






(Supplementary Material)
**ATTIQA: Generalizable Image Quality Feature
Extractor using Attribute-aware Pretraining**

Daekyu Kwon¹, Dongyoung Kim¹, Sehwan Ki², Younghyun Jo²,
Hyong-Euk Lee², and Seon Joo Kim¹

¹ Yonsei University

² Samsung Advanced Institute of Technology

A Details on Prompt Selection

This section explains the prompts and techniques used in the prompt selection strategy. Our prompt selection strategy generates prompt candidates from GPT-4 [4] through the following query that receives a set of adjectives: “*Suggest 50 positive/negative adjectives about {attribute} related to image quality*”. We create prompt candidates by changing *{attribute}* in the query to aligned attribute. The generated prompt candidates are reported in Table 14.

We also provide details on the proxy tasks stage. For the distortion intensity based proxy task, which measures attribute score by predicting the intensity of the corresponding distortion, we pair image attributes and corresponding distortions as follows:

- Sharpness : Gaussian Blur, ZoomBlur, LensBlur
- Contrast : Contrast adjustment multiplying factors to RGB value
- Brightness : V adjustment in HSV space
- Colorfulness : Saturation adjustment in HSV space
- Noisiness : Gaussian Noise, ISO Noise

In attributes where multiple distortions are described, we execute proxy tasks for each distortion and compute the final result by averaging each output. In Table 8, we note selected prompts that only utilize a single proxy task whose results are reported in Table 2.

B Details on Experimental Configuration

During pretraining, we randomly cropped the image at a resolution of 224×224 . We train our network for 100 epochs, using AdamW optimizer with batch size 256 and learning rate $1e-4$.

Table 8: Results of the prompt selection with single proxy task. We only denote adjective for this table.

Proxy task Attribute	Distortion intensity		Human perception	
	Positive	Negative	Positive	Negative
Sharpness	"Unambiguous"	"Vague"	"High-definition"	"Out-of-focus"
Contrast	"Enhanced"	"Bleak"	"Splendid"	"Blurred"
Brightness	"Clear"	"Starless"	"Dark"	"High-key"
Colorfulness	"Multicolored"	"Grayish"	"Lively"	"Blurred"
Noisiness	"Clutter-free"	"Spattered"	"Peerless"	"Blurry"

C Additional Experiments

C.1 Impact of dataset scale

To verify the scalability of our pretraining scheme, we conduct experiments changing the ratio of the used dataset. In this experiment, we pretrain our model only utilizing 20% and 50% of the ImageNet dataset. As shown in Table 9, the performance of ATTIQA increases with the amount of available datasets. This result indicates that the performance of ATTIQA can be improved when we can train it on larger datasets (e.g., ALIGN, JFT-300M).

Table 9: Performance comparison of ATTIQA using various ratios of datasets.

Methods	CLIVE [2]		KonIQ [3]		SPAQ [1]	
	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC
20%	0.887	0.904	0.935	0.946	0.923	0.928
50%	0.889	0.913	0.938	0.949	0.924	0.928
100%	0.898	0.916	0.942	0.952	0.926	0.930

C.2 Impact of attribute head

To evaluate the impact of each attribute head, we conduct experiments that only utilize the head’s feature space. To implement this experiment, we only adopt a single attribute head for fine-tuning instead of concatenating the five attributes feature map. As shown in Table 10, although each attribute head shows similar performance since they share a backbone, we observe a slightly better performance in the sharpness attribute head. In contrast, the performance of other attribute heads varies depending on the dataset. It can be seen that this result follows an analysis of SPAQ [1], which demonstrates that sharpness is the most highly related attribute to image quality.

Moreover, as we denoted at Sec 4.5, we conduct an additional experiment to verify that each attribute head accurately captures its corresponding attributes. To provide a detailed explanation, we manipulate each attribute by sequentially

Table 10: performance comparison of fine-tuning which only utilizes ATTIQA’s each attribute head.

Methods	CLIVE		KonIQ		SPAQ	
	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC
Sharpness	0.890	0.910	0.939	0.950	0.925	0.929
Contrast	0.870	0.905	0.937	0.947	0.923	0.928
Brightness	0.880	0.905	0.936	0.948	0.923	0.926
Colorfulness	0.881	0.906	0.934	0.945	0.924	0.927
Noisiness	0.878	0.901	0.939	0.950	0.924	0.928
ATTIQA	0.898	0.916	0.942	0.952	0.926	0.930

attaching relevant distortions, described in Sec A, to a given image and assess the significance of features by measuring the Grad-CAM from the MOS prediction to each feature map. Furthermore, to preserve the integrity of each feature map, we conduct this experiment using a linear probing setup. As shown in Figure 5, we first analyze the tendency of MOS variation in response to changes in image attribute. For sharpness and noisiness, we can observe that MOS decreases as more noise and blur are applied to the image. For contrast and brightness, the highest MOS is achieved when the attributes are at an optimal medium level, with the MOS decreasing when the attributes become either too high or too low. A similar result is observed for colorfulness, but the pattern of overall variation exhibits a slight difference. Interestingly, a negative correlation between the Mean Opinion Score (MOS) and Grad-CAM is observed for every attribute. It suggests that each attribute has a more significant impact on MOS prediction when a particular attribute varies to the extent that compromises the image quality. This observation indicates that each attribute head of ATTIQA can effectively capture changes in specific attributes from the original image.

Table 11: Performance comparison with LIQE.

Methods	Seen Dataset						Unseen Dataset			
	LIVE	CSIQ	KADID	BID	LIVEC	KonIQ	TID	SPAQ	PIPAL	
SR	LIQE	0.970	0.936	0.930	0.875	0.904	0.919	0.811	0.881	0.478
CC	ATTIQA	0.974	0.936	0.923	0.891	0.901	0.920	0.796	0.891	0.521
PL	LIQE	0.951	0.939	0.931	0.900	0.910	0.908	-	-	-
CC	ATTIQA	0.977	0.950	0.927	0.918	0.917	0.932	-	-	-

C.3 Comparison with LIQE

In the main experiments, we only carried out single-dataset based experiments with LIQE to ensure a fair comparison with other baselines. To further assess the

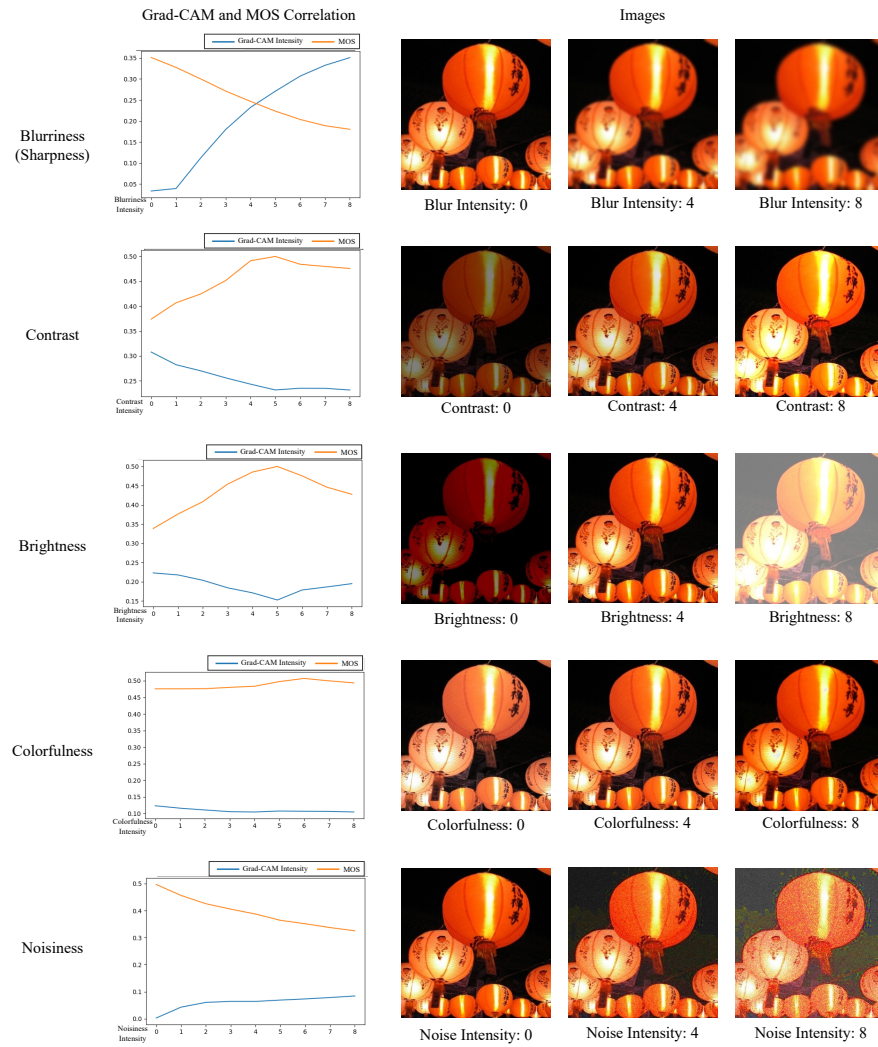


Fig. 5: Correlation graph illustrating the relationship between MOS and Grad-CAM across varying image attributes. On the right side, examples are provided to showcase how image attributes have been manipulated.

efficacy of our model in multiple-dataset environments, we measure ATTIQA’s performances utilizing a dataset tailored for LIQE and its training recipe. As shown in Table 11, though ATTIQA is trained only using MOS, unlike LIQE which utilizes additional prompt-based annotations, it exhibits overall superior performances on given datasets. Exceptions are observed with two datasets, TID and KADID, which primarily focus on specific distortions. These cases benefit

Table 12: Performances on Synthetic Distortion IQA Dataset

Methods	LIVE		CSIQ		TID2013		KADID-10k	
	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC
CONTRIQUE	0.962	0.964	0.945	0.955	<u>0.885</u>	<u>0.883</u>	0.919	0.919
Re-IQA	0.966	0.968	<u>0.946</u>	0.955	0.877	<u>0.883</u>	0.912	0.910
CLIP	<u>0.968</u>	<u>0.970</u>	0.944	0.934	0.858	0.858	0.898	0.897
CLIP-IQA+	0.919	0.922	0.886	0.899	0.856	0.862	0.792	0.790
ATTIQA	0.970	0.972	0.948	<u>0.953</u>	0.898	0.903	<u>0.917</u>	<u>0.917</u>

from LIQE’s approach of utilizing an additional dataset that contains additional annotations on synthetic distortions.

C.4 Experiments on Synthetic Distortion-based Dataset

To further validate the robustness of ATTIQA, we propose an additional experiment with synthetic distortion-based datasets. We utilize LIVE, CSIQ, TID2013, and KADID-10k datasets, which measure MOS based on synthetic distortions. As shown in Table 12, ATTIQA exhibits the best or second-best performance across all datasets. On the KADID-10k dataset, we note that our method shows comparable results to CONTRIQUE, which uses the KADID-700k dataset at the pretraining stage. This experiment supports the robustness of ATTIQA, demonstrating that its representation contains information about both synthetic and authentic distortions.

D Details on Application

D.1 Metrics for Generative Model

This section provides details on the image generation pipeline used in a user study and more examples of results. For the generative model backbone, we utilize Stable Diffusion-2.1 [5], a widely used text-to-image synthesis model. To create diverse images, we establish prompt template as “*masterpiece, best quality, photorealistic photography, crystal clear, 8K UHD, {action} {object} {place}*” and generate images by replacing “*action*”, “*object*”, and “*place*” respectively with various candidates reported in Table 13. Figure 7 shows additional qualitative comparison results of a user study. Note that this user study aims to evaluate the quality of generated images solely by showing only the two images to the subject without any information on the corresponding prompts.

D.2 Details on User Study

We conduct a user study using Amazon Technical Turk (AMT), gathering subjects in an anonymous setting without bias about gender or nationality. For the survey, the descriptions were presented as follows:

Table 13: Prompt candidates utilized in image generation.

Target	Prompt candidates
Action	<i>"playing", "running", "sleeping", "eating", "walking", "standing", "sitting", "jumping", "dancing"</i>
Object	<i>"a dog", "a cat", "a clothed man", "a dressed woman", "a bear"</i>
Place	<i>"in the grass", "in the room", "in the forest", "in the water", "in the snow", "in the desert"</i>

- 1) Participate in the image quality preference voting for tuning images. You can vote for the image you prefer between the two provided images
- 2) choose a better one with good image quality(color, sharpness, expressiveness..)

D.3 Image Enhancement

In this section, we explain another application in ATTIQA’s MOS that is used as a reward for reinforcement learning to find optimal parameters for the ISP [6]. The ISP pipeline consists of 14 modules, of which 16 tuning parameters of 7 modules were auto-tuned from the perspective of maximizing the ATTIQA’s MOS. The tuned seven modules are as follows, and the numbers in parentheses of each module represent the number of tuning parameters: Tone-mapping (3), UV channel Denoising (3), Y channel Denoising (2), Brightness/Contrast control (2), Hue/Saturation control (2), Sharpening (3), and Texture enhancing (1). The agent of the reinforcement learning was trained using 730 raw images that were 4000x3000-sized. In the user survey, we uniformly sampled 200 raw images among 5000 raw images for fair comparisons. Figure 6 shows additional qualitative comparison results. It can also be seen that our pipeline retouches images to make them more colorful and vivid compared to both retouching by expert C and the default settings.



Fig. 6: More qualitative comparisons about image enhancement.

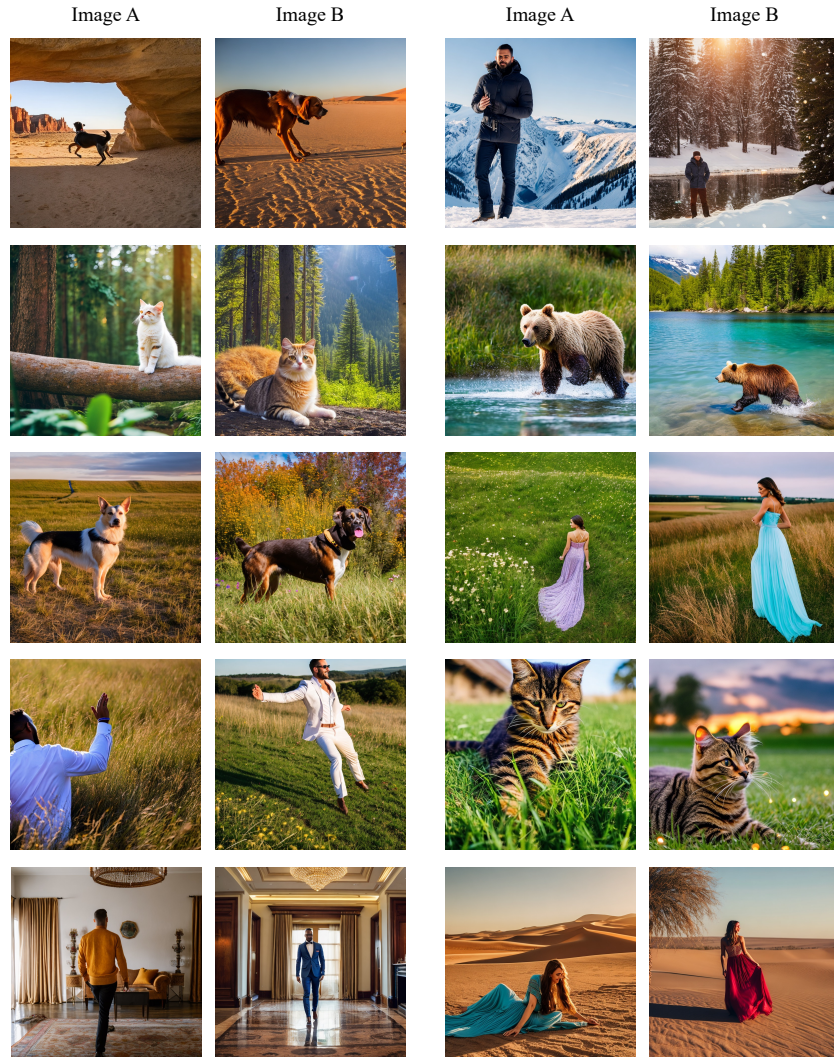


Fig. 7: Examples of generated images. Each paired image is synthesized using the same prompt. Image A is the image that CONTRIQUE and Re-IQA predict the high quality score. In contrast, Image B is preferred by humans and assigned a high MOS by ATTIQA.

Table 14: Prompt candidates for each image attribute.

Attribute	Prompt type	Prompt candidates
Sharpness	positive	<i>"crisp", "sharp", "defined", "clear", "distinct", "vivid", "bright", "detailed", "refined", "pristine", "flawless", "lucid", "exact", "polished", "pure", "radiant", "sleek", "smooth", "resolute", "immaculate", "brilliant", "vibrant", "rich", "clean", "meticulous", "unblemished", "sublime", "superior", "splendid", "exquisite", "true-to-life", "tactile", "textured", "illuminated", "lustrous", "glossy", "granular", "pinpoint", "spot-on", "focused", "unambiguous", "concise", "intense", "high-definition", "lifelike", "bold", "harmonious", "stunning", "undistorted"</i>
	negative	<i>"blurry", "fuzzy", "hazy", "vague", "indistinct", "muddled", "obscured", "smudged", "cloudy", "dull", "muted", "out-of-focus", "pixilated", "jagged", "noisy", "grainy", "mottled", "muddy", "murky", "dim", "foggy", "shadowy", "bleary", "washed-out", "weak", "gloomy", "clouded", "patchy", "shrouded", "veiled", "lacking clarity", "rough", "distorted", "tarnished", "ill-defined", "ambiguous", "flat", "listless", "pale", "insipid", "smeared", "streaked", "stained", "splotchy", "spotty", "blotchy", "dingy", "drab", "tainted"</i>
Contrast	positive	<i>"crisp", "defined", "vivid", "sharp", "clear", "distinguished", "bold", "pronounced", "high-contrast", "distinct", "lucid", "striking", "intense", "robust", "dynamic", "stark", "rich", "deep", "emphasized", "highlighted", "vibrant", "solid", "notable", "prominent", "enhanced", "powerful", "contrastive", "standout", "bright", "conspicuous", "discernible", "marked", "potent", "compelling", "dramatic", "forceful", "illuminated", "brilliant", "radiant", "meticulous", "articulate", "impressive", "splendid", "magnified", "amplified", "accentuated", "divergent", "outstanding", "captivating"</i>
	negative	<i>"muddy", "flat", "blurred", "faded", "indistinct", "washed-out", "lackluster", "dull", "weak", "subdued", "undistinguished", "low-contrast", "ambiguous", "pale", "muted", "obscure", "vague", "insipid", "clouded", "nebulous", "tenuous", "confusing", "feeble", "diminished", "indiscernible", "unnoticeable", "slight", "unemphasized", "shadowed", "doubtful", "hazy", "unsaturated", "ill-defined", "unremarkable", "bleak", "insignificant", "bland", "monotone", "uniform", "muddled", "equivocal", "unaccentuated", "listless", "understated", "unimpressive", "nondescript", "faint", "impotent", "inaudible", "discreet"</i>
Brightness	positive	<i>"luminous", "bright", "vivid", "brilliant", "radiant", "gleaming", "illuminated", "clear", "sparkling", "shiny", "glowing", "light-filled", "dazzling", "luculent", "resplendent", "shimmering", "lustrous", "beaming", "crisp", "vibrant", "intense", "well-lit", "brilliant", "glittering", "glistering", "blazing", "effulgent", "reflective", "aglow", "incandescent", "high-key", "fiery", "lambent", "twinkling", "opulent", "sunlit", "burnished", "pristine", "flashing", "undimmed", "sunny", "spotlit", "blinding", "flawless", "translucent", "glossy", "crystal-clear", "immaculate", "gleamy"</i>
	negative	<i>"dim", "dull", "dark", "shadowy", "obscured", "faint", "gloomy", "pale", "muted", "clouded", "bleak", "underexposed", "drab", "faded", "murky", "shaded", "veiled", "flat", "dusky", "tenebrous", "sombre", "gray", "lackluster", "washed-out", "overcast", "smoky", "subdued", "muffled", "eclipsed", "sullen", "unlit", "opaque", "low-key", "blurred", "darkened", "blackened", "shadowed", "misty", "flightless", "moonless", "starless", "inky", "twilight", "foggy", "overclouded", "cimmerian", "umbrous", "pitch-dark"</i>
Colorfulness	positive	<i>"vibrant", "rich", "vivid", "saturated", "brilliant", "lively", "radiant", "bold", "bright", "colorful", "intense", "resplendent", "lush", "deep", "dazzling", "varied", "dynamic", "electric", "illuminated", "vibrantly-hued", "multicolored", "kaleidoscopic", "strong", "fluorescent", "fiery", "prismatic", "stunning", "flashy", "beaming", "pigmented", "chromatic", "glistening", "spectacular", "polychromatic", "sunny", "iridescent", "opulent", "rainbow-like", "effulgent", "color-laden", "invigorating", "gorgeous", "lustrous", "gleaming", "dramatic", "bursting", "captivating", "energetic"</i>
	negative	<i>"drab", "dull", "washed-out", "muted", "faded", "pale", "flat", "monochrome", "grayish", "uninspiring", "lifeless", "bleak", "tarnished", "insipid", "blurred", "cloudy", "dim", "neutral", "colorless", "lackluster", "subdued", "murky", "dusty", "dusky", "undistinguished", "muddy", "unsaturated", "shadowed", "overcast", "veiled", "bland", "indistinct", "unvaried", "uniform", "faint", "anaemic", "vague", "wan", "stale", "ashen", "pastel", "watered-down", "sallow", "obscured", "indeterminate", "discolored", "ill-defined", "tinged", "hazy"</i>
Noisiness	positive	<i>"clear", "crisp", "smooth", "pure", "sharp", "pristine", "flawless", "noiseless", "unblemished", "intact", "clean", "polished", "immaculate", "well-defined", "impeccable", "spotless", "untainted", "sleek", "neat", "unspoiled", "intact", "refined", "clutter-free", "lucid", "undisturbed", "unmarred", "untarnished", "perfect", "refreshing", "distinct", "vivid", "bright", "detailed", "accurate", "faithful", "exquisite", "superb", "top-notch", "first-rate", "matchless", "peerless", "masterful", "skilful", "uncompromised", "optimal", "optimum", "superior", "prime", "finest"</i>
	negative	<i>"grainy", "speckled", "mottled", "patchy", "dirty", "blemished", "marred", "flecked", "spotty", "noisy", "blurry", "fuzzy", "hazy", "cloudy", "splotchy", "streaky", "dotted", "gauzy", "scratchy", "scattered", "dappled", "distorted", "messy", "smudged", "lackluster", "gloomy", "dim", "overcast", "pockmarked", "crackled", "choppy", "erratic", "spattered", "discolored", "inconsistent", "irregular", "shoddy", "subpar", "mediocre", "unrefined", "vague", "ambiguous", "dull", "dreary", "stained", "blemish-ridden", "defective", "imperfect", "second-rate", "tarnished", "degraded"</i>

References

1. Fang, Y., Zhu, H., Zeng, Y., Ma, K., Wang, Z.: Perceptual quality assessment of smartphone photography. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3677–3686 (2020) [2](#)
2. Ghadiyaram, D., Bovik, A.C.: Massive online crowdsourced study of subjective and objective picture quality. *IEEE Transactions on Image Processing (TIP)* **25**(1), 372–387 (2015) [2](#)
3. Hosu, V., Lin, H., Sziranyi, T., Saupe, D.: Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing (TIP)* **29**, 4041–4056 (2020) [2](#)
4. OpenAI: Gpt-4 technical report (2023) [1](#)
5. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10684–10695 (2022) [5](#)
6. Shin, U., Lee, K., Kweon, I.S.: Drl-isp: Multi-objective camera isp with deep reinforcement learning. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 7044–7051 (2022) [6](#)