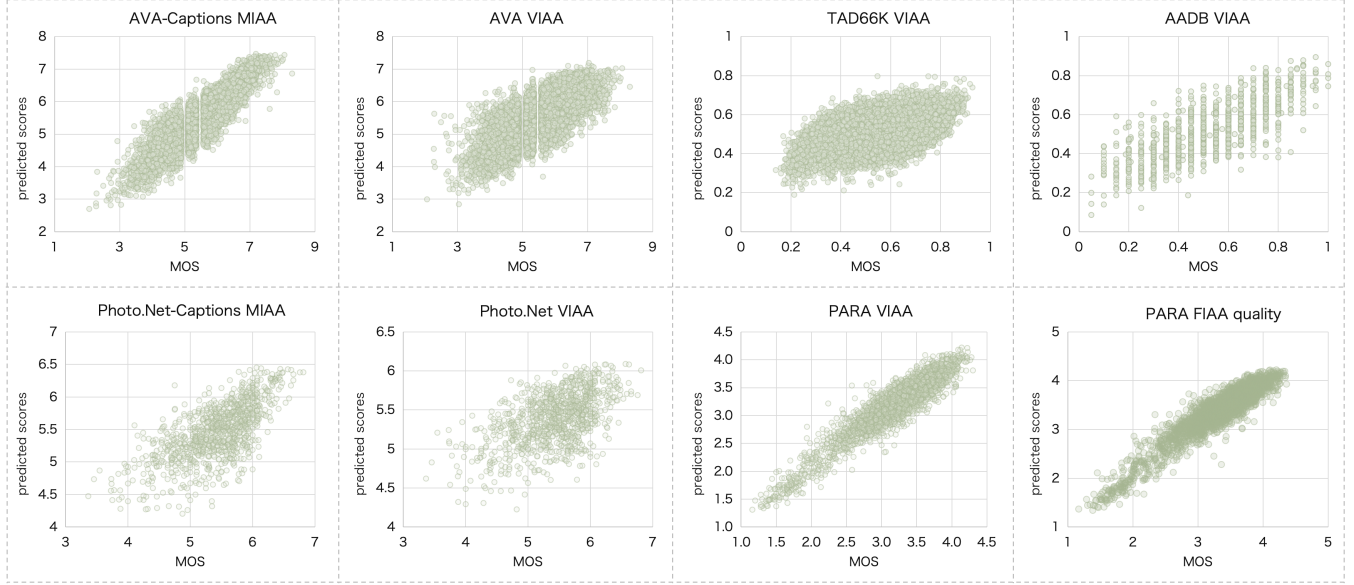


# Supplementary Materials: AesMamba: Universal Image Aesthetic Assessment with State Space Models

Anonymous Authors



**Figure 1: Scatter plots of predicted scores vs. MOSs, w.r.t. (1) the MIAA tasks on AVA-Captions and Photo.Net-Captions, (2) the VIAA tasks on AVA, Photo.Net, TAD66K, AADB, and PARA, and (3) the *quality* prediction in the FIAA task on PARA.**

## 1 ADDITIONAL EXPERIMENTS

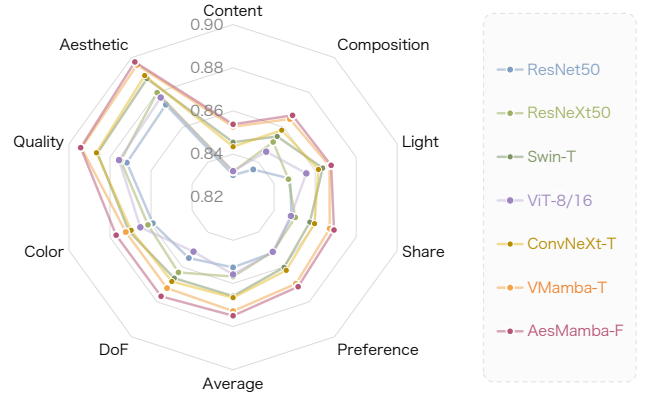
We first provide more results to visualize the superiority of the proposed AesMamba models, on diverse *Image Aesthetic Assessment* (IAA) tasks, across different benchmark datasets. Afterward, we provide an additional ablation study on the FIAA task. All these experiments are conducted on AVA [6], TAD66K [3], PARA [7], AADB [4], Photo.Net [1], AVA-Captions [2], and Photo.Net-Captions [1], following standard settings.

### 1.1 Scatter Plots on Each dataset

To illustrate the consistency between the predicted results with subjective evaluations, we show the scatter plots of predicted scores vs. MOSs. As shown in Fig. 1, the predicted scores are consistent with MOSs, across all the tasks, on diverse datasets. Besides, AesMamba-M achieves distinctly better consistency (in the MIAA task on the corresponding \*-Captions datasets) than AesMamba-V (in the VIAA task), on both AVA and Photo.Net. Such superiority demonstrate the significant role of text comments in representing image aesthetic. Besides, we visualize the scatter plots of quality prediction, in the FIAA task on PARA. Similarly, the predicted quality scores show high consistency with subjective evaluations.

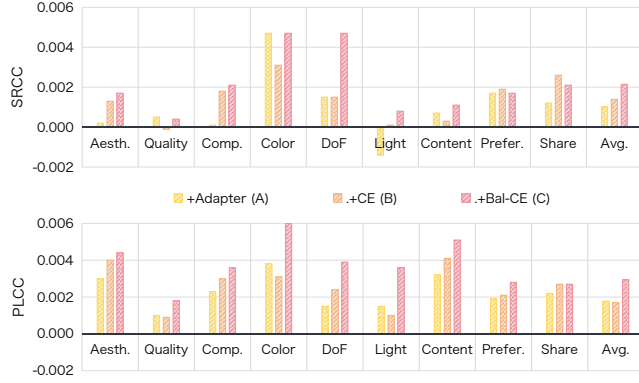
### 1.2 Analysis of visual backbones on FIAA

In addition, Fig. 2 visualize the comparison between different visual backbones, in the FIAA task. Each branch of the radar chart, shows



**Figure 2: Comparison between AesMamba-F and advanced visual backbones on the FIAA task, in terms of SRCC.**

the SRCC value of an aesthetic-related attribute. Obviously, our model consistently achieves the and SRCC values, across all the attributes. Besides, our base model, i.e. VMamba-T, consistently outperforms all the other advanced visual backbones, either CNNs or Transformers. Such superiority demonstrates the effectiveness of VMamba, in capturing both local or global representations, for representing diverse visual attributes.



**Figure 3: Impact of MBA for FIAA.** Each bar shows the relative change of PLCC/SRCC, compared to our base model, i.e. VMamba-T [5] with the MSE loss  $\mathcal{L}_{\text{MSE}}$ .

### 1.3 Ablation Study on FIAA

To visualize the impact of each module, we calculate the *relative* performance of the other models, compared to our base model, VMamba-T [5] with the MSE loss  $\mathcal{L}_{\text{MSE}}$  (Fig. 3).

**Effectiveness of task-adaptation.** First, multitask adaptation consistently boosts the PLCC values across all the attributes; and boots both PLCC and SRCC in average (Model-A). Such improvements demonstrate the necessity of learning adaptive features for diverse attributes.

**Effectiveness of using auxiliary categorization task.** Second, Model-B achieves slightly higher SRCC and comparable PLCC, compared to Model-B. This superiority indicates the potential of classification loss for performance boosting.

**Effectiveness of balanced learning.** Finally, using  $\mathcal{L}_{\text{Bal-CE}}$  consistently boosts both PLCC and SRCC values, across all the attributes (Model-C). Such significant superiority of our full model demonstrates the crucial role of balanced learning in multi-attribute evaluation.

## 2 CROSS-DATASET GENERALIZATION

In addition, we conduct cross-dataset VIAA and MIAA experiments, to evaluate the generalization ability of our AesMamba models. Specifically, we train AesMamba on the training set of one dataset, and apply the learned model to the testing set of all datasets. Table 1 and Table 2 show the corresponding experimental results. Besides, we adopt the inner-dataset performance, i.e. when the model is trained and test on the same dataset, as the benchmark (denoted by 100%); and calculate the quotient of performance achieved under the cross-dataset train/test settings (Fig. 1 and Fig. 5). For example, AesMamba-V achieves an accuracy of 84.60, when it is trained and test both on AVA. In contrast, AesMamba-V achieves an accuracy of 79.84, if it is trained on Photo.Net but test on AVA. The corresponding relative performance is computed by  $79.84/84.60 = 94.4\%$ .

**Strong Generalization Ability of AesMamba.** Fig.1 shows that AesMamba-V relatively achieve accuracy values (over 83%), in all the cross-dataset settings. Besides, the cross-dataset performance exceeds 90% of the benchmark performance, in terms of all the tree metrics, between PARA and AADB; and exceeds or approaches

**Table 1: VIAA Performance, of AesMamba-V, in cross-dataset experiments. In each setting, our AesMamba-V model is trained on one dataset, but applied to all benchmark datasets.**

Accuracy (%)		<i>test</i>				
		AVA	Photo.net	PARA	AADB	TAD66K
<i>train</i>	AVA	84.60	79.33	74.83	72.00	64.20
	Photo.net	79.84	80.30	74.87	71.40	64.61
	PARA	75.46	77.48	88.70	78.10	61.72
	AADB	75.09	77.38	83.83	82.90	62.22
	TAD66K	80.13	79.65	74.07	72.80	72.00

PLCC		<i>test</i>				
		AVA	Photo.net	PARA	AADB	TAD66K
<i>train</i>	AVA	0.760	0.537	0.620	0.529	0.422
	Photo.net	0.605	0.547	0.629	0.475	0.408
	PARA	0.415	0.288	0.936	0.703	0.287
	AADB	0.367	0.260	0.865	0.774	0.258
	TAD66K	0.629	0.491	0.607	0.475	0.511

SRCC		<i>test</i>				
		AVA	Photo.net	PARA	AADB	TAD66K
<i>train</i>	AVA	0.751	0.502	0.598	0.509	0.396
	Photo.net	0.591	0.518	0.616	0.461	0.385
	PARA	0.384	0.281	0.902	0.702	0.272
	AADB	0.335	0.261	0.839	0.768	0.240
	TAD66K	0.619	0.452	0.605	0.457	0.483

**Table 2: MIAA Performance, of AesMamba-M, in cross-dataset experiments. In each setting, our AesMamba-M model is trained on one dataset, but applied to both benchmark datasets.**

Accuracy (%)		<i>test</i>	
		AVA-Captions	Photo.net-Captions
<i>train</i>	AVA-Captions	89.6	83.3
	Photo.net-Captions	84.8	82.0

PLCC		<i>test</i>	
		AVA-Captions	Photo.net-Captions
<i>train</i>	AVA-Captions	0.899	0.684
	Photo.net-Captions	0.800	0.685

SRCC		<i>test</i>	
		AVA-Captions	Photo.net-Captions
<i>train</i>	AVA-Captions	0.892	0.676
	Photo.net-Captions	0.796	0.664

80% of the benchmark performance, among AVA, TAD66K and Photo.Net. In the MIAA task, AesMamba-M achieves over 89% of the benchmark performance, when it is training on AVA-Captions but test on Photo.Net-Captions. Inspiringly, AesMamba-M surpasses the benchmark performance, if it is trained on AVA-Captions but

		test					test					test				
train		AVA	TAD66K	Photo.Net	PARA	AADB	AVA	TAD66K	Photo.Net	PARA	AADB	AVA	TAD66K	Photo.Net	PARA	AADB
	AVA	100	89.2	98.8	84.4	86.9	100	82.6	98.2	66.3	68.3	100	81.9	96.9	66.3	66.3
	TAD66K	94.7	100	99.2	83.5	87.8	82.8	100	89.7	64.9	61.4	82.4	100	87.2	67.1	59.5
	Photo.Net	94.4	89.7	100	84.4	86.1	79.6	79.8	100	67.2	61.4	78.7	79.7	100	68.3	60.0
	PARA	89.2	85.7	96.5	100	94.2	54.6	56.1	52.7	100	90.8	51.1	56.3	54.2	100	91.4
	AADB	88.8	86.4	96.4	94.5	100	48.3	50.5	47.4	92.4	100	44.6	49.8	50.4	93.0	100
		Accuracy					PLCC					SRCC				

**Figure 4: Relative VIAA performance (%), of AesMamba-V, in the cross-dataset experiments. The inner-dataset performance, i.e. when the model is trained and test on the same dataset, is adopted as the benchmark (denoted by 100%). We calculate the quotient of performance achieved under the cross-dataset train/test settings.**

		test		test		test	
train		AVA-Captions	Photo.Net-Captions	AVA-Captions	Photo.Net-Captions	AVA-Captions	Photo.Net-Captions
	AVA-Captions	100	101.5	100	99.9	100	101.8
	Photo.Net-Captions	94.6	100	89.0	100	89.2	100
		Accuracy		PLCC		SRCC	

**Figure 5: Relative MIAA performance (%), of AesMamba-M, in the cross-dataset experiments. The inner-dataset performance, i.e. when the model is trained and test on the same dataset, is adopted as the benchmark (denoted by 100%). We calculate the quotient of performance achieved under the cross-dataset train/test settings.**

applied to Photo.Net-Captions. This might due to the fact that, AVA-Captions includes about 250K instances, while Photo.Net-Captions only include 10K instances.

**Cross-dataset Correlations.** The diversities among the cross-dataset results, under different settings, also imply the correlations between existing benchmark datasets. For example, there might be strong similarities among datasets in the group {AVA, TAD66K, Photo.Net}, or in the group {PARA, AADB}. In contrast, there might be great divergences between these two groups. It’s meaningful to develop robust IAA methods, by using a joint training set derived from diverse datasets; and to consider such correlations during the design of learning strategies.

## REFERENCES

- [1] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang. 2006. Studying aesthetics in photographic images using a computational approach. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 288–301.
- [2] Koustav Ghosal, Aakanksha Rana, and Aljosa Smolic. 2019. Aesthetic Image Captioning From Weakly-Labelled Photographs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*.
- [3] Shuai He, Yongchang Zhang, Rui Xie, Dongxiang Jiang, and Anlong Ming. 2022. Rethinking image aesthetics assessment: Models, datasets and benchmarks. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*. 942–948.
- [4] Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charless Fowlkes. 2016. Photo aesthetics ranking network with attributes and content adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 662–679.

- [5] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, and Yunfan Liu. 2024. Vmamba: Visual state space model. *arXiv preprint arXiv:2401.10166* (2024), 1–14.
- [6] Naila Murray, Luca Marchesotti, and Florent Perronnin. 2012. AVA: A large-scale database for aesthetic visual analysis. In *2012 IEEE conference on Computer Vision and Pattern Recognition*. IEEE, 2408–2415.
- [7] Yuzhe Yang, Liwu Xu, Leida Li, Nan Qie, Yaqian Li, Peng Zhang, and Yandong Guo. 2022. Personalized image aesthetics assessment with rich attributes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19861–19869.