# What Makes for a Good Stereoscopic Image?

Netanel Y. Tamir* [1,2] Shir Amir* [1] Ranel Itzhaky [1] Noam Atia [1]
Shobhita Sundaram [3] Stephanie Fu Ron Sokolovsky [1] Phillip Isola [3]
Tali Dekel [2] Richard Zhang Miriam Farber [1]

[1] Apple [2] Weizmann Institute of Science [3] MIT



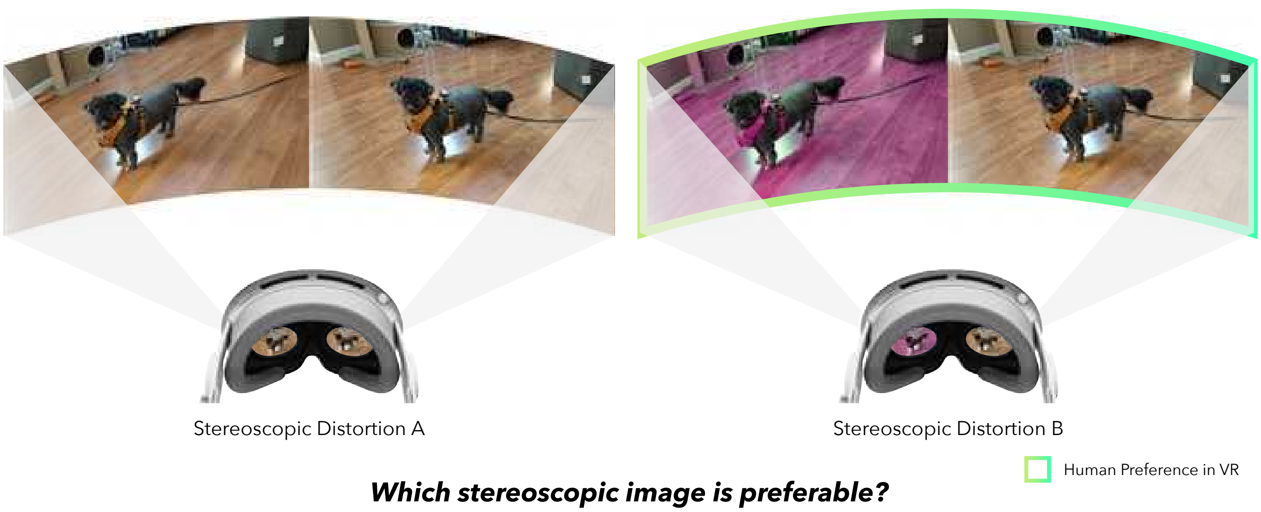


Figure 1. **Capturing human preference in VR.** We present a stereoscopic dataset created by showing participants two distorted versions of a stereoscopic image through a VR headset and asking which one they prefer. Our dataset encompasses a variety of distortion types. In the example above, the left image of each stereo pair is distorted: (left stereo image) SDEdit [53] is applied resulting in texture changes, such as on the floor and dog; (right stereo image) a shift in hue is used as the distortion. Interestingly, we find that preferences in VR often differ from preferences on screen. We leverage this dataset to train a stereo quality prediction model.

## Abstract

*With rapid advancements in virtual reality (VR) headsets, effectively measuring stereoscopic quality of experience (SQoE) has become essential for delivering immersive and comfortable 3D experiences. However, most existing stereo metrics focus on isolated aspects of the viewing experience such as visual discomfort or image quality, and have traditionally faced data limitations. To address these gaps, we present SCOPE (Stereoscopic COntent Preference Evaluation), a new dataset comprised of real and synthetic stereoscopic images featuring a wide range of common perceptual distortions and artifacts. The dataset is labeled with preference annotations collected on a VR headset, with our findings indicating a notable degree of consistency in user preferences across different headsets. Additionally, we present iSQoE, a new model for stereo quality of experience assessment trained on our dataset. We show that iSQoE aligns better with human preferences than existing methods when comparing mono-to-stereo conversion methods.*

## 1. Introduction

*"What is to come of the stereoscope and the photograph we are almost afraid to guess, lest we should seem extravagant." – Oliver Wendell Holmes, 1859*

Stereoscopy, commonly referred to as stereo imaging, is a technique in which a slightly varied image is displayed separately to each eye, creating the illusion of a 3D scene. For over a century, this method has been used with a diverse array of viewing tools, starting from the revolutionary Victorian-era stereoscopes, to the red and cyan anaglyph glasses popularized in the 20$^{th}$ century [21], and now rapidly advancing virtual reality (VR) headsets. With recent improvements in the quality of VR headsets, stereoscopic image quality is more important than ever to evaluate. Various factors may effect the final quality, including 2D aesthetics, depth sensation (stereopsis) and viewing comfort; making this task particularly challenging. While stereo imaging has a long history, the current tools available for assessing these images are still quite limited.

Recent advancements in neural rendering have paved the way for advancing immersive 3D content creation. Neural Radiance Fields (NeRFs) [5–7, 22, 55] and 3D Gaussian Splats (3DGS) [23, 36, 99], in particular, have emerged as powerful scene representation techniques, allowing for efficient and realistic scene rendering. In addition, many text-to-3D models leverage recent progress in text-to-image generation [33, 47, 63, 89, 98], often using 2D diffusion priors to generate realistic 3D content. Moreover, the recent boom in text-to-video generation [4, 9, 87, 96] has paved the way for advances in stereoscopic video synthesis. All of these inspired recent stereoscopic images and video generation algorithms [17, 50, 76, 103].

This progress in 3D content creation, which serves as a source for stereo content, is occurring alongside improvements in stereo image capturing technology. Dual or multi-camera systems are standard in today's smartphones, providing the necessary hardware for stereo photography [71, 73]. Some models leverage these multiple cameras to enable capturing true stereo images [2, 70]. In parallel, VR headsets have become more accessible than ever, with millions of users and a multi-billion dollar market [32, 58]. These complementary trends lead us to a moment ripe for reassessing our methods for evaluating stereo imagery.

Despite the surge in stereo content, evaluation methods for it remain limited. While single-image quality assessment (IQA) tools exist [56, 86, 93, 95], they do not capture the complex relations between the two monocular images of a stereo pair, such as viewing comfort or realistic depth, when viewed together in stereoscopic 3D. On the other hand, existing methods for evaluating stereo content suffer from a lack of annotated training data [59, 78, 100, 105], and focus predominantly on a few low-level artifacts (e.g. noising and excessive horizontal disparity). Furthermore, these methods have not been trained or tested on more complex artifacts (e.g. results of inpainting or depth algorithms).

To address these bottlenecks, we present a new benchmark for stereo imagery. Our dataset, SCOPE – Stereoscopic COntent Preference Evaluation – is comprised of pairs of distorted stereo images. Some of these images are created by distorting physically-captured stereo images, while others are created via generative methods applied to monocular images. The distortions are selected from a wide range of common image augmentations (e.g. photometric and spatial distortions) and generative methods including 3D Gaussian splatting [36] and MotionCtrl [90], to encompass the variety of artifacts that may appear when generating images. Human annotators then participate in a two-alternative forced choice (2AFC) test, where they vote on which version of a given stereo image they prefer when viewed through a VR headset. Using this dataset, we train a stereo quality of experience (SQoE) assessment model, meant to capture a holistic sense of the overall visual experience. We demonstrate its practicality in assessing different mono-to-stereo generation methods and its ability to extrapolate to both distortion types and strengths that are not present in its training data. To the best of our knowledge, this is the largest data-driven effort to develop an SQoE evaluator, addressing a broad range of stereo artifacts.

To summarize, our key contributions are as follows:

- SCOPE - A two-alternative forced choice (2AFC) stereoscopic dataset, containing 2400 samples annotated by 103 participants;
- iSQoE - An SQoE model trained using our dataset;
- Demonstrating our model's effectiveness in assessing off-the-shelf stereo synthesis methods.

## 2. Related Work

### 2.1. Stereo Image Assessment

Stereo Quality of Experience (SQoE) encompasses the user's overall viewing experience of the stereoscopic 3D content. Compared to traditional monoscopic image assessment, stereoscopic evaluation presents unique challenges, as factors like visual discomfort and depth perception may influence a user's overall impression or satisfaction [48, 67].

Some works evaluate SQoE through *stereo image quality assessment* (SIQA), an extension of single image quality assessment, which aims to evaluate the degree of distortion in images. No-reference quality assessment methods are widely seen as the most practical and adaptable, with many no-reference stereo image quality assessment (NR-SIQA) methods having been proposed over the years. These include both hand-crafted feature-based methods [1, 15, 45, 49, 74] and deep learning-based techniques [24, 34, 75, 78, 100, 102, 104].

Other studies have concentrated on evaluating *discomfort* experienced when viewing stereoscopic images, which are not addressed by traditional image quality assessment methods. Several factors contributing to viewer discomfort have been highlighted in the literature [3, 40, 43, 77, 83, 97]. Visual discomfort predictors developed for this purpose include approaches based on hand-crafted features [12, 37, 62, 66] as well as deep learning-based models [38, 59, 105]. Like the aforementioned SIQA models, these models rely on small-scale datasets, limiting their overall effectiveness.

Our work aims to capture the overall impression of a stereo image, not only image quality or viewer experience. We provide a single data-driven model that implicitly considers both image quality and user comfort based on user annotations.

Additionally, there are IQA models developed specifically for VR (VR-IQA). These mainly focus on evaluating monoscopic and stereoscopic *omnidirectional* images, commonly referred to as 360° images [14, 46, 65, 80]. Evaluating such images presents unique challenges, such as projection distortions, that differ from the focus of our work.

### 2.2. Stereoscopic Datasets

Numerous stereoscopic image datasets are publicly accessible, offering a variety of sizes, resolutions, and camera baselines [10, 29, 52, 54, 72, 88]. The viewer's depth perception and overall comfort while viewing stereo images are affected by how closely the camera baseline matches their interpupillary distance, which averages approximately 63 mm [20]. Several of these datasets include human-annotated preferences regarding quality or comfort, and as such were used to develop automated assessment methods [57, 62, 84, 85]. However, the limited size and diversity of these annotated datasets has become the main bottleneck in advancing automated assessment techniques.

### 2.3. Psychophysics in Virtual Reality

Numerous psychophysical studies have investigated human perception and cognition within VR environments to understand how users experience and interact with these systems. Krajancich *et al*. [41] demonstrated that adjusting rendering based on user gaze enhances depth perception and realism, while Guan *et al*. [26] highlighted users' sensitivity to small errors in camera positioning. Additionally, work by Chen *et al*. [13] examined the effects of power-saving techniques on perceptual quality in VR, and Matsuda *et al*. [51] found that current VR headsets often fall short of user expectations for display brightness. Thomas [82] found that users can accurately perceive the size of virtual objects within arm's reach in VR, with height and width judgments closely matching actual dimensions. Collectively, this body of research underscores the critical role of understanding human perception to guide the development of effective and user-centric VR technologies. In our work we implicitly capture characteristics of human stereoscopic perception within our dataset, and train a model to predict SQoE accordingly.

## 3. Stereoscopic Dataset Collection & Learning

### 3.1. SCOPE Dataset

We collect human preferences for stereoscopic 3D experiences by creating different variations of stereo images. Each stereo image undergoes two distinct distortions among those listed in Table 1. The stereo images are then compared by five annotators specifying which version they prefer.

Each stereo image in SCOPE is created in one of the following manners:

(i) **Single image to stereoscopic image.** We use one of two distinct approaches: (a) MotionCtrl [90], an image-to-video model which enables camera control, while keeping the scene generally static. We simulate interocular distance by imposing horizontal camera motion to the input images. (b) we create a *2D lifting* pipeline, using one of several off-the-shelf monocular depth estimators [35, 69, 94] to estimate disparity and forward warp the image accordingly. Then, we use LaMa inpainting [81] to fill in the dis-occluded regions.

(ii) **Multiview images to stereoscopic image.** We use 3D Gaussian splatting [36] fitted to multi-view scenes from several datasets [6, 27, 39]. We rendered each stereo image as two 2D images with a horizontal offset between them. For each scene we used two different 3DGS representations optimized at different levels, resulting in differing amounts of artifacts.

(iii) **Distorting *existing* stereoscopic images.** We leverage physically acquired stereoscopic images from Holopix50k HD [29] and apply a diverse set of image editing techniques as distortions. These include noise injection, photometric and spatial transformations, and SDEdit [53] to modify the high-frequency components of the images.

| Sub-type | Distortion type |
|---|---|
| Novel-view synthesis | 2D Lifting, MotionCtrl, 3D Gaussian splatting |
| Noise | Uniform white noise, Gaussian white noise, Checkerboard artifact |
| Blur | Average blur, Gaussian blur |
| Compression | JPEG compression |
| Photometric | Hue shift, Saturation shift, Brightness shift, Contrast shift |
| Spatial | Magnification, Rotation, Keystone effect, Warping, Chromatic aberration |
| Diffusion-based editing | SDEdit |

Table 1. **Distortion Pool.** The diverse set of distortions applied to stereoscopic images generates a wide variety of artifacts.

The dataset is hence comprised of two subsets to avoid accumulating distortions: The first uses distortion types (i) and (iii) on images from Holopix50k HD [29], which contains various in-the-wild physically captured stereo images. The second applies distortion types (i) and (ii), and is based on 13 multi-view scenes from several datasets: Tanks and Temples [39], Deep Blending [27] and Mip-Nerf [6].

We create 2400 data samples of resolution $1280 \times 720$, 2000 examples from the first subset and 400 from the second. Figures 1 and 2 contain dataset examples. Each unique image, drawn from existing datasets, serves as the basis for two distorted versions, and is used only once. For every

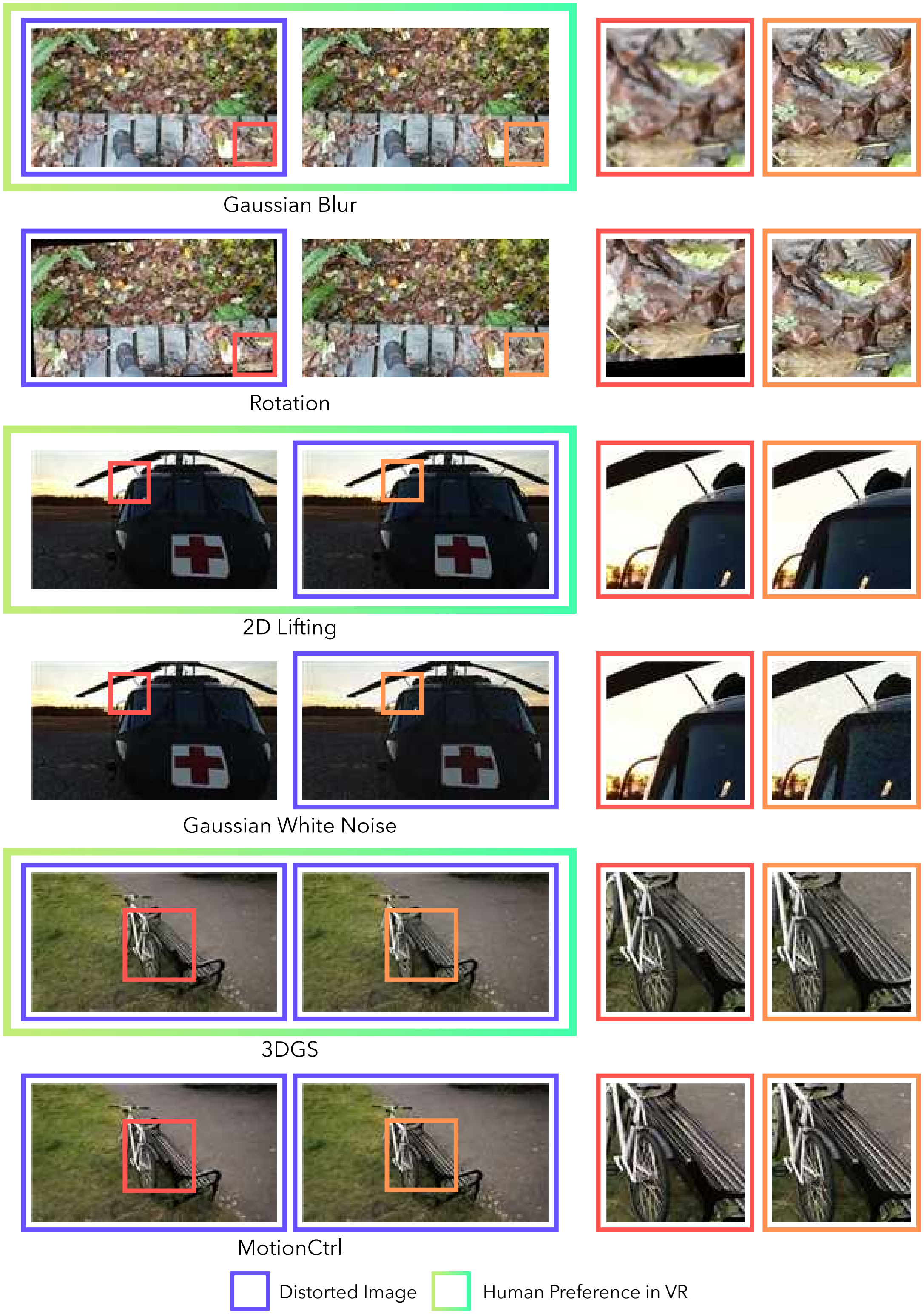


Figure 2. **Dataset examples.** Each stereo image was subjected to two different distortions, applied consistently to either the left, right, or both images. Participants in a VR-based user study were then asked to choose their preferred version. On the right of each sample, we zoom to highlight the differences between the images. Some distortions are more easily visible in 2D (*e.g.* Gaussian White Noise, Rotation) while others are more visible on VR devices (*e.g.* disparity differences cause increased depth sensation in the 2D lifting example).

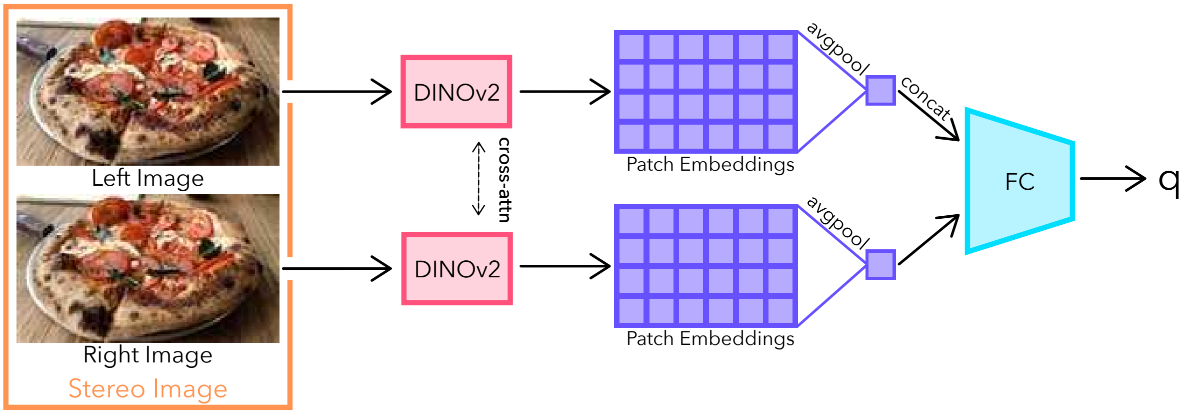

Figure 3. **Model architecture.** The left and right images of a stereo pair are processed by a modified DINOv2 [60] network with cross-attention between images. The resulting spatial tokens are pooled, concatenated, and passed through a small fully-connected network, outputting a single value indicating quality (lower is better). We train the model with a hinge loss and LoRA [28] for the DINOv2 network.

data sample we ensure the distortions are applied consistently on either the left, right, or both images uniformly, reducing noise caused by ocular dominance [64] (see Figure 2). The variety of distortion types results in a diverse array of artifacts which can lead to viewer discomfort, reduce image quality, and/or alter depth perception. More information is available in Section 6 of the supplementary material (SM). Table 5 of the SM compares our dataset to previous stereoscopic preference datasets, indicating it is has significantly more samples than previous datasets. Furthermore, it is the first dataset with 2AFC annotations, that can be directly used to train a SQoE model.

### 3.2. VR Annotations in SCOPE

We conduct a large scale 2AFC study, similar to LPIPS [101] and DreamSim [25] where participants are asked to annotate our dataset from Section 3.1 on an Apple Vision Pro headset. Specifically, for each real stereoscopic image pair $x = (x_l, x_r)$, users compare two modified versions. These versions are defined as $(x_l^m, x_r^m), (x_l^n, x_r^n)$, where $l$ and $r$ are left and right views in a stereo image respectively; and $m$ and $n$ are distinct distortions from Table 1. These are applied to at least one view in each stereo pair, and are applied to the same views across both stereo pairs.

We collect judgments $y \in \{m, n\}$ by asking participants which version is *preferable*, as depicted in Figure 1. We pose the question to the participants in this manner to collect majority-based binary labels, which can then be used to train a holistic SQoE model. We also ask participants to avoid closing one eye at a time and instead view the left and right images simultaneously. The user study was conducted in batches of 25 examples, with breaks between batches to mitigate visual fatigue, a condition characterized by Lambooij *et al.* [43] as eye strain and reduced visual performance after prolonged viewing. Our dataset, $D = \big\{\big((x_l^m, x_r^m), (x_l^n, x_r^n)\big), y\big\}$, ultimately contains 2400 examples, each labeled with 5 annotations and collected from 103 participants. We find that a third of the examples are unanimously agreed upon, with a $5/0$ split, another third have a $4/1$ split, and the remaining third have a $3/2$ split. We randomly partition our data into train (80%), validation (10%), and test (10%) sets. We name our dataset SCOPE – Stereoscopic COntent Preference Evaluation – and make it publicly available.

### 3.3. Training an SQoE Model

After collecting our dataset, we use it to train an SQoE model which given a stereoscopic image as input, can then produce a single score that takes into account quality, comfort, and depth sensation. We name our model iSQoE - immersive Stereoscopic Quality of Experience assessor. The architecture is described in Figure 3 and is inspired by LPIPS [101] and DreamSim [25] which also train perceptual models on 2AFC annotations.

Since stereoscopic images are composed of two 2D images, we leverage existing pretrained image backbones to process each image in a stereoscopic pair. The resulted features are pooled, concatenated, and passed through a lightweight fully-connected network, which generates a quality score normalized by a sigmoid function. We train the model in a Siamese fashion [8] with a hinge loss between the two stereo pairs using the participants' preferences as ground truth. To enable information flow between the 2D backbones of each image in the stereo pair we pass information between the attention modules of the pretrained backbones.

We conduct several ablations presented in Table 2. We report total mean accuracy, as well as on subsets with unanimous (5-0), majority (4-1) and ambiguous (3-2) annotations. We consider several backbone features, such as those extracted from foundation models like DINOv2 [60] and CLIP [68], and also backbones with task specific knowledge

| | Model | | | | Mean Accuracy | | | |
|---|---|---|---|---|---|---|---|---|
| **Ablation type** | **Resolution** | **Backbone** | **Attention Fusion [Layers]** | **LoRA** | **3–2 Split** | **4–1 Split** | **5–0 Split** | **Total** |
| **Image Resolution** | 224 × 224 | DINOv2 S/14 [60] | Concat [2, 5, 8, 11] | ✓ | 60.4 | 68.4 | 83.5 | 70.8 |
| **Backbone** | 1280 × 720 | CLIP L/14 [68] | Concat [2, 5, 8, 11] | ✓ | 59.5 | 66.7 | 78.2 | 68.2 |
| | 1280 × 720 | OpenCLIP L/14 [30] | Concat [2, 5, 8, 11] | ✓ | 61.4 | 68.1 | 78.4 | 69.3 |
| | 1280 × 720 | Croco [92] | Unmodified | Dec. | 59.0 | 67.5 | 80.1 | 69.0 |
| | 1280 × 720 | StereoQA-Net [104] | – | – | 62.5 | 64.6 | 77.3 | 68.1 |
| **Attention Module** | 1280 × 720 | DINOv2 S/14 | Swap [2, 5, 8, 11] | ✓ | **62.6** | **72.2** | 83.6 | 73.0 |
| | 1280 × 720 | DINOv2 S/14 | Concat [0 − 11] | ✓ | 61.7 | 69.3 | 84.4 | 71.8 |
| | 1280 × 720 | DINOv2 S/14 | Concat [11] | ✓ | 60.2 | 72.1 | 83.9 | 72.2 |
| | 1280 × 720 | DINOv2 S/14 | Unmodified | ✓ | 60.6 | 71.7 | 82.6 | 71.9 |
| **Optimization** | 1280 × 720 | DINOv2 S/14 | Concat [2, 5, 8, 11] | – | 60.4 | 68.7 | 82.2 | 70.5 |
| **iSQoE (Ours)** | 1280 × 720 | DINOv2 S/14 | Concat [2, 5, 8, 11] | ✓ | 62.1 | 72.0 | **84.8** | **73.1** |

Table 2. **Model ablations.** Our chosen variant performs the best on the entire test set as well as the 5-0 uniform split, with the other attention fusion variants being close.

like Croco [92] which was trained for Novel View Synthesis and StereoQA-Net [104] trained for SIQA. We found DINOv2 to yield the best results, with Croco not falling much behind, indicating 2D image understanding is important for SQoE. Similar to DreamSim we found that finetuning DINOv2 with LoRA yields an additional improvement.

Furthermore, we examine early information fusion between the 2D image backbones by manipulating the attention layers in the feature extractors, to enable information sharing within the left and right parts of the stereo image. We considered two fusion strategies: (i) *swapping* the keys and values between left and right images and (ii) *concatenating* the keys and values from left and right images. Queries are kept in place in both methods. Inspired by Croco [91] we examined fusion on all layers [0-11], last layer [11], and alternating fusing and non-fusing layers [2, 5, 8, 11]. We found *concatenating* fusion on alternating layers to perform best, and all fusion forms to perform better than no fusion at all. Intuitively, this could imply sharing information across left and right images is beneficial for stereo understanding.

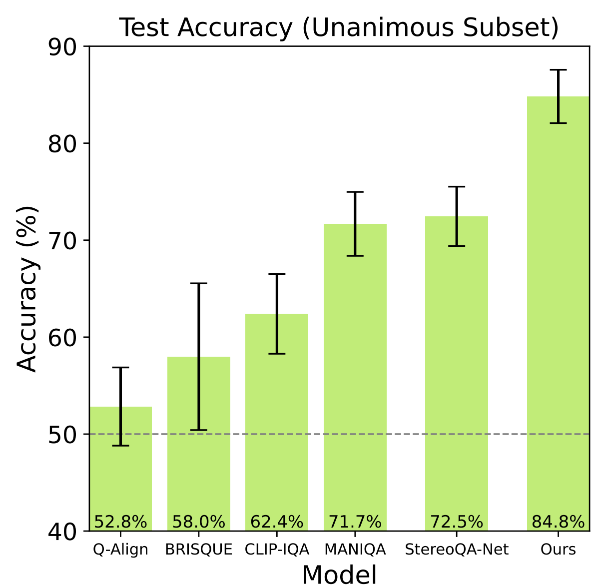


Figure 4. **Test accuracy on SCOPE.** We report the mean and standard deviation of the unanimous cases in the test set over several splits. Our model outperforms the other SIQA and IQA models.

## 4. Results & Analysis

### 4.1. Performance on the SCOPE Dataset

We benchmark several models on SCOPE in Figure 4. We consider the known IQA metrics BRISQUE [56], MANIQA [95], CLIP-IQA [86] and Q-Align [93]. For each stereo image we calculate the IQA score for the left and right views individually and use their mean as the score for the entire stereo image. We also consider StereoQA-Net [104], an existing NR-SIQA model, based on a CNN [42, 44] architecture and trained on the LIVE 3D Phase I dataset [57]. Further comparisons are limited due to availability of code or pretrained models.

We partition the dataset into train, validation, and test subsets several times. When evaluating model performance, we report the accuracy specifically on the subset of the test data that has unanimous human annotations. These unanimously-annotated examples are likely to contain the least amount of annotation noise and represent the most cognitively impenetrable cases [11, 79]. Figure 4 shows our model successfully outperforms the other methods by a margin. Moreover, we find that MANIQA performs best within the IQA metrics, corroborating results from Zhu *et al*. [106]. Additional information is available in Sections 8 and 9 of the SM.

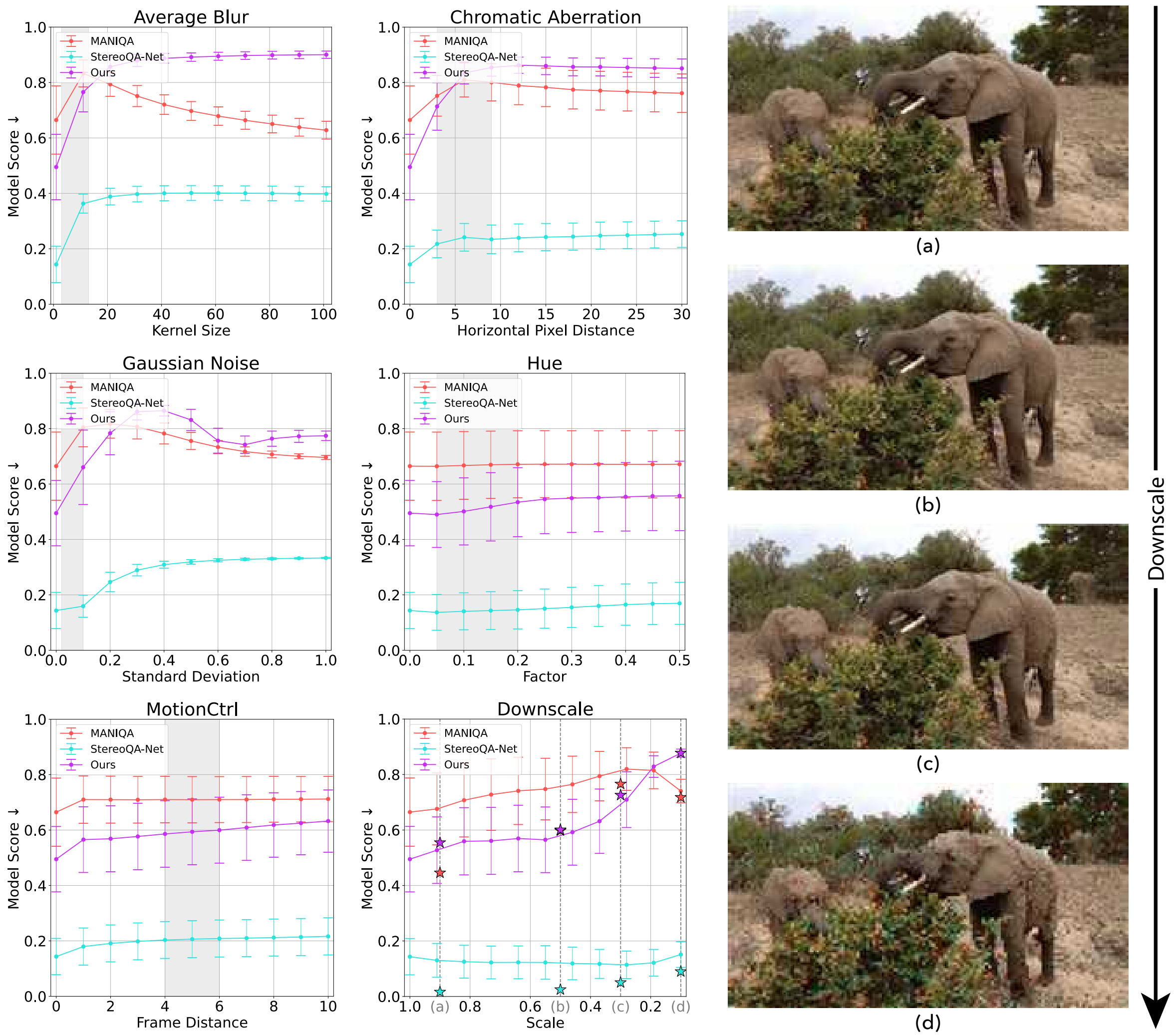


Figure 5. **Progressive degradation evaluation.** We report model scores on 200 stereo images for six different distortions. The gray regions indicate the distortion intensity used in SCOPE. Stereo images (a) – (d), presented as anaglyphs, exhibit progressive downscaling and are represented by stars.

### 4.2. How do models respond to unseen distortions with varying strengths?

In Figure 5 we apply a series of degradations to stereo images with increasing severity and evaluate the distorted images using different models. The selected distortions may effect quality, comfort, and depth perception in VR. We also include downscaling, a distortion absent from SCOPE. We find that in all cases our model is able to extrapolate, behaving monotonically for a wider distortion range than it was trained on. We observe that MANIQA [95] frequently rates an image as higher quality the more degraded it becomes, once an initial threshold is passed. StereoQA-Net [104] generally maintains monotonic behavior, however it tends to plateau quickly. Contrastingly, our model exhibits mostly monotonic behavior and a larger dynamic range, with gaussian noise being the primary exception.

### 4.3. How do viewing mediums influence stereo 3D perception?

The underlying assumption that led us to collect annotations for SCOPE on VR, rather than a 2D screen, is that there is limited correlation between 2D perception of stereo images and their perception on VR devices. In Figure 6 we validate this hypothesis by randomly selecting 50 samples from SCOPE and having 10 participants repeatedly annotate them using different mediums: Apple Vision Pro, Meta Quest Pro, anaglyph images, and toggling between the left and right images on a monitor. We compute Cohen's kappa coefficient for each participant's responses across different mediums, then average these values across participants.

Evidently, the correlation between VR devices (Meta Quest Pro, Apple Vision Pro) and non-VR devices (anaglyph, toggle) is quite low. Therefore, these simpli-

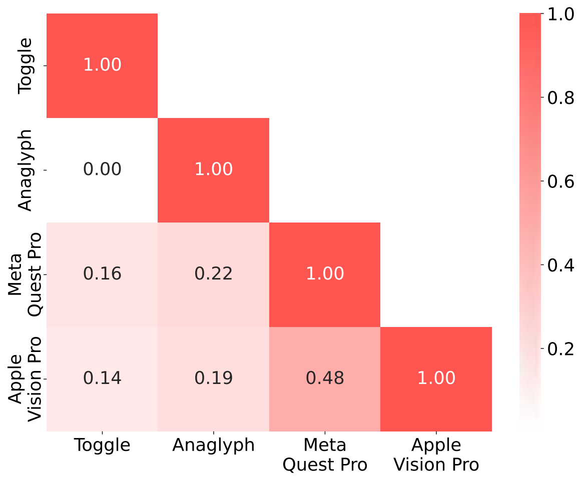


Figure 6. **Viewing medium comparison.** We measure the correlation between human preferences across viewing mediums by calculating Cohen's kappa coefficient averaged across participants.

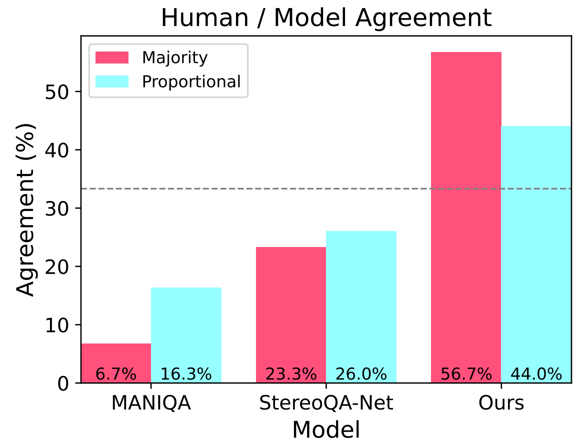


Figure 7. **Alignment of stereo image preference between our participants and models.** We calculate the agreement in two different ways. (i) *Majority*, where the model receives a binary score based on agreement with the majority human vote. (ii) *Proportional*, in which the model score is proportional to the fraction of human votes for its preferred image.

fied setups do not effectively reflect user experience on VR headsets, and thus are less suitable for SQoE data annotation. On the other hand, human preferences on Apple Vision Pro and Meta Quest Pro are have non-negligible correlation, which is a positive indication for the SCOPE's generalization across VR devices. In addition, we find that there is nontrivial agreement between participants across any specific viewing medium, as shown in Figure 13 in the SM. Notably, due to the cognitively penetrable nature of some examples in our dataset, perfect correlation would not be expected even for responses from the same user on the same device across two separate annotation sessions. See Section 10 of the SM for more details.

### 4.4. Do models & humans evaluate off-the-shelf stereo generators similarly?

To assess our model's practicality for real use cases, we correlate human opinion with model ranking on off-the-shelf mono-to-stereo conversion techniques (see Figure 7): Depthify.ai [18], Immersity AI [31], and Owl3D [61]. We applied these to 30 monoscopic images from the Spring dataset [52], with the main resulting artifacts being inaccurate depth, flawed inpainting, and jagged foreground edges. Importantly, our model did not see any images from Spring during training, and its animated artistic style is out of SCOPE's distribution.

We conduct a user study using the Apple Vision Pro headset with 10 participants, asking them to select their preferred stereo image from the three options generated for each original image. To minimize bias, we randomize both the order of the stereo image triplets and the arrangement of images within each triplet for every participant. Our findings indicate that participants generally favored the stereo images produced by Immersity AI, followed by Owl3D. Subsequently, we compare the alignment of user preferences with the assessments from MANIQA [95] - the top-performing IQA method, Stereo-IQA [104], and our proposed model. Figure 7 exhibits our model has the highest correlation with human opinion. More information about this study appears in Section 11 of the SM.

## 5. Discussion

A major challenge in SQoE evaluation is the scarcity of annotated data. To address this, we curated a large dataset and developed a novel SQoE predictor to capture stereoscopic quality nuances. We validated its effectiveness by comparing novel-view synthesis methods, demonstrating superior alignment with human preference over existing methods.

However, our work has some inherent limitations, mainly resulting from our data being sourced from few existing sources. 41% of our stereo images originate from the Holopix50k-HD [29] dataset and are distorted with a non-NVS distortion. This dataset was primarily captured using the RED Hydrogen One smartphone, which limits the horizontal disparity between the left and right images and affects depth perception. Moreover, our NVS subset is comprised of 400 images taken from only 13 scenes. Furthermore, our model inherits certain limitations from DINOv2 [60], which we chose as our backbone.

Despite these limitations, we believe that our work represents a significant step forward in SQoE evaluation. As stereoscopic content increases in prevalence, robust automated assessment is increasingly essential. We hope that our contributions lay the groundwork for more sophisticated and accurate evaluation methods.

**Acknowledgements** We thank all the participants in the user studies presented in this paper. We thank Bahjat Kawar, Yaron Ostrovsky Berman, Dror Simon and Matan Hugi for their early feedback. This work was partially supported by the Sagol Weizmann-MIT Bridge Program and an NSF GFRP Fellowship to SS.

## References

[1] Roushain Akhter, ZM Parvez Sazzad, Yuukou Horita, and Jacky Baltes. No-reference stereoscopic image quality assessment. In *Stereoscopic Displays and Applications XXI*, pages 271–282. SPIE, 2010. 2

[2] Apple. Apple introduces spatial video capture on iphone 15 pro. https://www.apple.com/il/newsroom/2023/12/apple-introduces-spatial-video-capture-on-iphone-15-pro/, 2023. [Accessed 16-10-2024]. 2

[3] Takehiko Bando, Atsuhiko Iijima, and Sumio Yano. Visual fatigue caused by stereoscopic images and the search for the requirement to prevent them: A review. *Displays*, 33 (2):76–83, 2012. 2

[4] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Yuanzhen Li, Tomer Michaeli, et al. Lumiere: A space-time diffusion model for video generation. *arXiv preprint arXiv:2401.12945*, 2024. 2

[5] Jonathan T. Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P. Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. *ICCV*, 2021. 2

[6] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *CVPR*, 2022. 3

[7] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Zip-nerf: Anti-aliased grid-based neural radiance fields. *ICCV*, 2023. 2

[8] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a" siamese" time delay neural network. *NeurIPS*, 6, 1993. 5

[9] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. 2

[10] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In *ECCV*, 2012. 3

[11] Patrick Cavanagh. The cognitive impenetrability of cognition. *Behavioral and Brain Sciences*, 22(3):370–371, 1999. 6

[12] Jianyu Chen, Jun Zhou, Jun Sun, and Alan Conrad Bovik. Visual discomfort prediction on stereoscopic 3d images without explicit disparities. *Signal Processing: Image Communication*, 51:50–60, 2017. 2

[13] Kenneth Chen, Thomas Wan, Nathan Matsuda, Ajit Ninan, Alexandre Chapiro, and Qi Sun. Pea-pods: Perceptual evaluation of algorithms for power optimization in xr displays. *ACM Transactions on Graphics (TOG)*, 43(4):1–17, 2024. 3

[14] Meixu Chen, Yize Jin, Todd Goodall, Xiangxu Yu, and Alan Conrad Bovik. Study of 3d virtual reality picture quality. *IEEE Journal of Selected Topics in Signal Processing*, 14(1):89–102, 2019. 2

[15] Ming-Jun Chen, Lawrence K Cormack, and Alan C Bovik. No-reference quality assessment of natural stereopairs. *IEEE Transactions on Image Processing*, 22(9):3379–3391, 2013. 2, 3

[16] Ming-Jun Chen, Che-Chun Su, Do-Kyoung Kwon, Lawrence K Cormack, and Alan C Bovik. Full-reference quality assessment of stereopairs accounting for rivalry. *Signal Processing: Image Communication*, 28(9):1143–1155, 2013. 3, 4

[17] Peng Dai, Feitong Tan, Qiangeng Xu, David Futschik, Ruofei Du, Sean Fanello, Xiaojuan Qi, and Yinda Zhang. Svg: 3d stereoscopic video generation via denoising frame matrix. *arXiv preprint arXiv:2407.00367*, 2024. 2

[18] Depthify.ai. Depthify.ai — convert 2d videos to 3d spatial videos. http://depthify.ai, 2024. 8

[19] Yong Ding, Ruizhe Deng, Xin Xie, Xiaogang Xu, Yang Zhao, Xiaodong Chen, and Andrey S Krylov. No-reference stereoscopic image quality assessment using convolutional neural network for adaptive feature extraction. *IEEE Access*, 6:37595–37603, 2018. 3

[20] Neil A Dodgson. Variation and extrema of human interpupillary distance. In *Stereoscopic displays and virtual reality systems XI*, pages 36–46. SPIE, 2004. 3

[21] Eric Dubois. A projection method to generate anaglyph stereo images. pages 1661 – 1664 vol.3, 2001. 1

[22] Daniel Duckworth, Peter Hedman, Christian Reiser, Peter Zhizhin, Jean-François Thibert, Mario Lučić, Richard Szeliski, and Jonathan T. Barron. Smerf: Streamable memory efficient radiance fields for real-time large-scene exploration, 2023. 2

[23] Zhiwen Fan, Wenyan Cong, Kairun Wen, Kevin Wang, Jian Zhang, Xinghao Ding, Danfei Xu, Boris Ivanovic, Marco Pavone, Georgios Pavlakos, et al. Instantsplat: Unbounded sparse-view pose-free gaussian splatting in 40 seconds. *arXiv preprint arXiv:2403.20309*, 2024. 2

[24] Yuming Fang, Jiebin Yan, Xuelin Liu, and Jiheng Wang. Stereoscopic image quality assessment by deep convolutional neural network. *Journal of Visual Communication and Image Representation*, 58:400–406, 2019. 2, 3

[25] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *NeurIPS*, 2023. 5

[26] Phillip Guan, Eric Penner, Joel Hegland, Benjamin Letham, and Douglas Lanman. Perceptual requirements for world-locked rendering in ar and vr. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–10, 2023. 3

[27] Peter Hedman, Julien Philip, True Price, Jan-Michael Frahm, George Drettakis, and Gabriel Brostow. Deep

blending for free-viewpoint image-based rendering. 37(6): 257:1–257:15, 2018. 3

[28] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *ICLR*, 2022. 5

[29] Yiwen Hua, Puneet Kohli, Pritish Uplavikar, Anand Ravi, Saravana Gunaseelan, Jason Orozco, and Edward Li. Holopix50k: A large-scale in-the-wild stereo image dataset. In *CVPR Workshop on Computer Vision for Augmented and Virtual Reality, Seattle, WA, 2020.*, 2020. 3, 8

[30] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. If you use this software, please cite it as below. 6

[31] Immersity AI. Immersity ai — convert image and video to 3d. https://www.immersity.ai, 2024. 8

[32] Fortune Business Insights. Virtual Reality [VR] Market Size, Growth, Share — Report, 2032 — fortunebusinessinsights.com. https://www.fortunebusinessinsights.com/industry-reports/virtual-reality-market-101378, 2024. [Accessed 16-10-2024]. 2

[33] Ajay Jain, Ben Mildenhall, Jonathan T Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. In *CVPR*, pages 867–876, 2022. 2

[34] Sen Jia and Yang Zhang. Saliency-based deep convolutional neural network for no-reference image quality assessment. *Multimedia Tools and Applications*, 77:14859–14872, 2018. 2

[35] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *CVPR*, 2024. 3

[36] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):1–14, 2023. 2, 3

[37] Donghyun Kim and Kwanghoon Sohn. Visual fatigue prediction for stereoscopic image. *IEEE transactions on circuits and systems for video technology*, 21(2):231–236, 2011. 2

[38] Hak Gu Kim, Hyunwook Jeong, Heoun taek Lim, and Yong Man Ro. Binocular fusion net: Deep learning visual comfort assessment for stereoscopic 3d. *IEEE Transactions on Circuits and Systems for Video Technology*, 29:956–967, 2019. 2

[39] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics*, 36 (4), 2017. 3

[40] Frank L Kooi and Alexander Toet. Visual comfort of binocular and 3d displays. *Displays*, 25(2-3):99–108, 2004. 2

[41] Brooke Krajancich, Petr Kellnhofer, and Gordon Wetzstein. Optimizing depth perception in virtual and augmented reality through gaze-contingent stereo rendering. *ACM Trans. Graph.*, 39(6), 2020. 3

[42] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *NeurIPS*, 2012. 6

[43] Marc Lambooij, Wijnand IJsselsteijn, Marten Fortuin, Ingrid Heynderickx, et al. Visual discomfort and visual fatigue of stereoscopic displays: A review. *Journal of imaging science and technology*, 53(3):30201–1, 2009. 2, 5

[44] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 6

[45] Yafei Li, Feng Yang, Wenbo Wan, Jun Wang, Min Gao, Jia Zhang, and Jiande Sun. No-reference stereoscopic image quality assessment based on visual attention and perception. *IEEE Access*, 7:46706–46716, 2019. 2, 3

[46] Heaun-Taek Lim, Hak Gu Kim, and Yang Man Ra. Vr iqa net: Deep virtual reality image quality assessment using adversarial learning. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6737–6741. IEEE, 2018. 2

[47] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *CVPR*, 2023. 2

[48] Tsung-Jung Liu, Kuan-Hsien Liu, and Kuan-Hung Shen. Learning based no-reference metric for assessing quality of experience of stereoscopic images. *Journal of Visual Communication and Image Representation*, 61:272–283, 2019. 2, 3, 4

[49] Yun Liu, Chang Tang, Zhi Zheng, and Liyuan Lin. No-reference stereoscopic image quality evaluator with segmented monocular features and perceptual binocular features. *Neurocomputing*, 405:126–137, 2020. 2, 3

[50] Zhen Lv, Yangqi Long, Congzhentao Huang, Cao Li, Chengfei Lv, Hao Ren, and Dian Zheng. Spatialdreamer: Self-supervised stereo video synthesis from monocular input. *IEEE VR*, 2024. 2

[51] Nathan Matsuda, Alex Chapiro, Yang Zhao, Clinton Smith, Romain Bachy, and Douglas Lanman. Realistic luminance in vr. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–8, 2022. 3

[52] Lukas Mehl, Jenny Schmalfuss, Azin Jahedi, Yaroslava Nalivayko, and Andrés Bruhn. Spring: A high-resolution high-detail dataset and benchmark for scene flow, optical flow and stereo. In *CVPR*, 2023. 3, 8, 4

[53] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *ICLR*, 2022. 1, 3

[54] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *CVPR*, 2015. 3

[55] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf:

Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 2

[56] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12): 4695–4708, 2012. 2, 6

[57] Anush Krishna Moorthy, Che-Chun Su, Anish Mittal, and Alan Conrad Bovik. Subjective evaluation of stereoscopic image quality. *Signal Processing: Image Communication*, 28(8):870–883, 2013. 3, 6, 4

[58] NRG. Beyond reality: Is the long-awaited vr revolution finally on the horizon?, 2022. 2

[59] Heeseok Oh, Sewoong Ahn, Sanghoon Lee, and Alan Conrad Bovik. Deep visual discomfort predictor for stereoscopic 3d images. *IEEE Transactions on Image Processing*, 27(11):5420–5432, 2018. 2

[60] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *TMLR*, 2024. 5, 6, 8

[61] Owl3D. Owl3d — ai-powered 2d to 3d conversion software. https://www.owl3d.com, 2024. 8

[62] Jincheol Park, Heeseok Oh, Sanghoon Lee, and Alan Conrad Bovik. 3d visual discomfort predictor: Analysis of disparity and neural activity statistics. *IEEE transactions on image processing*, 24(3):1101–1114, 2014. 2, 3

[63] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *ICLR*, 2023. 2

[64] Clare Porac and Stanley Coren. The dominant eye. *Psychological bulletin*, 83(5):880, 1976. 5

[65] Ajay Kumar Reddy Poreddy, Raja Bharath Chandra Ganeswaram, Balasubramanyam Appina, Priyanka Kokil, and Ram Bilas Pachori. No-reference virtual reality image quality evaluator using global and local natural scene statistics. *IEEE Transactions on Instrumentation and Measurement*, 2023. 2

[66] Steven Poulakos, Rafael Monroy, Tunc Aydin, Oliver Wang, Aljoscha Smolic, and Markus Gross. A computational model for perception of stereoscopic window violations. In *2015 Seventh International Workshop on Quality of Multimedia Experience (QoMEX)*, pages 1–6. IEEE, 2015. 2

[67] Feng Qi, Tingting Jiang, Siwei Ma, and Debin Zhao. Quality of experience assessment for stereoscopic images. In *2012 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1712–1715, 2012. 2

[68] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 5, 6

[69] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3), 2022. 3

[70] Lars Rehm. RED HYDROGEN Holographic Smartphone - first look — dxomark.com. https://www.dxomark.com/red-hydrogen-holographic-smartphone-first-look/, 2018. [Accessed 25-11-2024]. 2

[71] Mike Rubini. Capturing the Future: Exploring the Ever-Increasing Popularity of Dual Camera Smartphones — treendly.com. https://treendly.com/blog/capturing-the-future-exploring-the-ever-increasing-popularity-of-dual-camera-smartphones, 2023. [Accessed 25-11-2024]. 2

[72] Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nešić, Xi Wang, and Porter Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *GCPR*, 2014. 3

[73] Neil Shah. Dual Camera Smartphones Are Going Mainstream: Explosive Growth in 2018 — counterpointresearch.com. https://www.counterpointresearch.com/insights/dual-camera-smartphones-going-mainstream-explosive-growth-2018/, 2018. [Accessed 16-10-2024]. 2

[74] Liquan Shen, Ruigang Fang, Yang Yao, Xianqiu Geng, and Dapeng Wu. No-reference stereoscopic image quality assessment based on image distortion and stereo perceptual information. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 3(1):59–72, 2018. 2, 3

[75] Lili Shen, Xiongfei Chen, Zhaoqing Pan, Kefeng Fan, Fei Li, and Jianjun Lei. No-reference stereoscopic image quality assessment based on global and local content characteristics. *Neurocomputing*, 424:132–142, 2021. 2, 3

[76] Jian Shi, Zhenyu Li, and Peter Wonka. Immersepro: End-to-end stereo video synthesis via implicit disparity learning. *arXiv preprint arXiv:2410.00262*, 2024. 2

[77] Takashi Shibata, Joohwan Kim, David M Hoffman, and Martin S Banks. The zone of comfort: Predicting visual discomfort with stereo displays. *Journal of vision*, 11(8): 11–11, 2011. 2

[78] Jianwei Si, Baoxiang Huang, Huan Yang, Weisi Lin, and Zhenkuan Pan. A no-reference stereoscopic image quality assessment network based on binocular interaction and fusion mechanisms. *IEEE Transactions on Image Processing*, 31:3066–3080, 2022. 2, 3

[79] Dustin Stokes. Cognitive penetrability of perception. *Philosophy Compass*, 8(7):646–663, 2013. 6

[80] Xiangjie Sui, Kede Ma, Yiru Yao, and Yuming Fang. Perceptual quality assessment of omnidirectional images as moving camera videos. *IEEE Transactions on Visualization and Computer Graphics*, 28(8):3022–3034, 2021. 2

[81] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *WACV*, pages 2149–2159, 2022. 3

[82] Bruce H Thomas. Examining user perception of the size of multiple objects in virtual reality. *Applied Sciences*, 10(11): 4049, 2020. 3

[83] Kazuhiko Ukai and Peter A Howarth. Visual fatigue caused by viewing stereoscopic motion images: Background, theories, and observations. *Displays*, 29(2):106–116, 2008. 2

[84] Jiheng Wang, Kai Zeng, and Zhou Wang. Quality prediction of asymmetrically distorted stereoscopic images from single views. In *2014 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2014. 3, 4

[85] Jiheng Wang, Abdul Rehman, Kai Zeng, Shiqi Wang, and Zhou Wang. Quality prediction of asymmetrically distorted stereoscopic 3d images. *IEEE Transactions on Image Processing*, 24(11):3400–3414, 2015. 3, 4

[86] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *AAAI*, 2023. 2, 6

[87] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023. 2

[88] Yingqian Wang, Longguang Wang, Jungang Yang, Wei An, and Yulan Guo. Flickr1024: A large-scale dataset for stereo image super-resolution. In *International Conference on Computer Vision Workshops*, pages 3852–3857, 2019. 3

[89] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *NeurIPS*, 36, 2024. 2

[90] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Tianshui Chen, Menghan Xia, Ping Luo, and Yin Shan. Motionctrl: A unified and flexible motion controller for video generation. 2024. 2, 3

[91] Philippe Weinzaepfel, Vincent Leroy, Thomas Lucas, Romain Brégier, Yohann Cabon, Vaibhav Arora, Leonid Antsfeld, Boris Chidlovskii, Gabriela Csurka, and Jérôme Revaud. Croco: Self-supervised pre-training for 3d vision tasks by cross-view completion, 2022. 6

[92] Philippe Weinzaepfel, Thomas Lucas, Vincent Leroy, Yohann Cabon, Vaibhav Arora, Romain Brégier, Gabriela Csurka, Leonid Antsfeld, Boris Chidlovskii, and Jérôme Revaud. Croco v2: Improved cross-view completion pre-training for stereo matching and optical flow, 2023. 6

[93] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023. 2, 6

[94] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024. 3

[95] 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. In *CVPR*, 2022. 2, 6, 7, 8

[96] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 2

[97] Sumio Yano, Masaki Emoto, and Tetsuo Mitsuhashi. Two factors in visual fatigue caused by stereoscopic hdtv images. *Displays*, 25(4):141–150, 2004. 2

[98] Taoran Yi, Jiemin Fang, Junjie Wang, Guanjun Wu, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Qi Tian, and Xinggang Wang. Gaussiandreamer: Fast generation from text to 3d gaussians by bridging 2d and 3d diffusion models. In *CVPR*, pages 6796–6807, 2024. 2

[99] Zehao Yu, Anpei Chen, Binbin Huang, Torsten Sattler, and Andreas Geiger. Mip-splatting: Alias-free 3d gaussian splatting. In *CVPR*, 2024. 2

[100] Huilin Zhang, Sumei Li, and Yongli Chang. Towards top-down stereoscopic image quality assessment via stereo attention. *arXiv preprint arXiv:2308.04156*, 2023. 2, 3

[101] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 5

[102] Wei Zhang, Chenfei Qu, Lin Ma, Jingwei Guan, and Rui Huang. Learning structure of stereoscopic image for no-reference quality assessment with convolutional neural network. *Pattern Recognition*, 59:176–187, 2016. 2, 3

[103] Sijie Zhao, Wenbo Hu, Xiaodong Cun, Yong Zhang, Xiaoyu Li, Zhe Kong, Xiangjun Gao, Muyao Niu, and Ying Shan. Stereocrafter: Diffusion-based generation of long and high-fidelity stereoscopic 3d from monocular videos. *arXiv preprint arXiv:2409.07447*, 2024. 2

[104] Wei Zhou, Zhibo Chen, and Weiping Li. Dual-stream interactive networks for no-reference stereoscopic image quality assessment. *IEEE Transactions on Image Processing*, 28 (8):3946–3958, 2019. 2, 6, 7, 8, 3

[105] Yang Zhou, Pingan Chen, Haibing Yin, Xiaofeng Huang, and Zhu Li. Stereoscopic image discomfort prediction using dual-stream multi-level interactive network. *Displays*, 78:102444, 2023. 2

[106] Xilei Zhu, Liu Yang, Huiyu Duan, Xiongkuo Min, Guangtao Zhai, and Patrick Le Callet. Esiqa: Perceptual quality assessment of vision-pro-based egocentric spatial images. *arXiv preprint arXiv:2407.21363*, 2024. 6

# What Makes for a Good Stereoscopic Image?

## Supplementary Material

## 6. Dataset Details

Our dataset is comprised of $2400$ examples, each containing a pair of stereo images, resulting in a total of $4800$ stereo images, all of which have undergone some form of distortion. Table 3 presents the total amount in which each distortion appears in the dataset. Figure 8 shows visual examples for several distortions, exaggerated for illustration purposes.

| Distortion type | Occurrence |
|---:|---|
| 2D lifting | 1364 |
| MotionCtrl | 1156 |
| 3D Gaussian splatting | 306 |
| SDEdit | 241 |
| Uniform White Noise | 152 |
| Chromatic Aberration | 144 |
| Rotation | 141 |
| Keystone | 138 |
| Average Blur | 127 |
| Gaussian Blur | 125 |
| JPEG Compression | 124 |
| Gaussian White Noise | 123 |
| Checkerboard | 117 |
| Warping | 115 |
| Brightness | 97 |
| Saturation | 95 |
| Contrast | 80 |
| Hue | 79 |
| Magnification | 76 |

Table 3. **Frequency of applied distortions in our proposed dataset.**

## 7. Training Details

Our model is trained on a single NVIDIA A100 using an Adam optimizer with a learning rate of $3e-5$ and batch size of $16$. We maintain the original $1280 \times 720$ resolution, only applying a center crop to $1274 \times 714$ for compatibility with DINOv2-S's patch size of $14$. During training, we finetune the DINOv2 backbone using LoRA, with a rank of $8$, alpha of $32$, and dropout of $0.1$. We use a with a margin of $0.05$ for the hinge loss. Additionally, the training data is weighted based on annotator consensus levels, with each epoch taking approximately $12$ minutes to train and $1$ minute to validate.

## 8. Detailed Performance on the SCOPE dataset

In Figure 9 we report test set accuracy across several different train, validation and test partitions, categorized by annotation consensus: unanimous ($5-0$ split), majority ($4-1$ split), and divided ($3-2$ split), with the latter being the noisiest and most cognitively penetrable. The sizes of these splits are similar, comprising $32.9\%$, $34.1\%$, and $32.9\%$ of the data respectively, confirming that our dataset contains a learnable signal. We train and evaluate our model on five $80\% - 10\% - 10\%$ dataset splits, using five different seeds for each split, and report the mean and standard deviation in Figure 9.

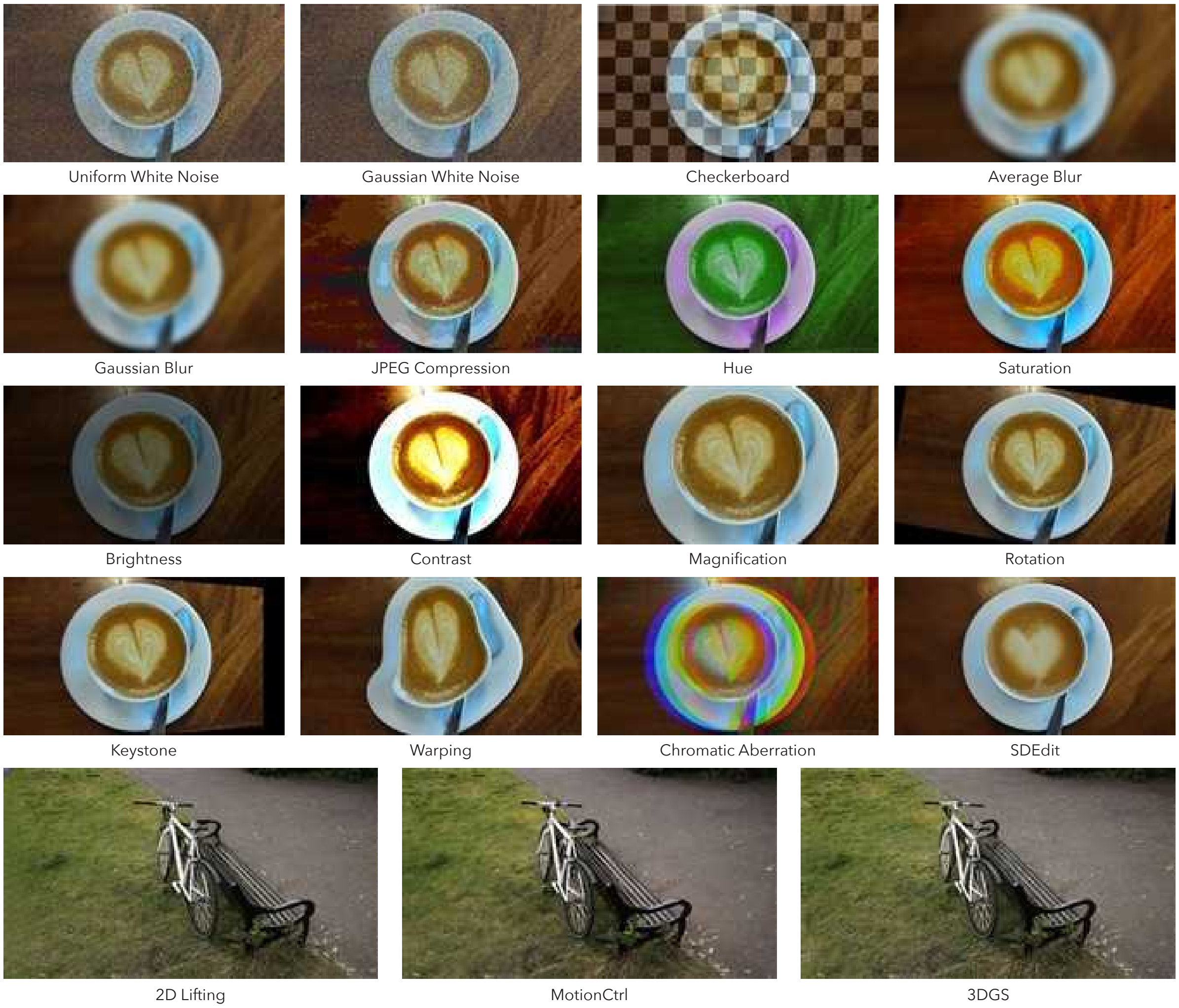


Figure 8. **Distortion examples.** We show examples of image distortions in our dataset, exaggerated for illustration purposes.

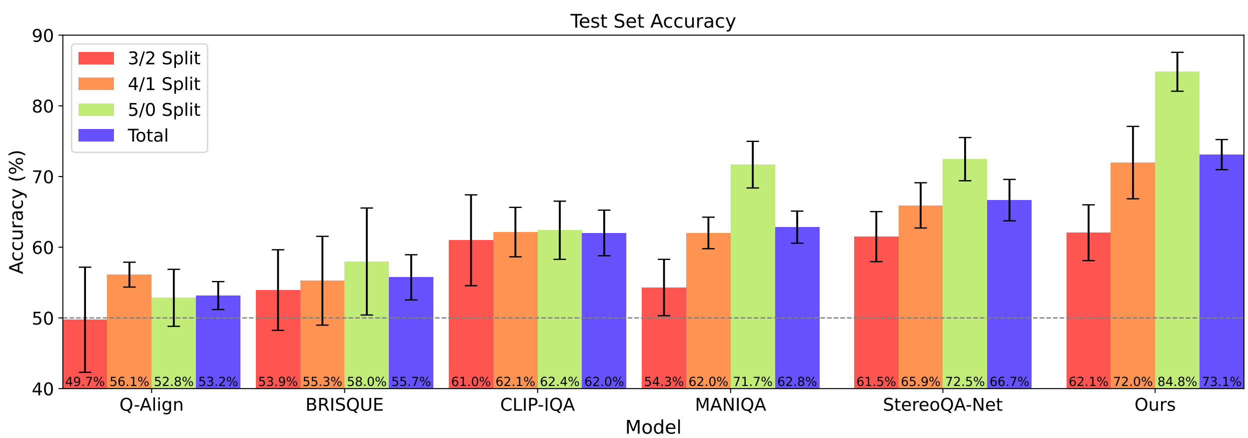


Figure 9. **Test set performance.** We test the performance of existing IQA and NR-SIQA models as well as our proposed model on a held out test sets, and show the results on different splits.


Figure 10. **Distortion strength comparison.** Comparing maximum distortion strengths across existing datasets and our proposed dataset, demonstrating that the distortions applied in our dataset exhibit significantly lower intensity compared to existing SIQA datasets.

| | | LIVE Phase I | | LIVE Phase II | | WIVC Phase I | | WIVC Phase II | |
|---|---|---|---|---|---|---|---|---|---|
| | **Method** | **SROCC ↑** | **PLCC ↑** | **SROCC ↑** | **PLCC ↑** | **SROCC ↑** | **PLCC ↑** | **SROCC ↑** | **PLCC ↑** |
| **Manual Feature Based** | Chen *et al.* [15] | 0.891 | 0.895 | 0.880 | 0.880 | – | – | – | – |
| | Shen *et al.* [74] | 0.932 | 0.936 | 0.927 | 0.932 | – | – | – | – |
| | Li *et al.* [45] | 0.953 | 0.965 | 0.946 | 0.955 | 0.937 | 0.949 | 0.952 | 0.960 |
| | Liu *et al.* [49] | 0.949 | 0.958 | 0.933 | 0.935 | 0.928 | 0.945 | 0.901 | 0.913 |
| **Deep Learning Based** | Zhang *et al.* [102] | 0.943 | 0.947 | 0.915 | 0.912 | – | – | – | – |
| | Ding *et al.* [19] | 0.942 | 0.940 | 0.924 | 0.930 | – | – | – | – |
| | Fang *et al.* [24] | 0.946 | 0.957 | 0.934 | 0.946 | – | – | – | – |
| | Zhou *et al.* [104] | 0.965 | 0.973 | 0.947 | 0.957 | – | – | – | – |
| | Shen *et al.* [75] | 0.962 | 0.972 | 0.951 | 0.953 | – | – | – | – |
| | Si *et al.* [78] | 0.966 | 0.978 | 0.953 | 0.972 | 0.960 | 0.969 | 0.950 | 0.958 |
| | Zhang *et al.* [100] | 0.972 | 0.977 | 0.962 | 0.964 | 0.972 | 0.973 | 0.972 | 0.973 |
| | iSQoE (Ours) | 0.774 | 0.758 | 0.763 | 0.767 | 0.627 | 0.687 | 0.542 | 0.536 |

Table 4. **Evaluation on existing datasets for stereoscopic image quality assessment.**

## 9. Performance on Existing SIQA Datasets

Table 5 provides a comparison between our dataset and existing stereo quality assessment datasets: LIVE 3D Phases I and II [16, 57], Waterloo IVC (WIVC) 3D Phases I and II [84, 85] and IEEE-SA [48]. SCOPE differs from them in several aspects:

1. **Image Quantity:** SCOPE is the largest of the datasets, with more than twice the amount of samples than IEEE-SA - the second largest dataset.
2. **Annotation medium:** The annotations in all these datasets were collected using passive stereoscopic displays or active shutter glasses, while ours were collected on a Vision Pro headset. In Section 4.3 and Figure 6 we demonstrate low correlation between preferences on VR devices and other stereo viewing methods.
3. **Annotation Protocol:** The other datasets collected Mean Opinion Score annotations, an absolute single-image protocol, while SCOPE collected 2AFC which are relative annotations.
4. **Distortion Strengths:** The other datasets applied significantly stronger distortions than SCOPE, see Figure 10.

We evaluate our model on LIVE 3D Phase I and II [15, 16, 57] and Waterloo IVC (WIVC) 3D Phase I and II [84, 85]. For these evaluations, we use standard performance metrics: Spearman rank order correlation coefficient (SROCC) and Pearson linear correlation coefficient (PLCC).

Table 4 shows there is a significant performance gap between our models and the state-of-the-art models reporting performance on these datasets. We attribute this to the difference in annotation mediums between these datasets and

| Dataset | Samples | Stereo Images | Clean Images | Annotation Type | Distortions |
|---|---|---|---|---|---|
| LIVE Phase I [57] | 365 | 365 | 20 | DMOS | Noise, Blur, Compression, Fast-fading |
| LIVE Phase II [16] | 360 | 360 | 8 | DMOS | Noise, Blur, Compression, Fast-fading |
| WIVC Phase I [84] | 330 | 330 | 6 | MOS | Noise, Blur |
| WIVC Phase II [85] | 460 | 460 | 10 | MOS | Noise, Blur, Compression, |
| IEEE-SA [48] | 800 | 800 | 160 | MOS | Horizontal disparity |
| SCOPE (Ours) | 2400 | 4800 | 2400 | 2AFC | 19 types, see Table 1 |

Table 5. **Stereoscopic Preference Datasets.** Prior datasets for stereo image evaluation vary in terms of size, the psychophysical experiment in which the annotations were collected, and the distortions they encompass.

SCOPE. Our model is trained and fitted to grade quality as it is perceived on a VR device, rather than on passive stereoscopic displays.

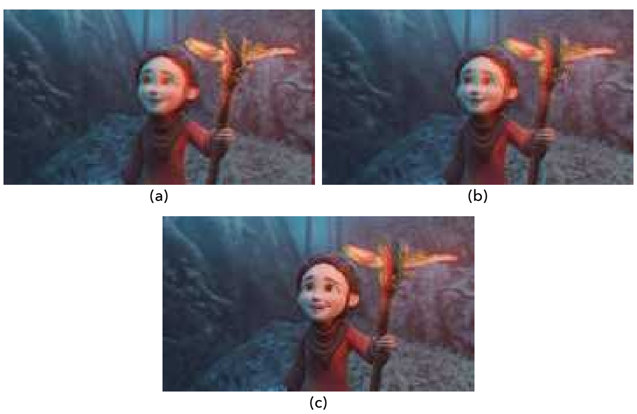


Figure 11. **Example from our off-the-shelf, mono-to-stereo experiment.** Three versions of the same stereo image are generated using different off-the-shelf, mono-to-stereo conversion methods. (a) Depthify.ai (b) Immersity AI (c) Owl3D. The stereo images are presented as anaglyph images for viewing purposes. We recommend viewing the images on a screen and zooming in to better observe the differences.

## 10. Cross-Medium User Study

Expanding on the user study outlined in Section 4.3, we detail the specific viewing setups for each device. Viewing stereoscopic images with the Apple Vision Pro was done through the native photos app in immersive mode. For the Meta Quest Pro, we employed a third-party application (Pegasus VR media player) due to the absence of a suitable first-party viewer. Both the toggling and anaglyph setups were presented via HTML pages, shown in Figure 12. We opted for full-color anaglyph images, as this convention provided the best stereoscopic 3D experience with our monitor and glasses combination, among all conventions tested.

In addition to Figure 6 that shows the mean correlation, Figure 13 we show the Cohen's kappa coefficient between each of the 10 participants for each viewing device.

## 11. Off-the-Shelf Mono-to-Stereo Evaluation

We evaluated alignment of human opinion with the different SQoE candidates on the Spring [52] dataset. Figure 11 shows an example from the user study.

## 12. Licenses

The models and datasets we use are provided under the licenses in Table 6.

| Dataset | License | Model | License |
|---|---|---|---|
| Tanks and Temples | CC BY 4.0 | MotionCtrl | Apache 2.0 |
| Deep Blending | Apache 2.0 | MiDaS | MIT |
| Mip-NeRF 360 | Apache 2.0 | Marigold | Apache 2.0 |
| Holopix50k | NC | Depth Anything | Apache 2.0 |
| SPRING | CC BY 4.0 | LaMa | Apache 2.0 |
| LIVE 3D Phase I | Custom Academic | 3DGS | NC |
| LIVE 3D Phase II | Custom Academic | DINOv2 | Apache 2.0 |
| WIVC 3D Phase I | Custom Academic | Q-Align | S-Lab 1.0 |
| WIVC 3D Phase II | Custom Academic | BRISQUE | Apache 2.0 |
| | | CLIP-IQA | S-Lab 1.0 |
| | | MANIQA | Apache 2.0 |
| | | StereoQA-Net | Custom Academic |
| | | CLIP | MIT |
| | | OpenCLIP | MIT |
| | | Croco | CC BY-NC-SA 4.0 |
| | | Depthify.ai | Custom |
| | | Immersity AI | Custom |
| | | Owl3D | Custom |

Table 6. **Dataset and model licenses.**

Figure 12. **User study setups: left/right image toggling (top) and anaglyph stereo (bottom)**

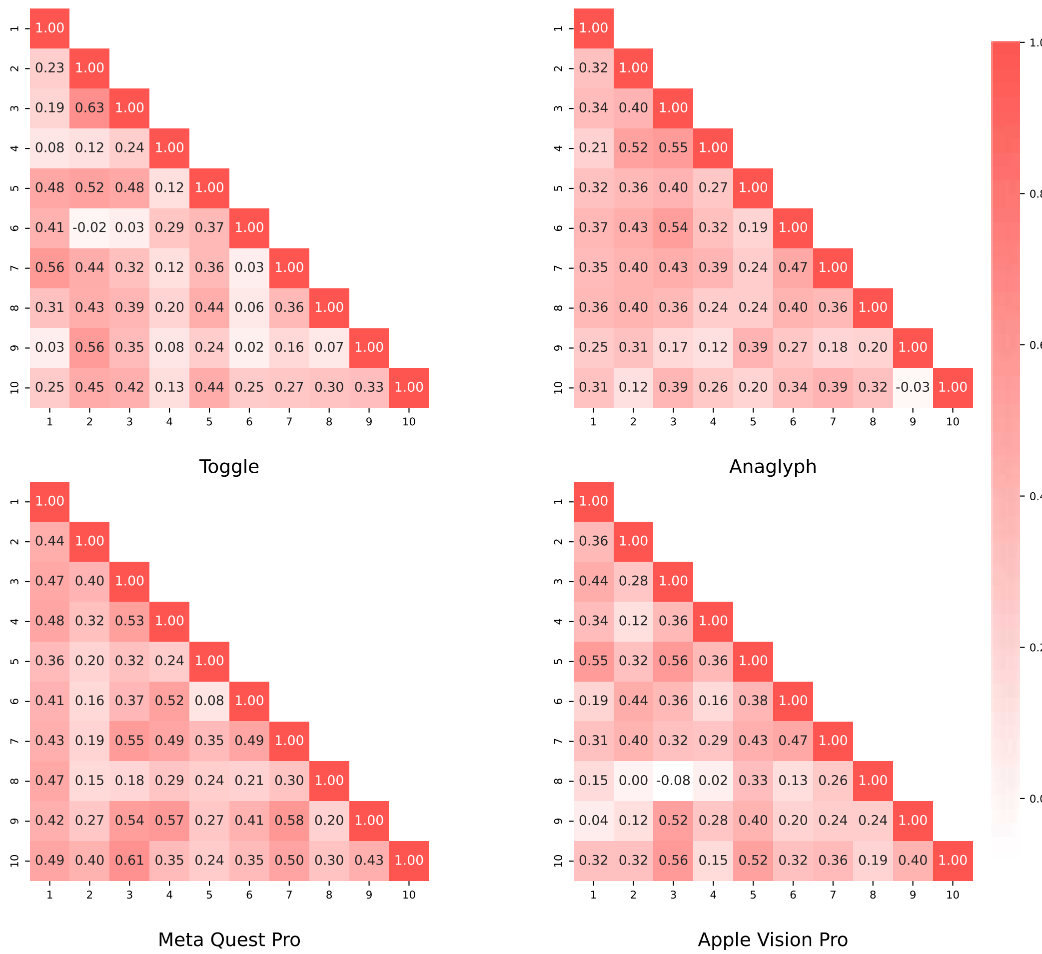


Figure 13. **Inter-rater agreement for each viewing medium, measured using Cohen's kappa coefficient.** The heatmap displays the agreement scores between all pairs of the 10 participants, highlighting the correlation in subjective evaluations across different viewing conditions.