

Supplementary Materials: G-Refine: A General Quality Refiner for Text-to-Image Generation

Anonymous Authors

1 TRAINING DETAIL OF THE PQ-MAP

In the training process of the PQ-Map, we aim to obtain a set of text embeddings to represent the general concept of ‘perceptual quality’. Therefore, we first trained on AIGQA-20K [3], followed the paradigm of CLIPQA [7], and set the initial text input of CLIP [6] to: [‘A good image’, ‘A bad image’]. Then train the subjective annotation of AGIN [1] in three dimensions, and set the initial text input of CLIP to: [‘Low distortion’, ‘High distortion’]; [‘High rationality’, ‘Low rationality’]; [‘High naturalness’, ‘Low naturalness’]. Note that in the first dimension, since words like ‘technical quality’ and ‘signal fidelity’ are not easily understood by text encoders, we take their antonyms and invert them. From this, four groups of text embeddings are obtained to provide accurate quality maps.

2 EVALUATION CRITERIA ANALYSIS

All evaluation criteria we choose in Table 2 - 5 are highly consistent with human subjective preference. Whether for perceptual or alignment quality, the SRoCC of each indicator we selected and the human subjective ratings are above 0.5 [3]. Therefore, it’s generally believed that taking these 13 quality indicators into consideration, models that lead in most objective indicators will also have better subjective quality.

There are two reasons why we do not use FID, a commonly used indicator. First, past IQA research [4] has fully confirmed that the correlation between FID and the human subjective evaluation is poor, usually not exceeding 0.2. For example, it can distinguish the quality of noisy images from natural images; but for fine details, and diverse textures, it does not give a better score. In addition to the low correlation, the most fatal problem of FID is that the change is too small. *Since this article is about image enhancement rather than image generation, although the optimizer can greatly improve image quality, it will not change its basic structure* (this is also shown in the subsequent visualization in Figure 1). Therefore, FID (i.e. The distance to an image group) will not change significantly. Taking GenImage [11] as an example, the original FID is 158.40, and the strongest optimizer only improves it to 157.94; for DiffusionDB [9], the two values are 179.43 and 177.81. Such a weak change is not convincing enough to reflect the performance of optimizers. In short, we admit that FID is a meaningful indicator in image generation tasks, but due to the above two reasons, it was not taken into consideration in the experiments of this article.

3 ADDITIONAL EXPERIMENTAL RESULT

Considering that the optimized objects in the main text are relatively mainstream and traditional models. Here we have also optimized some unpopular, but more advanced models with higher initial generated image quality. As in the main text, we use the four strongest optimizers, namely InstructIR [2], SDXL [5], StableSR [8], and PASD [10], as comparisons to verify the performance of G-Refine. Experimental data in Table 1 shows that G-Refine still

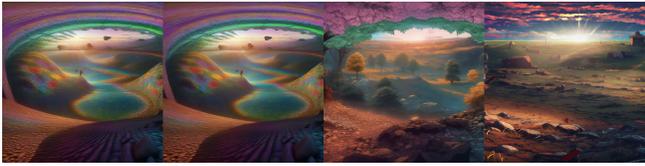
Table 1: Using different quality optimizers on DALLE 3, MJ 5.2, PG 2, PG 2.5, and SSD-1B. Abbreviations: MJ: MidJourney; PG: PlayGround. Other abbreviations and keys follow Table 2 in the main text.

DALLE3	UNIQUE	LIQE	DBCNN	MUSIQ	QALIGN	CLIPS	PicS
Original	1.3612	3.9102	0.6779	69.484	4.7557	0.9918	0.9885
InstructIR	0.4786	2.4663	0.4286	57.418	4.4993	0.9800	0.9534
PASD	1.3707	4.0405	0.6843	70.493	4.7609	0.9901	0.9667
SDXL	0.8187	3.5222	0.6638	67.290	4.6734	0.9954	0.9764
StableSR	0.6013	2.4745	0.4696	63.868	4.5092	0.9743	0.9070
G-Refine	1.3700	4.0490	0.6845	70.915	4.7447	0.9895	0.9473
MJ 5.2	UNIQUE	LIQE	DBCNN	MUSIQ	QALIGN	CLIPS	PicS
Original	0.9309	3.9316	0.6844	70.932	4.7380	0.9999	1.0833
InstructIR	0.3315	2.4198	0.3715	56.817	4.4336	0.9991	1.1121
PASD	1.2467	3.9384	0.6548	70.954	4.7899	0.9998	1.0632
SDXL	0.4841	3.2571	0.6541	67.507	4.6964	1.0000	1.0794
StableSR	0.5019	2.5586	0.4458	64.546	4.5992	0.9992	1.0311
G-Refine	1.2596	3.9454	0.6566	70.976	4.7821	0.9998	1.0476
PG 2.0	UNIQUE	LIQE	DBCNN	MUSIQ	QALIGN	CLIPS	PicS
Original	1.0388	3.6606	0.6788	70.568	4.6643	0.9906	0.7673
InstructIR	0.2855	2.2273	0.3712	54.418	4.3638	0.9892	0.7888
PASD	1.3150	4.0325	0.6776	72.538	4.7343	0.9913	0.7829
SDXL	0.6190	3.2248	0.6522	68.867	4.6783	0.9896	0.7643
StableSR	0.7287	2.7611	0.4815	66.293	4.6102	0.9891	0.7719
G-Refine	1.3364	4.0450	0.6777	72.560	4.6822	0.9880	0.7558
PG 2.5	UNIQUE	LIQE	DBCNN	MUSIQ	QALIGN	CLIPS	PicS
Original	1.1003	3.9623	0.6924	70.977	4.6882	0.9862	0.9688
InstructIR	0.2742	2.2716	0.3592	56.200	4.2638	0.9832	0.9815
PASD	1.2909	4.1251	0.6584	71.064	4.7306	0.9858	0.9568
SDXL	0.6856	3.5312	0.6568	68.194	4.7012	0.9887	0.9207
StableSR	0.6735	2.7393	0.4663	66.499	4.6015	0.9880	0.9650
G-Refine	1.3942	4.1848	0.6703	71.873	4.7324	0.9869	0.9524
SSD-1B	UNIQUE	LIQE	DBCNN	MUSIQ	QALIGN	CLIPS	PicS
Original	0.7569	3.8303	0.6636	71.381	4.4467	0.9960	0.9091
InstructIR	0.2961	2.2522	0.3854	58.124	4.0374	0.9914	0.9226
PASD	1.2137	4.1892	0.6819	73.268	4.5245	0.9946	0.8969
SDXL	0.8080	3.5678	0.6352	69.135	4.5663	0.9910	0.8912
StableSR	0.7626	2.9691	0.5051	67.470	4.3591	0.9912	0.8835
G-Refine	1.3784	4.2455	0.6878	73.711	4.5701	0.9882	0.8929

has an impressive ability in quality optimization, but compared with the experimental results in the main text, this advantage is not overwhelming. In terms of perceived quality, G-Refine maintains the first or second optimization effect and has satisfactory optimization capabilities; but for alignment, it almost keeps negative optimization. It is worth mentioning that negative optimization of alignment is not just a problem of G-Refine, but a common phenomenon. When the initial quality is high, although the model can adjust the intensity and avoid negative optimization for perceptual quality, at the alignment level, this intensity is not controlled perfectly. Therefore, optimizing the alignment quality of advanced models is a notable limitation for current optimizers to solve.

4 USER STUDY

To verify the practicality of G-Refine in industrial scenarios, we conducted subjective user studies in addition to objective indicators to analyze human preferences for compressed images. We focus on



(a) **AnimateDiff**: the beautiful, chilling, mundane panoramic view of a field after war filled with dead soldier calvary and rocks at **dusk**.



(b) **DALE 2**: an **a. i.**'s delusions of grandeur.



(c) **Dream**: 60s movie still of a soviet **Stalinist style** empty art museum with a soviet congress with yellow wall.



(d) **IF**: surly shaven **pudgy** British lad with short curly dark brown hair as a hobbit wearing a white men's sling chest bag and blue vest standing next to a giant rabbit.



(e) **MidJourney 5.2**: an impasto oil painting of **stick run**.



(f) **PixArt α** : an **ornate** illustration in the styles of mandalas and fractals, depicting a weasel staring deep into the heart of the impossible all.



(g) **PlayGround 2.0**: a komodo dragon with **dragon wings**, realistic painting, classical painting, high definition, digital art.



(h) **PlayGround 2.5**: a **hybrid robot elephant** on Socotra island, art germ, an epic fantasy, volumetric light, detailed, trending on art station, octane render, midsummer.



(i) **SD 1.5**: hunt showdown cowboys fighting for their life's against a **giant crocodile** in the bayou, by frank frazetta.



(j) **SD Cascade**: Microsoft Sam portrayed as a **person**, ultra realistic.



(k) **SDXL**: john wayne riding a **dolphin**. action figure by Hot Toys.studio lighting.



(l) **SSD-1B**: detailed, street photography, Moroccan, new york city, MTA subway entrance, public bus, bus stop, tree, car traffic, **traffic light**.

Figure 1: Visualization of different Text-to-Image generative model and corresponding prompt. From left-to-Right: Original, PASD, SDXL, G-Refine. G-Refine reached better perceptual quality while ensuring the alignment to keyword.

displaying the original image and two images processed by different optimizers on the 4,096*2,304 iMac monitor. The viewer needs to choose a preference between the two images on a perceptual

and congruent level. For a fair comparison, users can only see the above three images and corresponding prompts, and the specific optimizer source is invisible. The content optimized by GenImage

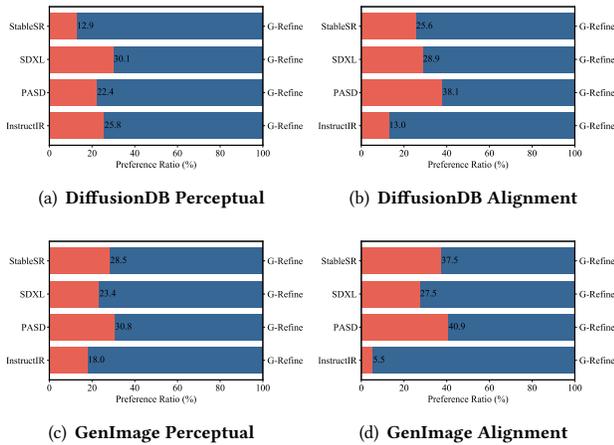


Figure 2: User preference for perceptual/alignment quality.

[11] and DiffusionDB [9] are annotated by 7 graduate students (4 males and 3 females). The objects chosen by the majority of people were identified as human preferences. The overall preference results are shown in Figure 2. G-Refine has demonstrated excellent performance on both data sets. Its perceptual quality optimization ability is the most outstanding. The proportion of human selection remains above 70%. Its T2I alignment optimization ability is also ahead of others.

5 VISUALIZATION RESULT

For 12 generative models, we visualize the original image and optimization results for PASD [10], SDXL [5], and G-Refine respectively. The images are center-cropped for better visualization. Figure 1 shows that when the original quality is poor, the optimization from PASD is limited; when the quality is good, SDXL will bring negative optimization. Only G-Refine can achieve substantial optimization in perception/alignment quality, regardless of initial quality. Therefore, with the widespread application of AIGI today, we believe that this optimization paradigm can promote various LMMs for a wider industrial application.

REFERENCES

- [1] Zijian Chen, Wei Sun, Haoning Wu, Zicheng Zhang, Jun Jia, Zhongpeng Ji, Fengyu Sun, Shangling Jui, Xiongkuo Min, Guangtao Zhai, and Wenjun Zhang. 2024. Exploring the Naturalness of AI-Generated Images. arXiv:2312.05476 [cs.CV]
- [2] Marcos V. Conde, Gregor Geigle, and Radu Timofte. 2024. InstructIR: High-Quality Image Restoration Following Human Instructions. arXiv:2401.16468 [cs.CV]
- [3] Chunyi Li, Tengchuan Kou, Yixuan Gao, Yuqin Cao, Wei Sun, Zicheng Zhang, Yingjie Zhou, Zhichao Zhang, Weixia Zhang, Haoning Wu, Xiaohong Liu, Xiongkuo Min, and Guangtao Zhai. 2024. AIGIQA-20K: A Large Database for AI-Generated Image Quality Assessment. arXiv:2404.03407 [cs.CV]
- [4] Chunyi Li, Zicheng Zhang, Haoning Wu, Wei Sun, Xiongkuo Min, Xiaohong Liu, Guangtao Zhai, and Weisi Lin. 2023. AGIQA-3K: An Open Database for AI-Generated Image Quality Assessment. *IEEE Transactions on Circuits and Systems for Video Technology* (2023).
- [5] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. arXiv:2307.01952 [cs.CV]
- [6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

- [7] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. 2023. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 2555–2563.
- [8] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin C. K. Chan, and Chen Change Loy. 2023. Exploiting Diffusion Prior for Real-World Image Super-Resolution. arXiv:2305.07015 [cs.CV]
- [9] Zijie J. Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. 2023. DiffusionDB: A Large-scale Prompt Gallery Dataset for Text-to-Image Generative Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 893–911. <https://doi.org/10.18653/v1/2023.acl-long.51>
- [10] Tao Yang, Rongyuan Wu, Peiran Ren, Xuansong Xie, and Lei Zhang. 2024. Pixel-Aware Stable Diffusion for Realistic Image Super-resolution and Personalized Stylization. arXiv:2308.14469 [cs.CV]
- [11] Mingjian Zhu, Hanting Chen, Qiangyu YAN, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. 2023. GenImage: A Million-Scale Benchmark for Detecting AI-Generated Image. In *Advances in Neural Information Processing Systems*, A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 77771–77782.

233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290

291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348