Not All That's Colorful Is Real: Rethinking Metrics for Image Colorization

Swarnim Maheswari[1]
cs25mtech02006@iith.ac.in

Syed Imam Ali[1]
ai24mtech14005@iith.ac.in

Panshul Jindal[1]
ai23btech11018@iith.ac.in

Arkaprava Majumdar[1]
ai24mtech02002@iith.ac.in

Vineeth N. Balasubramanian[1,2]
vineethnb@cse.iith.ac.in

[1]IIT Hyderabad [2]Microsoft Research, India

Abstract

Image colorization is the task of colorizing grayscale images. Unlike tasks with a well-defined ground truth, colorization is inherently ambiguous: a grayscale scene admits many plausible colorizations. Consequently, reference-based metrics are ill-suited for the problem. Distribution metrics such as FID cannot evaluate a single image and colorfulness scores often fail to reflect perceptual naturalness. We study how to evaluate image colorization at the single-image level. We benchmark 20+ no-reference IQA metrics and colorfulness variants across three datasets and >100k colorized images, and introduce a rank-based framework that compares how well each metric places the real image relative to its colorized variants. We find that NR-IQA metrics, especially HyperIQA and TOPIQ, consistently prefer real images over synthetic ones and align with distribution-level trends while providing per-image interpretability. Our study positions NR-IQA as a practical tool for evaluating colorization realism and offers a diagnostic benchmark for future methods.

1 Introduction

Image colorization—the process of assigning plausible colors to grayscale images—has long been a subject of interest in both computer vision and digital art. Formally, the task seeks to learn a mapping from intensity values to color channels under the constraint of semantic and contextual consistency. Beyond its technical appeal, image colorization has practical utility in restoring historical archives [17] or enabling creative re-colorization for media and design.

Recent advances in generative modeling, from convolutional architectures [**?**] to transformers [38, 19] and diffusion-based models [23, 25], have markedly improved the realism of colorized outputs. Yet, evaluating these outputs remains deeply problematic. Unlike tasks with a well-defined ground truth, colorization is inherently ambiguous: a grayscale scene admits many plausible colorizations. Consequently, reference-based metrics are ill-suited for the problem. Current practice relies on distribution-level measures such as Fréchet Inception Distance (FID) [13], or descriptors such as the Hasler colorfulness [12]. However, these metrics suffer from critical drawbacks: FID cannot

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The First Workshop on Generative and Protective AI for Content Creation.

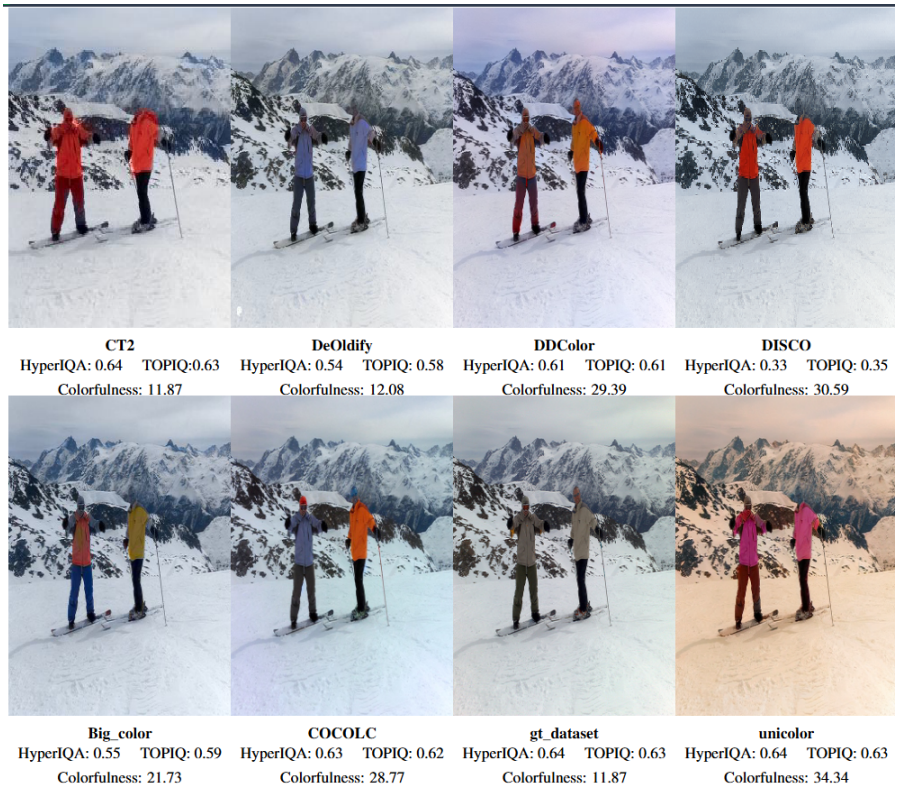

Figure 1: HyperIQA and TOPIQ prioritize realism: the real image (HyperIQA 0.64, TOPIQ 0.63, Colorfulness 11.87) scores higher than most synthetically colored image .

provide interpretable, per-image evaluations, and colorfulness scores often misrepresent perceptual naturalness, rewarding diverse, oversaturated colors but unrealistic outputs.

We propose that the evaluation bottleneck in colorization can be addressed by leveraging off-the-shelf image quality assessment (IQA) metrics. Though developed for perceptual quality tasks such as compression and distortion analysis, IQA methods are explicitly designed to measure naturalness in the absence of ground truth. We show that modern no-reference IQA metrics—particularly TOPIQ [8] and HyperIQA [32] —exhibit strong discriminatory power between natural photographs and their synthetically colorized variants. Moreover, these metrics consistently align with global measures such as FID while providing the additional benefit of per-image interpretability.

In this work, we present the first large-scale benchmark of IQA and colorfulness metrics for colorization. Using three datasets (COCO [6], Multi-Instance [39], ImageNet [29]) and several state-of-the-art methods, we generate over 100,000 recolorized images and evaluate them with 20+ classical and deep learning–based metrics. To enable fair comparison, we propose a rank-based evaluation framework, analyzing how well each metric aligns with ground-truth color distributions.

This study serves as both a benchmark and a diagnostic tool, guiding future work toward more reliable and perceptually consistent evaluation of colorization.

Our contributions are:

1. We demonstrate that existing colorfulness-based metrics fail to capture perceptual realism in colorized images.
2. Benchmark 30+ NR-IQA metrics on three datasets (>100k images) on various SOTA colorization methods, highlighting HyperIQA and TOPIQ as most consistent.
3. Introduce a simple real-rank framework to make heterogeneous metrics comparable across methods and datasets.

4. We propose IQA metrics as practical evaluation tools for colorization research, bridging the gap between human perception and automated benchmarking.

2 Related Works

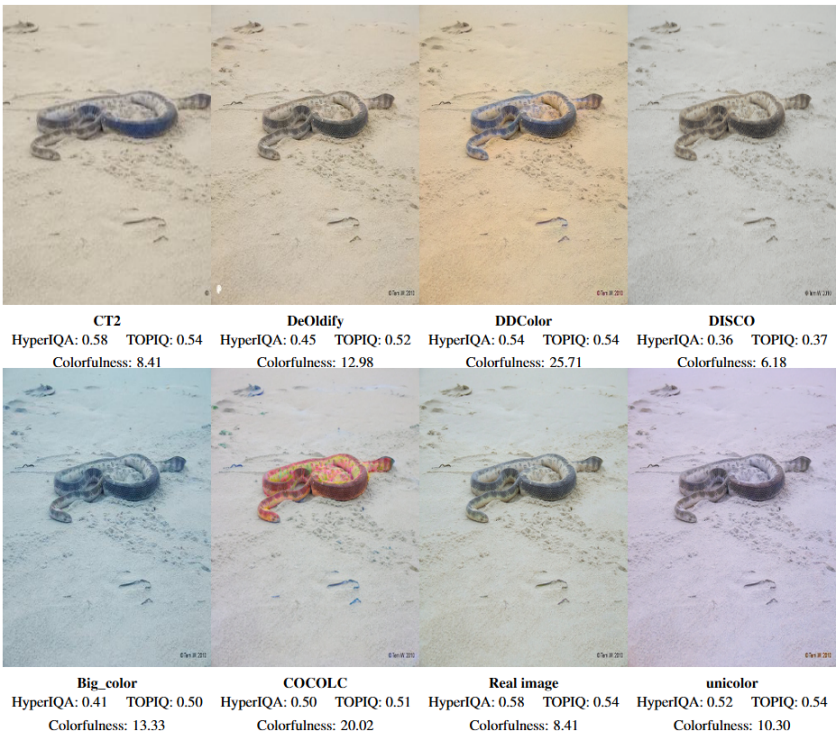

Figure 2: HyperIQA and TOPIQ prioritize realism: the real image scores higher (HyperIQA 0.58, TOPIQ 0.54, Colorfulness 8.41) while the edited 'red-snake' (COCOLC) has much higher colorfulness (20.02) but lower HyperIQA/TOPIQ — showing colorfulness can be fooled by vivid but unnatural edits

2.1 Image colorization

Early **CNN colorizers** fit a network to the color-image distribution: Cheng et al. introduced the first deep model but suffered from over-smoothing and limited data; CIC [47] framed colorization as classification in quantized CIELAB to avoid desaturated outputs; InstColor[31] colorizes instance crops with a fusion module (but can suffer from erroneous external priors and color overflow); and DISCO [41] uses a coarse-to-fine, anchor-based scheme to learn global affinities and reduce color ambiguity.

Diffusion models have shown strong generative capabilities for colorization by injecting grayscale structure into the pretrained diffusion model using controlnet [46, 25, 23, 24], midlayer injection using extended convolutional blocks [7] using grayscale image replacing the initial noise [44].

GAN-based colorization methods trade realism, speed, and cost: ChromaGAN [35] adopts a PatchGAN discriminator with a WGAN loss for realism, HistoryNet [17] augments generation with classification/segmentation modules and a large old-movie dataset, DeOldify [3] speeds training via asynchronous generator/discriminator updates, ToVivid [40] uses pretrained BigGAN [5] for inversion but suffers from inversion inaccuracies on grayscale inputs, and BigColor [21] embeds BigGAN's [5] generator/discriminator into an encoder–generator model at high computational cost and with occasional color artifacts.

Transformer-based colorization has advanced quickly: ColTran [22] proposed a multi-stage transformer colorizer but suffers from limited CNN-like inductive bias; ColorFormer [16] adds

a global–local hybrid self-attention and a color memory for efficient semantic–color mapping; ct2 [38] derives 313 meaningful color tokens from color-space statistics and uses adaptive attention to link them to luminance; DDColor [19] introduces learnable color tokens with cross-attention fusion of grayscale and color features; and MultiColor [10] builds on DDColor with a multi-branch design to better capture complementary and nuanced color cues.

2.2 Current metrics for image colorization

Methods similar to FID [13, 9][30] measure the feature distance between two datasets (real and recolored) to evaluate the naturalness of the entire dataset. This however cannot evaluate a single image.

Colorfulness [12] is widely used in image colorization methods for evaluation. It measures the diversity of color and color tint. It was created for evaluating goodness of color changes because of image compression and decompression.

2.3 IQA Metrics

IQA methods are classified into full-reference (FR) and no-reference (NR). FR metrics such as PSNR and SSIM [37] require ground truths and thus struggle with colorization's inherent ambiguity. NR metrics based on natural image statistics, e.g., BRISQUE [27], NIQE [28], and PIQE [34], capture distortions but ignore semantic plausibility. Learning-based NR-IQA methods, such as MUSIQ [20], MANIQA [42], and CLIPIQA [36], achieve strong benchmark performance but are not specifically trained on colorization artifacts. We review 21 no-reference IQA metrics on image colorization methods.

3 Methodology

To ensure a comprehensive evaluation, we employ three large-scale and diverse image datasets: COCO [6], Multi-Instance [39], and ImageNet [29]. For each dataset, we convert original color images to grayscale, which are then used as inputs to different colorization methods. For fair evaluation, we generated grayscale inputs using the same conversion procedure employed by each colorization method, since the real to grayscale step is part of the method and varies across approaches.

We compute over 25 evaluation metrics covering FID [13], sFID [9],FID-DINO [9], colorfulness hasler [12] and image quality assessment (IQA) metrics.

For each image, we decolorize and colorize it with all the methods, calcuate the metrics for each image and its colorized variants. For each image and its synthetically colored variants, every metric assigns a number scoring it. Which is followed by a real-rank-based evaluation framework for comparing goodness of various metrics.

3.1 Metric Comparison via Rank Normalization

Let $\mathcal{D} = \{x_i\}_{i=1}^n$ denote the set of test images. For each metric $M \in \{A, B\}$, let $v_i^{(M)} \in \mathbb{R}$ denote the raw score of image x_i.

Direction alignment. Suppose metric A is higher-is-better and metric B is lower-is-better. We define direction-aligned scores $\tilde{v}_i^{(M)}$ by

$$\tilde{v}_i^{(A)} = -\,v_i^{(A)}, \qquad \tilde{v}_i^{(B)} = v_i^{(B)},$$

so that in both cases, smaller values of $\tilde{v}$ correspond to better quality.

Rank transformation. For each metric M, compute ranks $r_i^{(M)} \in \{1, \ldots, n\}$ of the aligned scores $\tilde{v}^{(M)}$, where rank 1 corresponds to the smallest $\tilde{v}$ (the best image). Ties are resolved using average ranks.

We then scale the ranks to the unit interval:

$$s_i^{(M)} = \frac{r_i^{(M)} - 1}{n - 1}, \qquad s_i^{(M)} \in [0, 1].$$

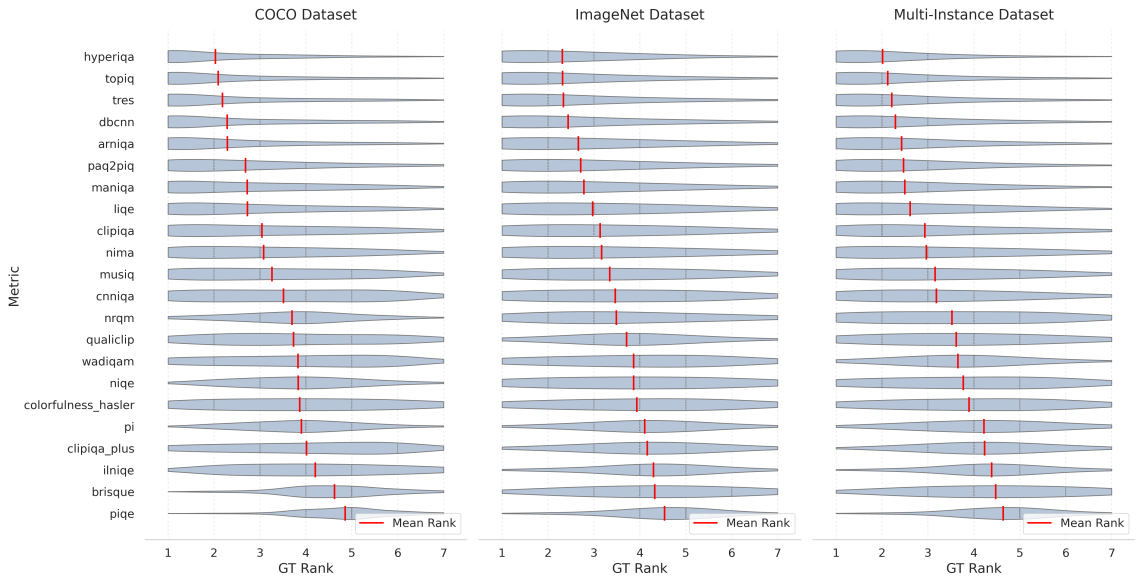

Figure 3: Real-rank distributions across COCO and ImageNet dataset. HyperIQA and topiq are consistently able to distinguish between real and synthetically colored images

Thus $s_i^{(M)} = 0$ for the best-ranked item and $s_i^{(M)} = 1$ for the worst.

Summary statistics. For each metric M, define the sample mean of the scaled ranks:

$$\bar{s}^{(M)} = \frac{1}{n} \sum_{i=1}^{n} s_i^{(M)},$$

Decision rule. We compare metrics A and B according to **mean:** Prefer the metric with the lower $\bar{s}^{(M)}$:

$$A \succ B \quad \Longleftrightarrow \quad \bar{s}^{(A)} < \bar{s}^{(B)}.$$

3.2 Interpretation

The mean $\bar{s}^{(M)}$ captures how highly a metric ranks images on average (lower is better). When two metrics have equal means, the standard deviation $\hat{\sigma}^{(M)}$ favors the more consistent metric. After ranking, we aggregate the results across all images and compute the distribution of ranks for each method and metric. This distribution provides insight into how consistently a method outperforms others across the dataset. This approach normalizes metric behaviors, avoids dependence on raw score magnitudes, and enables direct cross-metric comparability. The lower the rank of real image, the better the metric.

Limitation of this approach is that it cannot asses reference based metrics as they already take real image into account and score of real image will always be zero. Thus we do not study Δ Colorfulness [19], PSNR [15], SSIM [37] etc as they cannot be evaluated by this approach.

4 Experimental Results and Analysis

We evaluate a broad set of metrics across three datasets: the ImageNet 50K validation set (Table 2), the COCO test set (Table 1), and a multi-instance validation split (Table 3). For per-image realism, we use the GT-rank framework. For each image, the real photo (GT) is ranked among its colorized variants produced by competing methods; *lower GT rank is better*.

Figure 3 show, for each metric, the distribution across images of the (scaled) GT rank. Two consistent trends emerge. First, NR-IQA metrics substantially outperform colorfulness-based measures on all

datasets. Second, **HyperIQA** and **TOPIQ** are consistently among the top performers, attaining the lowest median (and mean) GT ranks across ImageNet, COCO, and the multi-instance split (see also Tables 1 to 3).

While distribution-level statistics such as FID and its variants are informative at the dataset level, they do not offer per-image interpretability. Our rank-based summaries complement these scores by revealing, for each image, how strongly a metric prefers the real photograph over its colorized counterparts.

Figures 1 and 2 present qualitative examples on two ImageNet images. NR-IQA metrics penalize artifacts characteristic of synthetic colorization—such as hue shifts, desaturation, and boundary bleeding—that simple colorfulness measures often miss, aligning better with perceived realism.

In summary, although most NR-IQA metrics were not designed specifically for colorization artifacts, our experiments indicate that they capture salient cues of naturalness and provide a practical, per-image signal for evaluating colorized outputs.

Table 1: Evaluation by various metrics on various methods on COCO test set. IQA metrics are given in blue, colorfulness metric is given in green, FID and variants are given in red.

	BigColor[21]	COCOLC[23]	DDColor[19]	DISCO[41]	unicolor[14]	DeOldify[3]	CIC_eccv16[47]	CIC_siggraph17[47]	InstColor[31]	Real
FID [13]	3.56	2.78	1.95	8.44	4.33	6.41				0.00
sFID [9]	1.67	2.99	1.21	2.28	1.36	2.35				0.00
FID-DINO [30]	30.45	15.08	13.76	48.60	28.63	34.62				0.00
colorfulness [12]	42.82	37.57	44.20	52.01	42.33	24.49	28.43	28.46	29.18	41.50
niqe [28]	3.77	3.77	3.89	4.85	3.87	3.83	3.87	3.87	5.39	3.89
musiq [20]	67.77	67.90	67.75	50.81	67.80	66.14	64.67	68.52	42.74	67.95
maniqa [42]	0.42	0.42	0.41	0.28	0.40	0.42	0.39	0.42	0.20	0.43
topiq [8]	0.55	0.58	0.58	0.34	0.58	0.55	0.53	0.58	0.28	0.60
piqe [34]	40.27	28.20	29.47	58.73	31.82	40.88	31.67	31.16	71.19	31.27
nima[33]	4.77	4.78	4.79	4.23	4.86	4.80	4.74	5.02	3.96	4.84
brisque[27]	19.34	14.31	14.40	37.92	15.26	19.96	15.35	15.22	43.68	15.34
ilniqe [45]	23.63	22.57	22.66	35.02	23.43	23.14	24.50	23.44	42.38	23.21
pi [50]	2.94	2.89	2.94	4.39	2.95	3.00	2.97	2.96	5.46	2.97
nrqm [26]	7.90	8.00	7.99	6.26	7.95	7.81	7.94	7.95	4.56	7.95
dbcnn [48]	0.58	0.61	0.61	0.34	0.61	0.58	0.57	0.59	0.35	0.62
liqe[49]	3.77	3.96	3.80	1.92	3.78	3.64	2.63	3.53	1.44	3.94
qualiclip [1]	0.73	0.76	0.74	0.61	0.75	0.73	0.61	0.71	0.47	0.74
arniqa [2]	0.67	0.68	0.69	0.56	0.69	0.67	0.64	0.66	0.48	0.70
tres [11]	76.46	78.75	77.61	45.62	78.79	76.44	72.36	78.64	35.92	80.23
clipiqa_plus [36]	0.64	0.64	0.63	0.51	0.65	0.64	0.58	0.63	0.50	0.63
paq2piq[43]	73.64	73.72	74.11	70.05	73.90	72.49	71.26	71.70	65.85	74.16
hyperiqa [32]	0.53	0.60	0.59	0.33	0.59	0.53	0.52	0.61	0.30	0.61
wadiqam [4]	-0.09	-0.11	-0.10	-0.80	-0.11	-0.10	-0.07	-0.10	-0.92	-0.11
cnniqa [18]	0.63	0.64	0.64	0.35	0.64	0.64	0.64	0.64	0.27	0.63
clipiqa [36]	0.56	0.55	0.53	0.42	0.56	0.56	0.47	0.52	0.35	0.57

Table 2: Evaluation by various metrics on various methods on ImageNet Validation set. IQA metrics are given in blue, colorfulness metric is given in green, FID and variants are given in red.

	BigColor [21]	COCOLC[23]	DDColor[19]	DISCO [41]	unicolor[14]	DeOldify[3]	CIC_eccv16[47]	CIC_siggraph17[47]	InstColor[31]	Real
FID [13]	2.62	–	1.35	5.71	3.65	4.77				0.00
sFID [9]	1.27	–	0.91	1.67	0.99	1.85				0.00
FID-DINO [30]	20.48	–	10.26	41.97	19.04	26.49				0.00
colorfulness [12]	42.83	35.48	44.70	51.33	42.28	22.64	28.10	27.58	27.70	41.17
niqe [28]	4.32	4.00	4.17	5.07	4.18	4.35	4.19	4.19	5.58	4.22
musiq [20]	65.67	66.27	66.12	53.21	66.33	65.26	63.68	66.75	47.02	66.16
maniqa [42]	0.41	0.40	0.40	0.30	0.39	0.41	0.38	0.40	0.23	0.41
topiq [8]	0.52	0.54	0.54	0.37	0.54	0.51	0.49	0.54	0.32	0.55
piqe [34]	43.38	29.58	32.64	58.74	35.06	43.25	34.98	34.56	69.10	34.70
nima [33]	4.72	4.74	4.74	4.26	4.80	4.74	4.66	4.92	4.03	4.79
brisque [27]	24.58	18.10	18.47	39.56	19.79	24.48	19.89	19.82	44.17	20.06
ilniqe [45]	28.58	26.07	26.63	38.55	27.62	27.68	28.44	27.20	44.74	27.32
pi [50]	3.43	3.18	3.26	4.68	3.30	3.45	3.31	3.32	5.56	3.33
nrqm [26]	7.53	7.72	7.72	5.92	7.67	7.48	7.65	7.65	4.57	7.64
dbcnn [48]	0.53	0.56	0.56	0.37	0.56	0.52	0.52	0.54	0.36	0.56
liqe [49]	3.63	3.83	3.66	2.25	3.73	3.46	2.72	3.52	1.70	3.81
qualiclip [1]	0.69	0.74	0.71	0.60	0.74	0.68	0.57	0.69	0.46	0.71
arniqa [2]	0.63	0.66	0.65	0.55	0.66	0.63	0.61	0.64	0.49	0.67
tres [11]	70.41	73.69	72.00	47.39	73.54	70.08	67.82	73.38	39.21	74.37
clipiqa_plus [36]	0.65	0.65	0.64	0.53	0.66	0.64	0.58	0.63	0.51	0.64
paq2piq [43]	72.91	73.10	73.42	70.16	73.44	71.46	70.67	71.17	66.06	73.58
hyperiqa [32]	0.50	0.56	0.54	0.36	0.55	0.49	0.49	0.56	0.33	0.56
wadiqam [4]	-0.20	-0.23	-0.22	-0.81	-0.23	-0.22	-0.20	-0.22	-0.91	-0.23
cnniqa [18]	0.59	0.59	0.59	0.34	0.59	0.59	0.59	0.59	0.28	0.59
clipiqa [36]	0.60	0.57	0.56	0.49	0.60	0.58	0.49	0.55	0.41	0.60

5 Conclusions, Limitations and Future Work

Our evaluation focuses solely on the naturalness of recolored images. We do not address the aesthetic quality of the chosen colors, which can vary widely depending on user preference or creative intent. Second, although IQA metrics capture perceptual quality at the per-image level, they were not specifically trained on distortions unique to colorization (e.g., hue bleeding, semantic color mismatches), which may limit their sensitivity in some cases. Finally, our analysis is restricted to still

Table 3: Evaluation by various metrics on various methods on multi-instance validation set. IQA metrics are given in blue, colorfulness metric is given in green, FID and variants are given in red.

method	BigColor [21]	COCOLC [23]	DDColor [19]	DISCO [41]	unicolor [14]	DeOldify[3]	CIC_eccv16[47]	CIC_siggraph17[47]	InstColor[31]	Real
FID [13]	8.37	6.69	5.09	13.15	8.86	11.18				0.00
sFID [9]	7.97	10.12	6.22	8.56	7.55	7.94				0.00
FID-DINO [30]	51.48	47.86	26.76	63.81	47.17	52.74				0.00
colorfulness [12]	43.77	38.86	45.30	52.35	43.16	24.17	28.71	28.25	30.79	44.31
niqe [28]	3.88	3.81	3.87	4.78	3.85	3.91	3.86	3.86	5.27	3.88
musiq [20]	67.58	68.74	67.69	50.54	67.84	67.32	64.61	68.73	42.83	68.25
maniqa [42]	0.41	0.39	0.40	0.27	0.39	0.41	0.38	0.40	0.19	0.41
topiq [8]	0.53	0.57	0.56	0.33	0.55	0.52	0.50	0.55	0.28	0.58
piqe [34]	42.58	32.60	32.37	60.60	34.98	42.85	34.96	34.48	71.33	34.60
nima [33]	4.72	4.77	4.74	4.20	4.82	4.74	4.70	4.98	3.96	4.78
brisque [27]	20.68	15.59	15.66	37.89	16.82	20.89	16.94	16.79	42.86	16.96
ilniqe [45]	23.32	22.59	22.01	34.46	22.86	22.77	23.78	22.91	41.27	22.61
pi [50]	3.02	2.94	2.95	4.39	2.97	3.07	2.98	2.98	5.37	2.98
nrqm [26]	7.80	7.87	7.92	6.13	7.87	7.70	7.86	7.87	4.58	7.86
dbcnn [48]	0.55	0.61	0.59	0.34	0.59	0.55	0.55	0.57	0.34	0.61
liqe [49]	3.66	4.06	3.74	1.86	3.71	3.47	2.53	3.45	1.42	3.96
qualiclip [1]	0.72	0.70	0.74	0.61	0.75	0.72	0.61	0.72	0.48	0.74
arniqa [2]	0.66	0.71	0.68	0.57	0.69	0.66	0.64	0.66	0.48	0.70
tres [11]	74.20	79.62	76.03	44.33	77.30	73.99	70.91	77.46	34.92	79.19
clipiqa_plus [36]	0.63	0.66	0.63	0.51	0.64	0.63	0.58	0.63	0.50	0.63
paq2piq [43]	73.74	74.37	74.33	70.10	74.11	72.54	71.52	71.90	66.32	74.58
hyperiqa [32]	0.50	0.60	0.57	0.32	0.57	0.50	0.49	0.58	0.30	0.59
wadiqam [4]	-0.12	-0.13	-0.13	-0.82	-0.14	-0.13	-0.10	-0.13	-0.92	-0.14
cnniqa [18]	0.62	0.62	0.62	0.35	0.62	0.62	0.62	0.62	0.28	0.62
clipiqa [36]	0.55	0.54	0.52	0.40	0.54	0.54	0.46	0.51	0.34	0.55

images; extending these insights to video colorization introduces temporal consistency challenges that are not covered in this work. We propose a visual Turing test that measures the fooling rate—the proportion of colorized images judged as real by humans—and hypothesize that NR-IQA scores will strongly correlate with this rate across methods.

Looking forward, several directions merit exploration. One avenue is fine-tuning IQA metrics on distortions characteristic of colorization models, potentially yielding better alignment with human perception. Another is the development of localizable IQA metrics that can highlight where color distortions occur within an image, providing more diagnostic feedback to model developers.

References

[1] Lorenzo Agnolucci, Leonardo Galteri, and Marco Bertini. Qualiclip: Quality-aware image-text alignment for real-world image quality assessment. *arXiv preprint arXiv:2403.11176*, 2024. Opinion-unaware, CLIP-based IQA (QualiCLIP and QualiCLIP+ variants).

[2] Lorenzo Agnolucci, Leonardo Galteri, Marco Bertini, and Alberto Del Bimbo. Arniqa: Learning distortion manifold for image quality assessment. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 189–198, 2024.

[3] Jason Antic. Deoldify: A deep learning based project for colorizing and restoring old images (and video!), 2019.

[4] Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek. Deep neural networks for no-reference and full-reference image quality assessment (diqam & wadiqam). *IEEE Transactions on Image Processing*, 27(1):206–219, 2018.

[5] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

[6] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218, 2018.

[7] Chang, Zheng and Weng, Shuchen and Zhang, Peixuan and Li, Yu and Li, Si and Shi, Boxin. L-cad: Language-based colorization with any-level descriptions using diffusion priors. In *Advances in Neural Information Processing Systems*, 2023.

[8] Chaofeng Chen, Jiadi Mo, Jingwen Hou, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin. Topiq: A top-down approach from semantics to distortions for image quality assessment. *arXiv preprint arXiv:2308.03060*, 2023. also published / available as IEEE TIP / conference paper (see linked source).

[9] Xin Ding, Yongwei Wang, Zuheng Xu, William J Welch, and Z Jane Wang. Continuous conditional generative adversarial networks for image generation: Novel losses and label input mechanisms. *arXiv preprint arXiv:2011.07466*, 2020.

[10] Xiangcheng Du, Zhao Zhou, Yanlong Wang, Zhuoyao Wang, Yingbin Zheng, and Cheng Jin. Multicolor: Image colorization by learning from multiple color spaces. *arXiv preprint arXiv:2408.04172*, 2024.

[11] S. Alireza Golestaneh, Saba Dadsetan, and Kris M. Kitani. No-reference image quality assessment via transformers, relative ranking, and self-consistency (tres). In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3209–3218, 2022.

[12] David Hasler and Sabine E Suesstrunk. Measuring colorfulness in natural images. In *Human vision and electronic imaging VIII*, volume 5007, pages 87–95. SPIE, 2003.

[13] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, 2017.

[14] Zhitong Huang, Nanxuan Zhao, and Jing Liao. Unicolor: A unified framework for multi-modal colorization with transformer. *ACM Transactions on Graphics*, 2022.

[15] Quan Huynh-Thu and Mohammed Ghanbari. Scope of validity of psnr in image/video quality assessment. *Electronics Letters*, 2008.

[16] Xiaozhong Ji, Boyuan Jiang, Donghao Luo, Guangpin Tao, Wenqing Chu, Zhifeng Xie, Chengjie Wang, and Ying Tai. Colorformer: Image colorization via color memory assisted hybrid-attention transformer. In *Proc. European Conf. Computer Vision*, 2022.

[17] Xin Jin, Zhonglan Li, Ke Liu, Dongqing Zou, Xiaodong Li, Xingfan Zhu, Ziyin Zhou, Qilong Sun, and Qingyu Liu. Focusing on persons: Colorizing old images learning from modern historical movies. In *Proc. ACM Int'l Conf. Multimedia*, 2021.

[18] Le Kang, Peng Ye, Yi Li, and David Doermann. Convolutional neural networks for no-reference image quality assessment (cnniqa). In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1733–1740, 2014.

[19] Xiaoyang Kang, Tao Yang, Wenqi Ouyang, Peiran Ren, Lingzhi Li, and Xuansong Xie. Ddcolor: Towards photo-realistic image colorization via dual decoders. In *Proc. Int'l Conf. Computer Vision*, 2023.

[20] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-scale Image Quality Transformer. In *IEEE/CVF international conference on computer vision*, 2021.

[21] Geonung Kim, Kyoungkook Kang, Seongtae Kim, Hwayoon Lee, Sehoon Kim, Jonghyun Kim, Seung-Hwan Baek, and Sunghyun Cho. Bigcolor: Colorization using a generative color prior for natural images. In *Proc. European Conf. Computer Vision*, 2022.

[22] Manoj Kumar, Dirk Weissenborn, and Nal Kalchbrenner. Colorization transformer. In *Proc. Int'l Conf. Learning Representations*, 2021.

[23] Yifan Li, Yuhang Bai, Shuai Yang, and Jiaying Liu. Coco-lc: Colorfulness controllable language-based colorization. In *Proc. ACM Int'l Conf. Multimedia*, 2024.

[24] Zhexin Liang, Zhaochen Li, Shangchen Zhou, Chongyi Li, and Chen Change Loy. Control color: Multimodal diffusion-based interactive image colorization. *arXiv:2402.10855*, 2024.

[25] Hanyuan Liu, Jinbo Xing, Minshan Xie, Chengze Li, and Tien-Tsin Wong. Improved diffusion-based image colorization via piggybacked models. *arXiv preprint arXiv:2304.11105*, 2023.

[26] Chao Ma, Chih-Yuan Yang, Xiaokang Yang, and Ming-Hsuan Yang. Learning a no-reference quality metric for single-image super-resolution. *Computer Vision and Image Understanding*, 158:1–16, 2017.

[27] Anish Mittal, Anush K. Moorthy, and Alan C. Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, 2012.

[28] Anish Mittal, Radhakrishnan Soundararajan, and Alan C. Bovik. Making a "completely blind" image quality evaluator. *IEEE Signal Processing Letters*, 20(3):209–212, 2013.

[29] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

[30] George Stein, Jesse Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Ross, Valentin Villecroze, Zhaoyan Liu, Anthony L Caterini, Eric Taylor, and Gabriel Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. *Advances in Neural Information Processing Systems*, 36:3732–3784, 2023.

[31] Jheng-Wei Su, Hung-Kuo Chu, and Jia-Bin Huang. Instance-aware image colorization. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2020.

[32] 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 (hyperiqa). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[33] Hossein Talebi and Peyman Milanfar. NIMA: Neural Image Assessment. In *IEEE Transactions on Image Processing*, 2018.

[34] N. Venkatanath, D. Praneeth, M. C. Maruthi Chandrasekhar Bh, S. S. Channappayya, and S. S. Medasani. Blind image quality evaluation using perception based features (piqe). In *2015 Twenty First National Conference on Communications (NCC)*, pages 1–6. IEEE, 2015.

[35] Patricia Vitoria, Lara Raad, and Coloma Ballester. Chromagan: Adversarial picture colorization with semantic class distribution. In *Proc. IEEE Winter Conf. Applications of Computer Vision*, March 2020.

[36] Jianyi Wang, Kelvin C.K. Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images (clip-iqa). In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023. zero-shot CLIP-based IQA; code: IceClear/CLIP-IQA.

[37] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 2004.

[38] Shuchen Weng, Jimeng Sun, Yu Li, Si Li, and Boxin Shi. Ct2: Colorization transformer via color tokens. In *Proc. European Conf. Computer Vision*, 2022.

[39] Shuchen Weng, Hao Wu, Zheng Chang, Jiajun Tang, Si Li, and Boxin Shi. L-code: Language-based colorization using color-object decoupled conditions. In *Proc. AAAI Conference of Artifical Intelligence*, 2022.

[40] Yanze Wu, Xintao Wang, Yu Li, Honglun Zhang, Xun Zhao, and Ying Shan. Towards vivid and diverse image colorization with generative color prior. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2021.

[41] Menghan Xia, Wenbo Hu, Tien Tsin Wong, and Jue Wang. Disentangled image colorization via global anchors. *ACM Transactions on Graphics*, 2022.

[42] Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment. *arXiv preprint arXiv:2204.08958*, 2022. NTIRE/CVPRW submissions and code available.

[43] Zhenqiang Ying, Haoran Niu, Praful Gupta, Dhruv Mahajan, Deepti Ghadiyaram, and Alan C. Bovik. From patches to pictures (paq-2-piq): Mapping the perceptual space of picture quality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3572–3582, 2020.

[44] Nir Zabari, Aharon Azulay, Alexey Gorkor, Tavi Halperin, and Ohad Fried. Diffusing colors: Image colorization with text guided diffusion. In *SIGGRAPH Asia 2023 Conference Papers*, 2023.

[45] Lin Zhang, Lei Zhang, and Alan C. Bovik. A feature-enriched completely blind image quality evaluator. *IEEE Transactions on Image Processing*, 24(8):2579–2591, 2015.

[46] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2023.

[47] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *Proc. European Conf. Computer Vision*, 2016.

[48] 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, 2020.

[49] Weixia Zhang, Guangtao Zhai, Ying Wei, et al. Blind image quality assessment via vision-language correspondence: A multitask learning perspective (liqe). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. Authors name the model LIQE (Language-Image Quality Evaluator).

[50] Zicheng Zhang, Chunyi Li, Wei Sun, Xiaohong Liu, Xiongkuo Min, and Guangtao Zhai. A Perceptual Quality Assessment Exploration for AIGC Images. In *IEEE International Conference on Multimedia and Expo Workshops*, 2023.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes, abstract and introduction reflect the paper's contributions and scope

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: A section on limitation is in included

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: No proof per say are presented

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All the details are included

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: Error bars are not reported because of computational and presentation reasons.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper analyses the current methods, and does not present its own method

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification:

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification:

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We simply analyze the current methods on current metric. No ethics guidelines were breached.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification:

    Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Yes.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: No new assests are provided

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

Justification:

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification:

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification:

    Guidelines: LLM was used editing, finding catchy title, sentence structure, gammer check etc.

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.