Charm: The Missing Piece in ViT fine-tuning for Image Aesthetic Assessment

Supplementary Material

6. Advantages of *Charm* over existing approaches

Charm is not the first approach to focus on processing images at their original resolution. A prominent method in this area is AnyRes [1]. To show the added value of *Charm*, we analyzed the Hugging Face implementation of AnyRes (LlavaNextImageProcessor). While Charm can process images of any size, AnyRes resizes and **pads** images to a resolution that is a multiple of the patch embedding module's input. Then, the images are divided into smaller sub-images, which, along with a downscaled version of the original image, are **independently** encoded by patch embedding modules. This prevents the model from capturing relationships between smaller sub-images. Additionally, AnyRes does not account for cross-scale relationships and treats tokens from different scales equally. In contrast, Charm leverages position and scale embeddings to effectively capture image composition and cross-scale relationships. For batch processing in AnyRes, **additional padding** is required as images produce different numbers of sub-images. AnyRes achieved PLCC/SRCC/ACC scores of 0.637/0.619/0.697 on the AADB dataset, which are lower than the standard Dinov2-small tokenizer (0.695/0.682/0.754, respectively). This is likely due to **excessive** padding at various stages, which significantly impacts IAA (see Table 2 in the paper).

Closet to Charm are mixed-resolution [15] and mixed-scale tokenization [4]. To understand the difference between charm and these approaches, consider two versions of a 1024×1024 image: one sharp and the other unsharp. This difference clearly affects their aesthetic scores, requiring distinct representations for the network to differentiate them. Refs. [15] and [4] first downscale both images to a small fixed size, then retain resolution in some regions while further downsizing others. This produces identical representations for both versions, sufficient for classification but inadequate for aesthetics. In contrast, Charm incorporates high-frequency information from the original image that gets lost with downsampling. Furthermore, while other methods rely on patches from 2 fixed resolutions, Charm is more flexible, learning from patches across varying resolutions. Table 3 in the paper shows Charm significantly outperforms these methods (identified as MS).

7. 3-scale Charm

Charm tokenization prepares a sequence of image patches, each with a size of $p \times p \times c$, where p is the patch size of ViT's patch encoding module and c is the number of channels. For low-resolution images, we directly tokenize the image using the patch size of p (as mentioned in Section 3.2.1).

For high-resolution images, we employ a multiscale approach. When dealing with 3-scales, we first define our patch sizes as follows:

$$patch\_sizes = \{\alpha p, \beta p, \gamma p\} \tag{1}$$

where $\gamma > \beta > \alpha, \{\gamma, \beta, \alpha\} \in \mathbb{N}$. The maximum downscaling is $f = \alpha/\gamma$.

The image is initially tokenized using the largest patch size (γp). A subset of patches is then selected and further tokenized using the base patch size (p). We then select another subset of patches from unselected regions for the second scale. These selected regions are downscaled to the intermediate patch size (βp) and then tokenized using p. Finally, the remaining regions are downscaled to the smallest patch size (αp) and then further tokenized using p.

As discussed in Section 4.4.5, a scaling factor of $f = 0.5$ yields optimal results. Consequently, we employ patch sizes of 2p, 3p, 4p for our 3-scale version. Table 6 indicates that the 2-scale *Charm* tokenization yields the best performance on AVA and AADB datasets.

Dataset	Charm	Scale	PLCC	SRCC	ACC
AVA	-	1	0.710	0.706	0.802
	✓	2	**0.779**	**0.779**	**0.826**
	✓	3	0.759	0.757	0.817
AADB	-	1	0.695	0.682	0.754
	✓	2	**0.767**	**0.754**	0.767
	✓	3	0.753	0.745	**0.775**

Table 6. Performance of Dinov2-small on AVA and AADB datasets across different scales.

8. Patch selection strategies

To identify important areas for patch selection, we explore various strategies. The goal is to preserve visually interesting regions and sharp details in high resolution.

We initially consider using saliency maps generated by the SAM-HQ model [8] with different prompts. These prompts include "figure ground reversal," "figure ground separation," "figure ground segmentation," "camo object," "salient object," and "hidden object." All of these prompts represent the figure-ground organization, where humans simplify a scene into the main object (the figure) and everything else (the background) [14, 21]. The results show that the 'salient object' is the most effective prompt. After creating saliency maps, we randomly select patches within the salient area in each epoch.

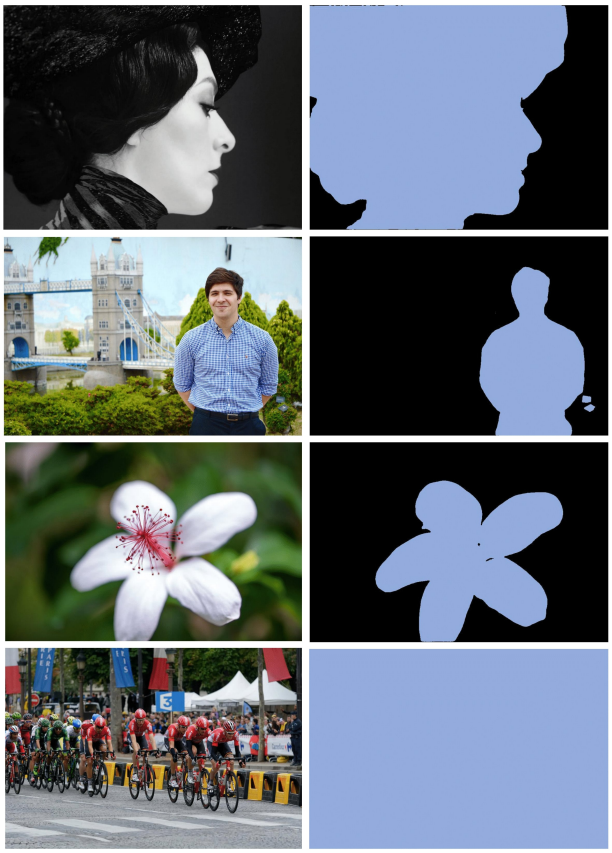

Figure 7. Examples of saliency maps generated by SAM-HQ with the prompt of "salient object". The final row shows a failure case where the Sam-HQ model did not detect any salient object.

Figure 7 shows examples of the salient objects identified by the model. In rare cases, the model fails to identify any salient objects. In such instances, we simply consider the entire image as salient, ensuring that patches are selected from all areas. The rate of model failures is negligible, **less than 0.14%** in the AVA dataset.

We also consider other methods, such as frequency, gradient, and entropy, to identify important regions. These metrics are calculated using the Fast Fourier transform, Sobel filter, and Shannon entropy, respectively. We start with a fully deterministic approach, selecting only patches with the highest frequency, entropy, or gradient. We then investigate whether introducing a degree of randomness can improve performance. To balance determinism and randomness, we introduce a threshold parameter (t). To select patches, we first rank them based on their frequency, entropy, or gradient values. We then select a larger number of patches (t times the desired number) with the highest scores. Finally, we randomly select the required number of patches from this pool in each epoch. As shown in Figure 9, smaller values of t result in lower variability and higher correlation between selected patches.

Tokenization	PLCC	SRCC	ACC
Standard	0.695	0.682	0.754
Charm + Random (first epoch)	0.705	0.693	0.756
Charm + Random (each epoch)	**0.767**	**0.754**	**0.767**

Table 7. Performance of Dinov2-small on the AADB dataset using different tokenization approaches. *Charm* with random patch selection in each epoch achieves the best performance.

We fine-tune Dinov2-small on the AADB dataset using different thresholds. Figure 10 indicates that increasing the threshold t generally improves validation performance. However, we use $t = 2$ to balance the inclusion of high-frequency, entropy, and gradient-based patches while maintaining diversity in the selected regions.

We also consider random patch selection. Table 7 demonstrates that randomly selecting patches in the first epoch and keeping them fixed throughout training yields better results than the standard Dinov2-small (around 1 % improvement). However, by randomly selecting patches in each epoch, we can achieve further performance improvements, surpassing other approaches (Table 8). Randomly sampling patches in each epoch exposes the ViT model to different regions of the image in high resolution. This diversity in training data helps prevent overfitting. Red squares in Figure 8 demonstrate the selected patches using our patch selection strategies.

9. Fine-tuning Dinov2-small for IAA

The original Dinov2 code downscales images to the shortest edge of 256 and applies center cropping to create fixed input sizes (224×224). Additionally, images are normalized, a common practice in deep learning to help the network learn faster and better. However, we observed that normalization can negatively impact IAA performance (Table 9) due to significant changes in the images (Figure 11). As a result, we remove normalization. Also, after downscaling images to the shortest edge of 256, we use random cropping instead of center cropping. As shown in Table 10, our approach outperforms the original Dinov2-small setting for IAA. Throughout this paper, the 'Standard approach' refers to our settings.

10. Data augmentation methods

Some argue that existing global data augmentation techniques, which alter the entire image, can potentially change the aesthetic labels and should thus be avoided in IAA [19]. However, ViTs are prone to overfitting and often require large datasets for fine-tuning, which can be challenging in IAA. Among existing global data augmentation methods, horizontal flipping, random rotation (at angles of 90, 180,

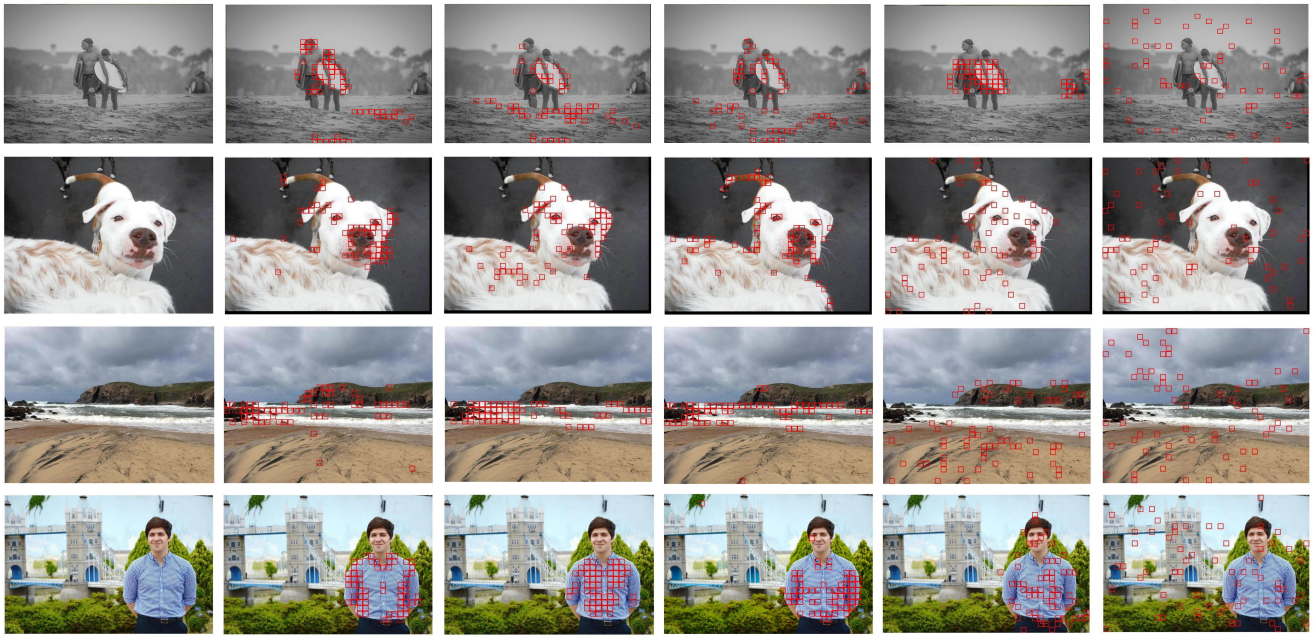

Figure 8. Visualization of patch selection strategies. Red squares highlight areas selected by different methods. Frequency, gradient, and entropy approaches are fully deterministic in these examples.

Patch selection	Charm	PLCC		SRCC		ACC	
		train	test	train	test	train	test
Random	✓	0.800	**0.767**	0.794	**0.754**	0.801	**0.767**
Frequency	✓	0.831	<u>0.756</u>	0.824	<u>0.747</u>	0.815	0.761
Entropy	✓	0.891	0.726	0.887	0.714	0.823	0.761
Gradient	✓	0.838	0.734	0.845	0.721	0.825	<u>0.766</u>
Saliency	✓	0.870	0.751	0.862	0.738	0.835	0.756
Standard	-	0.768	0.695	0.753	0.682	0.773	0.754

Table 8. The performance of Dinov2-small on the AADB dataset using different patch selection approaches. **Bold** and <u>underlined</u> numbers represent the best and second-best results. This table is the extended version of Table 4 in the paper. We use $t = 2$ to ensure a diverse set of patches with high frequency, entropy, and gradient.

Normalization	PLCC	SRCC	ACC
True	0.474	0.446	0.664
False	**0.488**	**0.458**	**0.794**

Table 9. The effect of image normalization on the performance of Dinov2-small + *Charm* on the TAD66k dataset. Image normalization negatively affects the IAA.

Implementation	PLCC	SRCC	ACC
Standard	0.697	0.694	0.799
Our settings	**0.710**	**0.706**	**0.802**

Table 10. Performance of Dinov2-small on the AVA dataset with two different data preprocessing approaches.

or 270 degrees), and grayscale augmentation preserve the composition of elements. We apply these methods to evaluate their impact on model performance. Figure 12 shows examples of these augmentations, which are applied with a probability of 50%.

While these augmentations may affect human aesthetic judgment, they consistently improve the generalizability of ViTs (as shown in Table 11). Grayscale augmentation slightly decreases performance, highlighting the importance of color in IAA. Therefore, we only use random horizontal flipping and random rotation in our experiments.

Figure 9. The impact of different thresholds (t) on the frequency, gradient, and entropy-based patch selection. Increasing the value of t introduces more diversity to the selected patches.

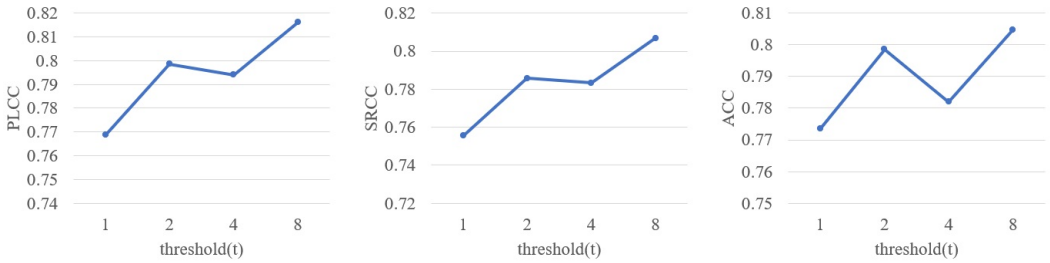

Figure 10. The performance of Dinov2-small + *Charm* on the AADB dataset using gradient-based patch selection with different thresholds (t). The results are averaged over 5 runs on the validation set. Generally, increasing the threshold (t) leads to improved performance.

11. Comparing dataset resolutions

We compare the image resolutions within both IAA and IQA datasets. Resolution represents the number of pixels in the image and is calculated by multiplying the width and height of images. As shown in Figure 13, PARA and SPAQ have the highest variety and resolution compared to the others. Among the other datasets, AVA has the lowest resolution, and all datasets exhibit a normal distribution of image resolutions. As all images in KonIQ10k have a fixed resolution of 1024 × 768 pixels, it appears as a single line in the box plot. Our analysis shows no significant differences in image resolution between the training and test sets of the datasets.

12. Importance of scale embedding

A key focus of this paper is to preserve the compositional relationships between elements. To achieve this, we avoid cropping and use position and scale embeddings to capture the relationships between tokens both across and within different scales. Adding scale embedding to

Augmentation	PLCC		SRCC		ACC	
	train	test	train	test	train	test
No augment	0.833	0.768	0.829	0.766	0.859	0.822
HF	0.818	0.776	0.812	<u>0.775</u>	0.851	<u>0.824</u>
HF + G	0.823	0.774	0.817	0.772	0.855	<u>0.824</u>
HF + R	0.801	**0.779**	0.791	**0.777**	0.843	**0.826**
HF + R + G	0.783	<u>0.778</u>	0.772	**0.777**	0.834	<u>0.824</u>

Table 11. The performance of Dinov2-small + *Charm* on the AVA dataset using different data augmentation methods. HF, G, and R represent horizontal flipping, grayscale augmentation, and random rotation, respectively. **Bold** and <u>underlined</u> numbers represent the best and the second-best results.

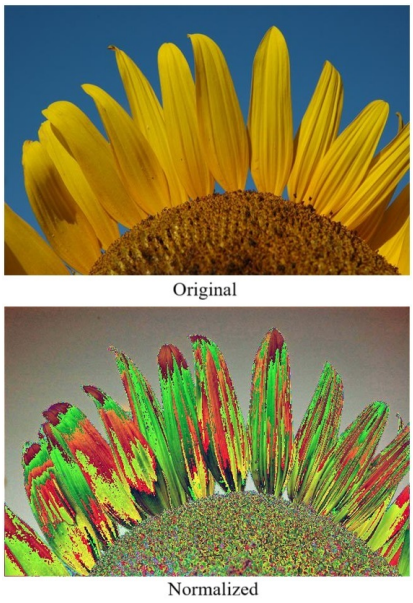

Figure 11. The effect of normalization on an image. Normalization can introduce visual artifacts and distort the aesthetic quality of an image.

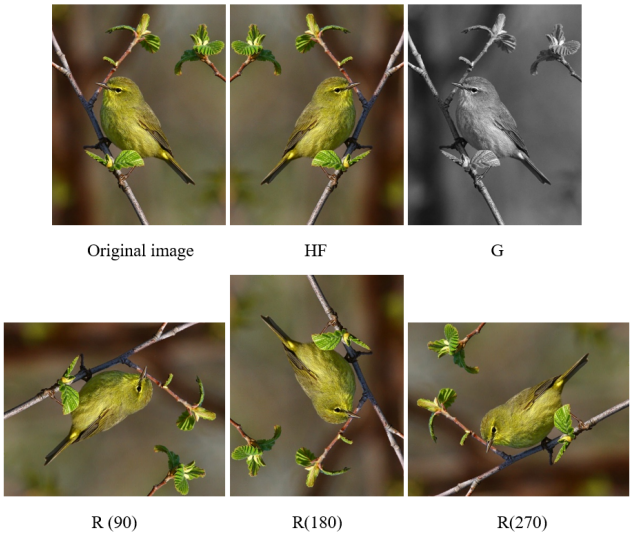

Figure 12. Data augmentation methods that preserve the composition of elements in the image. HF, G, and R represent horizontal flipping, grayscale augmentation, and random rotation.

Dinov2-small with the Charm tokenizer increases PLCC/SRCC/ACC from 0.748/0.739/0.750 to 0.767/0.754/0.767 on the AADB dataset. This 2% improvement surpasses the scale embedding gains reported in Ref. [7]. This is likely because we leverage scale embedding alongside the pre-trained position embeddings of the ViT rather than introducing new position embeddings.

13. The maximum number of patches (l)

We conduct ablation studies with different input lengths (l) during training. The optimal value for l should be chosen based on the image resolutions in the dataset. Setting l larger than the average number of patches (average of s across images) can lead to excessive padding while setting it smaller can result in excessive cropping. Both scenarios can negatively impact performance in IAA. For example, in the AADB dataset, the average number of patches after preprocessing using *Charm* is 1090. As shown in Table 12, an input length (l) of 1024, which is closest to the average number of tokens, yields the best performance for predicting the aesthetic score.

Input length(l)	PLCC	SRCC	ACC
768	0.743	0.739	0.770
1024	**0.767**	**0.754**	0.767
1500	0.736	0.727	**0.775**

Table 12. The impact of input length (l) on performance of Dinov2-small + *Charm* on the AADB dataset. Selecting the optimal l value is crucial for achieving the best results.

14. *Charm*'s influence on various ViTs

In this section, we demonstrate that incorporating Charm enhances the performance of various ViT models. We evaluated its effectiveness on different backbones, including

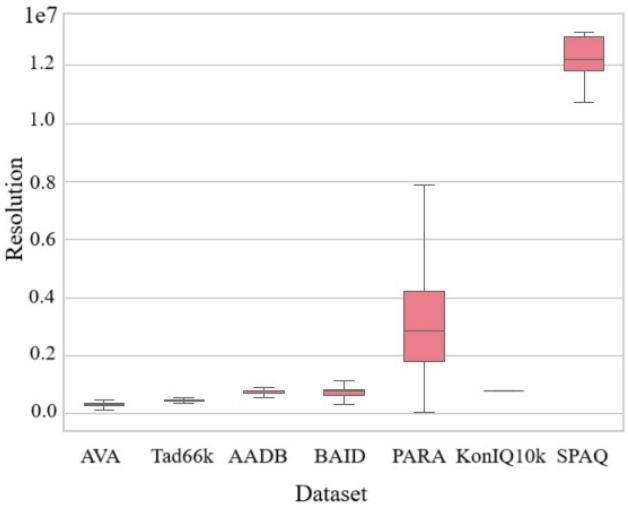

Figure 13. Distribution of image resolutions across datasets.

Model	*Charm*	PLCC	SRCC	ACC
Dinov2 -small	-	0.710	0.706	0.802
	✓	0.779 (↑ 6.9%)	0.777 (↑ 7.1%)	0.826 (↑ 2.4%)
ViT-small	-	0.687	0.679	0.794
	✓	0.762 (↑ 7.5%)	0.760 (↑ 8.1%)	0.827 (↑ 3.3%)
Dinov2 -large	-	0.734	0.732	0.808
	✓	0.783 (↑ 4.9%)	0.781 (↑ 4.9%)	0.828 (↑ 2%)

Table 13. Performance improvement across different models on the AVA dataset by replacing their standard tokenization with *Charm*. All experiments using *Charm* employ a random patch selection strategy.

ViT-small, Dinov2-small, and Dinov2-large. As shown in Table 13, *Charm* significantly improves performance on the AVA dataset across all tested models. These results indicate that Charm is not dependent on a specific architecture and performs consistently well across both smaller and larger models.

15. Super high-resolution images

As shown in Figure 6 and Table 5 in the paper, increasing the resolution can increase the computational costs of our method. This is especially challenging when dealing with extremely large images (e.g., 3k by 4k pixels). However, our experiments on the PARA dataset demonstrate that there is a threshold for performance improvement due to preserving high-resolution information. Beyond this threshold, further performance gains are limited. By downscaling images to a maximum edge of 1024, we can significantly reduce computational costs without sacrificing much performance. Table 14 shows that the performance difference between downscaling to the maximum edge of 1024 and 1500 is less than 0.3% in PLCC and SRCC and only 0.8% in ACC. This suggests that processing images at excessively high resolutions may not provide significant benefits, especially considering the increased computational costs.

Image size	PLCC	SRCC	ACC
Standard	0.904	0.855	0.863
Maximum edge = 1024	0.938	0.905	0.892
Maximum edge = 1500	**0.940**	**0.908**	**0.900**

Table 14. Performance of Dinov2-small on PARA dataset with different resolution. "Standard" refers to the approach explained in Appendix 9. The others use Charm. Processing images in high resolution positively affects the performance of the IAA model.

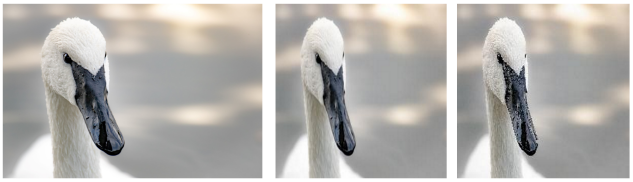

Figure 14. From left to right: the original image, downscaled image by bilinear interpolation, and downscaled image by Muller. While effective in image classification, the Muller method can introduce distortions that negatively impact aesthetics.

Figure 15. Comparison of Dinov2-small standard input (Section 9) and padded input (Section 16). While padding preserves the aspect ratio, it can negatively impact the performance of IAA models.

16. Dinov2 + Muller / Padding

Muller [20] is a learnable resizer that aims to boost details in certain frequency subbands during downscaling. While effective in image classification, Muller can introduce distortions (Figure 14) that negatively affect aesthetics (Table 2).

A straightforward approach to preserving the aspect ratio is to add padding to images. For this approach, we resize

images to a maximum edge of 512 and add padding (Figure 15). While this approach preserves the aspect ratio and maintains higher-resolution images compared to the standard method, it yields worse results (Table 2), highlighting the significant negative impact of padding in IAA. Moreover, padding not only fails to add valuable information but also significantly increases the training costs of the standard Dinov2-small model.

17. Detailed comparison of our approach with state of the art in IAA and IQA

Table 15 and 16 illustrate the detailed comparison of our approach with state-of-the-art models in IAA and IQA. *Charm* focuses on crucial image-based factors to improve the performance of ViTs, which are the foundation of many state-of-the-art IAA models. While the use of multimodal models is orthogonal to *Charm*'s contribution, it is likely that integrating *Charm* in a multimodal method will lead to improved performance. Unfortunately, none of these methods (refs. [6],[16], [11], and [23]) have released their code, preventing us from testing this hypothesis. Other methods (refs. [9], [3], and [10]) are CNN-based.

Additionally, our approach achieves comparable performance while having considerably fewer parameters (Figure 4). Our approach adds only 1152 parameters to the Dinov2-small model for the scale embedding (1, 2, 384) and mask token (1, 1, 384). The mask token is a learnable parameter used in the masking process of the input. In batch training, masks are used to identify effective inputs while ignoring padding tokens that may be present in some images.

18. The effect of varying number of tokens in Table 3

We evaluated the standard tokenizer with 384×384 input images (vs. 224×224 used in Table 3), producing 729 tokens — comparable to *Charm*'s 768 on the Tad66k dataset. Using the Dinov2-small backbone, this setting achieved PLCC/SRCC/ACC of 0.429/0.404/0.653 on TAD66k, while *Charm* reached 0.488/0.458/0.794. *Charm*'s strong IAA performance is therefore not due to a higher number of tokens but rather its ability to preserve aspect ratio, composition, and high-resolution details, along with its patch selection strategy that alleviates overfitting.

19. Downscaling factor

Figure 2 illustrates 2-scale *Charm*, where important areas of the image are further tokenized using a patch size of p while others are downscaled to p. As described in Section 3.2.1, the amount of downscaling is defined by f. As shown in Table 18, a large f negatively impacts model performance due to increased information loss from downscaling unselected regions.

20. Integrating *Charm* with the Swin transformer

Charm is incompatible with the Swin transformer [13] due to Swin's reliance on relative position embeddings and patch merging to capture the hierarchy. Swin transformer requires a specific token order and a fixed grid of patches, which are not guaranteed by *Charm*'s tokenization process. While integrating *Charm* with the Swin Transformer presents challenges, it remains a promising direction for future research.

References

[1] Lucy Chai, Michael Gharbi, Eli Shechtman, Phillip Isola, and Richard Zhang. Any-resolution training for high-resolution image synthesis, 2022. 1

[2] Chaofeng Chen, Jiadi Mo, Jingwen Hou, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin. Topiq: A top-down approach from semantics to distortions for image quality assessment. *IEEE Transactions on Image Processing*, 2024. 8

[3] Hangwei Chen, Feng Shao, Baoyang Mu, and Qiuping Jiang. Image aesthetics assessment with emotion-aware multibranch network. *IEEE Transactions on Instrumentation and Measurement*, 73:1–15, 2024. 7, 8

[4] Jakob Drachmann Havtorn, Amélie Royer, Tijmen Blankevoort, and Babak Ehteshami Bejnordi. Msvit: Dynamic mixed-scale tokenization for vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 838–848, 2023. 1

[5] Shuai He, Anlong Ming, Shuntian Zheng, Haobin Zhong, and Huadong Ma. Eat: An enhancer for aesthetics-oriented transformers. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1023–1032, 2023. 8

[6] Yipo Huang, Leida Li, Pengfei Chen, Jinjian Wu, Yuzhe Yang, Yaqian Li, and Guangming Shi. Coarse-to-fine image aesthetics assessment with dynamic attribute selection. *IEEE Transactions on Multimedia*, 26:9316–9329, 2024. 7, 8

[7] 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*, pages 5148–5157, 2021. 5, 8

[8] Lei Ke, Mingqiao Ye, Martin Danelljan, Yu-Wing Tai, Chi-Keung Tang, Fisher Yu, et al. Segment anything in high quality. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[9] Leida Li, Yipo Huang, Jinjian Wu, Yuzhe Yang, Yaqian Li, Yandong Guo, and Guangming Shi. Theme-aware visual attribute reasoning for image aesthetics assessment. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(9):4798–4811, 2023. 7, 8

[10] Leida Li, Tong Zhu, Pengfei Chen, Yuzhe Yang, Yaqian Li, and Weisi Lin. Image aesthetics assessment with attribute-assisted multimodal memory network. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(12):7413–7424, 2023. 7, 8

Algorithm	MM	AVA		AADB		TAD66k		PARA		BAID		# params
		PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	
[9]	✓	0.736	0.725	0.763	0.761	-	-	0.940	0.911	-	-	48.84 M
[6]	✓	0.753	0.751	0.770	0.768	-	-	-	-	-	-	87 M
[3]	✓	0.754	0.752	-	-	-	-	0.928	0.895	-	-	76.7 M
[16]	✓	0.779	0.771	-	**0.79**	-	-	**0.951**	**0.926**	-	-	149.6 M
[11]	✓	0.785	0.776	-	0.771	-	-	-	-	-	-	3.149 B
[10]	✓	0.83	0.816	-	-	-	-	-	-	-	-	158.8
[23]	✓	**0.834**	**0.819**	-	-	-	-	-	-	-	-	135.5 B
[17]	-	0.758	0.758	-	-	**0.553**	0.530	-	-	**0.558**	**0.508**	56 M
[12]	-	0.777	0.764	**0.772**	0.760	0.539	0.496	0.943	0.912	-	-	3 B
[5]	-	0.814	0.803	-	-	0.546	**0.57**	-	-	-	-	87 M
[22]	-	-	-	-	-	-	-	-	-	0.473	0.467	27.3 M
Dinov2-small + *Charm*	-	0.779	0.777	0.767	0.754	0.488	0.458	0.940	0.908	0.439	0.368	**21.53 M**

Table 15. Detailed comparison of our approach with existing IAA models. MM represents using multimodal data like text and attributes. Our approach achieves comparable performance to state-of-the-art IAA models while using significantly fewer parameters. **Bold** and underlined numbers represent the best and the second-best methods.

Algorithm	SPAQ		KonIQ10k		# params
	train	test	train	test	
[18]	**0.928**	**0.923**	**0.945**	**0.934**	30.97 M
[2]	0.924	0.921	0.939	0.926	25.6 M
[7]	0.921	0.917	0.928	0.916	27 M
Ours	0.919	0.915	0.944	0.930	**21.53 M**

Table 16. Detailed comparison of our approach with existing IQA models. Our approach falls slightly behind state-of-the-art methods in IQA, with a difference of less than 1% in SPAQ and 0.4% in KonIQ10k. While other studies have reported the median of 10 runs, we conducted 5 runs and found consistent results, with standard deviations below 0.006 for SPAQ and 0.002 for KonIQ10k. **Bold** and underlined numbers represent the best and the second-best methods.

[11] Leida Li, Xiangfei Sheng, Pengfei Chen, Jinjian Wu, and Weisheng Dong. Towards explainable image aesthetics assessment with attribute-oriented critiques generation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 7, 8

[12] Limin Liu, Shuai He, Anlong Ming, Rui Xie, and Huadong Ma. Elta: An enhancer against long-tail for aesthetics-oriented models. In *Forty-first International Conference on Machine Learning*. 8

[13] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 7

[14] Mary A Peterson. Low-level and high-level contributions to figure-ground organization. 2014. 1

[15] Tomer Ronen, Omer Levy, and Avram Golbert. Vision transformers with mixed-resolution tokenization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4613–4622, 2023. 1

[16] Xiangfei Sheng, Leida Li, Pengfei Chen, Jinjian Wu, Weisheng Dong, Yuzhe Yang, Liwu Xu, Yaqian Li, and Guangming Shi. Aesclip: Multi-attribute contrastive learning for image aesthetics assessment. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1117–1126, 2023. 7, 8

[17] Tengfei Shi, Chenglizhao Chen, Zhenyu Wu, Aimin Hao, and Yuming Fang. Improving image aesthetic assessment via multiple image joint learning. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2024. 8

[18] Nyeong-Ho Shin, Seon-Ho Lee, and Chang-Su Kim. Blind image quality assessment based on geometric order learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12799–12808, 2024. 8

[19] Ombretta Strafforello, Gonzalo Muradas Odriozola, Fatemeh Behrad, Li-Wei Chen, Anne-Sofie Maerten, Derya Soydaner, and Johan Wagemans. Backflip: The impact of local and global data augmentations on artistic image aesthetic assessment. *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2024. 2

[20] Zhengzhong Tu, Peyman Milanfar, and Hossein Talebi. Muller: Multilayer laplacian resizer for vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6877–6887, 2023. 6

[21] Johan Wagemans, James H Elder, Michael Kubovy, Stephen E Palmer, Mary A Peterson, Manish Singh, and Rüdiger Von der Heydt. A century of gestalt psychology in visual perception: I. perceptual grouping and figure–ground organization. *Psychological bulletin*, 138(6):1172, 2012. 1

[22] Ran Yi, Haoyuan Tian, Zhihao Gu, Yu-Kun Lai, and Paul L Rosin. Towards artistic image aesthetics assessment: a large-scale dataset and a new method. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22388–22397, 2023. 8

[23] Tong Zhu, Leida Li, Pengfei Chen, Jinjian Wu, Yuzhe Yang,

Model	Input size	Charm	#tokens	ms	GMACs	MB
ViT-small	224 x 224	-	196	**5.6**	**4.58**	**168.4**
	640 x 640	-	1600	23.6	58.11	1352.1
		✓	2-scale:512	<u>7.1</u>(↓ 69.9%)	<u>13.49</u>(↓ 76.8%)	<u>328.3</u>(↓ 75.7%)
		✓	3-scale:700	9.1(↓ 61.4%)	19.65(↓ 66.2%)	469.3(↓ 65.3%)

Table 17. ViT-small inference cost breakdown for processing one single image: number of tokens (#tokens) based on varying input sizes, runtime in milliseconds (ms), Giga multiply accumulation (GMACs), and GPU memory in Megabytes (MB). **Bold** and <u>underlined</u> values highlight the most and second-most computationally efficient configurations. Percentages indicate the reduction in computational cost compared to processing the image in its original size.

p'	f	PLCC	SRCC	ACC
28	0.5	**0.767**	**0.754**	0.767
42	0.66	0.732	0.725	**0.774**
56	0.75	0.739	0.729	0.749

Table 18. Performance of Dinov2-small + 2-scale *Charm* on AADB dataset with different scaling factors. p' and f represent the initial patch size and the downscaling factor, respectively. The patch size p is set to 14 to match the patch size of the Dinov2-small patch encoding module. Increasing f negatively affects performance due to increased information loss during the downscaling of unselected regions.

and Yaqian Li. Emotion-aware hierarchical interaction network for multimodal image aesthetics assessment. *Pattern Recognition*, 154:110584, 2024. 7, 8