Supplementary Material

1. Effect of Hyper-parameter α, β

Tab. 6 presents the ablation study on the use of different hyper-parameter α, β ratios in Eq. (2) during the pretraining of VILA. The results demonstrate that the setting of $\alpha : \beta = 1 : 2$ yields the best performance, which is consistent with the recommended setting in the CoCa [10] paper.

$\alpha : \beta$	SRCC	PLCC
2 : 1	0.645	0.651
1 : 1	0.656	0.661
1 : 2	**0.657**	**0.663**

Table 6. Effect of α and β ratios in aesthetic pretraining, comparing zero-shot IAA performance with an ensemble of prompts on the AVA dataset.

2. Effect of Margin Hyper-parameter m

Tab. 7 presents an ablation study for using different margin hyper-parameter m in Eq. (7) when finetuning VILA-R. The results show that a margin of $m = 0.1$ achieves the best performance, and we adopt this value as the default for all other experiments.

Margin	SRCC	PLCC
$m = 0.01$	0.769	0.769
$m = 0.05$	0.770	0.770
$m = 0.1$	**0.774**	**0.774**
$m = 0.15$	0.772	0.771
$m = 0.2$	0.772	0.770

Table 7. Ablation for different margin hyper-parameter m for the proposed rank-based adapter tuning on AVA dataset.

3. Effect of Random Sampling Comments

During training, we create image-comment pairs by randomly selecting one comment from the available list of comments for an image, if there are multiple comments associated with the same image. Tab. 8 shows the effect of such random sampling during aesthetic pretraining. When a fixed comment is used for each image, the AVA ZSL performance drops from 0.663 PLCC to 0.596. Random sampling is an effective approach since different comments may cover different aesthetic aspects of the same image, allowing the model to fully expose itself to diverse and rich aesthetic information in the noisy dataset. This strategy enables the mining of open-set aesthetic concepts automatically.

Comment Sampling	SRCC	PLCC
Random	**0.657**	**0.663**
Fixed	0.585	0.596

Table 8. Effect of random sampling comment in aesthetic pretraining, comparing zero-shot IAA performance with an ensemble of prompts on the AVA dataset.

4. Per-class Evaluation on AVA-Style

We show the per-class evaluation on AVA-Style in Tab. 9, comparing to the same baselines as in our main paper.

5. Details on ZSL for AVA-Style Classification

Single prompt. In this approach, we use the 14 photographic style names as the language prompts: {"complementary colors", "duo tones", "hdr", "image grain", "light on white", "long exposure", "macro", "motion blur", "negative image", "rule of thirds", "shallow dof", "silhouettes", "soft focus", "vanishing point"}. The cosine similarity between the prompt text embedding and the image embedding is used as the prediction score.

Ensemble of prompts. In this approach, we manually curate five sentences/phrases that are frequently mentioned in the AVA-Caption user comments, for each of the styles. These prompts either use synonyms (*e.g.* "color" and "colors") of the styles or add more text contexts (*e.g.*, "i like the lines and fading or vanishing"). Tab. 10 shows these prompts.

6. Details on ZSL for IAA

To effectively perform zero-shot learning for IAA, we use a pair of prompts with opposite meanings ("good" v.s. "bad").

Single prompt. In this approach, we use {"good image", "bad image"} as input prompts. Let $\boldsymbol{p}_g$ and $\boldsymbol{p}_b$ be the normalized unimodal text embedding for the "good" and "bad" prompts respectively, $\boldsymbol{v}$ be the normalized image contrastive embedding. We compute the cosine similarity and use the softmax normalized score for "good image" as the final score r for IAA.

$$r = \frac{e^{\boldsymbol{v}^\top \boldsymbol{p}_g}}{e^{\boldsymbol{v}^\top \boldsymbol{p}_g} + e^{\boldsymbol{v}^\top \boldsymbol{p}_b}}$$

Ensemble of prompts. In this approach, we similarly construct six pairs of "good" v.s. "bad" prompts for {"image", "lighting", "content", "background", "foreground", "composition"). The second group in Tab. 11 shows these pairs of prompts. For each pair, we can obtain a score $r_i, i = 1, ..., 6$. Then we use the average ensemble of the scores to get the final score r for IAA.

	Compl. Colors	Duo tones	HDR	Image Grain	Light On White	Long Expos.	Macro	Motion Blur	Negative Image	Rule of Thirds	Shallow DOF	Silhouet.	Soft Focus	Vanish. Point	mAP
Murray *et al*. [32]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	53.9
Karayev *et al*. [18]	46.9	67.6	66.9	64.7	90.8	45.3	47.8	47.8	59.5	35.2	62.4	79.1	31.2	68.4	58.1
Lu *et al*. [29]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	64.1
MNet [42]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	65.5
Sal-RGB [10]	61.4	87.6	72.9	82.2	83.0	61.9	66.6	62.0	87.7	41.7	82.4	93.1	46.4	76.8	71.8
Zero-shot Learning															
General Pretraining (single prompt)	36.5	21.0	23.9	7.1	37.0	34.6	49.9	32.8	12.7	14.3	33.5	67.9	15.6	23.4	29.3
General Pretraining (ensemble prompts)	36.0	51.3	36.9	8.1	30.5	40.4	55.0	33.4	13.8	14.5	27.0	64.9	18.0	27.3	32.6
VILA-P (single prompt)	48.1	55.8	76.6	76.0	72.9	66.1	70.8	67.6	34.9	25.8	77.9	81.6	**51.2**	67.3	62.3
VILA-P (ensemble prompt)	**53.6**	**81.8**	**79.3**	**86.7**	**75.4**	**69.2**	**72.9**	**74.1**	**58.6**	**30.9**	**78.6**	**85.4**	51.0	**67.8**	**69.0**

Table 9. AVA-Style per-class evaluation results. Supervised baselines are shown in gray color.

Style	Prompts	Style	Prompts
Complementary_Colors	*"complementary colors"* *"complementary color"* *"great complementary colors"* *"great use of complementary colors"* *"good use of complementary colors"*	Motion_Blur	*"motion blur"* *"nice motion blur"* *"great use of the motion blur"* *"i love the motion blur"* *"cool motion blur"*
Duotones	*"duo tones"* *"duotone"* *"nice duotone"* *"duotone works very well"* *"use of duotone"*	Negative_Image	*"negative image looks good"* *"love the negative images"* *"the negative image is captivating"* *"use of the negative image is interesting"* *"fan of the negative image"*
HDR	*"hdr"* *"i like the hdr"* *"great job with the hdr"* *"hdr done well"* *"love the hdr shot"*	Rule_of_Thirds	*"rule of thirds"* *"benefited from the rule of thirds"* *"followed the rule of thirds nicely"* *"use of the rule of thirds is fantastic"* *"great use of rule of thirds"*
Image_Grain	*"image grain"* *"i like the image grain"* *"nice use of image grain"* *"a good job with the image grain"* *"excellent use of image grain"*	Shallow_DOF	*"shallow dof"* *"nice shallow dof"* *"i love the shallow DOF"* *"lovely use of shallow DOF"* *"shallow dof works perfect here"*
Light_On_White	*"light on white"* *"great for the light on white"* *"nice light on white"* *"love the light on white"* *"like the light on white"*	Silhouettes	*"silhouettes"* *"like the silhouettes"* *"great silhouettes"* *"i really like silhouettes"* *"silhouettes are lovely"*
Long_Exposure	*"long exposure"* *"nice long exposure"* *"nice use of long exposure"* *"enjoy these long exposure shots"* *"look great with the long exposure"*	Soft_Focus	*"soft focus"* *"love the soft focus"* *"love the effect of soft focus"* *"excellent use of soft focus"* *"lovely soft focus"*
Macro	*"macro"* *"excellent detailed macro"* *"nice macro"* *"good macro shot"* *"great macro"*	Vanishing_Point	*"vanishing point"* *"i like the lines and fading or vanishing"* *"i love the lines and vanishing point"* *"nice to see the vanishing point off of center"* *"the background with the vanishing point is nice"*

Table 10. Text prompts used in the ensemble approach for AVA-Style ZSL.

7. Results on KonIQ-10k

Table 12 presents additional results on the image quality dataset KonIQ-10k [2]. We adopt the same data split as [2] and and employ a batch size of 32 to finetune the rank-based adapter for 30k steps, with a learning rate of 5e-4 and linear decay to zero, and 0.04 weight decay. Our proposed VILA-R outperforms CLIP-IQA$^+$ [8] which trains a prompt tuning module on top of CLIP features. While CLIP features only use general pretraining, VILA-R benefits from the aesthetic pretraining which learns rich perceptual quality information, highlighting the importance of the proposed aesthtic pretraining. Remarkably, with only 0.1% tunable parameters, VILA-R's performance is competitive with KonCept512 [2] and MUSIQ [3], which rely on much

	Prompts	
	p_g	p_b
Single Prompt	*"good image"*	*"bad image"*
Ensemble of Prompts	*"good image"* *"good lighting"* *"good content"* *"good background"* *"good foreground"* *"good composition"*	*"bad image"* *"bad lighting"* *"bad content"* *"bad background"* *"bad foreground"* *"bad composition"*

Table 11. Text prompts used in ZSL for IAA.

Method	SRCC	PLCC
BRISQUE [6]	0.665	0.681
ILNIQE [12]	0.507	0.523
HOSA [9]	0.671	0.694
BIECON [4]	0.618	0.651
WaDIQaM [1]	0.797	0.805
PQR [11]	0.880	0.884
SFA [5]	0.856	0.872
DBCNN [13]	0.875	0.884
MetaIQA [14]	0.850	0.887
BIQA [7]	0.906	0.917
CLIP-IQA$^+$ [8]	0.895	0.909
KonCept512 [2]	0.921	0.937
MUSIQ [3]	0.924	0.937
VILA-R	0.919	0.932

Table 12. Results on KonIQ-10k [2] dataset. We take numbers from [3, 8] for results of the reference methods.

larger resolutions. It is worth noting that KonIQ-10k [2] is not solely focused on aesthetics quality, and it includes images with technical quality problems such as compression and blur. There is limited user comments mentioning such aspects on the AVA-Captions dataset. Despite the gap, our model demonstrates competitive performance on KonIQ-10k, showcasing its robustness in capturing the visual appeal of the image across different datasets.

8. More Qualitative Examples

Fig. 6 displays additional style retrieval results (top-5) on KonIQ-10k [2] using AVA-style names as the query. In order to provide clear attribution to the image sources, we have opted to showcase images from the KonIQ-10k dataset instead of the AVA dataset. Attribution to the images are provided in Table 13. Overall, the retrieved results align with our aesthetic perspective. Notably, VILA accurately captures the lighting or color related information. For example, images retrieved for "Silhouettes" and "Complementary colors" accurately depict the corresponding concepts. Additionally, VILA recognizes concepts aesthetic concepts like "Motion blur" with high accuracy. However, there are also some failure cases where improvements are possible. For example, among the images retrieved using

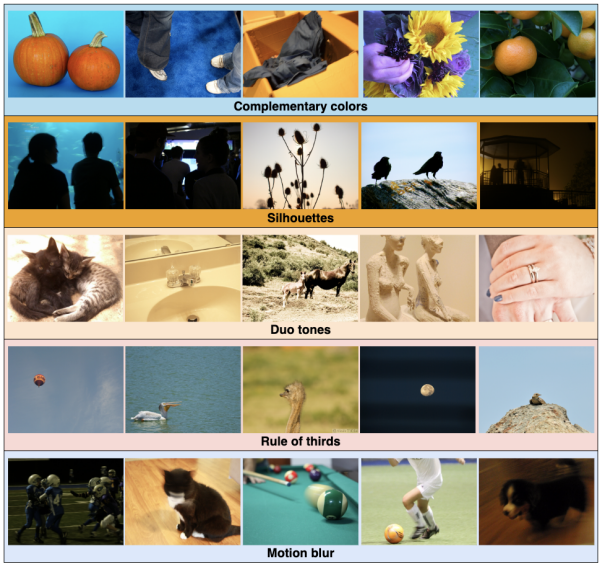

Figure 6. More examples for the top-5 images retrieved using style name query on KonIQ-10k [2]. The source of the displayed images are provided in Table 13.

the query "Rule of thirds", the last three images are centered rather than following the rule of thirds, which may be attributed to the random cropping augmentation during training. Augmentation improvement may help mitigate this issue. For "Duo tones", the top retrieved images have a yellowish tone, possibly due to training data bias in the AVA-Captions dataset. Thus, using a more diverse aesthetic pretraining dataset may further enhance the model's performance.

9. KonIQ-10k Images Attribution

In this paper, we display several images from KonIQ-10k [2]. The Flickr links and the license information for these images can be found in Table 13. We extend our gratitude to the original photographers for sharing their images.

Flickr Link	User	License
Figure 3 (from left to right, top to bottom)		
http://www.flickr.com/photos/43437767@N00/7499578096/	43437767@N00	CC BY-SA 2.0
http://www.flickr.com/photos/12708857@N00/228617373/	12708857@N00	CC BY-SA 2.0
http://www.flickr.com/photos/39443895202@N01/4295525241/	39443895202@N01	CC BY-NC 2.0
http://www.flickr.com/photos/43343993@N00/6814873580/	43343993@N00	CC BY-NC-SA 2.0
http://www.flickr.com/photos/93656595@N00/5212354067/	93656595@N00	CC BY-NC 2.0
http://www.flickr.com/photos/19761391@N06/6190643783/	19761391@N06	CC BY-NC-SA 2.0
http://www.flickr.com/photos/8490344@N04/5805954950/	8490344@N04	CC BY-NC-SA 2.0
http://www.flickr.com/photos/28577026@N02/4732322374/	28577026@N02	CC BY 2.0
http://www.flickr.com/photos/40595948@N00/4125943270/	40595948@N00	CC BY 2.0
http://www.flickr.com/photos/8397802@N05/6782627736/	8397802@N05	CC BY-NC-SA 2.0
Figure 4 (from left to right, top to bottom)		
http://www.flickr.com/photos/28081633@N00/3388712525/	28081633@N00	CC BY-SA 2.0
http://www.flickr.com/photos/69078621@N00/2501256504/	69078621@N00	CC BY-NC 2.0
http://www.flickr.com/photos/21657526@N00/8460154333/	21657526@N00	CC BY-NC-SA 2.0
http://www.flickr.com/photos/31990116@N03/8746457937/	31990116@N03	CC BY-SA 2.0
http://www.flickr.com/photos/61585804@N00/4639050491/	61585804@N00	CC BY-NC-SA 2.0
http://www.flickr.com/photos/78135748@N00/3824485636/	78135748@N00	CC BY 2.0
http://www.flickr.com/photos/86381710@N00/163977327/	86381710@N00	CC BY-NC-SA 2.0
http://www.flickr.com/photos/11152520@N03/5700224418/	11152520@N03	CC BY 2.0
http://www.flickr.com/photos/77175355@N07/10862487886/	77175355@N07	CC BY-NC 2.0
http://www.flickr.com/photos/76042652@N00/9467289840/	76042652@N00	CC BY-NC-SA 2.0
http://www.flickr.com/photos/74167788@N00/2746983219/	74167788@N00	CC BY-NC 2.0
http://www.flickr.com/photos/16409072@N08/7107776883/	16409072@N08	CC BY-NC-SA 2.0
http://www.flickr.com/photos/13447407@N00/205662428/	13447407@N00	CC BY-NC 2.0
http://www.flickr.com/photos/60635600@N08/6070993297/	60635600@N08	CC BY-NC 2.0
http://www.flickr.com/photos/30626457@N00/9629668489/	30626457@N00	CC BY 2.0
http://www.flickr.com/photos/31916492@N02/9702908989/	31916492@N02	CC BY-NC 2.0
http://www.flickr.com/photos/24742305@N00/3561351919/	24742305@N00	CC BY 2.0
http://www.flickr.com/photos/19072679@N00/8216243918/	19072679@N00	CC BY-NC-SA 2.0
http://www.flickr.com/photos/50795598@N02/8532549490/	50795598@N02	CC BY-NC-SA 2.0
http://www.flickr.com/photos/40573754@N04/4162338388/	40573754@N04	CC BY-NC-SA 2.0
Figure 6 (from left to right, top to bottom)		
http://www.flickr.com/photos/33602849@N00/1348094685/	33602849@N00	CC BY-NC-SA 2.0
http://www.flickr.com/photos/32842313@N00/4435718106/	32842313@N00	CC BY-NC-SA 2.0
http://www.flickr.com/photos/37306288@N02/6788530155/	37306288@N02	CC BY 2.0
http://www.flickr.com/photos/32535586@N07/8120735257/	32535586@N07	CC BY-NC 2.0
http://www.flickr.com/photos/62528187@N00/8596361580/	62528187@N00	CC BY 2.0
http://www.flickr.com/photos/36543005@N00/242395156/	36543005@N00	CC BY 2.0
http://www.flickr.com/photos/47100034@N08/8968242975/	47100034@N08	CC BY-NC-SA 2.0
http://www.flickr.com/photos/8397802@N05/6856483689/	8397802@N05	CC BY-NC-SA 2.0
http://www.flickr.com/photos/40355539@N00/4948021193/	40355539@N00	CC BY-NC-SA 2.0
http://www.flickr.com/photos/10734170@N08/4049047636/	10734170@N08	CC BY-NC-SA 2.0
http://www.flickr.com/photos/83670786@N03/8184338593/	83670786@N03	CC BY-NC-SA 2.0
http://www.flickr.com/photos/24328811@N00/5091406011/	24328811@N00	CC BY-NC-SA 2.0
http://www.flickr.com/photos/46318514@N06/4911877298/	46318514@N06	CC BY-NC-SA 2.0
http://www.flickr.com/photos/16482030@N00/5593982816/	16482030@N00	CC BY-NC 2.0
http://www.flickr.com/photos/68683191@N00/7915207374/	68683191@N00	CC BY-SA 2.0
http://www.flickr.com/photos/51963363@N00/5921291430/	51963363@N00	CC BY-NC-SA 2.0
http://www.flickr.com/photos/52713160@N00/6422846525/	52713160@N00	CC BY-NC-SA 2.0
http://www.flickr.com/photos/28658116@N02/7914158128/	28658116@N02	CC BY-SA 2.0
http://www.flickr.com/photos/10957255@N08/9365989820/	10957255@N08	CC BY-NC-SA 2.0
http://www.flickr.com/photos/38439215@N06/4809080599/	38439215@N06	CC BY-NC-SA 2.0
http://www.flickr.com/photos/92755733@N00/3017896883/	92755733@N00	CC BY 2.0
http://www.flickr.com/photos/33455872@N05/7443887988/	33455872@N05	CC BY-NC-SA 2.0
http://www.flickr.com/photos/8833673@N05/4462477090/	8833673@N05	CC BY-NC-SA 2.0
http://www.flickr.com/photos/54852753@N05/5766806996/	54852753@N05	CC BY 2.0
http://www.flickr.com/photos/90088957@N00/6434876963/	90088957@N00	CC BY-NC 2.0

Table 13. Flickr links to the KonIQ-10k [2] images shown in the paper.

References

[1] Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on image processing*, 27(1):206–219, 2017. 3

[2] 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, 3, 4

[3] 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. 2, 3

[4] Jongyoo Kim and Sanghoon Lee. Fully deep blind image quality predictor. *IEEE Journal of Selected Topics in Signal Processing*, 11(1):206–220, 2016. 3

[5] Dingquan Li, Tingting Jiang, Weisi Lin, and Ming Jiang. Which has better visual quality: The clear blue sky or a blurry animal? *IEEE Transactions on Multimedia*, 21(5):1221–1234, 2018. 3

[6] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708, 2012. 3

[7] Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3667–3676, 2020. 3

[8] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *AAAI*, 2023. 2, 3

[9] Jingtao Xu, Peng Ye, Qiaohong Li, Haiqing Du, Yong Liu, and David Doermann. Blind image quality assessment based on high order statistics aggregation. *IEEE Transactions on Image Processing*, 25(9):4444–4457, 2016. 3

[10] 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. 1

[11] Hui Zeng, Lei Zhang, and Alan C Bovik. A probabilistic quality representation approach to deep blind image quality prediction. *arXiv preprint arXiv:1708.08190*, 2017. 3

[12] Lin Zhang, Lei Zhang, and Alan C Bovik. A feature-enriched completely blind image quality evaluator. *IEEE Transactions on Image Processing*, 24(8):2579–2591, 2015. 3

[13] Weixia Zhang, Kede Ma, Jia Yan, Dexiang Deng, and Zhou Wang. Blind image quality assessment using a deep bilinear convolutional neural network. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(1):36–47, 2018. 3

[14] Hancheng Zhu, Leida Li, Jinjian Wu, Weisheng Dong, and Guangming Shi. Metaiqa: deep meta-learning for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14143–14152, 2020. 3