NVRQ: A Benchmark for Object-Level Quality in 3D Neural Reconstruction

Wangren Xu* Bidur Khanal* Naveen Kumar Rai Bjoern Haefner Sean Pieper
Nick Schneider

NVIDIA

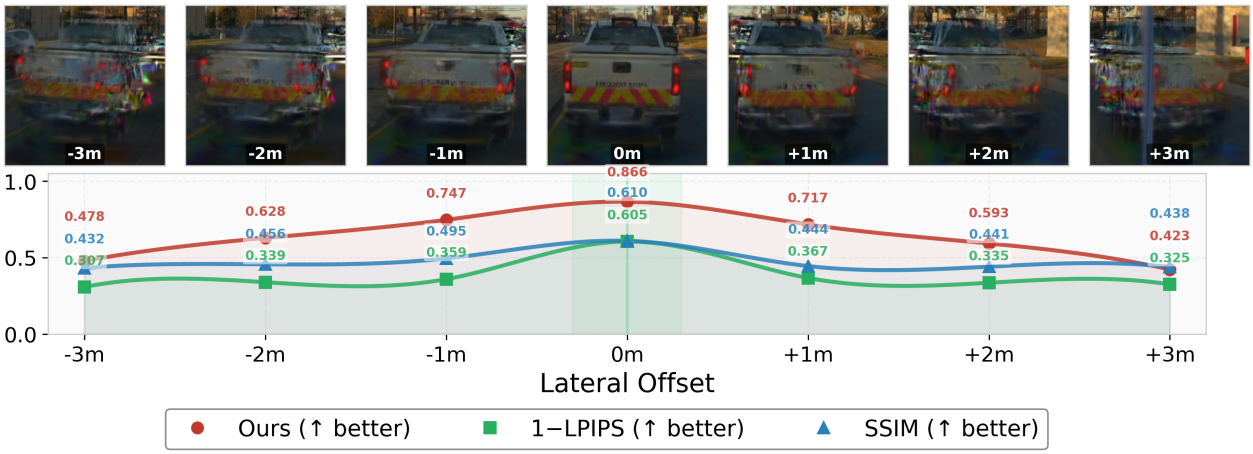

Figure 1. Our Curve Adjusted Novel View Reconstruction Quality (*NVRQ-CA* ↑) metric is able to better quantify rendering degradation in dynamic objects with increasing lateral offsets. Compared to two standard baselines like LPIPS (↑, we show 1 − LPIPS here for easier comparison) and SSIM (↑), our proposed *NVRQ-CA* metric is more sensitive to reconstruction artifacts, demonstrating a much sharper decrease in score as visual degradation intensifies.

Abstract

Realistic real-world sensor simulations are key to the development of modern end-to-end models for Physical AI. Neural reconstruction has emerged as a promising solution, but it often shows degradations when rendering novel views. Evaluating the quality of novel views remains challenging because standard scene-level image quality metrics often average away localized artifacts on safety-critical actors such as pedestrians, cyclists, and vehicles. We address this gap by introducing a framework for object level evaluation that establishes strict 3D-to-2D correspondences between original and novel rendered views which are then used to compute a semantic similarity metric. Additionally, we release a dataset for novel view synthesis quality evaluation based on the Physical AI NuRec dataset. The dataset contains pre-reconstructed 3D driving scenes, novel rendered views, corresponding 3D semantic bounding boxes and subjective human quality ratings. Experiments show that our Novel View Reconstruction Quality (NVRQ) metric consistently outperforms both classical image quality metrics and recent Novel View Synthesis (NVS)-specific baselines, achieving the strongest correlation with downstream task degradation and human judgments.

1. Introduction

The evolution of 3D neural reconstruction – from explicit geometric primitives to implicit representations like Neural Radiance Fields (NeRF) [23] and 3D Gaussian Splats (3DGS) [17] – has made real-time, photorealistic novel view synthesis (NVS) a reality. A critical, high-stakes application for this technology is the development of autonomous vehicles (AVs). The industry has recently shifted toward integrated end-to-end policy models [28, 37], which require massive amounts of driving data for effective training. Because collecting rare, safety-critical edge cases in the real world is fundamentally limited, NVS has emerged as a highly promising solution. It facilitates the scalable conversion of sensor data into interactive real-world replicas, bridging the persistent "reality gap" of conventional simulators and enabling the synthesis of novel driving

*Equal contribution.

scenes from varied viewpoints [3, 5, 11, 27, 37].

However, as these NVS technologies transition from academic benchmarks to safety-critical domains, methodologies for evaluating their reconstruction quality have reached a significant bottleneck [8, 32, 53]. Current image quality assessment (IQA) for NVS largely relies on scene-level comparisons. This approach can be highly misleading in dynamic driving environments, where scene-level metrics dilute or completely mask localized artifacts on safety-critical actors – such as pedestrians, cyclists, and vehicles. Furthermore, evaluating novel views rendered from large spatial offsets is inherently treated as a non-reference problem, complicating accurate and reliable quality assessment.

In this work, we address these challenges by formulating a reference-based, object-level evaluation pipeline tailored for autonomous driving scenarios. By establishing strict spatial correspondences across views, we transform the historically ill-posed non-reference evaluation of extrapolated poses into a mathematically grounded, reference-based problem. Rather than relying on scene-level metrics that obscure localized errors, our approach isolates and evaluates the rendering quality of critical dynamic actors.

In summary, our main contributions are as follows:

- We propose a reference-based, object-level image quality assessment pipeline for evaluating NVS reconstructions, explicitly focusing on dynamic actors (e.g., cars, pedestrians, bicycles) that are critical for driving scenes but frequently overlooked by traditional scene-level metrics.
- We introduce a comprehensive, real-world driving dataset featuring extensive metadata, paired ground truth, and rendered novel views to establish fine-grained, object-level NVS image quality benchmarks.
- We demonstrate that our Novel View Reconstruction Quality (*NVRQ*) assessment pipeline, which leverages semantic features from foundation models like DINOv2, effectively isolates reconstruction artifacts. This is validated through both objective downstream tasks (e.g., object detection and classification scores) and rigorous human subjective evaluations.

2. Related Works

Novel View Synthesis Quality Evaluation: Image and video quality assessment spans reference-based and non-reference methods [20, 59]. While non-reference metrics are widely used for in-the-wild distortions [16, 52] and AI-generated content [18, 57, 58], lacking ground-truth anchors limits their semantic fidelity assessment. In novel view synthesis (NVS), alternative viewpoint ground truth is inherently available but frequently underutilized [32, 42]. Existing NVS evaluations predominantly rely on traditional full-reference metrics (e.g., PSNR [15], SSIM [47], LPIPS [54]) or blind statistical evaluators (e.g., BRISQUE [24], PIQE [45], NIQE [25]). Regardless of type, these scene-level metrics struggle to capture localized artifacts on safety-critical dynamic actors. To address NVS-specific flaws, specialized non-reference metrics like NeRF-NQA [32], NVS-SQA [33], and NIQSV+ [42] have emerged; however, because they lack a ground-truth anchor, they inherently evaluate perceptual plausibility rather than measuring true reconstruction fidelity. Consequently, for assessing exact scene reproduction, perceptual reference metrics provide a more reliable measure of semantic preservation, with DreamSim [9] outperforming classical metrics [48]. Closest to our work is NOVA [12], which computes semantic feature distances between non-aligned views using a finetuned DINOv2 backbone. However, NOVA performs strictly scene-level evaluation, diluting actor-specific degradation. Our pipeline projects 3D bounding boxes to establish strict 2D correspondences, transforming extrapolation into a mathematically grounded, object-level evaluation for dynamic objects.

Novel View Synthesis Evaluation Datasets: Early NVS quality datasets [19, 21, 22, 29, 40, 51] primarily focused on static NeRF reconstructions, featuring limited real-world viewpoints and lacking dynamic actors. While recent efforts like Zhang *et al*. [56] introduced subjective datasets for dynamic scenes, they continue to rely on conventional IQA/VQA metrics rather than NVS-specific evaluations. A critical drawback of existing benchmarks is the absence of extreme spatial view shifts (e.g., lateral lane changes), which are paramount for autonomous driving simulation. Furthermore, existing datasets evaluate quality strictly at the scene level and underrepresent diverse conditions (urban, night, highway). To address these gaps, we introduce a dataset specifically designed for large lateral view shifts, object-centric assessment of dynamic actors, and real-world driving diversity.

3. Methods

3.1. 3D-to-2D Projection in Novel Views

To compute object-level quality metrics, we project 3D semantic bounding boxes from world space into both the 2D ground-truth view and the synthesized view (Figure 2), so the same object is aligned across views.

The global pose of a camera sensor is obtained by composing two rigid-body transforms in $SE(3)$. The vehicle body defines a fixed reference frame called the *rig frame*, to which all onboard sensors are calibrated. The transform $\mathbf{T}_{\text{sensor}\to\text{rig}} \in SE(3)$ encodes the camera's static mounting position and orientation relative to the rig frame, while $\mathbf{T}_{\text{rig}\to\text{world}} \in SE(3)$ captures the vehicle's time-varying global pose. Their composition $\mathbf{T}_{\text{rig}\to\text{world}}\,\mathbf{T}_{\text{sensor}\to\text{rig}}$ yields the sensor's pose in the world frame.

To render novel views, we left-multiply the sensor-to-rig mounting transform by a pure lateral translation

Figure 2. Overall framework for computing the Novel View Reconstruction Quality (*NVRQ*). We first establish strict spatial correspondence by projecting 3D bounding boxes into both the original ground truth view and the synthesized novel view. The resulting 2D crops are processed by a shared image encoder to extract semantic embeddings. The base *NVRQ* is computed via cosine similarity between these embeddings. The more robust *NVRQ-CA* metric is computed via a non-linear curve adjustment to expand the discriminative range.

$\mathbf{T}_{\text{offset}} \in \text{SE}(3)$, defined with identity rotation and translation $[0,\, y_{\text{offset}},\, 0]^T$:

$$\mathbf{T}_{\text{sensor}'\to\text{rig}} = \mathbf{T}_{\text{offset}}\, \mathbf{T}_{\text{sensor}\to\text{rig}} \tag{1}$$

This effectively defines a *virtual camera* frame (sensor′) that is laterally displaced by y_{offset} meters from the physical sensor in the rig frame, while retaining the original orientation. The world-to-sensor mapping used to render the novel view from this virtual camera is obtained by inverting the composed transform:

$$\mathbf{T}_{\text{world}\to\text{sensor}'} = \left(\mathbf{T}_{\text{rig}\to\text{world}}\, \mathbf{T}_{\text{sensor}'\to\text{rig}}\right)^{-1}. \tag{2}$$

Setting $y_{\text{offset}} = 0$ reduces $\mathbf{T}_{\text{offset}}$ to the identity, recovering the original ground-truth camera pose ($\mathbf{T}_{\text{world}\to\text{sensor}}$).

The Physical AI NuRec dataset [26] annotates each object with an oriented 3D bounding box and a tracked pose in world coordinates. We project its eight corners onto the image plane of the virtual camera via

$$\mathbf{p}^{(i)}_{\text{img}} = \pi\big(\mathbf{T}_{\text{world}\to\text{sensor}'}\, \mathbf{P}^{(i)}_{\text{world}};\ \mathbf{K},\, \mathcal{D}\big), \tag{3}$$

where $\pi(\cdot)$ applies the intrinsic calibration $\mathbf{K}$ and distortion model $\mathcal{D}$ [14, 49]. The axis-aligned 2D bounding box enclosing the projected corners defines the object crop in the rendered image.

3.2. Bounding Box Extraction

The projection yields eight 2D image coordinates $\mathbf{p}_i = (p_i^x, p_i^y)$ for $i \in \{1, \ldots, 8\}$. To obtain 4-parameter bounding box coordinates, we take the spatial extrema across all points and round outward to ensure that the discrete pixel window fully encapsulates the object $B_p = \{\lfloor \min_i p_i^x \rfloor, \lfloor \min_i p_i^y \rfloor, \lceil \max_i p_i^x \rceil, \lceil \max_i p_i^y \rceil\}$

Before extracting image crops for metric evaluation, we apply strict geometric filtering to ensure semantic integrity. First, we enforce an *8-corner visibility criterion*: a bounding box is only retained if all eight $\mathbf{p}_i$'s fall strictly within the 2D image boundaries. Second, we apply a *size criterion*: any bounding box whose 2D projected area falls below a user-defined minimum threshold is discarded, preventing the evaluation of objects that are too distant or degraded by resolution limits.

Crucially, for any given object, we require that the bounding box passes these filtering criteria in *both* the ground truth view and the rendered novel view. The valid paired crops, framing the same physical object from different viewpoints (see Figure 2), are then forwarded to the object-level metric pipeline.

3.3. Object-Level Metric Computation

For valid crop pairs, we extract robust semantic representations using a pre-trained Vision Transformer [6] backbone such as DINOv2 [30]. We obtain a single 1D feature embedding from the backbone's final layer by applying Global Average Pooling (GAP). Self-supervised models like DINOv2 encode features robust to scale and viewpoint changes, and applying GAP further collapses spatial dimensions, making the representation invariant to 2D shifts and bounding-box jitter; together, this ensures the metric

captures actual reconstruction errors rather than penalizing pixel-level misalignment.

To quantify semantic fidelity, we compute the cosine similarity between the pooled feature vectors, rescaled to $[0, 1]$. We call this the Novel View Reconstruction Quality (*NVRQ*) metric, denoted as $\mathcal{S}$:

$$\mathcal{S} = \frac{\cos(\mathbf{f}_{\text{gt}}, \mathbf{f}_{\text{rendered}}) + 1}{2} \tag{4}$$

where $\mathcal{S} \in [0, 1]$, and $\mathbf{f}_{\text{gt}}$ and $\mathbf{f}_{\text{rendered}}$ are the pooled feature embeddings of the ground-truth and synthesized crops, respectively.

Curve-Adjusted Metric (***NVRQ-CA***): While mathematically sound, the raw $\mathcal{S}$ score suffers from a narrow, encoder-dependent discrimination range. For instance, using DINOv2, completely unrelated object pairs typically yield a similarity score around 0.65. This value effectively acts as a noise floor, meaning the entire meaningful quality range is compressed into just 35% of the $[0, 1]$ spectrum. Furthermore, directly comparing raw scores across different encoders is misleading due to variations in their respective noise floors.

We propose a custom adjustment function to mitigate this limitation and standardize the metric. Let τ_{base} represent the empirically determined baseline threshold, defined as the noise boundary for unrelated objects, and w_{low} denote a scaling weight assigned to the below-baseline range. The *curve adjusted* score, which we coin *NVRQ-CA*, is computed via a continuous, piecewise mapping from the raw score $\mathcal{S}$:

$$\mathcal{S}_{\text{CA}} = \begin{cases} \left(\frac{\mathcal{S}}{\tau_{\text{base}}}\right) w_{\text{low}}, & \text{if } \mathcal{S} < \tau_{\text{base}} \\ w_{\text{low}} + (1 - w_{\text{low}}) \left(\frac{\mathcal{S} - \tau_{\text{base}}}{1 - \tau_{\text{base}}}\right)^2, & \text{if } \mathcal{S} \geq \tau_{\text{base}} \end{cases} \tag{5}$$

When $\mathcal{S} < \tau_{\text{base}}$, the function applies a simple linear scaling that aggressively compresses the noisy regime. For example, if we set $\tau_{\text{base}} = 0.65$ and $w_{\text{low}} = 0.1$, then the bottom 65% of the raw DINOv2 similarity scale is mapped into just 10% of the final adjusted metric. This correctly reflects that scores in this regime carry minimal discriminative information; they simply represent varying degrees of "noise". Conversely, when $\mathcal{S} \geq \tau_{\text{base}}$, the quadratic expansion non-linearly stretches the meaningful similarity range, ensuring that subtle semantic differences in high-fidelity renderings are heavily amplified for evaluation. At the boundary condition $\mathcal{S} = \tau_{\text{base}}$, both branches evaluate strictly to w_{low}, ensuring the adjustment function remains continuous across the entire domain.

4. Dataset

We utilize the public Physical AI NuRec dataset [49] as the foundation for our novel view synthesis evaluations. This dataset provides 3D semantic bounding box annotation and pre-reconstructed 3D driving scenes generated using 3DGUT [49], an extension of 3DGS [17]. 3DGUT [49] provides a mathematically robust unscented transform projection that natively handles complex optics such as distorted fisheye lenses and windshield refractions, making the explicitly bounded 3D Gaussian parameterization highly effective for real-world autonomous driving data.

The learned scene is represented as an unstructured volumetric cloud of 3D Gaussians [17]. Because the NuRec dataset provides these fully reconstructed 3D worlds, we do not need to train the scenes ourselves. Instead, we directly render novel views by perturbing the original sensor poses along controlled spatial degrees of freedom, as shown in Figure 2 and explained in Section 3.1.

Human Subjective Rating: We conducted a psychophysical experiment with 5 human raters for subjective quality assessment. Since rating all crops across offsets, objects, and frames is infeasible, we systematically select 12 representative object tracks spanning 5 classes (automobile, person, bus, heavy truck, rider) from diverse scenes, subsampled to an average of 12 frames per track, yielding 1,002 crops across seven lateral offsets (-3m, -2m, -1m, 0m, 1m, 2m, 3m).

We show the ground-truth and rendered versions of each crop side by side and ask raters to provide a binary label: "good" or "bad". Since each crop is rated by up to 5 raters, we compute the Good Vote Fraction, $\text{GVF} = n_{\text{good}}/n_{\text{raters}}$, representing the proportion of raters who judged the crop as visually acceptable.

5. Experiments

5.1. Implementation Details

Preprocessing Details: First, we discard any crop whose shorter side is below 30 pixels to avoid low-resolution, noisy regions. To prepare crops for metric evaluation, each rectangular crop is padded with black pixels to form a square, preserving the aspect ratio, and then resized via bicubic interpolation. The crops are resized to 224×224 pixels for our *NVRQ* variants that use DINOv2 [30], DINOv2-Reg [4], CLIP [35] and SigLIP [43], while for SegFormer [50] we resize to its required input size of 1024×1024. For reference baselines, DreamSim [9] and HyperIQA [39] process 224×224 inputs, while NOVA [12] uses 518×518. Other metrics, such as PSNR [15], SSIM [47], LPIPS [54], and MUSIQ [16]), impose no fixed resolution and operate directly on the raw rectangular crops that pass the 30-pixel minimum side-length filter.

***NVRQ-CA* Hyperparameter setting:** We compute the

Table 1. Quality metrics correlation across evaluation strategies. PLCC and SRCC are shown as absolute values ($|r|$); ↑ = higher is better.

	Obj. Detection Confidence		**ImageNet ΔConfidence**		**CLIP Zero-Shot ΔConfidence**		**Human Subjective Rating**		
Metric	|PLCC| ↑	|SRCC| ↑	|PLCC| ↑	|SRCC| ↑	|PLCC| ↑	|SRCC| ↑	|PLCC| ↑	|SRCC| ↑	ROC-AUC ↑
Non-Reference									
BRISQUE [24]	0.241	0.259	0.042	0.034	0.116	0.132	0.265	0.230	0.625
NIQE [25]	0.395	0.438	0.090	0.080	0.168	0.191	0.255	0.241	0.663
NIQSV+ [42]	0.306	0.292	0.055	0.036	0.168	0.185	0.078	0.026	0.542
Entropy [36]	0.322	0.285	0.052	0.029	0.156	0.169	0.101	0.143	0.549
MUSIQ [16]	0.369	0.369	0.101	0.096	0.169	0.193	0.212	0.181	0.619
HyperIQA [39]	0.178	0.115	0.056	0.042	0.074	0.007	0.258	0.296	0.621
CLIP-IQA+ [46]	0.088	0.076	0.053	0.052	0.011	0.019	0.134	0.127	0.567
TOPIQ-NR [2]	0.048	0.068	0.003	0.004	0.003	0.001	0.133	0.095	0.548
DBCNN [55]	0.029	0.028	0.016	0.009	0.031	0.023	0.003	0.089	0.539
NVS-SQA [33]	0.070	0.168	0.019	0.047	0.041	0.082	0.421	0.495	0.762
Reference									
PSNR [15]	0.084	0.097	0.110	0.136	0.049	0.032	0.525	0.529	0.751
SSIM [47]	0.171	0.178	0.126	0.149	0.106	0.097	0.547	0.516	0.762
LPIPS [54]	0.057	0.117	0.112	0.114	0.062	0.008	0.267	0.244	0.590
DreamSim [9]	0.239	0.198	0.236	0.251	**0.310**	0.265	0.574	0.589	0.799
NOVA [12]	0.224	0.276	0.195	0.222	0.186	0.194	0.437	0.470	0.704
NVRQ (Ours)	0.542	0.597	0.307	0.325	0.291	0.290	0.642	0.706	0.884
NVRQ-CA (Ours)	**0.574**	**0.597**	**0.316**	**0.325**	0.291	**0.290**	**0.675**	**0.706**	**0.884**

curve-adjusted score based on an encoder-specific baseline threshold, τ_{base} that reflects each model's innate noise floor. We empirically chose the following values for different encoder variants: $\tau_{\text{base}}^{\text{DINOv2}} = \tau_{\text{base}}^{\text{DINOv2-Reg}} = 0.65$, $\tau_{\text{base}}^{\text{CLIP}} = 0.70$, $\tau_{\text{base}}^{\text{SigLIP}} = 0.55$, $\tau_{\text{base}}^{\text{SegFormer}} = 0.50$. We set $w_{\text{low}} = 0.1$ for all the models.

Metrics Computation Framework: To ensure standardized and reproducible evaluations, we utilized the open-source IQA-PyTorch (PyIQA) toolbox [1] to compute the majority of the reference and non-reference quality metrics used as baselines. For any remaining metric not implemented within PyIQA, we utilized the method's respective official GitHub repository [10, 13, 34, 41].

5.2. Evaluation Strategy

To validate the effectiveness of each metric, we evaluate them on scenes rendered at seven distinct lateral offsets and correlate their values with the respective downstream object detection and classification performance and with human subjective evaluations on a representative sample set.

Object Detection: We run YOLOv8 [44] with three-scale features to extract bounding boxes and class confidence scores from rendered frames. Only predictions with class confidence ≥ 0.1 are retained, and per-class non-maximum suppression with an IoU threshold of 0.45 is applied to obtain the final detection boxes. Our dataset already contains projected object bounding boxes with class information, but the classes are not mapped to COCO categories.We map each class to its corresponding COCO category, where generic labels such as *automobile* are associated with multiple specific COCO classes like *car*, *truck*, and *bus*. Given a rendered frame I_{rend}, YOLOv8 produces a set of N detections:

$$\mathcal{D} = \{(b_k, s_k, y_k)\}_{k=1}^{N}, \quad s_k \geq 0.1, \tag{6}$$

where b_k is the bounding box, $s_k = \max_c \sigma(z_{k,c})$ is the class confidence score, and $y_k = \arg\max_c \sigma(z_{k,c})$ is the predicted class for the k-th detection.

We first define the subset of class-valid detection indices as $K_v = \{k \mid y_k \in \mathcal{C}_{\text{valid}}\}$. For each projected bounding box B_p, we define the best matching detection index k^* using a piecewise fallback mechanism:

$$k^* = \begin{cases} \arg\max\limits_{k \in K_v} \text{IoU}(B_p, b_k) & \text{if } \max\limits_{k \in K_v} \text{IoU}(B_p, b_k) \geq 0.3, \\ \arg\max\limits_{k} \text{IoU}(B_p, b_k) & \text{otherwise.} \end{cases} \tag{7}$$

The detection confidence is then assigned based on this optimal match, enforcing the strict Intersection over Union (IoU) threshold to rule out misses:

$$s_{\text{det}}(B_p) = \begin{cases} s_{k^*} & \text{if IoU}(B_p, b_{k^*}) \geq 0.3, \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

We then compute the Pearson linear correlation coefficient (PLCC) [31] and the Spearman rank correlation coefficient (SRCC) [38] between s_{det} and each of the metrics, across all projected bounding boxes.

ImageNet ΔConfidence: Since we already have projected ground truth bounding boxes, we can directly extract the crops from ground truth and rendered views and classify

the objects. The confidence drop in classification can serve as a proxy for measuring the degradation of rendering quality. Let $\mathbf{x}^{\text{gt}}$ and $\mathbf{x}^{\text{r}}$ denote the ground truth and rendered crop, respectively, and let $f(\mathbf{x}, c)$ denote the softmax confidence for class c. We first classify the ground truth crop using an ImageNet-pretrained model f_{IN} to obtain $\hat{c} = \arg\max_c f_{\text{IN}}(\mathbf{x}^{\text{gt}}, c)$, then compute the confidence drop:

$$\Delta p_{\text{IN}} = f_{\text{IN}}(\mathbf{x}^{\text{gt}}, \hat{c}) - f_{\text{IN}}(\mathbf{x}^{\text{r}}, \hat{c}) \tag{9}$$

A higher Δp_{IN} indicates greater semantic degradation, as poor reconstruction weakens the features the classifier relies on. We use EVA-02 Base [7] as f_{IN}. We rely on the predicted ImageNet class $\hat{c}$ from ground truth crops because our dataset classes do not map exclusively to ImageNet categories. Although this gives a good estimate of relative performance, a drawback is that confidence is estimated over 1K classes, of which only a few are relevant to our classes, potentially diluting the estimates.

CLIP Zero-Shot ΔConfidence: To address this, we employ zero-shot CLIP confidence with domain-specific prompts. We merge all classes into $K = 5$ categories (automobile, person, animal, large vehicle, and rail vehicle) and encode each with a descriptive text prompt $\mathbf{t}_j$ (*e.g.*, for automobile: *"a photo of a car, sedan, SUV, pickup truck, minivan, or hatchback on the road"*). The zero-shot confidence is computed as the softmax over cosine similarities between the CLIP image embedding and all K text embeddings $\{\mathbf{t}_j\}_{j=1}^{K}$. Since the ground truth class c^* is known from the ground truth annotations, the confidence drop is:

$$\Delta p_{\text{CLIP}} = f_{\text{CLIP}}(\mathbf{x}^{\text{gt}}, c^*) - f_{\text{CLIP}}(\mathbf{x}^{\text{r}}, c^*) \tag{10}$$

Unlike Eq. 9, which relies on the predicted class $\hat{c}$ (which may itself be incorrect), Eq. 10 uses the actual ground truth class c^* for both crops. If the rendered crop yields higher confidence, $\Delta p_{\text{CLIP}} < 0$, effectively rewarding faithful reconstructions. This makes CLIP-based evaluation fairer to the non-reference metrics, as compared the ImageNet-based approach that directly penalizes deviation from ground truth class.

Human Subjective Evaluation: We used the collected human subjective ratings (Section 4) to compute PLCC and SRCC between each quality metric and the GVF. Additionally, we compute ROC-AUC by treating the task as binary classification: human labels are binarized by majority vote (GVF $\geq 0.5 \Rightarrow$ good), and each metric score is used as the confidence of the classifier, negating the sign for metrics where lower values indicate better quality so that higher scores consistently predict the positive (good) class. The ROC curve is obtained by scanning all metric thresholds, and we report the area under this curve (AUC).

5.3. Comparison Results

After per-frame 2D crop extraction following Sections 3.1 and 3.2, we compared *NVRQ-CA* against 10 non-reference and five reference metrics. For this, we assess correlation with downstream object detection and classification performance and human subjective ratings. Non-reference metrics are computed directly on the rendered crops, without using the ground truth reference, whereas reference metrics use the ground truth reference and compare it with the corresponding rendered crop to compute the metrics. Table 1 shows the overall results, comparing *NVRQ-CA* against all other metrics. *NVRQ-CA* consistently shows higher correlation with both downstream task and human evaluation. Figure 3 highlights that *NVRQ-CA* follows a similar trend to human subjective ratings across seven lateral offsets.

6. Additional Studies

Table 2. Correlation of *NVRQ-CA* against CLIP Zero-shot ΔConfidence. PLCC and SRCC are shown as absolute values ($|r|$); ↑ = higher is better. n = crop-level observations; k = clips. Results computed on all 919 clips (28.4M crop-level rows).

Category	Value	PLCC	SRCC	n	k
Overall		0.291	0.290	28.4M	919
Lighting	Daytime	**0.293**	**0.291**	25.2M	826
	Nighttime	0.234	0.223	275K	16
	Unspecified	0.253	0.245	1.25M	44
Road Type	Highways	**0.316**	**0.318**	9.21M	315
	Residential	0.272	0.275	9.11M	315
	Rural	0.256	0.261	474K	37
	Urban	0.293	0.285	8.35M	234
	Unspecified	0.283	0.282	1.88M	57
Object Class	Automobile	0.316	0.317	25.0M	918
	Large Vehicle	0.160	0.122	2.35M	602
	Person	0.049	0.016	1.08M	441
	Rail Vehicle	**0.452**	**0.331**	11.5K	7
	Animal	0.246	0.252	3.0K	12

Category Level Evaluation: While overall correlation indicates general reliability, assessment efficacy naturally varies across categories due to object scale, scene complexity, and specific artifact types. For instance, smaller objects like pedestrians yield fine-grained artifacts that are inherently harder to quantify than those of vehicles. To analyze this, we decompose the correlation of *NVRQ-CA* with CLIP Zero-shot Δconfidence across multi-label category assignments. In Table 2, daytime scenes correlate better than nighttime images due to clearer boundaries. Highways outperform dense urban roads, likely benefiting from reduced clutter and occlusion. Similarly, larger actors (rail vehicles, automobiles) demonstrate stronger correlation than smaller, complex classes (persons). These findings highlight a critical takeaway: metric efficacy is non-uniform. When benchmarking novel view synthesis, researchers must account for category-specific performance variations rather than relying solely on aggregate scores.

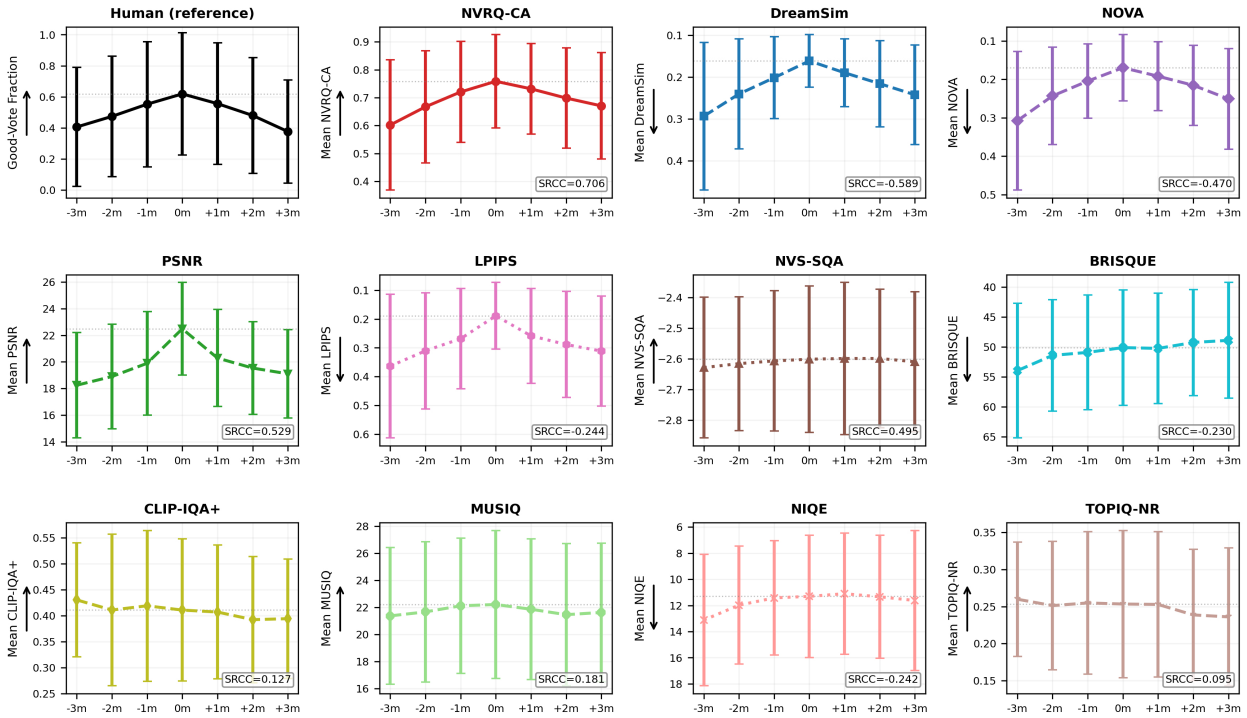

Figure 3. **Qualitative trend analysis of objective metrics versus human subjective ratings across lateral camera offsets**. The top-left panel (solid black line) shows the human reference baseline (Good-Vote Fraction), which exhibits a clear "tent" shape: highest perceived quality at the 0 m offset (original viewpoint), degrading symmetrically as the camera is shifted laterally to ± 3 m. A robust NVS quality metric should mirror this symmetrical degradation profile. Our proposed Curve-Adjusted *NVRQ* (*NVRQ-CA*) closely tracks human perceptual behavior and achieves the highest Spearman Rank Correlation (SRCC = 0.706) among all evaluated metrics. Other metrics such as DreamSim, PSNR, and NVS-SQA also capture the spatial degradation trend, though with lower correlation. Error bars denote ± 1 standard deviation across crops at each offset. A downward arrow (↓) indicates an inverted y-axis, applied so that opposite-polarity metrics share the same visual orientation as the human reference.

Impact of Encoder Backbone: We experimented with five different backbones: DINOv2 [30], DINOv2-Reg [4] (DinoV2 with registers), SegFormer [50], CLIP [35] and SigLIP [43] to assess the impact of different feature representations. Table 3 shows that DINOv2 produced the highest correlation with downstream tasks and subjective human ratings, with the DINOv2-Reg variant performing the best.

Visualization Results Figure 4 highlights the robustness of *NVRQ-CA* to occlusion and viewpoint shift, showing stronger correlation with human perception than pixel-based methods. This underscores why such methods fail at novel-view synthesis evaluation, where ground-truth comparisons inherently involve view changes and occlusions.

7. Conclusion

We presented a framework for object-level novel-view synthesis (NVS) evaluation that establishes strict 3D-to-2D spatial correspondences, transforming NVS assessment from an ill-posed scene-level problem into a mathematically grounded object-level one. By leveraging semantic features from foundation models such as DINOv2, *NVRQ* isolates true reconstruction artifacts from geometric misalignment. We further introduced *NVRQ-CA*, a scoring function that standardizes evaluations across encoders by compressing baseline noise and amplifying meaningful quality variations. To facilitate rigorous benchmarking, we built an evaluation benchmark upon the Physical AI NuRec dataset, providing paired ground-truth and novel-view crops across diverse lateral offsets, object classes, and environmental conditions. Extensive experiments demonstrate that *NVRQ-CA* consistently outperforms both full-reference and no-reference baselines, achieving superior correlation with downstream task performance and subjective human judgments.

A limitation of our current pipeline is its reliance on accurate 3D bounding box annotations, which may not be universally available in unconstrained NVS settings. Additionally, while our dataset emphasizes critical lateral view shifts, expanding the evaluation to encompass complex vertical and rotational camera perturbations remains a promis-

Table 3. Comparison of encoder backbones used in *NVRQ*. PLCC and SRCC are reported as absolute values ($|r|$); ↑ indicates higher is better. The curve-adjusted *CA* variant yields higher PLCC, while SRCC and ROC-AUC are largely unchanged, as they are rank-based and unaffected by monotonic curve adjustment.

	Obj. Detection Confidence		**ImageNet ΔConfidence**		**CLIP Zero-Shot ΔConfidence**		**Human Subjective Rating**		
Encoder Variant	|PLCC| ↑	|SRCC| ↑	|PLCC| ↑	|SRCC| ↑	|PLCC| ↑	|SRCC| ↑	|PLCC| ↑	|SRCC| ↑	ROC-AUC ↑
NVRQ-SegFormer[50]	0.080	0.045	0.151	0.157	0.173	0.127	0.299	0.245	0.609
NVRQ-CLIP [35]	0.040	0.039	0.175	0.173	0.193	0.115	0.281	0.257	0.606
NVRQ-SigLIP [43]	0.020	0.080	0.159	0.159	0.201	0.144	0.369	0.329	0.657
NVRQ-DINOv2 [30]	0.494	0.552	0.308	0.328	0.282	0.272	0.572	0.667	0.847
NVRQ-DINOv2-Reg [4]	0.542	0.597	0.307	0.325	0.291	0.290	0.642	0.706	0.884
NVRQ-CA-SegFormer [50]	0.050	0.045	0.141	0.157	0.156	0.127	0.275	0.245	0.609
NVRQ-CA-CLIP [35]	0.021	0.039	0.170	0.173	0.182	0.115	0.279	0.257	0.606
NVRQ-CA-SigLIP [43]	0.034	0.080	0.154	0.159	0.197	0.144	0.368	0.329	0.657
NVRQ-CA-DINOv2 [30]	0.527	0.552	**0.318**	**0.328**	0.282	0.272	0.603	0.667	0.847
NVRQ-CA-DINOv2-Reg [4]	**0.574**	**0.597**	0.316	0.325	**0.291**	**0.290**	**0.675**	**0.706**	**0.884**

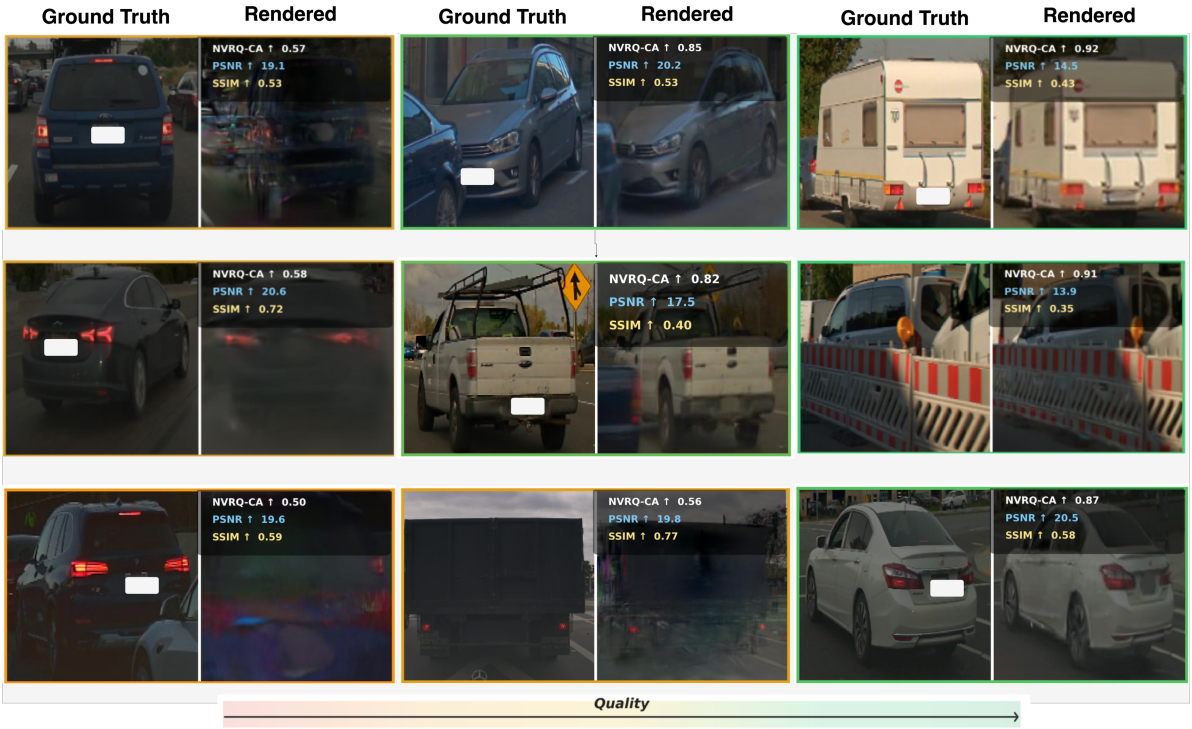

Figure 4. Comparison of *NVRQ-CA* (semantic) against pixel-based metrics (PSNR, SSIM). **Top (view shift)**: PSNR and SSIM drop when the rendered view is shifted relative to the reference, even when rendering quality is good. The right image has good semantic structure and looks visually strong but scores low under PSNR/SSIM because of this view shift. **Middle (occlusion)**: *NVRQ-CA* is tolerant to moderate occlusion. The center and right images contain occlusion but are visually preferable to the left; *NVRQ-CA* reflects this, while PSNR and SSIM are heavily penalized by the occluded regions. **Bottom (semantic)**: *NVRQ-CA* prioritizes semantic fidelity over pixel-level match. The first two images are visually poor but receive relatively high PSNR/SSIM; *NVRQ-CA* correctly assigns them lower scores by focusing on semantics.

ing direction for future work. Ultimately, we hope *NVRQ* and its accompanying benchmark will serve as a foundation for rigorous, object-centric quality assessment as NVS technologies are increasingly deployed in safety-critical applications.

References

[1] Chaofeng Chen and Jiadi Mo. IQA-PyTorch: Pytorch toolbox for image quality assessment. [Online]. Available: `https://github.com/chaofengc/IQA-PyTorch`, 2022.

[2] Chaofeng Chen, Jiadi Mo, Jingwen Hou, Haoning Wu, Liang Liao, Wenxiu Sun, Qiong Yan, and Weisi Lin. Topiq: A top-down approach from semantics to distortions for image quality assessment. *IEEE Transactions on Image Processing*, 33:2404–2418, 2024.

[3] Matt Cragun. Reconstructing the Real World in DRIVE Sim With AI. https://blogs.nvidia.com/blog/drive-sim-neural-reconstruction-engine/, 2022. Accessed: 2026-01-30.

[4] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *The Twelfth International Conference on Learning Representations*, 2024.

[5] Daniel Dauner, Marcel Hallgarten, Tianyu Li, Xinshuo Weng, Zhiyu Huang, Zetong Yang, Hongyang Li, Igor Gilitschenski, Boris Ivanovic, Marco Pavone, Andreas Geiger, and Kashyap Chitta. NAVSIM: Data-driven non-reactive autonomous vehicle simulation and benchmarking. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

[7] Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. EVA-02: A visual representation for neon genesis. *Image and Vision Computing*, 2024.

[8] Alejandro Fontan, Javier Civera, Tobias Fischer, and Michael Milford. Look ma, no ground truth! ground-truth-free tuning of structure from motion and visual slam. *arXiv preprint arXiv:2412.01116*, 2024.

[9] 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. *arXiv preprint arXiv:2306.09344*, 2023.

[10] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. DreamSim official code repository. https://github.com/ssundaram21/dreamsim, 2023.

[11] Junhao Ge, Zuhong Liu, Longteng Fan, Yifan Jiang, Jiaqi Su, Yiming Li, Zhejun Zhang, and Siheng Chen. Unraveling the effects of synthetic data on end-to-end autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

[12] Abhijay Ghildyal, Rajesh Sureddi, Nabajeet Barman, Saman Zadtootaghaj, and Alan Bovik. Non-aligned reference image quality assessment for novel view synthesis. *arXiv preprint arXiv:2511.08155*, 2025.

[13] Abhijay Ghildyal, Rajesh Sureddi, Nabajeet Barman, Saman Zadtootaghaj, and Alan Bovik. NOVA official project and code repository. https://stootaghaj.github.io/nova-project/, 2025.

[14] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[15] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pages 2366–2369. IEEE, 2010.

[16] 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*, pages 5148–5157, 2021.

[17] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

[18] Chunyi Li, Zicheng Zhang, Haoning Wu, Wei Sun, Xiongkuo Min, Xiaohong Liu, Guangtao Zhai, and Weisi Lin. Agiqa-3k: An open database for ai-generated image quality assessment. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(8):6833–6846, 2023.

[19] Hanxue Liang, Tianhao Wu, Param Hanji, Francesco Banterle, Hongyun Gao, Rafal Mantiuk, and Cengiz Öztireli. Perceptual quality assessment of nerf and neural view synthesis methods for front-facing views. In *Computer Graphics Forum*, page e15036. Wiley Online Library, 2024.

[20] Chengqian Ma, Zhengyi Shi, Zhiqiang Lu, Shenghao Xie, Fei Chao, and Yao Sui. A survey on image quality assessment: Insights, analysis, and future outlook. *arXiv preprint arXiv:2502.08540*, 2025.

[21] Pedro Martin, António Rodrigues, João Ascenso, and Maria Paula Queluz. Nerf view synthesis: Subjective quality assessment and objective metrics evaluation. *IEEE Access*, 2024.

[22] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7210–7219, 2021.

[23] 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. *Communications of the ACM*, 65(1):99–106, 2021.

[24] 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.

[25] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012.

[26] NVIDIA Corporation. Physicalai-autonomous-vehicles-nurec dataset. https://huggingface.co/datasets/nvidia/PhysicalAI-Autonomous-Vehicles-NuRec, 2025.

[27] NVIDIA Corporation. *NVIDIA Omniverse NuRec*. NVIDIA Developer Documentation, 2025. Accessed: 2026-01-30.

[28] NVIDIA Corporation. NVIDIA DRIVE Sim Archives. https://blogs.nvidia.com/blog/tag/nvidia-drive-sim/, n.d. Accessed: 2026-01-30.

[29] Chibuike Onuoha, Jean Atsumi Flaherty, Shihao Luo, Truong Thu Huong, and Truong Cong Thang. An evaluation of quality metrics for neural radiance field. In *2023*

IEEE 15th International Conference on Computational Intelligence and Communication Networks (CICN), pages 619–623. IEEE, 2023.

[30] 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. *arXiv preprint arXiv:2304.07193*, 2023.

[31] Karl Pearson. Mathematical contributions to the theory of evolution. iii. regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London. Series A*, 187:253–318, 1896.

[32] Qiang Qu, Hanxue Liang, Xiaoming Chen, Yuk Ying Chung, and Yiran Shen. Nerf-nqa: No-reference quality assessment for scenes generated by nerf and neural view synthesis methods. *IEEE Transactions on Visualization and Computer Graphics*, 30(5):2129–2139, 2024.

[33] Qiang Qu, Yiran Shen, Xiaoming Chen, Yuk Ying Chung, Weidong Cai, and Tongliang Liu. Nvs-sqa: Exploring self-supervised quality representation learning for neurally synthesized scenes without references. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[34] Qiang Qu, Yiran Shen, Xiaoming Chen, Yuk Ying Chung, Weidong Cai, and Tongliang Liu. NVS-SQA official code repository. `https://github.com/VincentQQu/NVS-SQA`, 2025.

[35] 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 *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[36] Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

[37] Gautham Sholingar and Katie Washabaugh. Accelerating AV simulation with neural reconstruction and world foundation models. `https://developer.nvidia.com/blog/accelerating-av-simulation-with-neural-reconstruction-and-world-foundation-models/`, 2025. Accessed: 2026-01-30.

[38] Charles Spearman. The proof and measurement of association between two things. *The American journal of psychology*, 15(1):72–101, 1904.

[39] Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3667–3676, 2020.

[40] Shaira Tabassum and Seyed Ali Amirshahi. Quality of nerf changes with the viewing path an observer takes: A subjective quality assessment of real-time nerf model. In *2024 16th International Conference on Quality of Multimedia Experience (QoMEX)*, pages 88–91. IEEE, 2024.

[41] Shishun Tian, Lu Zhang, Luce Morin, and Olivier Déforges. NIQSV and NIQSV+ official code repository. `https://github.com/wangxiaochaun/NIQSV`, 2017.

[42] Shishun Tian, Lu Zhang, Luce Morin, and Olivier Déforges. Niqsv+: A no-reference synthesized view quality assessment metric. *IEEE Transactions on Image Processing*, 27 (4):1652–1664, 2017.

[43] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

[44] Rejin Varghese and M Sambath. Yolov8: A novel object detection algorithm with enhanced performance and robustness. In *2024 International conference on advances in data engineering and intelligent computing systems (ADICS)*, pages 1–6. IEEE, 2024.

[45] Narasimhan Venkatanath, D Praneeth, S Channappayya Sumohana, S Medasani Swarup, et al. Blind image quality evaluation using perception based features. In *2015 twenty first national conference on communications (NCC)*, pages 1–6. IEEE, 2015.

[46] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI conference on artificial intelligence*, pages 2555–2563, 2023.

[47] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[48] Charith Wickrema, Sara Leary, Shivangi Sarkar, Mark Giglio, Eric Bianchi, Eliza Mace, and Michael Twardowski. Benchmarking image similarity metrics for novel view synthesis applications. *arXiv preprint arXiv:2506.12563*, 2025.

[49] Qi Wu, Janick Martinez Esturo, Ashkan Mirzaei, Nicolas Moenne-Loccoz, and Zan Gojcic. 3dgut: Enabling distorted cameras and secondary rays in gaussian splatting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26036–26046, 2025.

[50] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34: 12077–12090, 2021.

[51] Yuke Xing, Qi Yang, Kaifa Yang, Yiling Xu, and Zhu Li. Explicit-nerf-qa: A quality assessment database for explicit nerf model compression. In *2024 IEEE International Conference on Visual Communications and Image Processing (VCIP)*, pages 1–5. IEEE, 2024.

[52] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1191–1200, 2022.

[53] Tianxiang Ye, Qi Wu, Junyuan Deng, Guoqing Liu, Liu Liu, Songpengcheng Xia, Liang Pang, Wenxian Yu, and Ling Pei. Thermal-nerf: Neural radiance fields from an infrared camera. In *2024 IEEE/RSJ International Conference on Intel-*

ligent Robots and Systems (IROS), pages 1046–1053. IEEE, 2024.

[54] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[55] Weixia Zhang, Kede Ma, Jia Yan, Dexiang Deng, and Zhou Wang. Blind image quality assessment using a deep bilinear convolutional neural network. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(1):36–47, 2018.

[56] Yuhang Zhang, Joshua Maraval, Zhengyu Zhang, Nicolas Ramin, Shishun Tian, and Lu Zhang. Evaluating human perception of novel view synthesis: Subjective quality assessment of gaussian splatting and nerf in dynamic scenes. *arXiv preprint arXiv:2501.08072*, 2025.

[57] Zicheng Zhang, Ziheng Jia, Haoning Wu, Chunyi Li, Zijian Chen, Yingjie Zhou, Wei Sun, Xiaohong Liu, Xiongkuo Min, Weisi Lin, et al. Q-bench-video: Benchmark the video quality understanding of lmms. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3229–3239, 2025.

[58] Zhichao Zhang, Wei Sun, Li Xinyue, Jun Jia, Xiongkuo Min, Zicheng Zhang, Chunyi Li, Zijian Chen, Wang Puyi, Sun Fengyu, et al. Benchmarking multi-dimensional aigc video quality assessment: A dataset and unified model. *ACM Transactions on Multimedia Computing, Communications and Applications*, 21(9):1–24, 2025.

[59] Qi Zheng, Yibo Fan, Leilei Huang, Tianyu Zhu, Jiaming Liu, Zhijian Hao, Shuo Xing, Chia-Ju Chen, Xiongkuo Min, Alan C Bovik, et al. Video quality assessment: A comprehensive survey. *arXiv preprint arXiv:2412.04508*, 2024.