by multiple factors including composition, color, style, and high-level semantics. Existing image aesthetic assessment (IAA) methods primarily rely on human-labeled rating scores, which oversimplify the visual aesthetic information that humans perceive. Conversely, user comments offer more comprehensive information and are a more natural way to express human opinions and preferences regarding image aesthetics. In light of this, we propose learning image aesthetics from user comments, and exploring vision-language pretraining methods to learn multimodal aesthetic representations. Specifically, we pretrain an image-text encoder-decoder model with image-comment pairs, using contrastive and generative objectives to learn rich and generic aesthetic semantics without human labels. To efficiently adapt the pretrained model for downstream IAA tasks, we further propose a lightweight rank-based adapter that employs text as an anchor to learn the aesthetic ranking concept. Our results show that our pretrained aesthetic vision-language model outperforms prior works on image aesthetic captioning over the AVA-Captions dataset, and it has powerful zero-shot capability for aesthetic tasks such as zero-shot style classification and zero-shot IAA, surpassing many supervised baselines. With only minimal fine-tuning parameters using the proposed adapter module, our model achieves state-of-the-art IAA performance over the AVA dataset. [1]

1. Introduction

Image Aesthetic Assessment (IAA) aims to quantify the human perceived aesthetics of an image. It has many important applications, including photo recommendation, selection, and editing. IAA is challenging because it is inherently subjective, and depends on various factors including image composition, color usage, photographic style, and subject matter. In recent years, various learning-based IAA methods have been proposed by leveraging deep models such as convolutional neural networks (CNN) [2, 12, 15, 42] and transformers [19]. These approaches learn from human-labeled IAA datasets where images are paired with aesthetic ratings, and models are trained to regress towards the mean opinion scores (MOS).

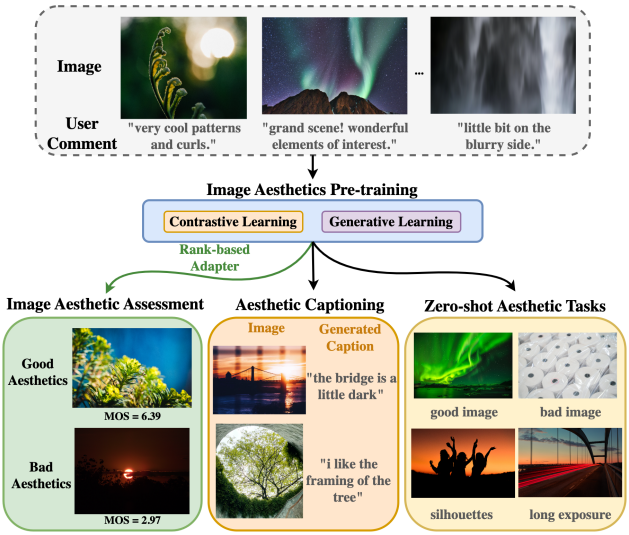

Figure 1. We present VILA, a vision-language aesthetics learning framework based on image and user comment pairs. By pretraining on a contrastive and generative target, it shows superior performance on aesthetic captioning as well as zero-shot aesthetic tasks, *e.g.*, IAA, and style classification. With a lightweight rank-based adapter, we can efficiently adapt the pretrained model to IAA.

Directly learning IAA models on human-labeled aesthetic ratings, such as MOS, can be suboptimal as it lacks context regarding why an image is aesthetically pleasing or not. To provide richer supervision, various methods have attempted to integrate external knowledge such as theme [12, 33], human eye fixation [9], and aesthetic attributes [4, 24], to enhance IAA performance. These approaches typically rely on multitask training or cascade

[1] Our model is available at https://github.com/google-research/google-research/tree/master/vila

score prediction with a frozen attribute network. However, obtaining additional labeled data or off-the-shelf models for such methods can be costly.

Compared to the aforementioned methods that require additional annotations, our approach utilizes the abundance of image-comment pairs available on aesthetic websites and photographic forums. These pairs can be easily obtained from the Internet and contain extensive aesthetic information (*e.g.* objects, themes, styles, and user emotions), since humans are better at expressing aesthetic preferences through natural language than through abstract scores. On image sharing platforms like Flickr and DPChallenge[2], user comments offer valuable insights into how they evaluate an image's aesthetics. For instance, as shown in Fig. 1 (top), comments such as "very cool patterns and curls" and "little bit on the blurry side" reflects users' positive and negative aesthetic opinions respectively. We aim to learn the diverse aesthetic semantics present in these image-comment pairs to establish a solid foundation for downstream IAA tasks.

Using image-comment pairs for aesthetics learning remains largely unexplored. While previous works have leveraged user comments to improve IAA, their approaches differ significantly from ours. For example, [14,57,58] proposed to aggregate visual and comment features, yet they require both the image and comment as inputs during inference. This requirement makes it difficult to use such methods in real-world settings where images may not always be accompanied by comments. To mitigate this, Niu *et al*. [33] proposed to use the LDA topics [1] from the comments as pseudo labels to guide image representation learning. However, the simplification of comments into topics may result in a loss of valuable contextual information. Therefore, we are motivated to explore other strategies for utilizing raw comments to extract richer aesthetic textual information.

In this paper, we present a novel two-stage **VI**sion-**L**anguage **A**esthetics (**VILA**) learning framework incorporating image-text pretraining. Our goal is to develop a model that can effectively generalize to multiple downstream aesthetic tasks (Fig. 1). In the first **P**retraining stage, we learn an image-text model (**VILA-P**) by employing contrastive and text sequence generation objectives, enbaling us to fully leverage fine-grained knowledge from aesthetic image-comment pairs. Our approach is motivated by recent advancements in vision-language models, such as CLIP [35], ALIGN [17], and CoCa [54], which exhibit impressive performance and generalization ability across multiple tasks. These models align vision and language feature spaces to capture the rich semantic information. However, these models are typically pretrained on general image-text pairs from the web, which can result in under-representation of aesthetic-related information. Our experimental results indicate that such generally pretrained vision-language models underperform on aesthetic tasks (Sec. 5.3). As a solution, we propose the adoption of vision-language pretraining on aesthetic image-comment pairs from photograph sharing websites. To the best of our knowledge, our work is the first to explore the use of image-comment pairs in vision-language pretraining for aesthetics learning.

After pretraining VILA-P on image-comment pairs, we finetune it for downstream score-based IAA tasks using a lightweight **R**ank-based adapter (**VILA-R**). This adapter involves adding feature residuals to the frozen image embeddings to move images with high aesthetic quality closer to the anchor text "good image," and images with low aesthetic quality away from it. This method can effectively rank images based on human rated preferences. With 0.1% tunable parameters, our model outperforms previous works on IAA correlation metrics over the AVA dataset [32].

Our proposed VILA is capable of tackling multiple aesthetic-related tasks beyond score-based IAA (Fig. 1). Not only can it generate high-quality aesthetic comments, but it also exhibits impressive zero-shot learning (ZSL) capabilities for aesthetic style classification and quality analysis. Using text queries such as "good image" and "bad image" to compare images, our ZSL model outperforms supervised learning models like NIMA [42] which requires labor-intensive ratings as ground truth. This highlights the potential of learning rich image aesthetic concepts without relying on human-labeled data, thereby significantly reducing data collection costs.

We summarize the contributions of our work as follows:

- We propose a vision-language aesthetic learning framework (VILA) for learning rich image aesthetic features using image-comment pairs.
- We design a novel rank-based module to adapt the model to downstream IAA tasks without perturbing the pretrained weights, effectively learning the aesthetic quality concepts with minimal additional parameters.
- Our pretrained aesthetic model outperforms prior works for aesthetic captioning on the AVA-Captions [10] dataset. Even without any supervised labels, our zero-shot model achieves 69% mAP on the AVA-Style [32] dataset and 0.657 SRCC on the AVA dataset [32], outperforming many supervised approaches. With the proposed adapter and a small number of tunable parameters, our method further achieves state-of-the-art performance on AVA.

2. Related Work

Image Aesthetic Assessment has a wide range of applications such as search, ranking, and recommendation. Unlike the technical quality assessment [6, 16, 53] which focuses on image distortion, cropping, or noise, IAA aims to measure the aesthetic quality. During the deep learning era,

[2] https://www.dpchallenge.com/

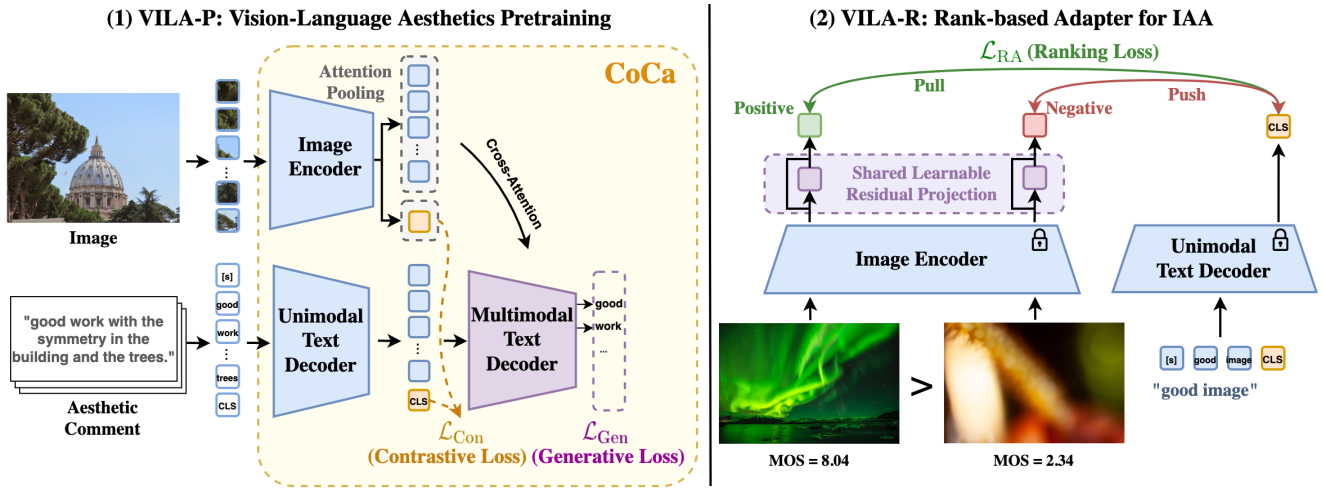

Figure 2. Our proposed vision-language aesthetic (VILA) framework contains two parts: (1) VILA-P: pretraining a vision-language model using images and user comments on aesthetics, and (2) VILA-R: a rank-based adapter that efficiently adapts the frozen pretrained model to score-based IAA with a small amount of tunable parameters (purple block).

works such as [12, 18, 26, 32, 36, 44, 51] focused on data-driven methods and collected large-scale datasets containing images and human ratings. Based on these datasets, [24] built a ranking-based model, while [31, 42, 55] proposed to approximate the groundtruth score distributions. Different from these works, our model benefits from the image-text pretraining framework that has rarely been explored in IAA.

Additional supervision in IAA has been explored in works such as [50, 58], where natural language annotations were introduced in their curated datasets. However, these methods either treat IAA as one of multiple parallel tasks [33, 50], do not generate quality related outputs [50, 58], or require both image and comment at inference time [14, 57, 58]. In contrast, our model leverages user comments to learn meaningful aesthetic representations using contrastive and generative targets, and the learned image model can be used independently without text input.

Moreover, various studies have focused on network design to preserve high-resolution aesthetic information for IAA, such as CNN-based methods [2, 15, 30] that reduce the negative effects of cropping and resizing, and transformer architectures [11, 19] that treat input image as visual tokens and support variable-length sequences, preserving image resolution and aspect ratios. Our method achieves state-of-the-art results with a fixed 224×224 input without considering original resolution and aspect ratios, and we believe that these related methods could further enhance our model and be incorporated in future work.

Image-Text Pretraining utilizes the fact that paired image and text are correlated. Initially, contrastive learning was used to draw image representation and aligned text representation closer [5, 8, 22]. Later, self-supervised learning objectives were explored, such as masked region reconstruction, masked object prediction, word region alignment [3, 27, 28, 40, 43]. These early models used off-the-shelf visual detectors, which limited their generalization to large-scale pretraining. The introduction of ViT [23] enabled end-to-end multimodal transformer-based methods [20, 49] for large-scale vision-language pretraining. Recently, several methods such as CLIP [35], ALIGN [17], and CoCa [54] have proposed image-text foundation models trained on large-scale image-text corpus [17, 56]. These methods adopted general pretraining using billions of image-text pairs from the web, and showed impressive results on various tasks such as retrieval, classification, and captioning. Concurrent works [13, 48] have shown the benefit of using such generally pretrained CLIP features for aesthetics learning. However, due to the sparsity of aesthetics-related image-text pairs on the web, aesthetic information gets diluted in such general pretraining process. To address this, we propose the aesthetics pretraining on image-comment pairs to further enhance aesthetics information. Our model is based on the CoCa [54] architecture, with a novel rank-based adapter module designed for IAA to learn relative aesthetic quality with minimal tunable parameters. The rank-based adapter optimizes only a small set of learnable parameters, avoiding catastrophic forgetting [7, 21] while retaining the rich knowledge from the pretrained model.

3. Image Aesthetics Pretraining using CoCa

In this section, we present our approach to pretrain the image aesthetic model VILA-P. Our goal in the pretraining stage is to learn powerful multimodal representations for image aesthetics in a self-supervised manner, using both images and their associated user comments.

Without loss of generality, we adopt the CoCa [54] archi-

tecture, which combines contrastive learning and image-to-caption generation in a single framework. Our approach is generally applicable to broader vision-language pretraining models. Fig. 2 (1) provides an overview of our pretraining architecture for VILA-P.

3.1. Preliminary of CoCa

CoCa contains an image encoder, a unimodal text decoder, and a multimodal text decoder. The image encoder produces an image representation, while the unimodal text decoder generates a text representation with an appended `[CLS]` token. These two representations are aligned using a contrastive objective. The multimodal text decoder generates captions by cross-attending to the image features.

Encoding Image: The image encoder is in the form of a Vision Transformer [23], which splits an image into patches and treats them as tokens. The patches are then projected to D-dimensional features and fed to the transformer blocks to generate a sequence of visual embeddings $\boldsymbol{V} = \{\boldsymbol{v}_1, ..., \boldsymbol{v}_K\}$, where K is the number of visual tokens.

Encoding Text: The text is first tokenized into a sequence of tokens, with each token mapped to a D-dimensional word embedding vector. A `[CLS]` token is appended to the sequence, and the sequence is passed through transformer layers to generate the unimodal text representation $\boldsymbol{W} = \{\boldsymbol{w}_1, ..., \boldsymbol{w}_L, \boldsymbol{w}_{cls}\}$, where $\boldsymbol{w}_{cls}$ is output of the `[CLS]` token, and L is the number of text tokens. The transformer text decoder layers are trained with causally-masked self-attention for the captioning objective, which prevents tokens from attending to future tokens. The learnable token $\boldsymbol{w}_{cls}$ is used as the contrastive text embedding.

Contrastive Learning Objective: The two unimodal encoding modules are jointly optimized by a contrastive target which tries to align the image-text pairs:

$$\begin{aligned}
\mathcal{L}^{i2t}_{\text{Con}} &= -\frac{1}{N}\Big(\sum_{i}^{N} \log \frac{\exp(\boldsymbol{x}_i^\top \boldsymbol{y}_i)/\tau}{\sum_{j=1}^{N} \exp(\boldsymbol{x}_i^\top \boldsymbol{y}_j/\tau)}\Big) \\
\mathcal{L}^{t2i}_{\text{Con}} &= -\frac{1}{N}\Big(\sum_{i}^{N} \log \frac{\exp(\boldsymbol{y}_i^\top \boldsymbol{x}_i)/\tau}{\sum_{j=1}^{N} \exp(\boldsymbol{y}_i^\top \boldsymbol{x}_j/\tau)}\Big) \\
\mathcal{L}_{\text{Con}} &= \mathcal{L}^{i2t}_{\text{Con}} + \mathcal{L}^{t2i}_{\text{Con}}
\end{aligned} \tag{1}$$

$\boldsymbol{x}_i$ and $\boldsymbol{y}_i$ are the normalized contrastive embeddings of the i-th image and text in the batch. L^{i2t}_{Con} is the image-to-text contrastive loss and L^{t2i}_{Con} is the text-to-image counterpart, τ is the learnable temperature, N is the batch size.

Generative Learning Objective: For captioning, the multimodal text decoder learns to maximize the likelihood of generating the paired text conditioned on visual features in an autoregressive manner:

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

Cotraining Contrastive and Generative Objective: To cotrain the two targets, two task-specific attentional pooling layers [25] are added on top of the image encoder to generate a contrastive image representation and a generative image representation. The pretraining objective is a weighted sum of the contrastive loss and the generative loss, using hyper-parameters α and β:

$$\mathcal{L} = \alpha \mathcal{L}_{\text{Con}} + \beta \mathcal{L}_{\text{Gen}}. \tag{2}$$

3.2. Vision-Language Pretraining for Aesthetics

Vision-language pretraining methods require large-scale data to learn the complex dynamics between visual and textual information. Many of these methods are trained on large proprietary datasets [17, 35] with image-text pairs crawled from the web. While this general pretraining strategy has proven useful for tasks such as image classification and retrieval, it is limited in its ability to represent aesthetic-related information due to the under-representation of such information on the web. Consequently, the aesthetic information gets diluted in the vast amount of pretraining data. To address this limitation, we propose a two-stage pretraining approach that involves initializing the model with a generally pretrained image-text model and then further pretraining it on aesthetic image-comment pairs. For general pretraining, we use a 650M filtered subset of the openly available LAION-5B-English [38] dataset. For aesthetic pretraining, we use the AVA-Captions dataset [10] which is currently the largest available dataset for aesthetic comments. Each image in AVA-Captions is associated with one or more user comments that provide informative insights into different aesthetic aspects of the image. We randomly sample one comment for each image to construct image-comment pairs during training.

In contrast to traditional supervised learning with predefined labels or categories, vision-language pretraining enables learning of open-set aesthetic concepts through noisy image-comment pairs. This results in visual and textual representations that encompass a wider range of aesthetic concepts, enhancing transferability to downstream tasks.

4. Adapting Vision-Language Model for IAA

The pretrained model VILA-P contains extensive multimodal aesthetic information, enabling it to perform zero-shot aesthetic tasks and to even outperform supervised models (Sec 5.3 and Sec 5.4). In this section, we aim to further enhance the model's performance for IAA tasks using the mean-opinion-score (MOS) labels. Finetuning the entire model is computationally expensive and can harm the pretrained model's zero-shot and captioning capability. Therefore, we propose a lightweight rank-based adapter module that adapts the pretrained vision-language model to

downstream IAA tasks while keeping the image and text backbone frozen with only a few tunable parameters. The adapter module allows the model to retain the benefits of the pretrained backbone, while leveraging the rich aesthetic textual information for IAA tasks. Fig. 2 (2) depicts the overview of the adapter module, and we refer to the resulting model as VILA-R.

4.1. Image Aesthetic Assessment Formulation

The goal of IAA is to predict the aesthetic score for a given image. We focus on the case where the image is represented by the frozen image embedding extracted by the image encoder in VILA-P. Formally,

$$\boldsymbol{v} = E(\boldsymbol{I}, \boldsymbol{\theta}_{frozen}), \tag{3}$$

$$r = F(\boldsymbol{v}, \boldsymbol{\gamma}), \tag{4}$$

where $\boldsymbol{I}$ is the input image, $\boldsymbol{v}$ is the image features extracted using image encoder E with its frozen pretrained weights $\boldsymbol{\theta}_{frozen}$. F is the IAA scoring model with parameters $\boldsymbol{\gamma}$, and r is the predicted aesthetic score.

During training, given two images represented by $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$, and their corresponding MOS labels l_i and l_j, the IAA model output r_i and r_j are trained to respect the order of l_i and l_j. The performance of the proposed model F is evaluated by the correlation between r and l.

To obtain an effective F with few parameters, we draw inspiration from the ZSL setting where no parameter tuning is required. Since the cosine similarity between paired image-text is maximized by the contrastive pretraining objective (Eq. 1), we can use the cosine similarity between the contrastive image embedding $\boldsymbol{v}$ and the text embedding $\boldsymbol{w}$ as a measure of how much the image aligns with the textual concept. By using text as "prompts", we can effectively score images for the textual concept (*e.g.*, whether they are "good image"). Our preliminary study shows that using text prompts for IAA scoring results in a correlation of over 0.6, suggesting that the text decoder in VILA-P contains useful information about what constitutes a visually pleasing image. We aim to utilize this information as an anchor to further enhance the model's IAA ranking capability by designing a lightweight rank-based adapter module.

4.2. Rank-based Adapter Module

The pretraining process, which includes contrastive and generative objectives, captures rich textual concepts related to aesthetically pleasing images in the text decoder, and embeds them in the same latent space as the image. Therefore, we can make slight adjustments to the image embedding to improve its alignment with these textual concepts. Concretely, we propose using the frozen text embedding of "good image" as an anchor to score images, and optimize the relative ranking between two images according to their MOS labels by adjusting their image representations. This is illustrated in Fig. 2 (2).

Let $\boldsymbol{v}$ represent the unnormalized contrastive image embedding from the frozen VILA-P image encoder. To obtain the rank-adjusted image embedding $\tilde{\boldsymbol{v}}$, we add a learnable residual represented by $\boldsymbol{H} \in \mathbb{R}^{D \times D}$ and normalize the output as follows:

$$\tilde{\boldsymbol{v}} = \text{normalize}(\boldsymbol{v}^\top \boldsymbol{H} + \boldsymbol{v}), \tag{5}$$

Next, we use "good image" as the prompt, and extract its normalized frozen text embedding $\boldsymbol{w}_p$ from the `[CLS]` position of the unimodal text decoder. The cosine similarity between the rank-adjusted image embedding $\tilde{\boldsymbol{v}}$ and the anchor $\boldsymbol{w}_p$ is used as the predicted IAA score for ranking:

$$r = \tilde{\boldsymbol{v}}^\top \boldsymbol{w}_p \tag{6}$$

To optimize the relative ranking between two images, we use $\boldsymbol{w}_p$ as the anchor and optimize the triplet ranking loss $\mathcal{L}_{\text{RA}}$ for a pair of input images:

$$\mathcal{L}_{\text{RA}} = \frac{1}{P} \sum_{i,j,i \neq j, l_i > l_j} \max\left(0, m - \tilde{\boldsymbol{v}}_i^\top \boldsymbol{w}_p + \tilde{\boldsymbol{v}}_j^\top \boldsymbol{w}_p\right) \tag{7}$$

m is the margin hyper-parameter with default value 0.1. The positive sample $\tilde{\boldsymbol{v}}_i$ corresponds to the image with a higher MOS label l_i, and the negative sample $\tilde{\boldsymbol{v}}_j$ corresponds to the image with a lower MOS label l_j. The ranking loss ensures that the similarity between the positive sample and the "good image" anchor is greater than that of the negative sample, effectively ranking the images according to its aesthetic ratings. The only tunable parameter is $\boldsymbol{H}$ with D^2 parameters, about 0.1% of the total parameters in VILA-P.

It is worth noting that the frozen text embedding $\boldsymbol{w}_p$ can be exported for training and inference without the text backbone. Therefore, the final IAA model has the same computational and storage as a single image-encoder-only model, and it only needs the image as input for IAA inference.

5. Experiments

5.1. Datasets

LAION-5B-English-Filtered is a 650M subset from the English split in LAION-5B [38], which is currently the largest publicly available dataset with 5B CLIP-filtered image-text pairs. The filtered subset is obtained by removing non-informative or bad data, such as poorly formatted text, bad image size or aspect ratio, and poor image content. We use this subset for general image-text pretraining.

AVA Dataset [32] is a widely-used IAA benchmark originating from the DPChallenge website. It consists of over 250,000 images with user voting scores ranging from 1 to 10. We evaluate the IAA performance of our model on the

available 19,928 AVA test images, reporting Spearman rank order correlation coefficient (SRCC) and Pearson linear correlation coefficient (PLCC) metrics.

AVA-Captions [10] dataset is a collection of user comments for the AVA images, crawled from the DPChallenge website, with basic text filtering applied. It contains 230k images and 1.5M captions, with an average of 5 comments per image. To avoid potential data leakage, we strictly follow the official data split of both AVA and AVA-Captions, excluding both test sets from training, resulting in a training dataset with 212,585 images paired with 1.2M captions. We evaluate the aesthetic comment generation quality of our model on 9,361 AVA-Captions test images, reporting BLEU [34], ROUGE [37], and CIDEr [47] scores.

AVA-Style [32] contains images with 14 photographic style labels. We use the 2,809 testing images to assess the zero-shot aesthetic style classification capability of our pretrained model.

5.2. Implementation Details

We use CoCa-Base, the smallest variant of CoCa [54]. It contains a ViT-B/16 [23] image encoder with 12 transformer [46] layers, hidden dimension $D = 768$, and MLP size 3072. The image resolution is set to 224×224 with a patch size of 16×16, resulting in $K = 196$ image tokens. Data augmentation during training includes random horizontal flipping and random cropping from 272×272. The unimodal text decoder consists of 6 transformer layers with the same hidden dimension and MLP size, while the multimodal text decoder consists of another 6 transformer layers. The maximum text length is set to 64 during training. For LAION pretraining, we train with 4096 batch size for 500k steps, using 5e-4 learning rate with linear decay to zero, and 0.01 weight decay. For image aesthetic pretraining on AVA-Captions, we train with 128 batch size for 500k steps, using 1e-5 learning rate with linear decay to zero, and 0.04 weight decay. We set contrastive loss weight $\alpha = 1$ and generative loss weight $\beta = 2$. A trainable temperature τ with an initial value of 0.07 is used for the contrastive loss, following [17, 54]. To finetune the rank-based adapter on AVA, we train with 128 batch size for 30k steps using 1e-5 learning rate with linear decay to zero, and 0.01 weight decay. All experiments use the Adafactor [39] optimizer with $\beta_1 = 0.9, \beta_2 = 0.999$, and are conducted on TPUv3.

5.3. AVA Image Aesthetic Assessment

Comparing to SOTA. Tab. 1 shows our results on the AVA dataset. The first group shows the baselines including the ranking method [24], distribution matching based approaches [31,42,55], customized neural networks [2,11,15, 19,45], and semantic-aware methods [12, 13, 33]. Our approach VILA-R achieves the best performance overall and outperforms the current SOTA GAT$_{\times 3}$-GATP [11] by 1.6% and 1.3% in terms of SRCC (0.774 vs 0.762) and PLCC (0.774 vs 0.764), respectively. Moreover, our method uses a lower resolution of 224×224 while other methods may benefit from the larger inputs. For example, MUSIQ [19] uses the full-size image and two additional resolutions, yet it underperforms our model. Hentschel *et al*. [13] utilize frozen CLIP features for learning image aesthetics, and VILA-R outperforms their approach, which shows the additional benefit of the proposed aesthetic pretraining.

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.

Zero-shot Learning (ZSL) for IAA. The second group in Tab. 1 shows the results of using our image-text pretrained model VILA-P (Sec. 3) for zero-shot IAA. We utilize the cosine similarity between the contrastive image and text embeddings for these experiments. In the single prompt setting, we compute the cosine similarity between the image and a single pair of prompts ("good image", "bad image"), and use the softmax normalized output for "good image" as the ZSL score for IAA. For ensemble prompts, we use an average ensemble of six pairs of prompts, each consisting of "good" or "bad" plus "image", "lighting", "composition", "foreground", "background", and "content" (see supplementary material). Notably, without any human label supervision, our ZSL model (SRCC 0.657, PLCC 0.663) has already outperformed several supervised baselines such as Kong *et al*. [24], NIMA [42], and AFDC + SPP [2]. These observations demonstrate the potential of leveraging unlabelled user comments for IAA, significantly reducing human labeling costs.

Effects of image-text pretraining. Tab. 2 presents an ablation study to validate the effectiveness of the proposed image-text pretraining. We conduct the general pretrain-

	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.

ing and aesthetic pretraining on the LAION [38] subset and AVA-Captions [10], respectively. With only the general pretraining, the model has suboptimal performance on the IAA task, verifying the assumption that image aesthetic information gets diluted by the vast amount of unrelated data from the web. Adding aesthetic pretraining greatly improves model performance in both zero-shot and finetuned settings. Both general and aesthetic pretraining have a significant positive impact on the final IAA task predictions. Regardless of the pretraining schema, the proposed rank-based adapter enhances the model's IAA performance with minimally tuned parameters.

Effectiveness of the proposed rank-based adapter. Tab. 3 shows an ablation study for the proposed rank-based adapter (Sec. 4). We compare different options for adapting the frozen VILA-P to downstream score-based IAA. The first group shows regression baselines that predict either the single MOS score using a L2 loss or the distribution of MOS scores using EMD loss [42]. VILA-R outperforms both of them, showing the effectiveness of a rank-based target. In the second group, we ablate the components in the proposed adapter. "w/o Text Anchor" denotes using a learnable projection to replace the frozen text prompt embedding $\boldsymbol{w}_p$. VILA-R performs better, showing the benefit of using the rich text embedding as a ranking anchor. For "w/o Residual", we use a simple learnable projection without the residual, *i.e.*, $\tilde{\boldsymbol{v}} = \text{normalize}(\boldsymbol{v}^\top \boldsymbol{H})$. Its sub-par performance confirms the intuition that we only need to slightly adjust the image embedding, thus learning the residual is easier. The final line shows that VILA-R can be further improved with finetuning the image encoder. However, its gain in performance comes at the cost of disturbing the generic pretrained weights, *e.g.* its ZSL performance on AVA-Style drops from 69.0% to 26.3% mAP. VILA-R enables effective IAA adaptation while inheriting the pretrained weights.

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**

Table 4. Results on AVA-Style dataset. We gray out supervised baselines as they are not directly comparable to our unsupervised model which is not exposed to the training labels.

5.4. AVA-Captions Image-Text Pretraining

In this section we aim to verify VILA-P model learns meaningful representations that are generalizable to other tasks. We evaluate its performance on zero-shot style classification and the quality of its generated aesthetic comments.

Zero-shot Style Classification. To demonstrate that VILA-P captures diverse aesthetic aspects such as composition, color, and style, we evaluate its ZSL performance on the AVA-Style test set. We manually curate text prompts based on the 14 class names, and use the cosine similarities to approximate the probability that an image involves specific styles (see supplementary material). Tab. 4 shows the results. The first group contains supervised methods trained on 11k images with style annotations. Without such supervision, VILA-P achieves 69.0% ZSL mAP, outperforming many supervised methods such as MNet [41] (65.5%) and Lu *et al.* [29] (64.1%). This demonstrates the ability of the proposed framework to learn open-set aesthetic information without human labelling. Tab. 4 also shows that the performance of the model trained only with general pretraining is much lower than that with aesthetic pretraining. This again verifies that the proposed aesthetic pretraining is necessary for capturing rich aesthetic information.

AVA Comments Generation. We evaluate the captioning performance of VILA-P on AVA-Captions test set, and the results are shown in Tab. 5. Our method outperforms CWS [10] and Yeo *et al.* [52] in terms of BLEU-2, BLEU-3, BLEU-4, ROUGE and CIDEr. Although our method has a slightly lower BLEU-1 than CWS, it is important to note that BLEU-1 only measures precision of unigram, while and higher order BLEU scores (BLEU-2, BLEU-3, BLEU-4) place more emphasis on the fluency of generated sentences. Moreover, our method's superior ROUGE and

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.

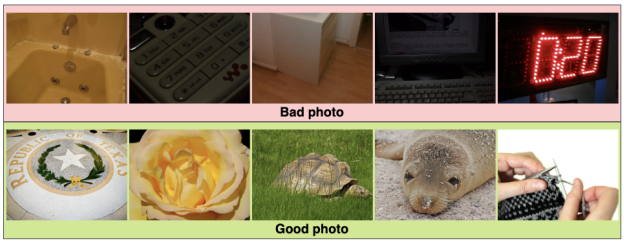

Figure 3. Top 5 images retrieved with "bad photo", "good photo" on KonIQ-10k [16]. See supplementary material for image sources.

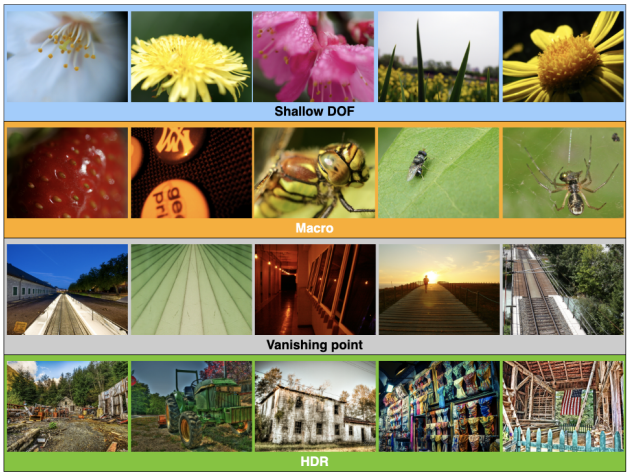

Figure 4. Top 5 images retrieved using AVA-Style class names on KonIQ-10k [16]. To give proper attribution to image sources, we choose to showcase images from the KonIQ-10k dataset instead of the AVA dataset. See supplementary material for image sources.

CIDEr scores indicates that our model generates more semantically similar sentences to the real user comments.

Qualitative Examples. To properly credit our image sources, we choose to display images from the KonIQ-10k [16] dataset instead of the AVA dataset for illustration in this section. The image sources are provided in supplementary material. Fig. 3 depicts the top-5 images retrieved by text queries "Bad photo" and "Good photo" on KonIQ-10k. For "Bad photo", the retrieved results exhibit poor lighting, bad composition and meaningless content. In contrast, the "Good photo" group has noticeably better aesthetic quality. These examples provide qualitative evidence of the aesthetic knowledge captured by the pretrained model.

Fig. 4 illustrates the AVA-Style predictions of VILA by visualizing the top-5 images retrieved using style class names on KonIQ-10k. This provides a qualitative demonstration of the aesthetic information captured by VILA. Results show that the aesthetic pretraining on image-comment pairs has helped the model to understand low-level aesthetic attributes quite well. For example, the learned model understands that "Macro" is a visual concept that captures finer details, regardless of the semantic objects, such as strawberry or insects. Another example is "HDR", for which all retrieved photos have high dynamic range while portraying different semantic objects such as buildings and cars.

Figure 5. Aesthetic comments generated by VILA.

Fig. 5 shows aesthetic comments generated by VILA. The model is capable of generating diverse captions conditioned on the images, mentioning attributes such as "color", "saturation" and "persepective". In addition, it even includes critiques about the cropping of the image, which aligns with our aesthetic perspective.

6. Conclusion

We propose a general framework for learning image aesthetics (VILA). By pretraining vision-language models on image-comment pairs from image sharing websites, we enable the model to learn rich aesthetic semantics in a self-supervised manner without the need for expensive labeled data. The resulting pretrained model, VILA-P, exhibits state-of-the-art performance on the AVA-Captions dataset and enables various interesting tasks, including zero-shot learning for IAA, style classification, and retrieval. Our experiments demonstrate that VILA-P surpasses many supervised baselines on these tasks with ZSL. To efficiently adapt the pretrained model for IAA without impairing its powerful zero-shot abilities or damaging the rich representation, we introduce a lightweight rank-based adapter module. By employing the text embedding as an anchor and explicitly modeling the ranking concept, we achieve state-of-the-art IAA performance on the AVA dataset with only a small amount of injected parameters. Although we design the rank-based adapter module for IAA, our method is generally applicable for adapting large-scale visual-language models to other ranking based tasks.

References

[1] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003. 2

[2] Qiuyu Chen, Wei Zhang, Ning Zhou, Peng Lei, Yi Xu, Yu Zheng, and Jianping Fan. Adaptive fractional dilated convolution network for image aesthetics assessment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 3, 6

[3] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations. In *European Conference on Computer Vision (ECCV)*, 2020. 3

[4] Sagnik Dhar, Vicente Ordonez, and Tamara L Berg. High level describable attributes for predicting aesthetics and interestingness. In *CVPR 2011*, pages 1657–1664. IEEE, 2011. 1

[5] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018. 3

[6] Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang. Perceptual quality assessment of smartphone photography. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[7] Robert M French. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135, 1999. 3

[8] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. *Advances in neural information processing systems*, 26, 2013. 3

[9] Koustav Ghosal, Mukta Prasad, and Aljosa Smolic. A geometry-sensitive approach for photographic style classification. *arXiv preprint arXiv:1909.01040*, 2019. 1, 7

[10] Koustav Ghosal, Aakanksha Rana, and Aljosa Smolic. Aesthetic image captioning from weakly-labelled photographs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 2, 4, 6, 7, 8

[11] Koustav Ghosal and Aljosa Smolic. Image aesthetics assessment using graph attention network. In *International Conference on Pattern Recognition (ICPR)*, 2022. 3, 6

[12] Shuai He, Yongchang Zhang, Rui Xie, Dongxiang Jiang, and Anlong Ming. Rethinking image aesthetics assessment: Models, datasets and benchmarks. In Lud De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 942–948. International Joint Conferences on Artificial Intelligence Organization, 7 2022. Main Track. 1, 3, 6

[13] Simon Hentschel, Konstantin Kobs, and Andreas Hotho. Clip knows image aesthetics. *Frontiers in Artificial Intelligence*, 5, 2022. 3, 6

[14] Yong-Lian Hii, John See, Magzhan Kairanbay, and Lai-Kuan Wong. Multigap: Multi-pooled inception network with text augmentation for aesthetic prediction of photographs. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 1722–1726. IEEE, 2017. 2, 3

[15] Vlad Hosu, Bastian Goldlucke, and Dietmar Saupe. Effective aesthetics prediction with multi-level spatially pooled features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1, 3, 6

[16] V. Hosu, H. Lin, T. Sziranyi, and D. Saupe. Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing*, 29:4041–4056, 2020. 2, 8

[17] 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 *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 2, 3, 4, 6

[18] Sergey Karayev, Matthew Trentacoste, Helen Han, Aseem Agarwala, Trevor Darrell, Aaron Hertzmann, and Holger Winnemoeller. Recognizing image style. In *Proceedings of the British Machine Vision Conference*. BMVA Press, 2014. 3, 7

[19] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5148–5157, October 2021. 1, 3, 6

[20] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021. 3

[21] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017. 3

[22] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *Transactions of the Association for Computational Linguistics*, 2015. 3

[23] Alexander Kolesnikov, Alexey Dosovitskiy, Dirk Weissenborn, Georg Heigold, Jakob Uszkoreit, Lucas Beyer, Matthias Minderer, Mostafa Dehghani, Neil Houlsby, Sylvain Gelly, Thomas Unterthiner, and Xiaohua Zhai. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. 3, 4, 6

[24] Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charless Fowlkes. Photo aesthetics ranking network with attributes and content adaptation. In *European conference on computer vision*, pages 662–679. Springer, 2016. 1, 3, 6

[25] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International conference on machine learning*, pages 3744–3753. PMLR, 2019. 4

[26] Jun-Tae Lee, Han-Ul Kim, Chul Lee, and Chang-Su Kim. Photographic composition classification and dominant geometric element detection for outdoor scenes. *Journal of Visual Communication and Image Representation*, 55:91–105, 2018. 3

[27] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 3

[28] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 3

[29] Xin Lu, Zhe Lin, Xiaohui Shen, Radomir Mech, and James Z Wang. Deep multi-patch aggregation network for image style, aesthetics, and quality estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 990–998, 2015. 7

[30] Long Mai, Hailin Jin, and Feng Liu. Composition-preserving deep photo aesthetics assessment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 3

[31] Naila Murray and Albert Gordo. A deep architecture for unified aesthetic prediction. *arXiv preprint arXiv:1708.04890*, 2017. 3, 6

[32] Naila Murray, Luca Marchesotti, and Florent Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2408–2415. IEEE, 2012. 2, 3, 5, 6, 7

[33] Yuzhen Niu, Shanshan Chen, Bingrui Song, Zhixian Chen, and Wenxi Liu. Comment-guided semantics-aware image aesthetics assessment. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022. 1, 2, 3, 6

[34] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 6

[35] 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 *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2, 3, 4

[36] Jian Ren, Xiaohui Shen, Zhe Lin, Radomir Mech, and David J. Foran. Personalized image aesthetics. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 3

[37] Lin CY ROUGE. A package for automatic evaluation of summaries. In *Proceedings of Workshop on Text Summarization of ACL, Spain*, 2004. 6

[38] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 4, 5, 7

[39] Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pages 4596–4604. PMLR, 2018. 6

[40] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations (ICLR)*, 2020. 3

[41] Tiancheng Sun, Yulong Wang, Jian Yang, and Xiaolin Hu. Convolution neural networks with two pathways for image style recognition. *IEEE Transactions on Image Processing*, 26(9):4102–4113, 2017. 7

[42] Hossein Talebi and Peyman Milanfar. Nima: Neural image assessment. *IEEE Transactions on Image Processing*, 27(8):3998–4011, 2018. 1, 2, 3, 6, 7

[43] Hao Tan and Mohit Bansal. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. 3

[44] Xiaoou Tang, Wei Luo, and Xiaogang Wang. Content-based photo quality assessment. *IEEE Transactions on Multimedia*, 15(8):1930–1943, 2013. 3

[45] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxvit: Multi-axis vision transformer. *European conference on computer vision*, 2022. 6

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. 6

[47] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 6

[48] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *AAAI*, 2023. 3

[49] Wenhui Wang, Hangbo Bao, Li Dong, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *arXiv preprint arXiv:2111.02358*, 2021. 3

[50] Wenshan Wang, Su Yang, Weishan Zhang, and Jiulong Zhang. Neural aesthetic image reviewer. *IET Computer Vision*, 13(8):749–758, 2019. 3

[51] Yuzhe Yang, Liwu Xu, Leida Li, Nan Qie, Yaqian Li, Peng Zhang, and Yandong Guo. Personalized image aesthetics assessment with rich attributes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19861–19869, June 2022. 3

[52] Yong-Yaw Yeo, John See, Lai-Kuan Wong, and Hui-Ngo Goh. Generating aesthetic based critique for photographs. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 2523–2527. IEEE, 2021. 7, 8

[53] Zhenqiang Ying, Haoran Niu, Praful Gupta, Dhruv Mahajan, Deepti Ghadiyaram, and Alan Bovik. From patches to pictures (paq-2-piq): Mapping the perceptual space of picture

quality. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[54] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 2, 3, 6

[55] Hui Zeng, Zisheng Cao, Lei Zhang, and Alan C Bovik. A unified probabilistic formulation of image aesthetic assessment. *IEEE Transactions on Image Processing*, 29:1548–1561, 2019. 3, 6

[56] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12104–12113, June 2022. 3

[57] Xiaodan Zhang, Xinbo Gao, Lihuo He, and Wen Lu. Mscan: Multimodal self-and-collaborative attention network for image aesthetic prediction tasks. *Neurocomputing*, 430:14–23, 2021. 2, 3

[58] Ye Zhou, Xin Lu, Junping Zhang, and James Z Wang. Joint image and text representation for aesthetics analysis. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 262–266, 2016. 2, 3