Supplementary Material

1. PLCC Performace Comparison in the Data-Efficient Setting

The Pearson's Linear Correlation Coefficient (PLCC) comparisons for our framework against other NR-IQA methods corresponding to the table in the main paper are provided in Tab. 6. We note that GRepQ_D outperforms all other methods in most cases of the data-efficient setting. However, on datasets such as LIVE [8] and CSIQ [4], GRepQ_D still achieves competitive performance.

2. Analyzing GRepQ's Zero-Shot Performance

We show the individual performances of the high and low-level models in Tab. 7. We observe that the low-level model is capable of capturing differing distortion levels conditioned on the same content. This can be validated from its higher performance compared to the high-level model on the synthetically distorted LIVE [8] and CSIQ datasets. On the other hand, the high-level model shows superior performance on in-the-wild datasets such as CLIVE [1] and KONIQ [3], validating its ability to capture content–based quality information. When images are similar in content, the low-level model is able to discriminate between levels of distortions. However, when images differ in content (as in the case of authentically distorted datasets), the high-level model performs better, demonstrating that the high and low-level models work in a complementary manner.

3. Analyzing GRepQ's Supervised Performance

To demonstrate the effectiveness of the learned features, we also show the performance of the high and low-level features in a fully supervised setting. The features from both models are concatenated and then regressed with MOS using support vector regression. We use $80-20$ train-test splits on the KonIQ and CLIVE datasets and report the median results over 100 runs each in Tab. 8. GRepQ_D performs the best and second-best among existing NR-IQA methods, indicating that the learned representations are good even in the fully supervised setting.

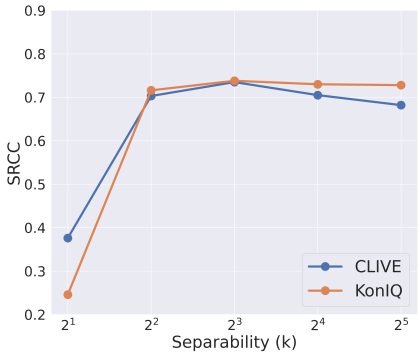

Figure 7. Analyzing the effect of separability between higher and lower quality groups of the high-level model.

4. Runtime Analysis

We show the evaluation times of GRepQ_Z along with other zero-shot benchmark algorithms for a 500×500 image in Fig. 8. We report the average times taken for 100 images. CNN-based methods are tested using an Nvidia RTX 2080Ti graphics card and an Intel(R) Core(TM) i7-9700F CPU.

5. Group Separability Analysis of High-Level Representations

While training the high-level feature encoder, every batch of images is divided into two groups based on their proximity to the two text prompt embeddings. Choosing the right group size is essential based on the number of images present in a batch of images. Two aspects are crucial for group contrastive learning to be beneficial based on text-prompt pairings: (1) The groups must be sufficiently separated in terms of the similarity to the prompt embeddings and (2) The groups must contain a sufficient number of images so that each anchor image contains a balanced number of positives within the group and negatives from the other quality group.

We experiment by varying the separability of the groups, parameterized by k. For a fixed batch size of N, the individual group sizes can be measured as $\texttt{round}(N/k)$. We vary k in powers of 2 as $k \in \{2^1, \cdots, 2^5\}$. For $N = 128$,

Method	CLIVE			KONIQ			CSIQ			LIVE			PIPAL		
Labels	50	100	200	50	100	200	50	100	200	50	100	200	50	100	200
TReS	0.702	0.776	0.813	0.740	0.748	0.824	0.820	0.863	0.915	0.916	**0.948**	0.960	0.177	0.370	0.515
Re-IQA	0.620	0.650	0.701	0.689	0.693	0.757	**0.905**	0.911	0.936	0.876	0.892	0.931	0.263	0.331	0.416
MANIQA	0.713	**0.823**	0.830	0.715	0.806	0.825	0.864	0.914	0.924	0.894	0.933	**0.962**	0.191	0.365	0.484
HyperIQA	0.689	0.755	0.806	0.650	0.758	0.807	0.851	0.869	0.935	0.903	0.922	0.931	0.138	0.325	0.403
DEIQT	0.695	0.739	0.818	0.670	0.707	0.778	0.828	0.889	0.944	0.916	0.942	0.957	0.334	0.412	0.423
LIQE	0.722	0.765	0.820	0.764	0.809	0.822	0.869	0.898	0.916	0.898	0.925	0.936	-	-	-
Resnet50	0.580	0.629	0.660	0.661	0.693	0.716	0.827	0.902	0.932	0.872	0.908	0.920	0.161	0.233	0.312
CLIP	0.676	0.739	0.758	0.749	0.790	0.802	0.881	0.913	0.744	0.891	0.924	0.942	0.266	0.313	0.374
CONTRIQUE	0.693	0.736	0.777	0.743	0.801	0.832	0.821	**0.944**	**0.957**	0.892	0.922	0.944	0.380	0.447	0.501
Re-IQA	0.620	0.650	0.701	0.689	0.693	0.757	**0.905**	0.911	0.936	0.876	0.892	0.931	0.263	0.331	0.416
GRepQ_D (LL)	0.542	0.581	0.639	0.578	0.618	0.654	0.817	0.831	0.853	0.867	0.884	0.886	0.410	0.413	0.439
GRepQ_D (HL)	0.748	0.790	0.822	0.789	0.811	0.834	0.886	0.918	0.945	0.907	0.931	0.949	0.413	0.417	0.440
GRepQ_D	**0.772**	0.798	**0.835**	**0.793**	**0.816**	**0.840**	0.896	0.927	0.947	**0.929**	0.936	0.957	**0.506**	**0.537**	**0.571**

Table 6. PLCC performance comparison of GRepQ_D with other NR-IQA methods trained using few labels on various IQA databases.

Model	CLIVE	KonIQ	CSIQ	LIVE	PIPAL
GRepQ (LL)	0.502	0.692	**0.711**	**0.784**	0.394
GRepQ (HL)	0.735	0.738	0.647	0.581	0.398
GRepQ_Z	**0.740**	**0.768**	0.693	0.741	**0.436**

Table 7. Performance comparison of the high and low-level models in the zero-shot setting.

NR-IQA	CLIVE	KonIQ
BRISQUE [6]	0.608	0.665
DB-CNN [11]	0.851	0.875
HyperIQA [9]	*0.859*	0.906
CONTRIQUE [5]	0.845	0.894
TReS [2]	0.846	**0.914**
Re-IQA [7]	0.840	**0.914**
GRepQ_D	**0.864**	*0.909*

Table 8. SRCC performance of NR-IQA methods on KonIQ-10K and CLIVE databases. The best and second-best performing methods are bolded and emphasized, respectively.

the corresponding group sizes are $64, 32, \cdots, 4$. As k increases, we identify smaller groups that are more similar to the respective text-prompt embeddings with better separation. From Fig. 7, we observe that a large group size (smaller k) is not favorable because they are not sufficiently distinguishable in terms of quality, necessitating a minimum separation between the groups. We observe that the performance is fairly stable for $k = 4, 8, 16$. However, as k increases further, the group sizes become extremely small for effective group contrastive learning.

6. Impact of Group Contrastive Learning With Respect to Choice of Text Prompts

We show that our contrastive learning over groups based on perceptually relevant text prompts offers better performance than a zero-shot application of CLIP (CLIP-IQA) irrespective of the choice of the prompts. We experiment with different text prompt pairs: `['a good photo.', 'a bad photo.']`, `['high definition photo.', 'low definition photo.']`, `['high quality photo.', 'low quality photo.']`, and `['pristine photo.', 'blurry photo.']`. The choice of prompts is motivated by CLIP-IQA [10]. The performance of the high-level model using these prompts is illustrated in Tab. 9. Since we consistently improve the performance of all prompt pairs, our method can leverage future improvements in prompt engineering to improve the performance of vision-language models for IQA.

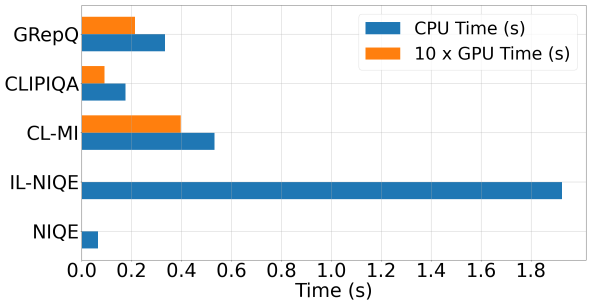

Figure 8. Runtime analysis.

References

[1] Deepti Ghadiyaram and Alan C Bovik. Massive online crowdsourced study of subjective and objective picture quality. *IEEE Transactions on Image Processing*, 25(1):372–387, 2015. 1

[2] S Alireza Golestaneh, Saba Dadsetan, and Kris M Kitani. No-reference image quality assessment via transformers, relative ranking, and self-consistency. In *Proceedings of the*

Prompt pair	CLIP-IQA		High-Level Model	
	CLIVE	KonIQ	CLIVE	KonIQ
['high quality photo.', 'low quality photo.']	0.462	0.524	0.475	0.688
['clean photo.', 'blurry photo.']	0.600	0.675	0.676	0.703
['high definition photo.', 'low definition photo.']	0.611	0.591	0.647	0.721
['a good photo.', 'a bad photo.']	**0.630**	**0.701**	**0.735**	**0.738**

Table 9. SRCC performance analysis on using different perceptual antonym text prompt pairs for fine-tuning and testing with the high-level model.

IEEE/CVF Winter Conference on Applications of Computer Vision, pages 1220–1230, 2022. 2

[3] Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe. Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing*, 29:4041–4056, 2020. 1

[4] Eric C Larson and Damon M Chandler. Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of Electronic Imaging*, 19(1):011006–011006, 2010. 1

[5] Pavan C Madhusudana, Neil Birkbeck, Yilin Wang, Balu Adsumilli, and Alan C Bovik. Image quality assessment using contrastive learning. *IEEE Transactions on Image Processing*, 31:4149–4161, 2022. 2

[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. 2

[7] Avinab Saha, Sandeep Mishra, and Alan C. Bovik. Re-iqa: Unsupervised learning for image quality assessment in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5846–5855, June 2023. 2

[8] Hamid R Sheikh, Muhammad F Sabir, and Alan C Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on Image Processing*, 15(11):3440–3451, 2006. 1

[9] 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. 2

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

[11] 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. 2