IQA-Adapter: Exploring Knowledge Transfer from Image Quality Assessment to Diffusion-based Generative Models

Supplementary Material

7. Contents

Here we briefly summarize the contents of all sections in this supplementary file:

- Section 8: Discussion of the possible use-cases of IQA-Adapter and Future Work;
- Section 9: A detailed summary of all IQA/IAA models used in this study;
- Section 10: Details on IQA-Adapter training;
- Section 11: Limitations of IQA-Adapter;
- Section 12: Ablation Study on adapter design;
- Section 13: More results regarding high-quality conditioning experiments, visual comparison with other methods;
- Section 14: Detailed results on generative capabilities of different methods;
- Section 15: Experiments regarding alignment with qualitative conditions;
- Section 16: Evaluation of image degradation with Full-Reference IQA metrics;
- Section 17: More details on Subjective Studies;
- Section 18: Miscellaneous experiments: time measurements, generation consistency, examples of quality modulation;
- Section 19: Some connections between quality optimisation and adversarial robustness;
- Section 20: More examples of Reference-based IQA-Adapter and comparison with IP-Adapter and StyleCrafter.

8. Discussion and Future Work

8.1. IQA-Adapter as a degradation model

As most IQA models are trained to assess distorted images, they can reliably detect noise, compression, blur, and other artifacts on images during IQA-Adapter training. Therefore, this knowledge is transferred to the generative model and such image attributes are connected with low-quality conditions. This allows IQA-Adapter to generate progressively more distorted images as input quality-condition decreases. The IQA-Adapter in Figure 4(b), for example, implicitly learned to simulate JPEG compression artifacts when conditioned on low quality (1st percentile of the training dataset). Figure 23 demonstrates more examples of similar artifacts appearing under low-quality guidance. As IQA models are mostly tailored to assess low-level quality attributes (in contrast with IAA methods), images produced with different quality levels usually retain similar content and composition, as illustrated in Figure 1 (bottom-to-top direction).

By applying appropriate filtering to exclude image pairs with unintended content differences, IQA-Adapter can generate large synthetic datasets of distorted and corresponding high-quality images. Such datasets can subsequently be used to pretrain models for image enhancement, deblurring, and other restoration tasks. While training such methods is a subject for future work, we additionally explore the distances between generated images with different target-quality conditions in Section 16.2. We also note that IQA-Adapter can be additionally fine-tuned with unpaired data containing specific distortions to simulate them during inference.

8.2. Exploring adversarial patterns and preferences of IQA models

When applied with a sufficiently high guidance scale, the gradient-based method can exploit vulnerabilities of the target IQA model, artificially inflating its values and shifting the generation towards an adversarial subdomain. This approach tends to produce images with distinct patterns specific to each IQA model. Figure 6(a) demonstrates adversarial patterns generated with different guidance models. For certain models, such as TRES and HYPER-IQA, these patterns form grid-like structures, and for others, like TOPIQ and DBCNN, they concentrate in smaller regions. We present more adversarial examples generated with gradient-based guidance and GradCAM [85] visualizations of corresponding IQA models in Section 19.

Our study further reveals that most IQA models exhibit distinct preferences when used with a high IQA-Adapter scale. For instance, TOPIQ often favors sharper images, while LAION-AES tends to enhance color saturation, producing more vibrant visuals. These effects can be compounded by using multiple IQA/IAA models simultaneously during adapter training, as illustrated in Figure 6(b).

9. Employed IQA/IAA methods

Table 3 provides a detailed summary of all IQA/IAA methods used in this study, along with their training datasets and architectural details. The column "PyIQA" lists model identifiers from the PyIQA library [86]. The column "Task" specifies supported tasks: most models are designed for IQA, while some (e.g., TOPIQ, MUSIQ) support both IQA and IAA, and others (e.g., NIMA) are exclusive to IAA. The column "Datasets" lists the datasets associated with

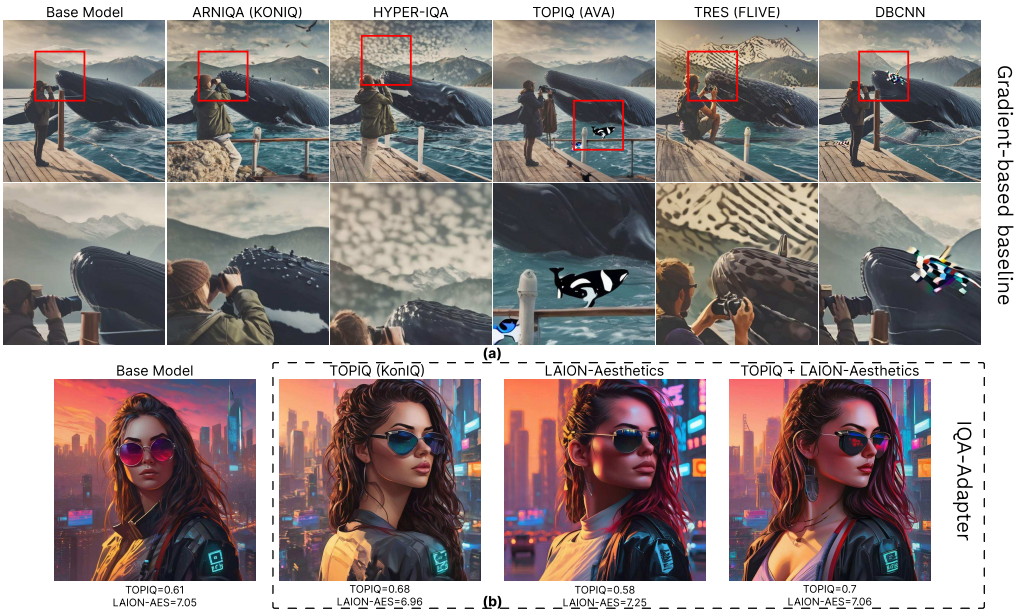

Figure 6. (a) Examples of adversarial patterns appearing under high **gradient-based** guidance scale. (b) Examples of images generated with the **IQA-Adapters** trained with different IQA models. Each IQA/IAA model has its stylistic preferences. All images in each line are generated with the same prompt and seed.

each model; note that the models were not trained on mixtures of datasets, except for LIQE-MIX, which was specifically trained on a dataset mixture. For models like TOPIQ, there are several variants, each trained on a distinct dataset. The column "Arch" outlines the backbone architecture of the models. Most models are trained using finetuning of a pretrained model; however, some, like MUSIQ, are trained from scratch. The final three columns, "Params," "FLOPs," and "MACs," highlight the performance metrics of the models. FLOPs and MACs were computed using the calflops package [87].

Table 4 provides a detailed overview of the datasets used for training the IQA and IAA models. The column "Type" categorizes the datasets: FR indicates the presence of a distortion-free reference image used for collecting subjective scores, whereas NR denotes datasets without such references. The column "Year" indicates the release year of each dataset. The column "# Ref" specifies the number of reference images used to generate distorted samples through augmentations. The column "# Dist" represents the total number of samples in the dataset. The column "Dist Type." describes how distorted images were created: "synthetic" refers to distortions introduced via augmentations such as JPEG compression or blurring, "algorithmic" applies to distortions generated by neural networks, such as GAN-based modifications, "authentic" denotes images captured in natural, real-world conditions, and "aesthetics" refers to high-quality images sourced from stock photography collections. The column "# Rating" indicates the number of ratings collected via crowdsourcing platforms. The column "Original size" details the resolution of images within the datasets.

10. IQA-Adapter training

The IQA-Adapters were trained on the CC3M dataset, which consists of approximately 3 million text-image pairs, for 24,000 steps, followed by fine-tuning on a subset of the LAION-5B dataset, containing 170,000 images, for 3,000 steps. During training on CC3M, the images were center-cropped to a resolution of 512×512. For fine-tuning on LAION, the resolution was increased to 1024×1024 to match SDXL's native resolution. We used the AdamW [92] optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and a weight decay of 1×10^{-2} for the IQA-Adapter parameters. All experiments utilized bf16 mixed precision to improve computational efficiency. Training was conducted using the PyTorch [93] and Accelerate [94] libraries, enabling efficient scaling across our hardware setup. We use batch_size=16 per GPU for 512×512 training resolution, and batch_size=4 for 1024×1024 fine-tuning. The learning rate was set to 10^{-4} during the primary training phase on CC3M and reduced to 10^{-5} for the fine-tuning on the LAION subset. For Reference-based IQA-Adapter, we apply series of degradations to training images with a probability $p = 0.1$ during training.

To ensure consistency and reproducibility, all experiments were conducted within Docker containers built from a shared image. The environment included Python 3.11, PyTorch 2.1, and other dependencies required for training

Model	PyIQA	Task	Datasets	Arch	Params	FLOPS	MACs
TOPIQ [1]	`topiq_nr`	IQA / IAA	KonIQ-10k [50], SPAQ [52], FLIVE [54], AVA [55]	ResNet50	45.2M	886 GFLOPS	441.5 GMACs
DBCNN [42]	`dbcnn`	IQA	KonIQ-10k [50]	VGG16	15.3M	2.1 TFLOPS	1 TMACs
HyperIQA [40]	`hyper_iqa`	IQA	KonIQ-10k [50]	ResNet50	27.4M	2.6 TFLOPS	1.3 TMACs
ARNIQA [43]	`arniqa`	IQA	KonIQ-10 [50], FLIVE [54], KADID [56]	ResNet50	23.5M	-	-
LIQE-Mix [75]	`liqe_mix`	IQA	Mixed (LIVE [88], CSIQ [89], KADID [56], CLIVE [53], BID [90], KonIQ-10k [50])	OpenAI CLIP ViT-B/32	151.3M	1.7 TFLOPS	850.7 GMACs
MANIQA [46]	`maniqa`	IQA	KonIQ-10k [50], PIPAL [51]	ViT-B/8	135.7M	56.4 TFLOPS	28.2 TMACs
CNN-IQA [47]	`cnniqa`	IQA	KonIQ-10k [50]	CNN	729.8K	49.4 GFLOPS	24.5 GMACs
LIQE [75]	`liqe`	IQA	KonIQ-10k [50]	OpenAI CLIP ViT-B/32	151.3M	1.7 TFLOPS	850.7 GMACs
MUSIQ [41]	`musiq`	IQA / IAA	KonIQ-10k [50], AVA [55], FLIVE [54]	Multiscale ViT	27.1M	400.6 GFLOPS	199.1 GMACs
CLIP-IQA+ [48]	`clip_iqa+`	IQA	KonIQ-10k [50]	OpenAI CLIP ResNet50	102.0M	981.1 GFLOPS	489.2 GMACs
NIMA [49]	`nima`	IAA	AVA [55]	InceptionResnetV2	54.3M	342.9 GFLOPS	171 GMACs
LAION-Aes [2]	`laion_aes`	IAA	Other	OpenAI CLIP VIT-L/14	428.5M	2 TFLOPS	1 TMACs
TReS [45]	`tres`	IQA	FLIVE [54]	ResNet50	152.5M	25.9 TFLOPS	12.9 TMACs
HPSv2 [91]	–	Human Preference	Human Preference Dataset v2 [91]	OpenAI CLIP VIT-L/14	428.5M	2 TFLOPS	1 TMACs

Table 3. List of employed metrics with their corresponding training datasets.

Type	Dataset	Year	# Ref	# Dist	Dist Type.	# Rating	Original size $W \times H$
FR	LIVE [88]	2006	29	779	Synthetic	25k	768×512 (typical)
	CSIQ [89]	2010	30	866	Synthetic	5k	512×512
	KADID-10k [56]	2019	81	10.1k	Synthetic	30.4k	512×384
	PIPAL [51]	2020	250	29k	Syth.+alg.	1.13M	288×288
NR	BID [90]	2010	120	6000	Synthetic	$\sim$ 7k	1K – 2K
	AVA [55]	2012	-	250k	Aesthetic	53M	< 800
	CLIVE [53]	2015	-	1.2k	Authentic	350k	500×500
	KonIQ-10k [50]	2018	-	10k	Authentic	1.2M	512×384
	SPAQ [52]	2020	-	11k	Authentic	–	4K (typical)
	FLIVE [54]	2020	-	160k	Auth.+Aest.	3.9M	Train< 640 | Test> 640

Table 4. Description of training datasets from Table 3.

and inference. We use adapter scale $\lambda = 0.5$ in all experiments, unless stated otherwise, and negative guidance scale $\delta = 0.3$, if IQA-Adapter name includes "+ Neg. G." ($\delta = 0$ otherwise). For Reference-based IQA-Adapter, we use adapter scale $\lambda = 0.65$.

11. Limitations

IQA-Adapter serves as a guiding mechanism for transferring knowledge from the IQA/IAA domain to generative models. However, the extent of this knowledge transfer is inherently constrained by the capabilities and limitations of current IQA/IAA models. Most existing IQA datasets, and the models trained on them, are designed to assess the quality of real images, focusing on aesthetical attributes and distortions common for human-generated images. These models often lack the ability to detect distortions specific to generated content, such as unnatural or anatomically incorrect features (e.g., distorted limbs or physically implausible scenes). As a result, these issues may not be adequately penalized in the quality estimates used for guidance, limiting the adapter's ability to address such generation defects. One possible direction of future work to address this limitation is to train a classifier for different kinds of generation artifacts and then attempt to utilize its logits as a conditioning factor.

Another limitation arises from biases in the training data. The IQA-Adapter can inadvertently learn and reproduce unintended relationships between image content and quality levels present in the dataset. For example, when conditioned on low aesthetic scores, the adapter may occasionally generate images with watermarks, likely because it encountered numerous stock photos with watermarks during training and associated them with lower-quality conditions. While some of these correlations may be considered genuine (e.g., watermarks generally reduce image aesthetics), such artifacts highlight the challenge of disentangling genuine quality attributes from dataset-specific correlations.

The training process itself introduces additional challenges. IQA-Adapter training occurs entirely in the latent space of the diffusion model, while the quality scores used for supervision are computed in pixel space. This discrepancy between the latent representations of images (compressed by the model's VAE encoder) and the pixel-level quality scores can introduce instability into the training process, as the adapter must work with imperfect representations of the input images. Furthermore, the VAE decoder used in the final generation step imposes inherent limita-

tions, as it may introduce artifacts (e.g., blurred text or texture inconsistencies) that the adapter cannot correct. In this work, we only cover existing quality assessment models; however, this limitation can be largely mitigated in the future by implementing a quality assessment model that operates in the latent space of the generative model.

12. Ablation Study

In this section, we report the results of our experiments with different architectural elements and hyperparameters of the IQA-Adapter. We compare our base design with a "simplified" model (Sec. 12.1) and a more sophisticated approach with Positional Encoding (Sec. 12.2). Furthermore, we evaluate the impact of the scaling hyperparameter λ of IQA-Adapter.

12.1. Impact of the Separate Qualitative Attention and Negative Guidance

Model	Quality Gain, %↑	SROCC w/ target↑	FID↓	FID (TOP-10%)↓	IS ↑	CLIP-T ↑	CLIP-I ↑
IQA-Adapter	8.95	0.97	**21.36**	**28.44**	**36.89**	**26.83**	**70.02**
IQA-Adapter + Neg. Guidance	**10.86**	**0.98**	22.16	29.25	36.33	26.80	69.82
IQA-Adapter w/o Separate Cross-Attn	8.31	0.26	29.04	39.91	30.22	26.34	67.9

Table 5. Comparison of IQA-Adapters with and without separate qualitative attention. Both adapters are trained with TOPIQ and LAION-Aesthtics IQA models. SROCC is calculated with target TOPIQ scores, and Quality Gain is evaluated similarly to Sec. 4.2 and averaged across all evaluation metrics.

To test the importance of the separate qualitative cross-attention operation, we test the ablated IQA-Adapter that simply concatenates qualitative tokens to the text ones and processes them within a single (textual) cross-attention operation. This simplified model functionally resembles "adaptive" Textual Inversion [27], controlled by a projection module.

In this setting, adapter loses the ability to control its impact via λ parameter, reducing its usability. As demonstrated in Table 5, the model partially retains the ability for qualitative improvements; however, qualitative prompt-following capabilities of the simplified model greatly diminish, as evidenced by reduced correlation between target and predicted quality of the generated images: it drops from 0.97 to 0.27 SROCC. Furthermore, simultaneous processing of the new tokens with contextual information reduces the textual prompt-following capabilities of the model, as evidenced by FID and CLIP scores. This emphasizes the importance of the attention separation for qualitative conditioning. It also demonstrates that the the disengagement of qualitative and contextual information is beneficial for learning content-independent relationships between quality-related image properties.

12.2. Positional Encoding

Given that the quality metrics used as input for the IQA-Adapter form a low-dimensional representation (e.g., a 2D space for quality and aesthetics, as shown in Figure 1), we explored the use of positional encoding to enrich these inputs. Inspired by the sinusoidal encoding strategy employed in NeRFs[95] and timestamp encoding in Stable Diffusion models[17], we applied the following transformation to each input IQA/IAA value independently:

$$\gamma(x) = \big(x, \sin(2^0\pi x), \cos(2^0\pi x), \ldots, \sin(2^{L-1}\pi x), \cos(2^{L-1}\pi x)\big),$$

where x is the input value, and L controls the number of additional components in the representation. All IQA/IAA inputs were normalized to zero mean and unit variance prior to this transformation.

We hypothesized that positional encoding would enhance the model's sensitivity to subtle quality variations, allowing for more fine-grained control over output quality without affecting behavior at the edges of the input range. However, our experiments demonstrated that positional encoding had minimal impact on the model's behavior.

To evaluate this, we conducted experiments where the IQA-Adapter was modulated on the input quality condition, as described in Sections 4.3 and 15. Using a dataset of user-generated prompts from Lexica.art, we compared IQA-Adapters with and without positional encoding across a range of evaluation metrics. The results, shown in Figure 7, indicate that positional encoding produced outcomes nearly identical to those of the baseline IQA-Adapter, regardless of the value of L.

Although our experiments did not reveal significant benefits from positional encoding for the quality-conditioning task, we believe there may be potential for improvement with alternative encoding strategies. For instance, rotary positional embeddings (RoPE)[96], which have shown success in recent large language models, could be a promising direction. We leave the exploration of such strategies for future research.

12.3. Impact of IQA-Adapter scaling factor

To evaluate the impact of the adapter scale parameter λ on the visual quality of generated images, we tested IQA-Adapters trained with various IQA/IAA models under both high- and low-quality input conditions. We evaluated 9 λ values ranging from 0.05 to 1.0. For each configuration, images were generated using 300 randomly sampled prompts from the Lexica.art dataset. The results are shown in Figure 9.

As λ increases, image quality scores deviate progressively from the base model's levels, aligning with the specified quality condition. Under high-quality conditions, the

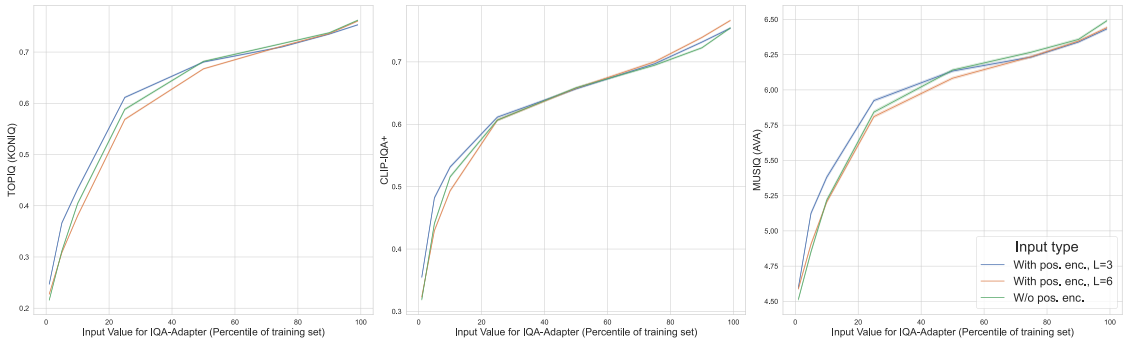

Figure 7. Results of the IQA-Adapter modulation on input quality-condition for different types of input preprocessing with positional encoding. For all evaluated types, adapter was trained with TOPIQ (KonIQ) model.

increase in quality is smooth and resembles a logarithmic curve for most adapters, reflecting diminishing returns as the base model already achieves relatively high-quality outputs. Beyond a certain threshold for λ, typically around 0.75, further increases cease to improve quality, with excessively high values ($\lambda > 0.9$) introducing artifacts that reduce both visual quality and IQA/IAA scores.

In low-quality conditions, the quality degradation progresses more rapidly, as the adapter has greater freedom to modify the image. The decrease in scores follows a sigmoidal trend: minimal change occurs for small λ values, but the effect accelerates significantly beyond $\lambda \sim 0.4$ and plateaus at the adapter's limits near $\lambda \sim 0.75 - 0.85$. This behavior highlights the non-linear relationship between adapter strength and its impact on image quality, with optimal performance generally observed for λ values in the range of [0.5, 0.75] for both low- and high-quality conditioning.

13. High-quality conditioning: more results

13.1. Gradient-based guidance

Figure 10(b) presents the relative gain in metric scores when using the gradient-based approach to optimize image quality during generation for prompts from PartiPrompts [77]. Unlike IQA-Adapter, direct optimization of the target metric improves that specific metric alone, while most other quality metrics tend to decline. This observation highlights the adversarial nature of gradient-based guidance, further confirmed by a closer examination of changes in generated images, which reveal adversarial patterns (as shown in Figure 24). Interestingly, certain metrics, such as ARNIQA (trained on KADID), LAION-AES, and LIQE MIX, show improvements even when unrelated quality metrics are targeted for optimization. This behavior points to their inherent instability and susceptibility to adversarial attacks, raising questions about their robustness as quality measures.

13.2. IQA-Adapter

Figure 11 presents detailed results for all tested IQA-Adapters on Lexica.art dataset, complementing Figure 3 (a) from the main paper. Figure 10 (a) provides additional results of high-quality conditioning with IQA-Adapter on PartiPrompts. The results on this dataset mirror the trends observed on the Lexica.art prompts, discussed in Section 4.2. Specifically, conditioning on the 99th percentile of target metrics not only boosts the target metrics themselves but also improves most other metrics, highlighting the strong transferability of IQA-Adapter. However, the average metric improvements on PartiPrompts are 1–2% lower than those observed on Lexica.art. This discrepancy can likely be attributed to the quality and completeness of the prompts. Unlike the more detailed and descriptive prompts in Lexica.art, PartiPrompts consists of shorter and more generic prompts. These simpler prompts impose fewer demands on the generation process, limiting the need for detailed generation, which is one of a key factors behind the significant metric improvements achieved by IQA-Adapter on Lexica.art.

Figure 8 demonstrates the comparison of IQA-Adapter with existing generation quality improvement methods on prompts sampled from Lexica.art dataset. IQA-Adapter conditioned on high quality usually results in sharper and more detailed results.

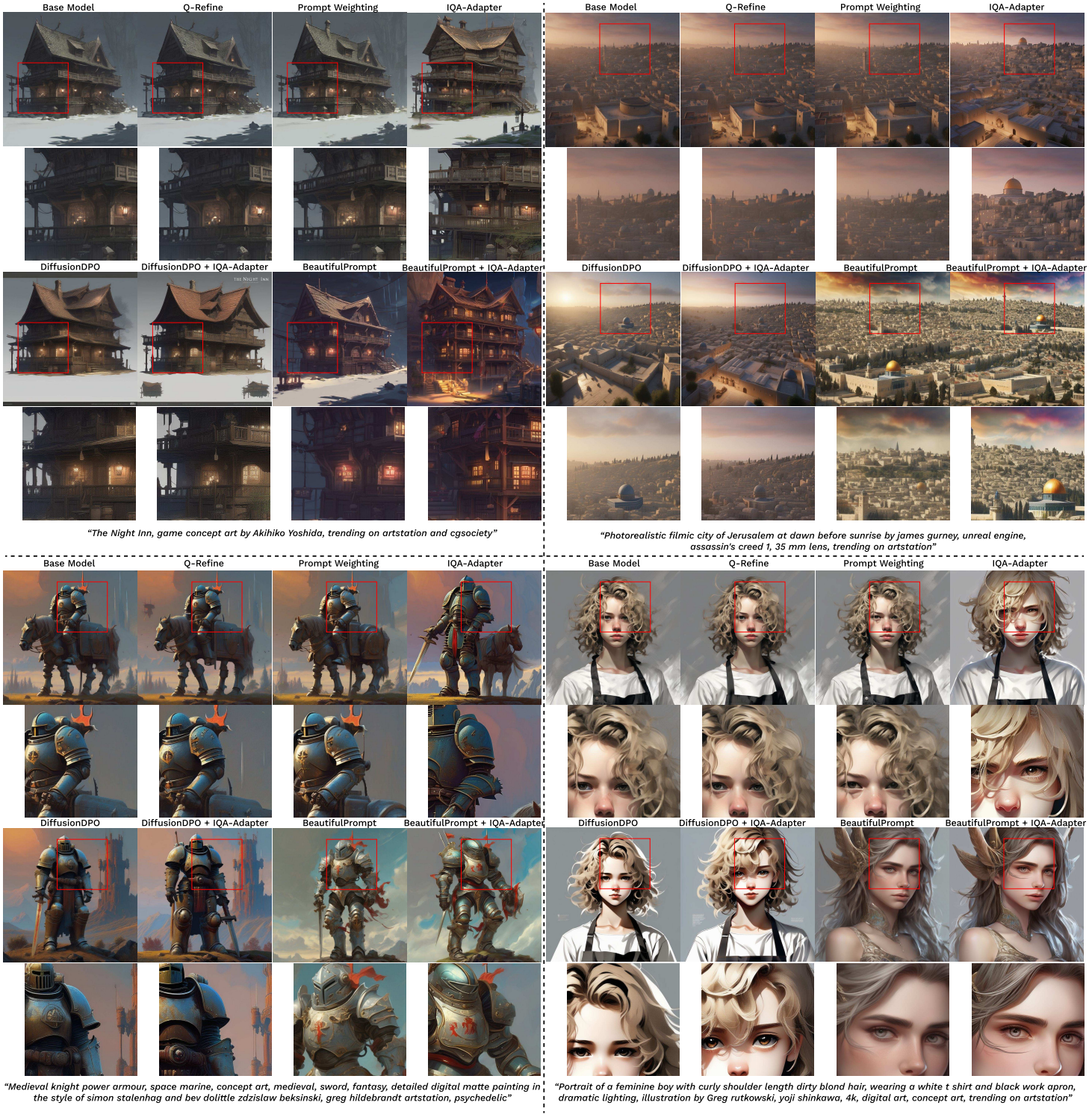

Figure 8. Comparison of different generation quality improvement methods.

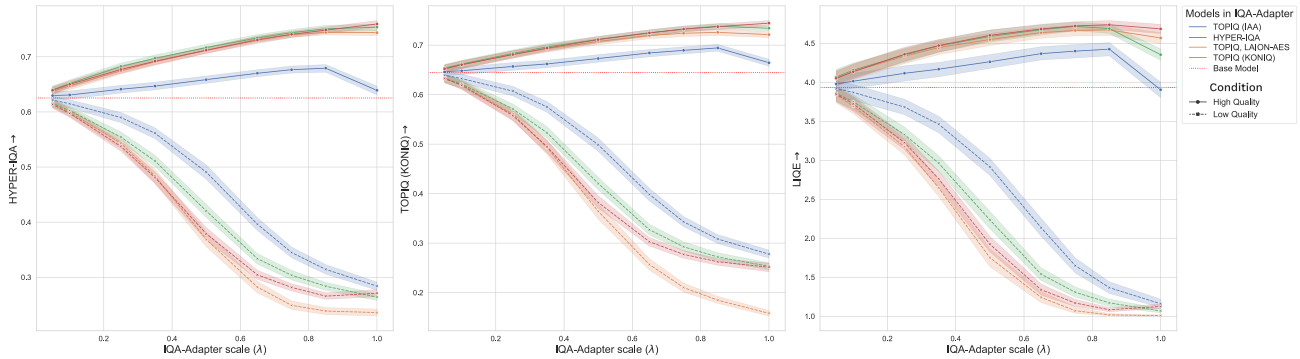

Figure 9. The relationship between image-quality scores (evaluated by the HYPER-IQA, TOPIQ and LIQE metrics) and the adapter scale parameter (λ) for the IQA-Adapters trained with different target IQA/IAA models and conditioned on low (dashed line) and high (solid line) target quality. For reference, the red dotted line indicates the quality level of the base model. The experiment utilized 300 random user-generated prompts from the Lexica.art dataset.

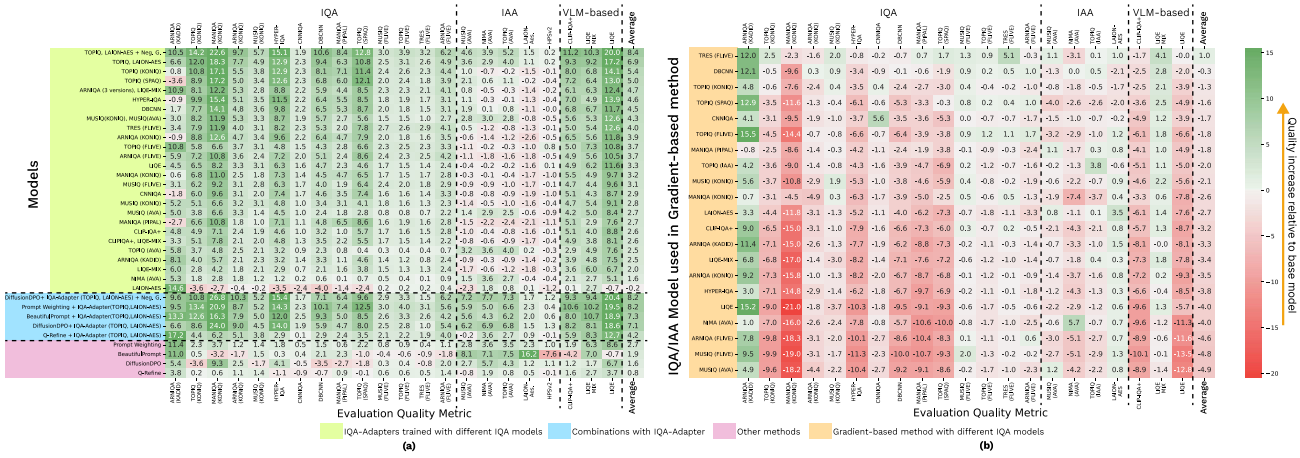

Figure 10. Quality improvement relative to base model (in %) for the IQA-Adapters trained on different IQA/IAA models and other generation quality improvement methods (a); and gradient-based method targeted on different IQA/IAA models (b). All IQA-Adapters are conditioned with high target quality (99th percentile of the training dataset) and use the same prompts and seeds. Prompts are taken from PartiPrompts dataset.

14. Evaluating Generative Capabilities: more results

Table 6 provides the complete results on the GenEval benchmark. Among the 25 evaluated IQA-Adapters, five outperform the Base Model in terms of the overall score. Notably, even the weakest IQA-Adapter surpasses the Base Model in the Counting and Position metrics. However, the best-performing IQA-Adapter underperforms the Base Model in the Two Object, Colors, and Single Object metrics. Overall, while all IQA-Adapters achieve performance levels comparable to the initial model, some manage to outperform it in specific areas.

Table 7 presents quantitative results for the FID, IS, and CLIP-similarity metrics. With a few exceptions, most IQA-Adapters exhibit slightly higher FID scores on the full MS COCO training dataset compared to the Base Model. This can be attributed to the diverse quality distribution of the dataset, which contains images of varying visual fidelity. Since IQA-Adapters are conditioned to prioritize high-quality generation, they naturally shift the output distribution toward a more specific subdomain characterized by higher visual quality. As a result, the distance to the broader, more heterogeneous image distribution of the full dataset increases. To address this domain shift, we also calculate FID scores on high-quality subsets of the MS COCO training dataset. These subsets include the top 10% and 25% of images, selected based on average quality scores from multiple IQA and IAA models. In this scenario, most IQA-Adapters consistently achieve lower FID scores than the Base Model, demonstrating superior alignment with the high-quality subsets.

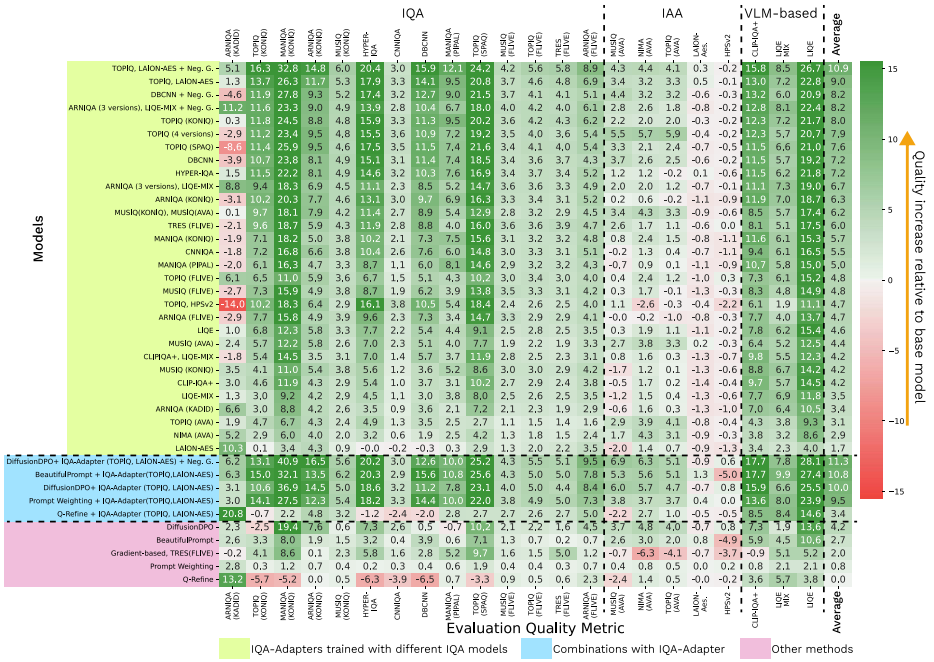

Figure 11. Quality improvement relative to base model (in %) for the IQA-Adapters trained on different IQA/IAA models and other generation quality improvement methods on Lexica.art dataset. This Figure complements the results reported in Figure 3 in the main paper.

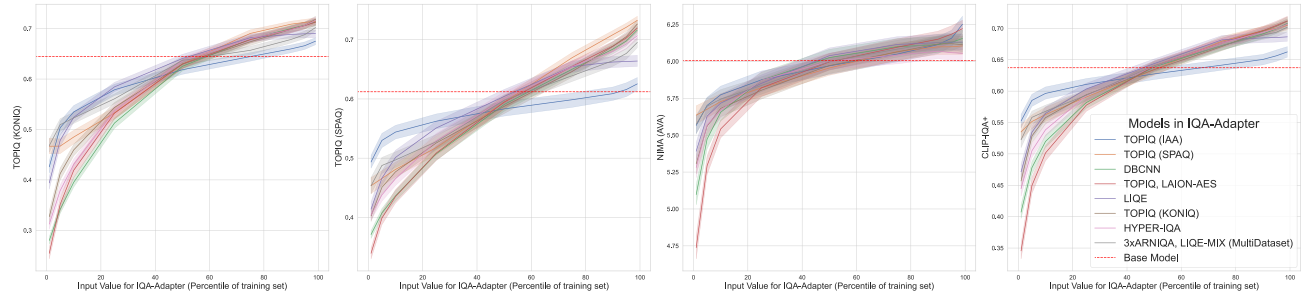

Figure 12. The relationship between input quality-condition (represented as a percentile of target IQA/IAA model on the training dataset) and image-quality scores evaluated by four different metrics (TOPIQ (KonIQ), TOPIQ (SPAQ), CLIP-IQA+, LIQE).

In addition to FID, we evaluate the Inception Score (IS) and CLIP-similarity metrics. CLIP-Text (CLIP-T) measures the similarity between generated images and their corresponding text prompts, using COCO captions as prompts in our experiment. CLIP-Image (CLIP-I) measures the distance between generated images and the real images corresponding to the captions. Results indicate that most IQA-Adapters achieve better CLIP scores than the Base Model, highlighting improved prompt-following capabilities. However, the Inception Score results are slightly lower compared to the Base Model. It is worth noting that the IS differences fall within the confidence interval. Additionally, IS is not well-suited for evaluating SDXL model, which is trained on large-scale internet datasets [97]. Furthermore, as IQA-Adapters generate more complex and detailed images, the classifier behind Inception Score struggles to identify the main object within the scene, further complicating its evaluation.

15. Alignment with qualitative condition: more results

To further evaluate the relationship between the input quality conditions provided to the IQA-Adapter during image generation and the quality of the resulting images, we analyzed correlations between the target quality and various metric scores. Figure 17 shows estimated correlations for each trained IQA-Adapter. Generally, the metrics demonstrate a strong alignment with the target quality, with the highest correlations observed when comparing different IQA models. In contrast, weaker correlations are noted

Models in IQA-Adapter	Two Object↑	Attribute Binding↑	Colors↑	Counting↑	Single Object↑	Position↑	Overall↑
LAION-AES	65.40%	16.75%	84.57%	45.00%	97.50%	12.25%	53.58%
MANIQA (PIPAL)	73.23%	20.25%	86.17%	36.56%	96.56%	10.50%	53.88%
ARNIQA (FLIVE)	69.70%	18.50%	84.04%	42.50%	97.81%	12.25%	54.13%
TOPIQ (KONIQ)	71.97%	18.75%	85.11%	38.75%	98.12%	13.75%	54.41%
CLIPIQA+, LIQE-MIX	71.72%	20.25%	85.64%	41.25%	97.81%	11.75%	54.74%
LIQE-MIX	68.43%	19.50%	87.50%	43.12%	98.12%	12.75%	54.91%
MUSIQ (FLIVE)	69.19%	23.25%	88.30%	39.38%	99.06%	12.50%	55.28%
TOPIQ (4 versions)	72.47%	21.75%	87.77%	40.31%	97.19%	12.25%	55.29%
TOPIQ, LAION-AES	69.70%	18.75%	85.90%	45.31%	99.38%	13.00%	55.34%
TOPIQ(KONIQ), HPSv2	71.21%	22.25%	85.64%	42.50%	98.44%	12.25%	55.38%
CNNIQA	71.72%	19.50%	87.50%	41.56%	98.12%	14.25%	55.44%
MUSIQ (AVA)	69.44%	24.25%	86.97%	40.94%	99.06%	12.50%	55.53%
MUSIQ(KONIQ), MUSIQ(AVA)	73.23%	22.75%	86.44%	40.94%	98.12%	12.50%	55.66%
TOPIQ (SPAQ)	73.48%	21.25%	86.70%	43.75%	97.50%	12.50%	55.86%
ARNIQA (3 versions), LIQE-MIX	73.99%	19.25%	**89.36%**	39.69%	**99.69%**	13.75%	55.95%
MANIQA (KONIQ)	73.48%	25.75%	88.30%	38.75%	96.88%	12.75%	55.98%
LIQE	72.73%	21.75%	86.97%	41.56%	98.75%	14.25%	56.00%
NIMA (AVA)	70.96%	23.00%	87.50%	44.69%	98.44%	11.50%	56.01%
MUSIQ (KONIQ)	73.74%	21.00%	86.44%	46.25%	97.50%	11.50%	56.07%
ARNIQA (KONIQ)	71.97%	22.00%	87.50%	44.38%	98.12%	12.75%	56.12%
CLIP-IQA+	72.73%	22.75%	88.03%	43.44%	98.44%	12.25%	56.27%
HYPER-IQA	73.99%	25.25%	85.90%	39.69%	98.75%	**14.75%**	56.39%
DBCNN	73.48%	22.75%	86.44%	44.38%	99.06%	13.00%	56.52%
ARNIQA (KADID)	72.98%	23.25%	86.97%	45.94%	98.75%	11.50%	56.56%
TOPIQ (AVA)	75.00%	22.50%	87.77%	42.81%	98.12%	13.50%	56.62%
TOPIQ (FLIVE)	72.73%	21.75%	87.77%	45.94%	99.38%	13.00%	56.76%
Base Model	73.74%	21.75%	88.30%	43.75%	99.69%	10.50%	56.29%
Gradient-based, TReS IQA model	61.87%	17.00%	81.91%	41.88%	95.31%	11.75%	51.62%
DiffusionDPO	**83.33%**	26.50%	87.77%	47.81%	99.69%	12.50%	59.60%
Q-Refine	70.96%	21.75%	88.83%	40.94%	99.06%	9.75%	55.21%
Prompt Weighting	71.21%	23.00%	87.23%	43.12%	99.38%	11.50%	55.91%
BeautifulPrompt	18.94%	1.00%	35.90%	9.38%	72.81%	4.75%	23.80%
DiffusionDPO + IQA-Adapter (TOPIQ, LAION-AES)	83.08%	26.50%	87.77%	45.94%	99.06%	13.75%	59.35%
DiffusionDPO + IQA-Adapter(TOPIQ, HPSv2)	80.30%	**31.00%**	86.97%	**50.62%**	99.06%	12.50%	**60.08%**
Q-Refine + IQA-Adapter (TOPIQ, LAION-AES)	68.94%	19.00%	86.70%	44.69%	98.44%	11.75%	54.92%

Table 6. GenEval, more results. The best results are **bold**, the second- and third-best are underlined. Table is sorted over "Overall" column.

when IQA models are compared with IAA models. Among the evaluated metrics, the poorest correlations are associated with images generated using the IQA-Adapter based on the IAA metric, LAION-Aes. Interestingly, even the metric's own values fail to exhibit significant correlation, which may be attributed to the IQA-Adapter training process, specifically the additional fine-tuning step. However, when LAION-Aes is paired with an IQA metric, the correlations with IAA models improves significantly. For example, the IQA-Adapter trained on the TOPIQ and LAION-Aes metrics achieves high correlations with both IQA and IAA models, making it an optimal choice for generating images with high visual quality.

Additionally, Figure 12 illustrates the relationship between the average scores of four metrics and the input-quality conditions across different IQA-Adapters. All metrics show a monotonic increase in their mean scores, reinforcing the strong correlations shown in Figure 17. This trend is consistent across all IQA-Adapter types, regardless of whether they are trained on IQA models, IAA models, or VLM-based approaches. Starting from a specific target percentile — typically around the 50th percentile — the mean metric scores surpass those of the base model.

16. IQA-Adapter as a degradation model

16.1. Examples of progressive quality degradation

Figures 19 and 20 illustrate the generation results for different percentiles of metric scores on the training dataset. As the percentile decreases, the generated images begin to exhibit various distortions, such as compression artifacts, noise, blurring, and others. These distortions are likely present in the corresponding training datasets for the metrics, causing them to become sensitive to these distortions and assign lower scores. By passing progressively lower scores to the adapter, we can approximate a continuous path in the image-space between low and high-quality images on the ends of the spectrum. This qualitatively monotonic

Models in IQA-Adapter	FID↓ Full	FID↓ (Top-25%)	FID↓ (Top-10%)	IS↑	CLIP-T↑	CLIP-I↑
LAION-AES	23.94	28.96	34.53	34.27±0.85	26.73	69.75
MUSIQ(KONIQ), MUSIQ(AVA)	22.48	24.96	29.68	37.00±1.43	26.79	69.47
NIMA (AVA)	22.32	25.65	30.55	37.72±1.08	26.70	69.80
TRES (FLIVE)	22.27	22.82	27.21	37.90±0.76	26.50	69.52
TOPIQ (AVA)	22.25	25.50	30.40	36.86±0.94	26.78	69.83
ARNIQA (3 versions), LIQE-MIX	21.95	22.92	27.58	37.55±1.02	26.69	69.62
TOPIQ (4 versions)	21.93	23.69	28.32	36.99±1.76	26.79	69.74
MANIQA (KONIQ)	21.74	23.85	28.57	37.63±1.23	26.91	69.61
CLIPIQA+, LIQE-MIX	21.43	22.45	27.02	38.33±1.83	26.70	69.65
TOPIQ, LAION-AES	21.36	23.53	28.44	36.89±1.33	26.83	<u>70.02</u>
MUSIQ (AVA)	21.20	24.92	30.08	36.42±1.39	26.93	69.96
ARNIQA (KONIQ)	21.13	22.70	27.53	37.32±0.87	26.86	69.53
TOPIQ (FLIVE)	21.04	**21.63**	**26.28**	37.93±0.70	26.64	69.54
HYPER-IQA	21.00	22.82	27.69	37.99±1.19	26.90	69.26
DBCNN	20.85	22.43	27.20	38.28±1.44	26.84	69.60
MUSIQ (KONIQ)	20.77	22.38	27.08	<u>38.57±1.12</u>	26.80	69.55
LIQE	20.76	22.34	27.21	37.72±1.46	26.82	69.81
CLIP-IQA+	20.45	21.89	<u>26.55</u>	37.66±1.05	26.80	**70.05**
ARNIQA (FLIVE)	20.44	<u>21.75</u>	<u>26.58</u>	38.25±1.20	26.85	<u>69.99</u>
ARNIQA (KADID)	20.35	22.50	27.56	37.67±1.31	26.76	69.32
LIQE-MIX	20.35	22.26	27.18	38.09±1.02	26.79	69.65
TOPIQ (SPAQ)	20.28	22.85	27.79	37.07±1.12	26.84	69.26
TOPIQ (KONIQ)	20.17	21.95	26.90	37.29±1.15	<u>26.96</u>	69.49
TOPIQ, HPSv2	<u>19.67</u>	22.08	27.40	36.71±1.45	27.00	69.12
CNNIQA	<u>19.61</u>	22.40	27.53	37.87±1.18	26.94	69.31
MANIQA (PIPAL)	**19.27**	<u>21.88</u>	27.19	37.98±1.46	26.77	69.48
Base Model	19.92	23.15	28.41	**39.44±1.66**	26.70	69.35
Gradient-based, TReS IQA model	25.02	29.97	35.84	33.34±1.28	24.88	64.77
BeautifulPrompt	30.92	35.64	40.83	33.30±1.12	21.23	58.01
DiffusionDPO	29.57	34.04	38.88	36.93±1.04	**27.10**	68.74
Prompt Weighting	24.14	26.02	30.50	38.44±2.15	26.42	68.78
Q-Refine	20.29	23.41	28.56	<u>39.05±1.11</u>	26.83	69.11

Table 7. FID, IS and CLIP scores of the IQA-Adapters trained with different IQA/IAA models on 10k subset of the MS COCO captions. FID-Full is calculated with the full MS COCO training dataset, and FID Top-n% measures FID to the highest-quality subset of MS COCO (as measured by the average score across all IQA/IAA metrics) of the corresponding size. The best results are **bold**, the second- and third-best are <u>underlined</u>. Table is sorted over "FID Full" column.

"path" (albeit with occasional local content changes) can potentially be used to train iterative image refinement algorithms.

This quality-modulation ability of IQA-Adapter enables leveraging diffusion models as degradation models to generate various distortions, including natural ones. To achieve this, the IQA-Adapter should be trained on a dataset containing the relevant distortions, using as guidance either subjective assessments or a specialized metric sensitive to these distortions. Exploring this approach will be the focus of our future research.

Figure 23 presents additional examples of generated distortions under low-quality conditioning. Furthermore, section 20 provides visualizations of Reference-based IQA-Adapter conditioning on different specific distortions.

16.2. Evaluating distances between high- and low-quality-conditioned generation

To investigate the differences between images generated with varying target quality levels, we estimated the distances between them using four FR IQA metrics: SSIM [33], LPIPS [98], DISTS [99], and PieAPP [100]. SSIM is a classical nonparametric method based on scene statistics, designed to assess structural similarity. LPIPS, on the other

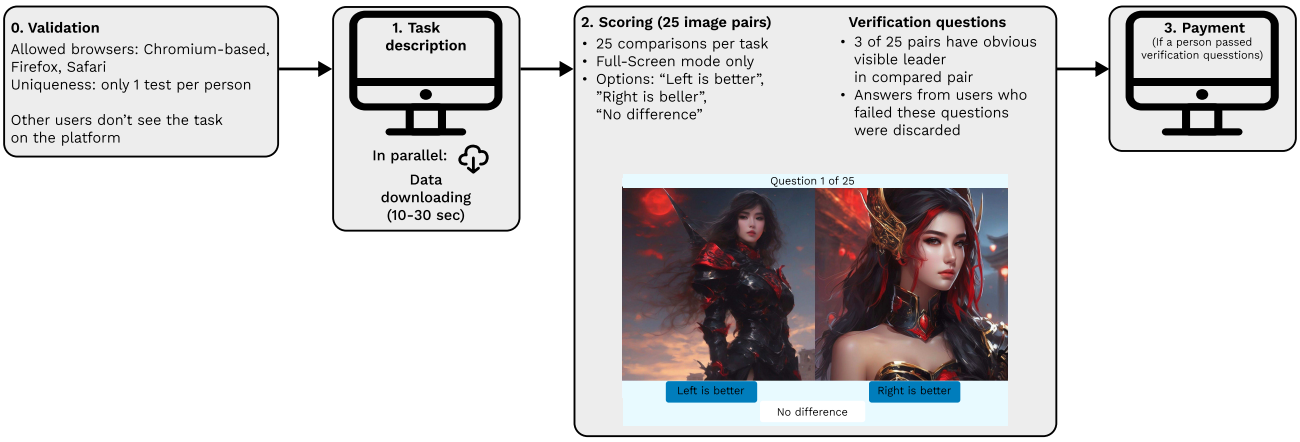

Figure 13. Overall scheme of the subjective study described in Sections 4.3 and 17.

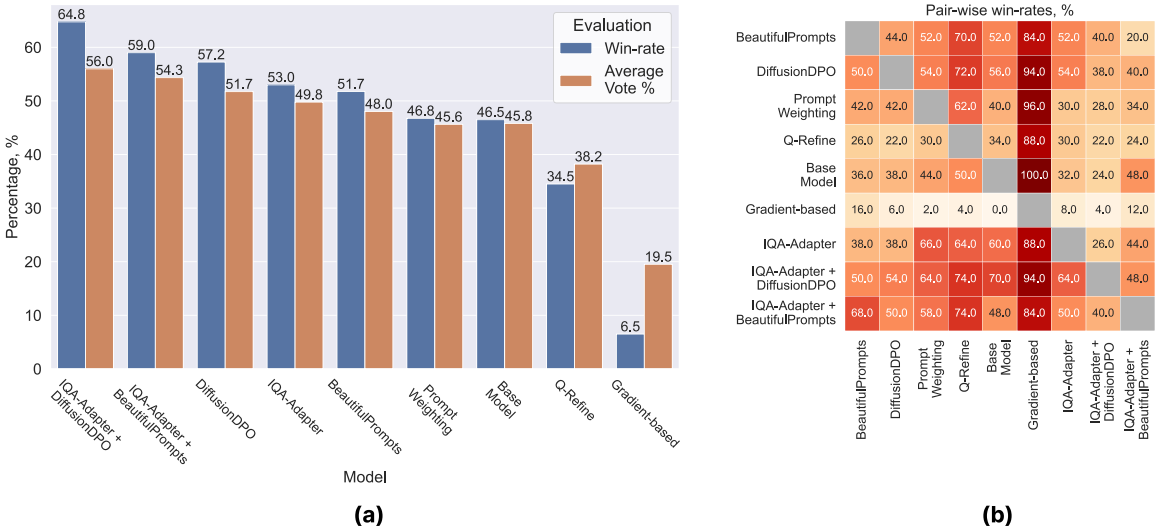

Figure 14. Results of the subjective comparison of generation quality improvement methods. (a) Win-rates and average vote % averaged across all image pairs. (b) Pair-wise win-rates between methods (percentage of wins of the model in the row against the model in the column). For more details, refer to Section 17.2.

hand, is a neural network-based metric that measures similarity as the cosine distance between the features extracted from a pre-trained convolutional network. DISTS refines LPIPS by incorporating additional insensitivity to small image shifts, making it more robust. Lastly, PieAPP demonstrates strong correlations with subjective scores, particularly for the super-resolution (SR) task [101].

We generated 8,200 images with user-generated prompts from the Lexica.art website for each target quality level (percentile of metric scores on the training dataset). Figure 15 shows the average distances between corresponding images across different percentiles, measured using the selected FR metrics. As the gap between percentiles increases, the distance between them grows consistently as well. High-quality percentiles (90, 95, 99) are the closest to each other, whereas distant percentiles (e.g., 1 and 99) differ significantly, mostly because of the introduced semantic variations. In contrast, the nearest 2–3 percentiles are quite similar, with differences primarily in small details. Notably, DISTS shows lower differences than LPIPS, suggesting the presence of minor content shifts between images in different percentiles.

17. Subjective Study

17.1. Alignment with qualitative condition

Subjective study described in Section 4.3 employed 300 randomly sampled user-generated prompts from the Lexica.art dataset. We used Subjectify.us platform for the evaluation. Overall scheme of the subjective study and the ex-

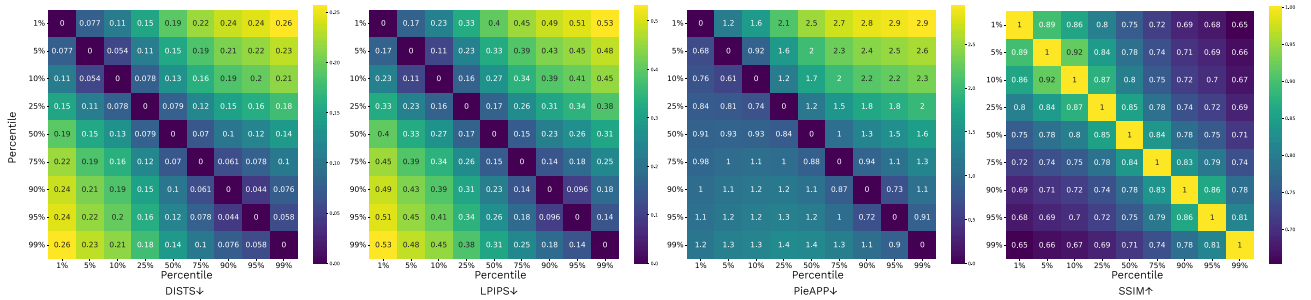

Figure 15. FR IQA metrics distances between images generated with the IQA-Adapter conditioned on different target-quality levels. The IQA-Adapter is trained for HYPER-IQA model.

ample of the user interface is demonstrated on Figure 13. During this study, we collected more than 22,300 valid responses of 1,017 unique users: each image-pair was independently assessed by at least 10 unique participants. As we compared 4 models (3 quality-conditions for the IQA-Adapter and the base model), total number of compared image-pairs was $\frac{4 \cdot 3}{2} \times 300 = 1800$. Participants were asked to evaluate the visual quality of the images generated from the same prompts and seeds across all models. Each participant was shown 25 pairs of images from which he had to choose which of them had greater visual quality. The respondent also had the option of "equal quality" in case he could not make a clear choice. Each participant could complete the comparison only once. Of the 25 pairs shown, 3 questions were verification questions and had a clear leader in visual quality. The answers of participants who failed at least one verification question were excluded from the calculation of the results. Comparisons were allowed only in full-screen mode and only through one of the allowed browsers. Before completing the comparison, each participant was shown the following instructions:

> Thank you for participating in this evaluation.
>
> In this study, you will be shown pairs of images generated by different neural networks from the same text prompt. From each pair, please select the image you believe has higher visual quality. The images may often look quite similar, so in addition to overall "aesthetic appeal," consider factors such as clarity, contrast, brightness, color saturation, and so on. Pay attention to generation defects, such as extra fingers or distorted bodies. If you cannot perceive any difference between the images, you may select "No difference."
>
> The text prompt used to generate the images will not be shown, as this study focuses on evaluating visual quality, and not textual alignment. Please note that the test includes verification questions! In these cases, the differences between the images will be clear, and selecting "indistinguishable quality" will not be considered a valid response."

17.2. Comparing IQA-Adapter with other methods in generation quality improvement task

To compare IQA-Adapter conditioned on high quality with other generation quality improvement methods mentioned in sec. 4.2, we conducted an additional pair-wise subjective study in a similar setting as described above. We evaluated 9 models in this experiment: IQA-Adapter (trained with TOPIQ and LAION-Aesthetics models), Gradient-based method (with TReS IQA model), Base Model (SDXL-Base), DiffusionDPO, BeautifulPrompts, Prompt Weighting, Q-Refine, and combinations of IQA-Adapter with DiffusionDPO and BeautifulPrompts. Since the complexity of the pair-wise study scales quadratically with the number of models, we used only 50 images per model, resulting in $50 * \frac{9*8}{2} = 1800$ image pairs for the comparison. This study involved 850+ unique users and 18,000+ valid responses.

Results are presented in Figure 14. Win-Rate denotes the share of side-by-side pairs where the given model was preferred over another one by most participants, and average vote % represents the consistency of user votes by averaging the share of user votes for this model across all image pairs involving it.

Overall, combining IQA-Adapter with DiffusionDPO and BeautifulPrompts shows the best results, improving upon these methods alone and confirming a similar observation from the objective evaluation (Figure 3 (a) and Section 4.2 in the main paper). IQA-Adapter alone demonstrates a better win-rate than most of the other methods and the base model, but slightly underperforms against DiffusionDPO. Gradient-based method is the least preferred, likely due to artifacts that often appear under direct quality optimization (see Figures 6 and 24).

18. Additional Experiments

18.1. Computational Overhead

In Table 8, we report time measurements for different generation methods used in this work. All evaluations were

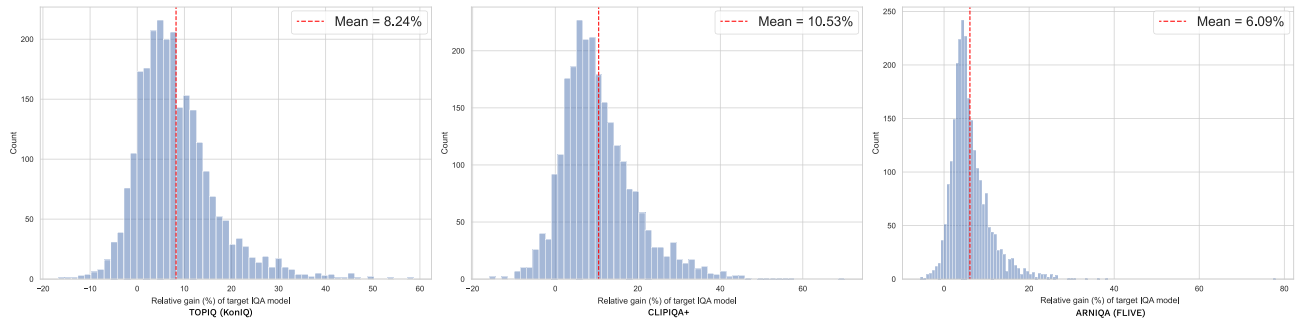

Figure 16. Distributions of relative gains defined in 4.2 across multiple generations with different seeds for IQA-Adapters trained with different IQA models. We use 25 random user-generated prompts and 100 seeds per prompt for this experiment.

Model	Time, s
Base Model (SDXL)	$3.83 \pm .04$
DiffusionDPO	$3.83 \pm .04$
IQA-Adapter w/o Separate Cross-Attn	$3.85 \pm .05$
Prompt Weighting	$3.93 \pm .04$
IQA-Adapter	$4.07 \pm .04$
BeautifulPrompt	$4.15 \pm .06$
Q-Refine	$(3.83 + 14.1) \pm .7$
IP-Adapter	$3.99 \pm .03$
Ref.-based IQA-Adapter	$4.11 \pm .04$
StyleCrafter	$7.66 \pm .08$

Table 8. Time complexity of different generative models and conditioning methods. See Section 18.1 for more details.

carried out in a similar environment on a dedicated server with a single Nvidia A100 GPU in float16 format and averaged across 1,000 generations. Images were generated in 1024x1024 resolution in 35 diffusion steps. We can see that the base model (SDXL) generates an image in $\sim$3.8s, and IQA-Adapter adds only $\sim$6% to the generation time. DiffusionDPO fine-tuning method does not add any inference-time computational overhead, and Q-Refine takes triple the time of the base model to refine an *already generated* image. Propmt refinement techniques generally do not add significant computational costs; however, BeatifulPrompt includes inference of a small Language Model, which adds few additional percents of computational overhead and memory use.

In the image-prompting scenario, Reference-based IQA-Adapter is a few milliseconds slower than IP-Adapter, mostly due to qualitative embedding extraction with the IQA model, and StyleCrafter is almost twice as slow as the other methods.

18.2. Consistency across different seeds

To evaluate the consistency of quality improvements across different seeds, we used 25 random user-generated prompts and sampled 100 random seeds for each, resulting in 2,500 generations per model. The same set of seeds was applied to both the base model and the IQA-Adapter. Figure 16 shows the distributions of relative gains (see Section 4.2) across all generations for adapters trained with different IQA/IAA metrics. Positive values indicate quality improvement relative to the base model for the same seed and prompt.

The results reveal that relative gains follow a unimodal distribution with a positive mean, indicating consistent quality improvement across generations. For some occasional seeds, the base model already achieves near-optimal quality scores and leaves limited room for improvement; in these instances, the adapter introduces negligible changes, resulting in gains close to zero.

Figure 22 illustrates images generated with the same prompt and different seeds, comparing the base model to the IQA-Adapter conditioned on high quality. For this demonstration, we used a strong adapter scale ($\lambda = 0.75$), which introduces noticeable stylization and detailing effects, particularly on high-frequency regions such as hair and textures.

18.3. Generation with different input quality-conditions

Figures 19 and 20 illustrate the effects of modulating the IQA-Adapter with progressively higher input quality conditions. From left to right, the target quality corresponds to increasing percentiles (1st to 99th) of the target model's scores on the training dataset. Different lines represent different IQA models used during adapter training. As the target quality increases, the generated images exhibit enhanced detail and clarity, demonstrating the adapter's ability to shift image quality in alignment with the specified condition.

19. Quality-conditioning and Adversarial Robustness of IQA models

Figure 24 presents a comparison of images generated by the base model (left column), the gradient-based method (middle column), and the IQA-Adapter (right column), alongside GradCAM visualizations of the target IQA model used for both gradient-based guidance and IQA-Adapter training. The gradient-based method often introduces artifacts that significantly alter the attention maps of the target model, inflating the quality score by exploiting architectural vulnerabilities. For instance, with the TOPIQ model (first row), new 'adversarial' objects are added to the image, capturing the model's attention and artificially boosting its scores. For TRES, grid-like patterns are generated that divert the model's focus away from the adversarial region. Similarly, with NIMA and HYPER-IQA, the method saturates the image with high-frequency details and color variations, dispersing the model's focus.

In contrast, the IQA-Adapter effectively preserves the target model's saliency maps, maintaining focus on relevant objects in the scene, even when the image undergoes structural modifications.

In summary, these findings underscore the potential negative impact of direct quality optimization, which can lead to the exploitation of the target quality estimator. Gradient backpropagation through the assessor model, either at inference time or during training (e.g., through the critic model in Reinforcement Learning-based approaches), can potentially exploit internal architectural vulnerabilities of the model. This makes the development of adversarially robust assessment models an important vector of future research.

IQA-Adapter largely avoids this problem by learning qualitative features across the entire quality spectrum during training instead of focusing on the optimizationtion of quality. However, we have also found out that under excessively large adapter scale ($\lambda \geq 1$) and strong negative guidance, IQA-Adapter can sometimes produce "over-stylized" images that are highly rated by many IQA/IAA models (Figure 21). This might indicate that the adapter identified qualitative preferences that are shared across multiple assessment models trained on different datasets and was able to exploit them.

20. Reference-based IQA-Adapter: more visualizations

Figure 25 demonstrates the comparison of Reference-based IQA-Adapter and IP-Adapter in image editing task. Figure 26 shows the results on Text-to-Image generation task with similar distortion references. It can be seen that other adapters copy objects and color palettes from the reference images and often fail to reproduce the distortion. We also note that we do not present the results of StyleCrafter in image editing since the official implementation of the adapter does not support SDXL Image-to-Image generation pipeline.

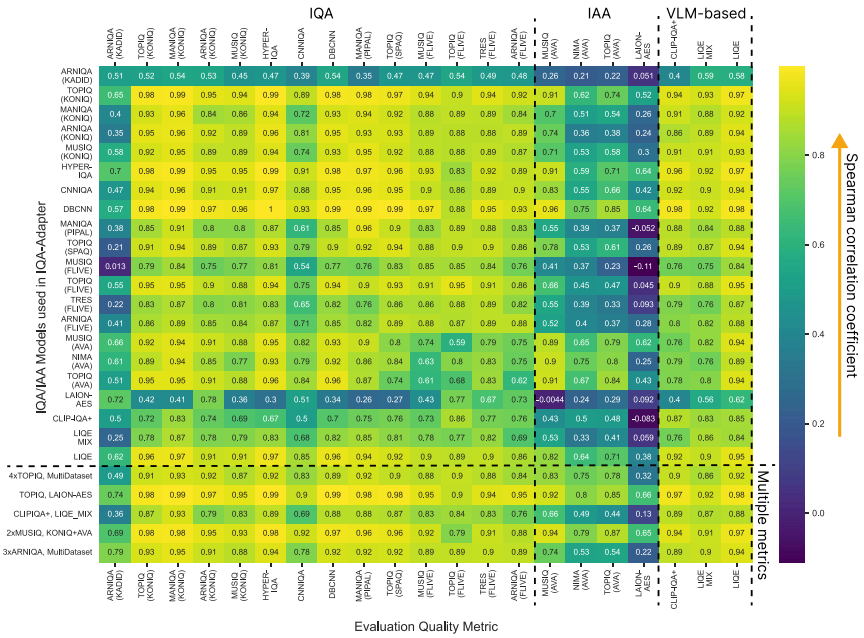

Figure 17. Correlations between input quality-conditions (represented as a percentile of target IQA/IAA model on the training dataset) and metric scores for the IQA-Adapters trained with different IQA/IAA models. Rows represent various IQA-Adapters, and columns indicate an IQA/IAA model used for SROCC calculation.

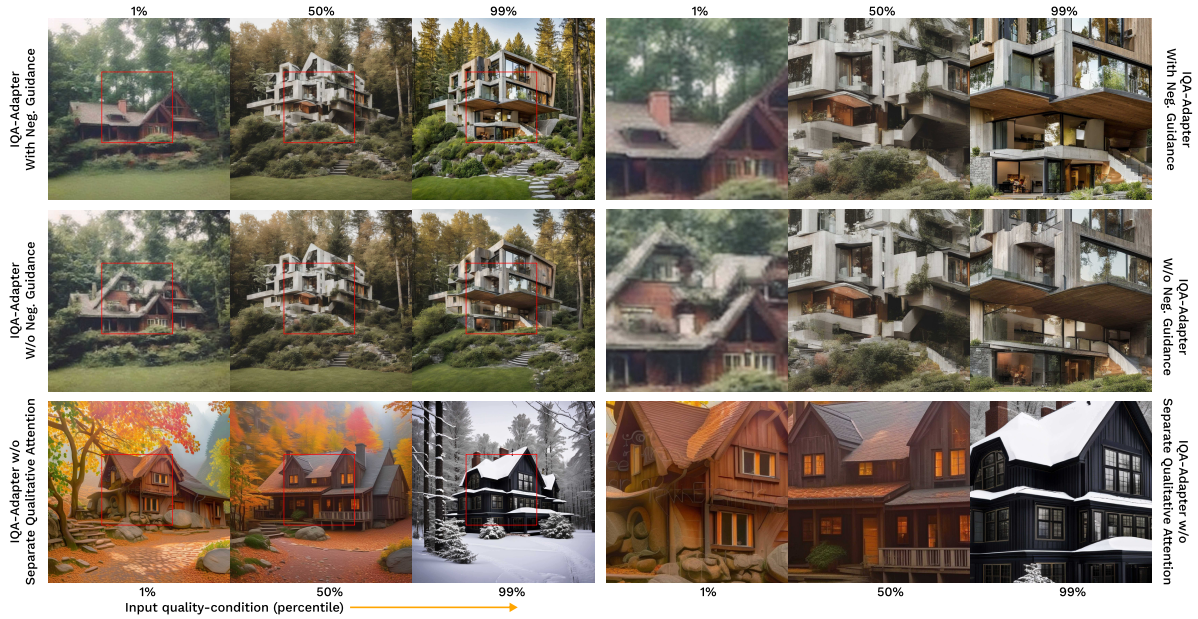

Figure 18. Ablation experiment: generations with IQA-Adapter with Neg. guidance enabled (1st row), without Neg. guidance (2nd row), and with a simplified IQA-Adapter without the Separate Qualitative Attention (3rd row). Simplified adapter exhibits poorer alignment with quality-condition and stronger content changes under different qualitative control signals. Negative guidance strengthens the effect of IQA-Adapter and magnifies the difference between low and high quality-conditions without significant content changes. Prompt: *'A beautiful house in the woods'*.

Figure 19. Visualization of generations with different target-quality conditions with IQA-Adapters trained with different IQA/IAA models. Input quality increases from left (1-st percentile of the training set) to right (99-th percentile).

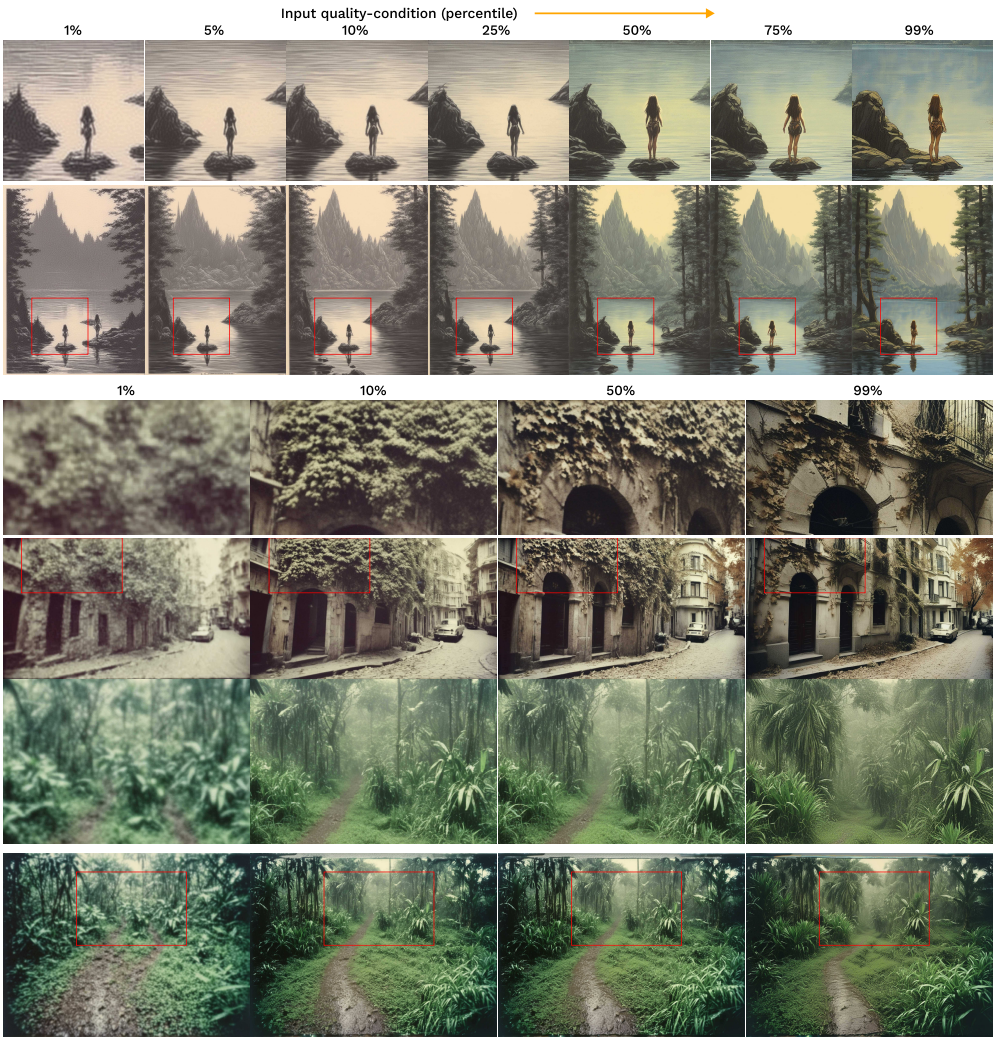

Figure 20. Additional visualizations of IQA-Adapter quality-modulation with different aspect ratios.

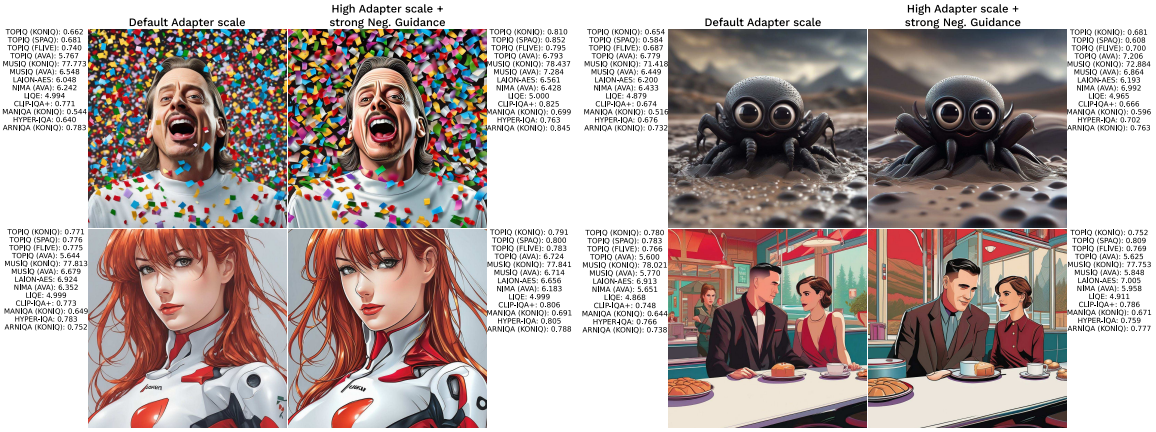

Figure 21. Example of images generated with and without strong negative guidance ($\delta = 1$) defined in Section 3.2.1 under high adaptive scale ($\lambda = 1$). Negative guidance magnifies the impact of the IQA-Adapter and occasionally results in the "over-stylisation" effect that is highly rated by most IQA/IAA models but usually does not reflect real quality improvement.

Figure 22. Examples of images generated with and without IQA-Adapter with the same prompt. The seeds are equal for corresponding images to the left and right. In this experiment, we employed the IQA-Adapter trained using the CLIP-IQA+ and LIQE-MIX models.

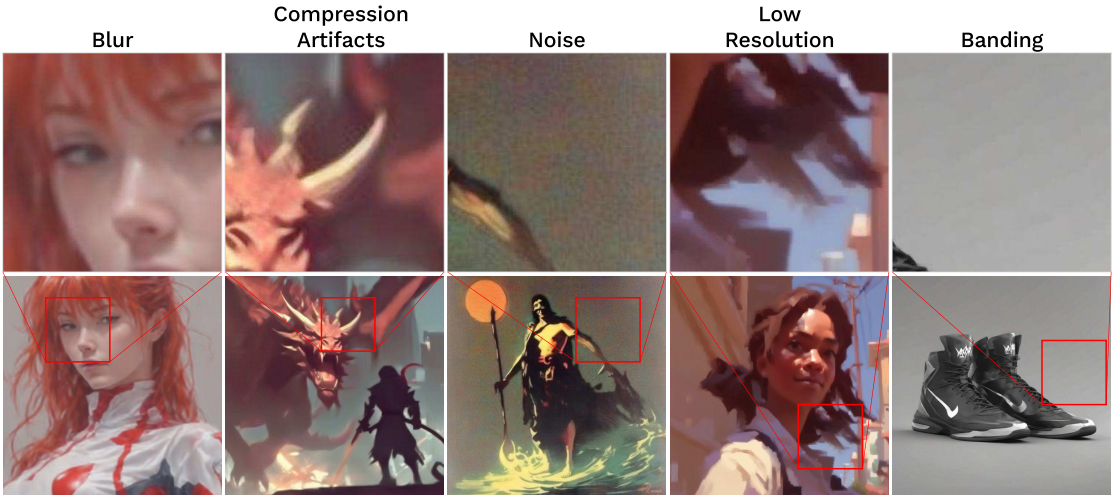

Figure 23. Examples of images generated with IQA-Adapter conditioned on **low** quality. IQA-Adapter is able to reproduce various distortions present in the training dataset.

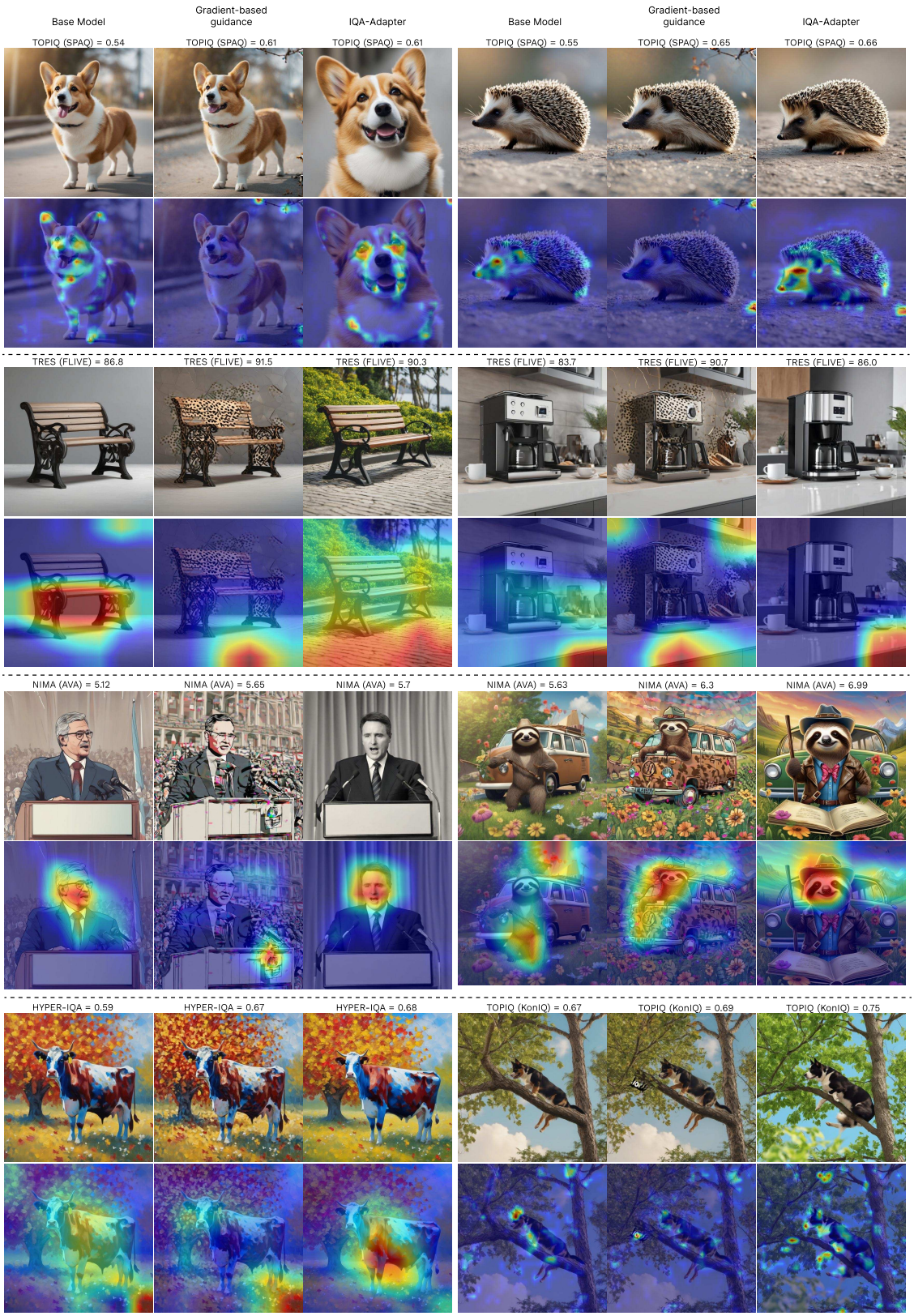

Figure 24. The comparison of adversarial examples generated with the gradient-based method (middle column) alongside outputs from the base model (left column) and the IQA-Adapter (right column), accompanied by their corresponding quality scores. Different rows represent different target IQA/IAA models in the gradient-based method and IQA-Adapter. Even-numbered rows display GradCAM visualizations of the target IQA model applied to the images in the respective columns. The prompts are taken from the PartiPrompts dataset.

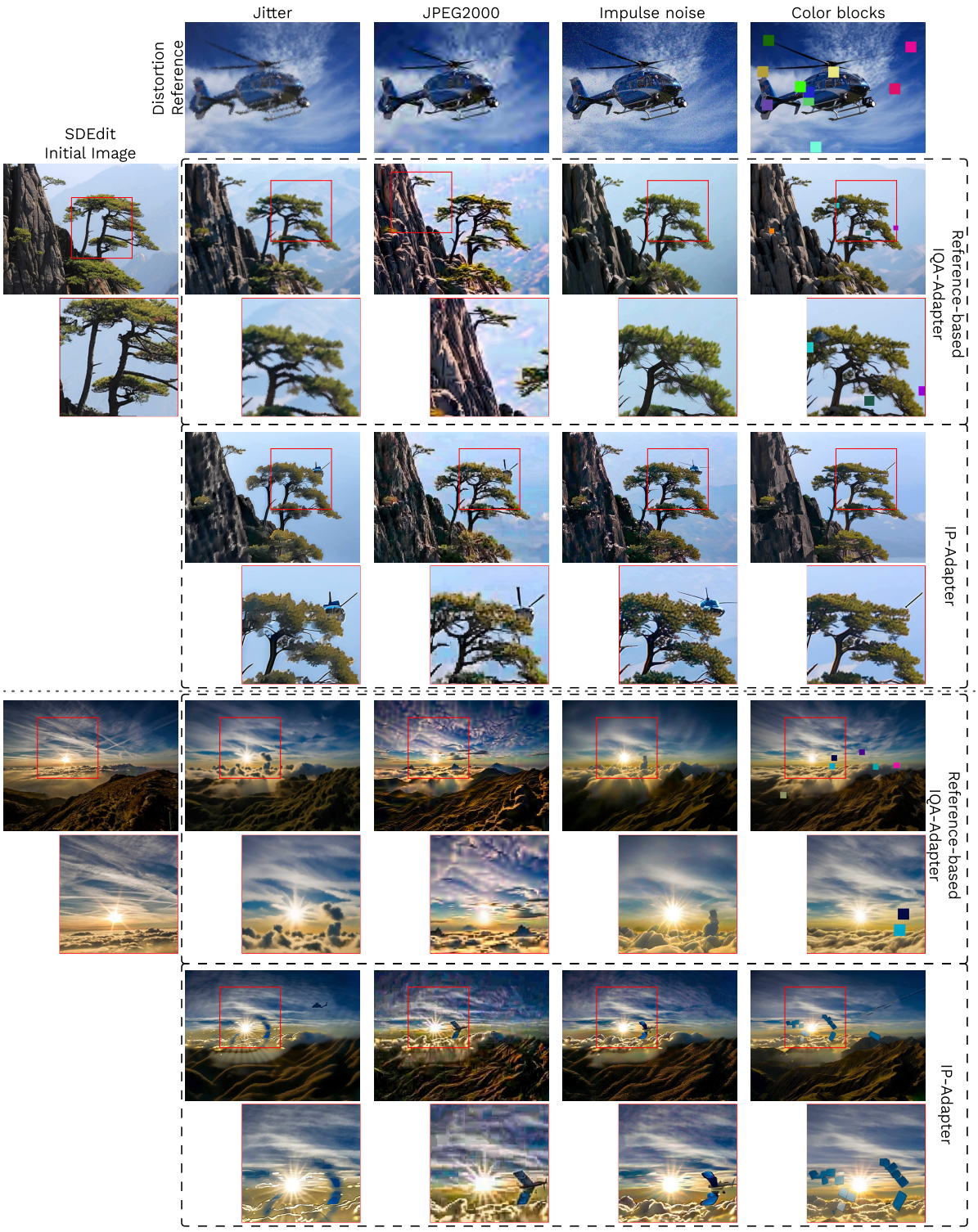

Figure 25. Reference-based Image Editing with SDEdit using a diffusion model equipped with Reference-based IQA-Adapter and IP-Adapter. IQA-Adapter transfers qualitative information more accurately, while IP-Adapter captures the semantics of the reference image.

Figure 26. Text-to-Image generation with qualitative reference. First row denotes generations with Reference-based IQA-Adapter and corresponding distortion reference, second — with IP-Adapter, and the last — with StyleCrafter adapter. Textual prompt for all generations: *"the sun rises over the clouds in the sky"*.