AI MEIMEI

Liu Maohan, Ng Min Teck, Zhang Zhiyuan, Zhu Xiaoyan, Tang Liqi

NUS-ISS, National University of Singapore, Singapore 119615

ABSTRACT

With the increasing global demand for high-quality photography and efficient image editing, our project introduces an AI-powered photo assistance system that enhances image capture and editing. Our system offers real-time photo-taking guidance, ensuring optimal composition, intelligent object repositioning for natural realignments, and real-world seamless outpainting that extends photo backgrounds using contextual data from real-world references. Unlike existing AI tools that generate generic or unrealistic edits, our approach integrates location-based references and advanced AI models to produce authentic and high-quality visual enhancements. Our solution significantly improves efficiency by automating complex editing tasks such as layer separation, object movement, and background extension, enabling users to achieve professional-grade results with minimal effort. By leveraging state-of-the-art deep learning techniques, including Smartphone Photography Attribute and Quality (SPAQ), Yolo11, U²-Net, Segment Anything Model (SAM), Real-ESRGAN and ControlNet our project aims to redefine photo editing with intelligent automation, contextual realism, and user-centric enhancements.

Keywords: AI Photo Enhancement, Real-Time Aesthetic Guidance, Intelligent Scene Manipulation, Image Dataset Integration

1 Introduction

Photography is a vital part of modern communication, self-expression, and content creation. However, many users struggle to capture and edit high-quality images due to poor composition, unnatural object placement, and the limitations of current AI tools. A 2020 study found over 40% of users lacked confidence in their photography skills, relying heavily on presets and automated suggestions [1]. Choi and Park (Researcher and CHI 2019 conference speaker) also found that users' limited knowledge of composition led to insecurity, but technical guidance improved confidence [2]. These findings underscore the need for real-time aesthetic guidance at the moment of capture. Traditional editors like Photoshop or Meitu require skill and time, while current AI tools often yield inconsistent or unrealistic results. Most assistive tools focus on post-processing, ignoring the importance of capturing strong composition, lighting, and color harmony from the start. With over 530 million photos taken daily, accessible, intelligent enhancement tools are more essential than ever. Our project aims to bridge the gap between AI automation and natural-looking edits through real-time, context-aware guidance requiring minimal effort.

Our system integrates state-of-the-art AI including SPAQ, Yolo11-l, U²-Net, SAM, Real-ESRGAN, and ControlNet to provide intuitive, high-fidelity image enhancement. Unlike current AI tools, it prioritizes realistic, context-aware edits that blend seamlessly with the scene. Users receive real-time feedback and enhancements via live camera input or uploaded images. The system automates complex tasks such as object detection, segmentation, background extension, and inpainting, delivering balanced, share-ready photos with minimal input.

This technology holds strong commercial appeal, particularly among mobile-first Gen Z users. A 2023 BMC Psychology report found over 50% of social media users edit photos before posting [3]. The global photo editing market, valued at USD 1.61 billion in 2022, is expected to grow to USD 3.21 billion by 2030 (CAGR: 8.1%) [4]. As smartphone camera tech advances, integrating intelligent editing features directly into camera software offers greater adaptability and B2B opportunities than standalone apps. This software-focused approach minimizes costs and risks while positioning us for collaboration with device manufacturers. With visual storytelling central to digital life, real-time AI tools are primed to meet rising demand and fuel market growth.

2 Literature review

2.1 Smartphone Photography Attribute and Quality

As smartphone photography becomes increasingly mainstream, researchers have focused on how to evaluate the aesthetic quality of these images in a way that reflects real user preferences. Recent studies emphasize multidimensional and deep learning-based approaches that assess factors such as brightness, color, contrast, and composition all at once[5] [6].

To make assessments more accurate and meaningful,

some models incorporate metadata and user behavior such as swipes or saves to provide context beyond the image itself [7] [8].This reflects a broader move toward personalized and context aware systems that adapt to how people actually interact with photos.

Subjectivity remains a challenge. Aesthetic judgment varies widely between individuals, so personalized models are gaining attention for their ability to account for different tastes[9]. Meanwhile, newer approaches are also exploring style-related aspects like color aesthetics using prompts[10] and comparing how professionals and amateurs compose images.

2.2 Detection with Yolo11-l

The field of AI-assisted photo editing has evolved rapidly in recent years, driven mainly by advancescomputer vision, deep learning, and human-computer interaction. The Yolo11-l model has emerged as a versatile framework for real-time object detection in various environments [11]. Luo et al. improved the model by integrating the BRA self-attention mechanism, creating Light-SA Yolo11-l to improve detection in complex natural settings [12]. Wang et al. adapted Yolo11-l for specialized agricultural applications such as estrus cow recognition, while also advancing rail surface defect detection techniques [13].

For industrial safety applications, Lin [14] developed Yolo11-l-SLIM-CA to improve helmet detection accuracy under varied conditions, while Fu et al. [15] proposed Yolo11-l-FADS specifically for the detection of miners' helmets in challenging underground environments, which achieved significant accuracy improvements. In the transportation sector, Tang et al. [16] combined transfer learning with Yolo11-l to enable fatigue detection in drivers wearing sunglasses, improving traffic safety technologies. These diverse applications demonstrate Yolo11-l's adaptability to addressmain-specific challenges while maintaining detection accuracy.

2.3 Advancements in Segmentation

Image segmentation is a crucial task in computer vision. Various models are developed to achieve accurate and efficient results. U2-Net, a deep learning-based semantic segmentation model, has gained popularity for its simplicity and effectiveness in background removal [17]. On the other hand, SAM has been identified as a strong encoder for U-shaped segmentation models, which showcased superior performance in vision understanding tasks [18]. Although U2-Net is known for its efficient background removal capabilities, SAM offers prompt segmentation for both images and videos, making it a versatile choice for a wide range of applications [19]. SAM's transformer architecture with streaming memory enables real-time video segmentation, distinguishing it from its predecessor SAM, which primarily focused on static images [19]. In addition, SAM has been praised for its increased segmentation accuracy. That would reduce the need for manual corrections and improve efficiency in labeling large datasets [20]. The choice between SAM and U2-Net for image segmentation depends on the specific requirements of the task at hand. While U²-Net excels in fast and efficient background removal for simple scenes, SAM provides promptable segmentation for both images and videos, offering higher accuracy and better performance in complex scenarios and dataset labeling tasks. Ultimately, the decision to choose SAM over U2-Net may be based on the need for real-time video segmentation and the versatility required for different applications [18].

2.4 AI-Powered Upscaling

AI-driven image upscaling has transformed resolution enhancement techniques. Michelini et al. [21] introduced edge-SR (eSR), bridging traditional methods with deep learning approaches through interpretable mechanisms. Real-ESRGAN, developed by xinntao [22], has become widely adopted for general image and video restoration, with integrations in platforms like Krita and Replicate API for face enhancement and flexible upscaling [23].

Despite its popularity, Real-ESRGAN has faced criticism regarding output quality issues, including artifacts appearing as random streaks and cartoon-like appearances when viewed up close. Some users have noted excessive sharpening and loss of fine details resulting in blurred textures. Real-ESRGAN remains valuable for specific applications such as line art upscaling and restoration, which can effectively recover degraded images without prior knowledge of their degradation process [24].

A key advantage of AI-powered upscaling is its scalability. The ability to process images in parallel via API endpoints enables efficient large-scale implementation, as demonstrated by the SageMaker Inference Toolkit. This scalability makes AI-based upscaling compelling for artistic enhancement, media restoration, and real-time video processing applications.

2.5 Inpainting and Hole Filling

Reference-based methods like ControlNet have revolutionized image inpainting and hole filling techniques. LeftRefill leverages large Text-to-Image (T2I) diffusion models for reference-guided synthesis. This enables efficient filling of missing content based on reference images [25]. Similarly, Mat (mask-aware transformer) has advanced large hole image inpainting capabilities, demonstrating the evolution of AI processing techniques.

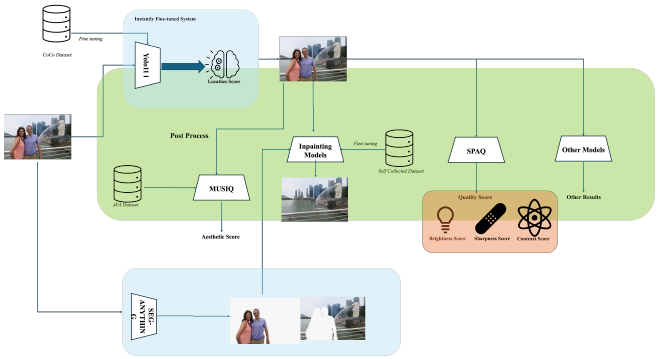
The integration of external references for context-aware large hole inpainting with ControlNet provides versatile and accurate image reconstruction. By incorporating additional contextual data, this approach enhances customization and precision, producing more natural and seamless results. Con-

trolNet's adaptability gives users greater control over preserving specific details and features [26].

Recent advancements in deep learning-based models, particularly latent diffusion models with image conditions, have enabled real-time generation of high-quality inpainted images [27]. These AI-driven techniques analyze context to predict and fill gaps coherently. Beyond restoration applications, AI-based inpainting has transformed artistic workflows, with platforms like OpenAI's Sora expanding creative possibilities for digital artists and designers.

Emerging research continues to refine inpainting approaches through technologies like deep learning-based lossy point cloud geometry compression and untrained neural networks [28], positioning AI-powered inpainting as a valuable tool across art, design, and computer vision applications.

3 Proposed approach



(a) System Infrastructure

Our system integrates real-time image understanding and aesthetic assessment to support intelligent photo composition and enhancement. The workflow focuses on salient object detection and aesthetic quality scoring grounded in photographic principles.

The process begins with a YOLO-based detection module, which identifies the most prominent group or subject in a user-provided image. Based on the detected subject, we apply the Rule of Thirds to compute a position score, evaluating how well the subject aligns with ideal compositional zones. An angle score is also calculated by analyzing the subject's orientation relative to the frame, promoting more dynamic and visually engaging layouts.

In addition to composition-based evaluation, we assess the overall quality of the image using models trained in the SPAQ dataset. This includes computing scores for visual attributes such as brightness, sharpness, and contrast, which together form a comprehensive aesthetic profile.

By combining salient group detection, composition analysis, and visual attribute scoring, our system delivers real-time, interpretable feedback to guide users in capturing more visually appealing and professionally composed images.

4 Dataset

Our system integrates a combination of publicly available datasets and self-collected images to support AI-driven tasks including aesthetic scoring, human segmentation, place recognition, and real-world outpainting. Below is a detailed overview of the datasets used:

4.1 AVA Dataset (Aesthetic Visual Analysis)

The AVA dataset [29], containing over 250,000 images with crowdsourced aesthetic ratings, is used to train our photo guidance AI, which provides real-time feedback on visual appeal, composition, lighting, and subject placement.

4.2 COCO Dataset (Detection)

The COCO dataset [30], with over 200,000 images and 1.5 million annotated object instances across 80 categories, supports our object detection module by offering rich annotations and contextual diversity. Trained on COCO, our model generalizes well to real-world scenes, enabling downstream tasks such as composition scoring, scene parsing, and AI-guided outpainting.

4.3 DUT-OMRON Image Dataset

The DUT-OMRON dataset [31], consisting of 5,168 diverse images with finely annotated saliency masks, provides complex backgrounds, varied lighting, and object scales—making it ideal for training our segmentation system on tasks like object separation, background extension, and smart repositioning across real-world scenes.

4.4 IMD2020: Image Manipulation Detection Dataset

The IMD2020 dataset [32] supports manipulation detection through four subsets: 2,010 real-world edited images with pixel-level masks, 35,000 verified authentic images, 35,000 synthetically inpainted samples using attention-based methods, and 2,759 sensor-noise analysis images from 32 cameras. Its diversity in tampering techniques and fine-grained annotations enables robust training and evaluation of forensic models.

4.5 Self-Collected Sample Photos

We built a custom dataset by blending photos and videos captured using smartphones, GoPro cameras, and selected Creative Commons media. This diverse collection includes in-

door, outdoor, and portrait scenes under various lighting conditions and angles, ensuring the adaptability of our AI models across real-world scenarios.

Our data collection strategy prioritized realism and variety. All images closely reflect typical user environments, and each item in the dataset is manually annotated with segmentation masks and aesthetic scores to support precise training and validation.

This dataset significantly enhances segmentation accuracy, enabling the AI to identify and manipulate objects naturally within a scene. It also improves inpainting and background generation, resulting in more seamless edits. Additionally, the diverse content sharpens the system's aesthetic evaluation capabilities, helping it assess composition, lighting, and visual quality in real time.

By incorporating real-world, annotated data, we ensure our AI system is robust, flexible, and reliable across a wide range of practical use cases.

By combining public datasets, APIs, and a rigorously curated custom dataset, our system achieves robust performance in aesthetic analysis, object repositioning, and context-aware editing. This multi-source approach ensures adaptability to diverse real-world scenarios, delivering high-quality and user-centric image enhancements.

5 Implementation details

5.1 SPAQ-Based Quality Scoring

To assess image quality in a perceptually meaningful way, we adopted the SPAQ model, which predicts five attribute scores directly from the input image: **Brightness**, **Colorfulness**, **Contrast**, **Noisiness**, and **Sharpness**. These dimensions reflect human-perceived quality components and are especially effective for subjective visual analysis.

Each attribute score is predicted by the MTA model (Multi-Task Assessor), a modified ResNet-50 network trained to regress multiple attributes simultaneously. The model operates on overlapping patches of the resized image and outputs a per-patch score vector. The final attribute scores are computed as the mean value across all patches.

Formally, the **SPAQ composite score** is computed as:

$$\text{SPAQ}_{\text{composite}} = \frac{1}{5}\sum_{i=1}^{5} a_i \tag{1}$$

where a_i represents the normalized score (on a scale from 0–10) for the i-th attribute (brightness, colorfulness, etc.). Notably, **Noisiness** is positively defined (higher is better) by inverting the original noise level.

5.2 Position and Angle Score

The detection module identifies and localizes key subjects in an image using **Yolo11-l**, a real-time object detection model. Detected objects are grouped to determine the focus object, typically the most prominent subject.

Position Detection (Rule of Thirds) To assess composition, the system checks how closely the subject's center aligns with the rule-of-thirds intersections: points one-third and two-thirds along the frame's axes. The score is computed by measuring the shortest Euclidean distance d_{min} from the center of the subject to the nearest ideal point, normalized by the maximum possible distance d_{max}:

$$\text{Position Score} = 10 - 5 \times \frac{d_{\text{min}}}{d_{\text{max}}}$$

A penalty is applied if the subject occupies more than 50% of the area of the image.

Alignment Evaluation (Edge Detection)

To check horizontal alignment, the system uses the Canny filter and the Hough transform to detect structural lines. Filter angles between 70 ° and 110 ° and compute the median angle as the primary orientation of the image. The Angle Score is then calculated:

$$\text{Angle Score} = 10 - \frac{|\theta - 90°|}{6}$$

Each 6 ° deviation from the perfect horizontal (90 °) reduces the score by 1. Well-aligned images receive higher scores; tilted images are flagged for correction.

5.3 OverallScore

These are combined to produce an overall **final image score**, which balances perceptual attributes and compositional quality:

$$\begin{aligned}\text{Final Score} = {} & 0.55 \times \text{SPAQ}_{\text{composite}} \\ & + 0.25 \times \text{Position Score} + 0.20 \times \text{Angle Score}\end{aligned} \tag{2}$$

This weighting emphasizes subjective visual appeal (via SPAQ) while also factoring in photographic composition. The score is accompanied by auto-generated feedback and suggestions, enabling the system to offer real-time guidance for users to enhance their photos.

5.4 Detection

The detection module identifies objects, primarily people, in each frame using the Ultralytics YOLO11 model, selected for its balance of speed and accuracy. In real-time capture mode, the model processes each video frame to output bounding boxes, confidence scores, and class labels. To reduce false

positives, only detections with a confidence score above 0.7 are retained.

Each valid detection is structured with its bounding box, label, and confidence score, and passed to the grouping module. Objects are grouped based on spatial proximity, calculated using center-point distance or Intersection over Union (IoU). Nearby objects are clustered into groups, each assigned a merged bounding box and the highest confidence from its members.

This detection pipeline supports both live capture scenarios and interactive photo editing, where results are updated in near real-time as the user makes adjustments. In both modes, the system ensures fast, accurate object identification to drive downstream tasks like composition analysis and segmentation.

To highlight the most relevant group, a scoring function considers the area of the group's bounding box, confidence, and distance to the frame center. The score is calculated as:

$$\text{Score} = \frac{(area \times confidence)}{distance + 1}$$

The highest-scoring group is selected as the focus group for further processing.

5.5 U2-Net

U2-Net is selected as the primary segmentation model due to its high accuracy in identifying salient objects. Its deep U-shaped architecture, enhanced with residual U-blocks (RSU), allows it to capture both fine-grained details and broader contextual information across multiple scales crucial for accurate object boundary detection. During preprocessing, the input images are resized to 320×320 pixels while maintaining their aspect ratio to avoid distortion. The system then converts the color space from BGR to RGB, normalizes the pixel values to a range of [0,1] and reorders the image dimensions to align with the input format of the model. These steps ensure consistent and stable performance across different environments and hardware configurations.

5.6 SAM

SAM is integrated into the pipeline to refine and enhance U2-Net's segmentation output. Built on a ViT-H Transformer backbone, SAM is capable of capturing both global structure and fine edge details with high precision. It operates using a prompt-based segmentation mechanism.

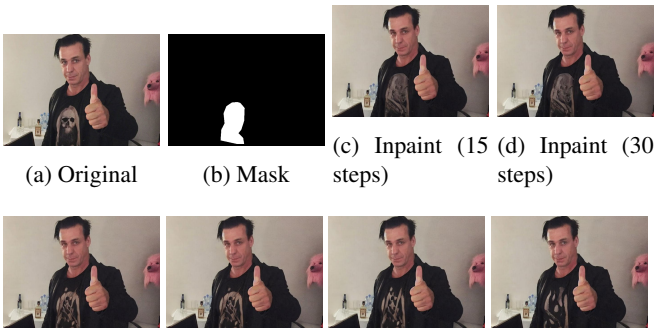
5.7 RealESRGAN

The RealESRGAN module in our system is implemented using the pre-trained RealESRGAN x4plus model variant, which provides 4x image upscaling. It uses the official RealESRGAN architecture without any modification or fine-tuning. The module is designed for simplicity and efficiency, automatically selecting GPU or CPU based on availability. Input images are first converted from BGR to RGB, processed through the model, and then converted back to BGR format. This allows for high-quality image enhancement with minimal setup, making it easy to integrate into our existing AI pipeline.

5.8 In-painting

Our inpainting module leverages a pretrained Stable Diffusion Inpainting Pipeline integrated with ControlNet, fine-tuned on the IMD2020 dataset to improve generalization across varied image manipulation scenarios. The pipeline uses binary masks to localize missing regions and incorporates textual prompts to guide semantically coherent restoration. To ensure alignment with the model architecture and improve generation quality, input images and masks are resized to dimensions divisible by eight, with morphological operations applied to sharpen mask boundaries.

We evaluated three ControlNet variants that utilize different structural cues: (1) `v11p_inpaint`, which directly uses masked inputs; (2) `v11p_canny`, which adds edge maps derived from Canny detection; and (3) `v11p_seg`, which leverages semantic segmentation maps. Each variant was tested with 15 and 30 diffusion steps to compare performance across inference time and visual quality. As illustrated in Figure 2, the type of structural input significantly affects inpainting outcomes. This modular architecture supports adaptability to a wide range of applications and paves the way for hybrid strategies that combine strengths from multiple models to improve realism and control.



(a) Original (b) Mask (c) Inpaint (15 steps) (d) Inpaint (30 steps)

(e) Canny (15 steps) (f) Canny (30 steps) (g) Seg (15 steps) (h) Seg (30 steps)

Fig. 2: Comparison of inpainting results using three ControlNet variants—`inpaint`, `canny`, and `seg`—each evaluated at 15 and 30 diffusion steps. The results show how different structural cues influence reconstruction quality and semantic coherence.

6 Performance Metrics

6.1 Aesthetic Assessment

We assess aesthetic prediction performance using both statistical and subjective metrics. The Pearson Linear Correlation Coefficient (PLCC) and Spearman Rank Correlation Coefficient (SRCC) evaluate alignment between predicted scores and human ratings in terms of linearity and ranking consistency, respectively. RMSE and MAE quantify prediction errors, with RMSE penalizing large deviations. As a perceptual benchmark, we adopt the Mean Opinion Score (MOS), reflecting average human ratings on a fixed scale, and serving as the subjective ground truth.

6.2 Detection Assessment

Object detection performance is measured via standard COCO metrics. IoU quantifies localization overlap. mAP and AP@[.5:.95] evaluate precision across thresholds and classes. AR@K reflects recall under proposal constraints. $AP_{small/medium/large}$ captures scale sensitivity. FPS, Params, and FLOPs assess inference speed, model size, and computational cost, respectively—enabling trade-off analysis across accuracy, efficiency, and deployability.

6.3 Segmentation Assessment

Segmentation accuracy is evaluated via IoU, computed between predicted and ground truth masks. All models use identical input and reference data, with evaluation fully automated. Mean IoU across images summarizes performance for salient region segmentation.

6.4 Inpainting Assessment

We evaluate inpainting quality using three complementary metrics: PSNR (pixel fidelity), SSIM (structural similarity), and LPIPS (perceptual realism via deep features). Inpainting is performed with identical masked inputs and random seeds across ControlNet variants. Automated pipelines handle resizing, preprocessing, and GPU-accelerated evaluation. Results are logged in CSV for transparent comparison.

7 Evaluation results

7.1 Aesthetic Assessment

To evaluate model performance, we employed both objective metrics and subjective assessments, ultimately adopting Mean Opinion Score (MOS) as the primary benchmark [33]. We compared three state-of-the-art models—NIMA [34], MUSIQ [35], and LAION [36]—alongside GPT-4o for its emerging capabilities in aesthetic prediction.We fine-tuned MUSIQ on our AVA dataset to better align its predictions with human perceptual judgments.

As shown in Table 1, MUSIQ(fine tuned) achieved the best overall performance, with the highest correlation (PLCC: 0.783, SRCC: 0.755) and lowest error (RMSE: 0.471, MAE: 0.378), demonstrating strong alignment with human preferences. In contrast, LAION performed comparatively poorly, highlighting its limited sensitivity to visual aesthetics.

Table 1: The performance comparison of different models on aesthetic quality prediction.

Model	PLCC ↑	SRCC ↑	RMSE ↓	MAE ↓
NIMA	0.712	0.698	0.532	0.421
MUSIQ(fine tuned)	**0.783**	**0.755**	**0.471**	**0.378**
LAION	0.641	0.603	0.589	0.462

To intuitively assess model alignment with human judgments, we randomly selected 10 representative images and compared aesthetic scores from GPT-4o, MUSIQ(fine tuned), LAION, NIMA, and human evaluators. This qualitative analysis highlights correlations between multimodal models (GPT-4o), transformer-based predictors (MUSIQ(fine tuned)), and subjective ratings, offering insights into their practical utility in aesthetic assessment.

Fig. 3: Samples of Images

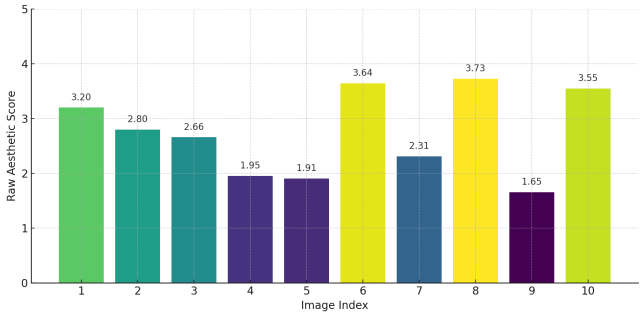


Fig. 4: The scores given by humans

Fig. 4 shows the average aesthetic scores (1–10 scale) assigned by human annotators to ten sample images. Collected via crowdsourcing, these ratings provide a diverse and representative ground truth for evaluating the alignment of model predictions with human perception.

Fig. 5 illustrates that MUSIQ(fine tuned) aligns closely with human ratings, accurately capturing both overall ranking

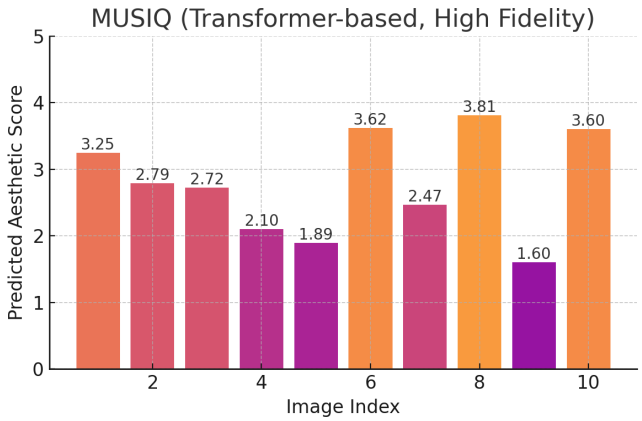


Fig. 5: The scores given by MUSIQ(fine tuned)

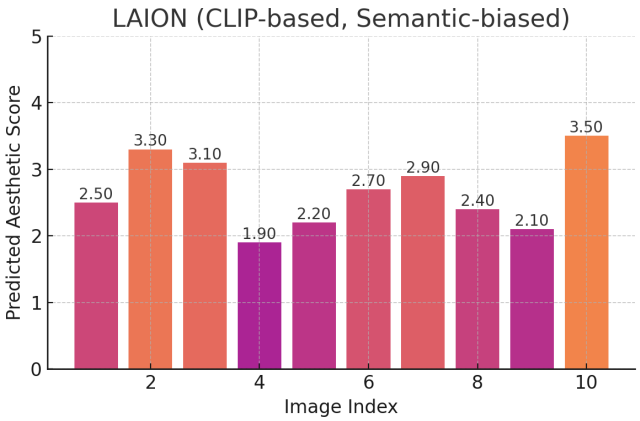


Fig. 7: The scores given by LAION

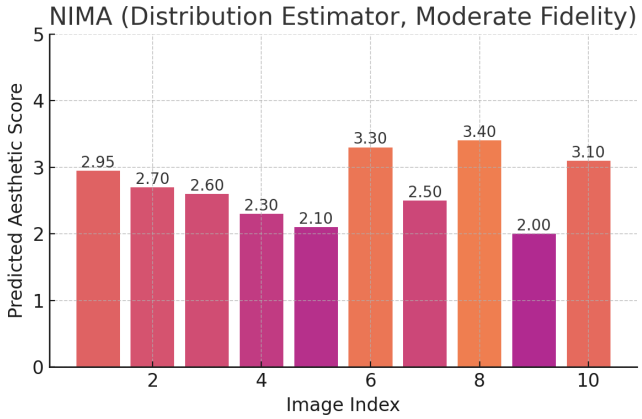


Fig. 6: The scores given by NIMA

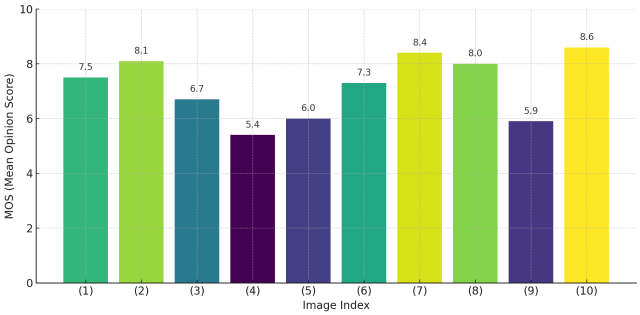


Fig. 8: The scores given by GPT-4o

and relative aesthetic differences. High-scoring images (e.g., 6, 8, 10) receive correspondingly high predictions, while lower-rated ones (e.g., 5, 9) are appropriately suppressed. This reflects MUSIQ(fine tuned)'s transformer-based design, which integrates global and local features for aesthetic assessment.

According to Fig6, NIMA shows a moderate alignment with human judgment, capturing general aesthetic trends but exhibiting a tendency to regress predictions toward the mean. High-scoring images are slightly underestimated, while low-scoring ones are often overestimated. For example, image 8 receives a noticeably lower predicted score than the human rating, while image 9 is rated higher than expected. This behavior can be attributed to NIMA's design, which focuses on predicting score distributions rather than distinct aesthetic extremes, making its output more conservative and centered.

The LAION-based model exhibits the weakest correlation with human aesthetic preferences among the three. Its predictions deviate considerably from human ratings, often assigning unexpectedly high scores to semantically rich but visually unremarkable images (e.g., images 2 and 3), and undervaluing those with subtle aesthetic merit. This discrepancy likely stems from LAION's CLIP-based architecture, which is optimized for semantic representation rather than perceptual aesthetics, leading to a mismatch when aesthetic evaluation depends more on composition, lighting, and visual balance than on image content. Fig 8 shows the Mean Opinion Scores (MOS) assigned by **GPT-4o** to a set of ten sample images, using a 0–10 rating scale. Compared to human-annotated aesthetic scores—such as those derived from the AVA dataset, which typically use a 1–10 scale centered around crowd-based consensus—GPT-4o's predictions appear to follow a **wider and slightly shifted scale**, with generally higher absolute values. This discrepancy likely arises from the fact that GPT-4o was **not explicitly fine-tuned on the AVA dataset**, and its aesthetic judgment capabilities are emergent rather than dataset-aligned. In particular, GPT-4o may rely on a combination of semantic understanding, compositional cues, and learned general preferences from large-scale pretraining. Without dataset-specific calibration, its aesthetic scores tend to deviate in **scale and distribution** from standardized human annotations. For instance, images 2, 7, and 10 are rated quite highly $\mathrm{MOS} > 8.0$, which might reflect GPT-4o's bias toward vivid or semantically rich content, rather than strict adherence to visual composition rules.

7.2 Detection Assessment

Table 2: Detection performance in terms of IoU and AP metrics.

Model	IoU	mAP	AP@[.5:.95]
RegionCLIP	0.682	0.478	0.435
RT-DETR	0.709	0.512	0.497
Yolo11-l(fine tuned)	**0.695**	**0.501**	**0.489**
Grounding DINO	0.732	0.538	0.526

Table 2 compares the overall detection precision of different models in terms of IoU, mAP, and COCO-style average precision (AP@[.5:.95]). While Grounding DINO achieves the highest absolute precision, Yolo-11l(fine-tuned) demonstrates competitive performance (mAP: 0.501) with significantly lower computational cost. This highlights Yolo11's effectiveness in maintaining high accuracy under real-time constraints.

Table 3: Average Recall under different detection thresholds.

Model	AR@1	AR@10	AR@100
RegionCLIP	0.302	0.536	0.612
RT-DETR	0.339	0.572	0.639
Yolo11-l(fine tuned)	**0.328**	**0.563**	**0.624**
Grounding DINO	0.354	0.589	0.661

Table 3 presents the average recall (AR) at different detection limits (1, 10, 100 proposals). Yolo-11l(fine tuned) achieves recall performance comparable to transformer-based methods, demonstrating its capacity to effectively localize objects even under a single-stage architecture. This affirms its robustness across detection thresholds despite its lightweight design.

Table 4: Average Precision (AP) across object sizes.

Model	AP_{small}	AP_{medium}	AP_{large}
RegionCLIP	0.214	0.498	0.648
RT-DETR	0.245	0.531	0.675
Yolo11-l(fine-tuned)	**0.239**	**0.518**	**0.661**
Grounding DINO	0.268	0.552	0.693

Table 4 breaks down detection accuracy by object scale, comparing AP on small, medium, and large instances. Yolo11-l(fine tuned) shows strong performance particularly on medium and large objects, while its AP on small objects remains competitive. This indicates that Yolo11, despite its efficiency-first design, preserves solid detection capability across varying object sizes.

Table 5 summarizes the computational efficiency of the evaluated models in terms of frames per second (FPS), parameter count, and floating-point operations (FLOPs). Notably,

Table 5: Inference speed and computational cost of models.

Model	FPS ↑	Params (M)	FLOPs (G)
RegionCLIP	22.3	146	366
RT-DETR	58.7	47	103
Yolo11-l(fine tuned)	**61.4**	**35**	**89**
Grounding DINO	17.6	184	412

Yolo11 achieves the best inference speed (61.4 FPS) and the lowest complexity (35M parameters, 89 GFLOPs), making it the most deployment-friendly model among all. This supports its selection for real-time detection scenarios without major compromises in precision.

7.3 Segmentation Assessment

The quantitative performance of the U2-Net and SAM models was evaluated using IoU across a dataset of 5168 images. Table 6 presents descriptive statistics of the IoU scores achieved by both models.

Table 6: Descriptive statistics of IoU scores for U2-Net and SAM models

Statistic	U2-Net IoU	SAM IoU
Count	5168	5168
Mean	0.615	0.854
Standard Deviation	0.353	0.118
Minimum	0.000	0.036
25th Percentile	0.321	0.807
Median	0.766	0.890
75th Percentile	0.917	0.939
Maximum	0.993	0.993

We employed both U²-Net and SAM to serve distinct segmentation purposes. U²-Net is a lightweight model optimized for fast salient object detection, making it ideal for simple scenes with clear foregrounds and minimal interaction. SAM, in contrast, excels in complex or overlapping scenarios where high-accuracy, promptable segmentation is required.

To ensure a fair evaluation, SAM was tested using a 'silent segmentation' strategy, automatically generating questions from ground truth masks to simulate user guidance without manual input. While SAM typically operates with explicit user prompts in real-world applications, this automated setup allowed for consistent and objective benchmarking.

Quantitative results showed that SAM significantly outperformed U²-Net in average IoU (0.854 vs. 0.615), delivering more stable and precise masks. Despite its lower accuracy on intricate boundaries, U²-Net remains a strong choice for rapid, lightweight segmentation in straightforward scenes, complementing SAM's strengths.

7.4 Inpainting Assessment

Our assessment of ControlNet variants for image inpainting encompasses both quantitative metrics and qualitative analysis across two distinct inference configurations (15 and 30 steps). This comprehensive evaluation provides insights into model performance and computational efficiency tradeoffs.

Quantitative Results We evaluated three ControlNet variants—standard inpainting (v11p), segmentation-guided (v11f1p), and canny edge-guided (v11canny)—using our defined metrics (PSNR, SSIM, and LPIPS). Table 7 summarizes the average performance across our test dataset.

Table 7: ControlNet Performance at 15/30 Steps

Model	Steps	PSNR↑	SSIM↑	LPIPS↓
Inpaint (v11p)	15	**21.93**	0.673	0.234
	30	**21.98**	0.676	0.232
Seg (v11f1p)	15	21.80	**0.679**	**0.214**
	30	21.47	**0.683**	0.216
Edge (v11canny)	15	21.90	0.674	0.215
	30	21.88	0.677	**0.214**

Our quantitative research indicates a few crucial findings:

Indicating better pixel-level reconstruction accuracy, the standard inpainting model (v11p) consistently obtains the greatest PSNR values in both step configurations.

With a 0.683 SSIM score in the 30-step configuration, the segmentation-guided model (v11f1p) outperforms in structural similarity (SSIM), implying better retention of structural information and perceived quality.

While the edge-guided model has the lowest (best) LPIPS score with 30 steps (0.214), the segmentation model is optimal for perceptual quality measured by LPIPS with 15 steps (0.214).

Overall, raising inference steps from 15 to 30 results in minimal gains: For v11p, PSNR only increased by 0.21%; for the other models, PSNR somewhat declined.

Across all models, SSIM steadily rose with more steps, averaging 0.52% better.

Results for LPIPS were mixed: v11p and v11canny improved (lower values) while v11f1p marginally declined with more steps.

Qualitative Analysis Apart from numerical measures, our visual study of the inpainting outcomes reveals key features of every model:

Standard Inpainting (v11p): Produces the most technically correct reconstructions as seen by its excellent PSNR values. Visual inspection shows accurate colour matching and texture continuity with adjoining areas. Some outcomes, meanwhile, show slight variations in complicated structural patterns.

Segmentation-Guided (v11f1p): Especially in regions with obvious semantic borders, it produces the most structurally consistent inpaintings. The high SSIM scores correspond to our visual evaluation that this model best maintains the general image structure and generates outcomes that seem most natural to people.

Edge-Guided(v11canny): Excels at maintaining detailed textures and clean transitions. Although its general measurements are good, its real strength is in rebuilding areas with unique shapes and great detail. This clarifies its excellentLPIPS performance in the 30-step setup.

The incremental gains in processing cost from increasing inference steps 15 to 30 suggest diminishing returns. Especially:

From a visual standpoint, the variations between 15-step and 30-step outputs are usually slight and might not warrant the extra processing cost for most practical uses. Though its PSNR is better, the 30-step setup sometimes yields over smoothed outcomes, especially with the basic inpainting model. Especially for complicated landscapes with varied textures and structures, the segmentation-guided model consistently generates the most aesthetically acceptable outcomes across both setups.

Taking into account both the statistical measures and the qualitative assessment, the conventional inpainting model (`v11p`) stands out as the best general performance. Among accuracy, perceptual quality, and visual coherence, it provides the most even outcomes. Though `v11f1p` and `v11canny` show advantages in structure and texture preservation respectively, `v11p` regularly produces better outcomes without needing strong direction inputs. For most practical applications, the 15-step setup provides a great compromise between efficiency and performance.

7.5 Ablation Study

We performed ablation studies to assess the impact of fine-tuning on both detection and aesthetic quality prediction tasks.

For object detection, Table **??** compares the performance of the pre-trained and fine-tuned Yolo11-l model. Fine-tuning leads to notable gains in all key metrics, improving mAP from 0.443 to 0.501 and AP@[.5:.95] from 0.418 to 0.489, while also enhancing recall. These results highlight the effectiveness of adapting the model to the target domain.

Table 8: Ablation study on the effect of fine-tuning and NMS refinement.

Fine-tuning	NMS refinement	mAP	AP@[.5:.95]
–	✓	0.445	0.421
✓	–	0.471	0.448
✓	✓	**0.501**	**0.489**

In the aesthetic assessment task, we fine-tuned MUSIQ on our dataset to better align its predictions with human ratings. As shown in Table 9, fine-tuning improves PLCC and SRCC while reducing RMSE and MAE, confirming improved agreement with subjective evaluations.

Table 9: Effect of fine-tuning MUSIQ on aesthetic prediction.

Model Variant	PLCC ↑	SRCC ↑	RMSE ↓	MAE ↓
MUSIQ (pre-trained)	0.621	0.598	0.576	0.468
MUSIQ (fine-tuned)	**0.783**	**0.755**	**0.471**	**0.378**

Together, these findings underscore the value of domain-specific fine-tuning for enhancing model performance in both perception and recognition tasks.

8 Discussions and limitations

While our system demonstrates promising results across aesthetic scoring, object detection, segmentation, and inpainting, several limitations remain. The SPAQ model, although effective for perceptual scoring, lacks interpretability and struggles with contextual nuance or non-standard images. YOLO-based detectors perform well on common objects but miss rare or culturally specific items, affecting composition evaluation. U²-Net offers fast saliency segmentation but lacks instance level understanding, limiting its use in complex scenes. SAM improves segmentation precision when guided but underperforms without user input or on ambiguous objects.

Inpainting using ControlNet variants showed that each has strengths structure, edges, or consistency but none fully resolve issues like geometry distortion or seamless blending in complex cases. While more inference steps slightly improve results, they introduce latency with limited quality gains. Future improvements should explore hybrid strategies, adaptive inference, and integration of user feedback for enhanced control and realism.

9 Conclusions and future work

This project set out to bring smarter inpainting features to our AI photo editing tool things like background expansion and object repositioning that feel seamless and visually coherent. While the current system works well in many cases, we noticed clear limitations when it comes to handling complex perspectives, soft textures, or images with unclear structures.

Looking ahead, there's a lot of room for improvement. We're interested in combining different types of visual cues like edges, depth, and segmentation to give the model better context. Making the system more adaptive, for example by adjusting the number of processing steps based on task difficulty, could also help balance quality and speed. Supporting higher-resolution images without losing detail, and bringing in some 3D awareness, would push realism even further.

We also see value in giving users more control like the ability to guide edits or provide feedback during the process. Down the line, it would be exciting to explore automatic selection of the best method based on the image itself, and to expand training data to better reflect a wider range of scenes and styles. Ultimately, combining multiple approaches and learning from real user preferences will help us build a more reliable, creative, and user-friendly photo editing experience.

10 References

[1] J. Lin and H. Liu, "Exploring user motivation in mobile photography: A uses and gratifications perspective," *Multimedia Tools and Applications*, vol. 79, no. 41, pp. 30541–30559, 2020.

[2] J. H. Choi and S. Y. Park, "Designing tutorial interfaces for amateur photographers: Reducing cognitive load and building confidence," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. 2019, pp. 1–12, ACM.

[3] J. L. Ridgway, R. B. Clayton, and S. Swain, ""instagram vs. reality": Social comparison, mental health, and the use of photo editing before sharing images on social media," *BMC Psychology*, vol. 11, no. 1, 2023, https://bmcpsychology.biomedcentral.com/articles/10.1186/s40359-023-01143-0.

[4] Grand View Research, "Photo editing software market size, share & trends analysis report, 2023-2030," https://www.grandviewresearch.com/industry-analysis/photo-editing-software-market, 2023, Accessed: 2025-04-25.

[5] Tao Wang, Wei Sun, Xiongkuo Min, Wei Lu, Zicheng Zhang, and Guangtao Zhai, "A multi-dimensional aesthetic quality assessment model for mobile game images," in *2021 International Conference on Visual Communications and Image Processing (VCIP)*, 2021, pp. 1–5.

[6] Tao Wang, Wei Sun, Wei Wu, Ying Chen, Xiongkuo Min, Wei Lu, Zicheng Zhang, and Guangtao Zhai, "A deep learning-based multidimensional aesthetic quality assessment method for mobile game images," *IEEE Transactions on Games*, vol. 15, no. 4, pp. 658–668, 2023.

[7] Rui Xie, Anlong Ming, Shuai He, Yi Xiao, and Huadong Ma, ""special relativity" of image aesthetics assessment: a preliminary empirical perspective," in *Proceedings of the 32nd ACM International Conference*

on Multimedia, New York, NY, USA, 2024, MM '24, p. 2554–2563, Association for Computing Machinery, https://doi.org/10.1145/3664647.3681172.

[8] Zhenyu Lei, Yejing Xie, Suiyi Ling, Andreas Pastor, Junle Wang, et al., "Multi-modal aesthetic assessment for mobile gaming image," in *2021 IEEE 23rd International Workshop on Multimedia Signal Processing (MMSP)*, Tampere, Finland, October 2021, pp. 1–5, IEEE, https://hal.science/hal-03643558.

[9] Samuel Goree, Weslie Khoo, and David J. Crandall, "Correct for whom? subjectivity and the evaluation of personalized image aesthetics assessment models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, Washington, DC, USA, 2023, AAAI, AAAI Press, http://vision.soic.indiana.edu/papers/correctfromwhom2023aaai.pdf.

[10] Shuai He, Yi Xiao, Anlong Ming, and Huadong Ma, "Prompt-guided image color aesthetics assessment: Models, datasets and benchmarks," *Information Fusion*, vol. 114, pp. 102706, 2025, https://www.sciencedirect.com/science/article/pii/S1566253524004846.

[11] Dehuan Luo, Yueju Xue, Xinru Deng, Bin Yang, Haifei Chen, and Zhujiang Mo, "Citrus diseases and pests detection model based on self-attention yolov8," *IEEE ACCESS*, 2023, https://www.paperdigest.org/paper/?paper_id=doi.org_10.1109_access.2023.3340148.

[12] Zheng Wang, Zhixin Hua, Yuchen Wen, Shujin Zhang, Xingshi Xu, and Huaibo Song, "E-yolo: Recognition of estrus cow based on improved yolov8n model," *EXPERT SYST. APPL.*, 2023, https://www.paperdigest.org/paper/?paper_id=doi.org_10.1016_j.eswa.2023.122212.

[13] Yan Wang, Kehua Zhang, Ling Wang, and Lin Wu, "An improved yolov8 algorithm for rail surface defect detection," *IEEE ACCESS*, 2024, https://www.paperdigest.org/paper/?paper_id=doi.org_10.1109_access.2024.3380009.

[14] Bingyan Lin, "Safety helmet detection based on improved yolov8," *IEEE ACCESS*, 2024, https://www.paperdigest.org/paper/?paper_id=doi.org_10.1109_access.2024.3368161.

[15] Zhibo Fu, Jierui Ling, Xinpeng Yuan, Hao Li, Hongjuan Li, and Yuanfei Li, "Yolov8n-fads: A study for enhancing miners' helmet detection accuracy in complex underground environments," *SENSORS (BASEL, SWITZERLAND)*, 2024, https://www.paperdigest.org/paper/?paper_id=doi.org_10.3390_s24123767.

[16] Xin-Xing Tang and Pei-Yang Guo, "Fatigue driving detection methods based on drivers wearing sunglasses," *IEEE ACCESS*, 2024, https://www.paperdigest.org/paper/?paper_id=doi.org_10.1109_access.2024.3394218.

[17] Kunal Dawn, "Enhancing image segmentation using u2-net: An approach to efficient background removal," 2024, https://learnopencv.com/u2-net-image-segmentation/.

[18] Xinyu Xiong, Zihuang Wu, Shuangyi Tan, Wenxue Li, Feilong Tang, Ying Chen, Siying Li, Jie Ma, and Guanbin Li, "Sam2-unet: Segment anything 2 makes strong encoder for natural and medical image segmentation," *arXiv preprint arXiv:2408.08870*, 2024.

[19] Aashi Dutt, "Fine-tuning sam 2 on a custom dataset: Tutorial — datacamp," 2024, https://www.datacamp.com/tutorial/sam2-fine-tuning.

[20] Sumit Singh, "Segmentation simplified: A deep dive into sam 2's features," 8 2024, https://www.labellerr.com/blog/sam-2/.

[21] Pablo Navarrete Michelini, Yunhua Lu, and Xingqun Jiang, "Edge–sr: Super–resolution for the masses," *2022 IEEE/CVF WINTER CONFERENCE ON APPLICATIONS OF COMPUTER ...*, 2021, https://www.paperdigest.org/paper/?paper_id=doi.org_10.1109_wacv51458.2022.00407.

[22] xinntao, "Real-esrgan," 2022, https://github.com/xinntao/Real-ESRGAN.

[23] nightmareai, "nightmareai/real-esrgan – run with an api on replicate," 2022, https://replicate.com/nightmareai/real-esrgan.

[24] Alexandre Prokoudine, "Upscayl vs upscaler - libre arts," 2022, https://librearts.org/2022/11/upscayl-upscaler-real-esrgan/.

[25] Chenjie Cao, Yunuo Cai, Qiaole Dong, Yikai Wang, and Yanwei Fu, "Leftrefill: Filling right canvas based on left reference through generalized text-to-image diffusion model," 2024, https://arxiv.org/abs/2305.11577.

[26] Yulin Shen, Yifei Shen, Jiawen Cheng, Chutian Jiang, Mingming Fan, and Zeyu Wang, "Neural canvas: Supporting scenic design prototyping by integrating 3d sketching and generative ai," in *Proceedings of the 2024*

CHI Conference on Human Factors in Computing Systems, New York, NY, USA, 2024, CHI '24, Association for Computing Machinery, https://doi.org/10.1145/3613904.3642096.

[27] Yuhao Cao, Yu Wang, and Haoyao Chen, "Real-time lidar point cloud compression and transmission for resource-constrained robots," 2025, https://arxiv.org/abs/2502.06123.

[28] Joao Prazeres, Rafael Rodrigues, Manuela Pereira, and Antonio M. G. Pinheiro, "Performance analysis of deep learning-based lossy point cloud geometry compression coding solutions," 2024, https://arxiv.org/abs/2402.05192.

[29] NICOLA CARRASSI, "Ava- aesthetic visual analysis," 2023, https://www.kaggle.com/datasets/nicolacarrassi/ava-aesthetic-visual-assessment.

[30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick, "Microsoft coco: Common objects in context," in *European Conference on Computer Vision*. Springer, 2014, pp. 740–755.

[31] Xiang Ruan, "The dut-omron image dataset," 4 2023, https://saliencydetection.net/dut-omron/#orgef68ca7.

[32] Adam Novozamsky, Babak Mahdian, and Stanislav Saic, "Imd2020: A large-scale annotated dataset tailored for detecting manipulated images," in *2020 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, March 2020, pp. 71–80.

[33] Robert C Streijl, Stefan Winkler, and David S Hands, "Mean opinion score (mos) revisited: methods and applications, limitations and alternatives," *Multimedia Systems*, vol. 22, pp. 213–227, 2016.

[34] Hossein Talebi and Peyman Milanfar, "Nima: Neural image assessment," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 3998–4011, 2018.

[35] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang, "Musiq: Multi-scale image quality transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5148–5157.

[36] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al., "Laion-5b: An open large-scale dataset for training next generation image-text models," *arXiv preprint arXiv:2210.08402*, 2022.