DEEPCR: DEEP CROSS-REFERENCE REASONING FOR ASSESSING 3D GAUSSIAN SPLATTING RENDERINGS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reliable quality assessment is crucial for identifying viewpoint-dependent artifacts and improving 3D Gaussian Splatting (3DGS) reconstruction, yet aligned ground truth is generally unavailable at inference. Cross-reference assessment uses non-aligned observations, but their relevance, representation, and spatial aggregation affect prediction accuracy. We propose DeepCR, a cross-reference method for dense structural quality assessment of 3DGS renderings. Pose-aware reference selection prioritizes nearby observations using normalized position and orientation distances, while multi-level cross-view reasoning over DINOv2 features integrates local distortion cues with higher-level scene semantics. A pixel decoder with auxiliary supervision predicts spatial quality variations from coarse patch representations, yielding a full-resolution structural similarity map and an image-level score. Experiments show strong agreement with ground-truth SSIM scores, exceeding the strongest competing CR-IQA method by 8.5 and 11.7 percentage points in average absolute PLCC and SRCC, respectively. Evaluation against human artifact annotations supports within-scene perceptual agreement, and active view selection experiments demonstrate the utility of DeepCR predictions for reconstruction.

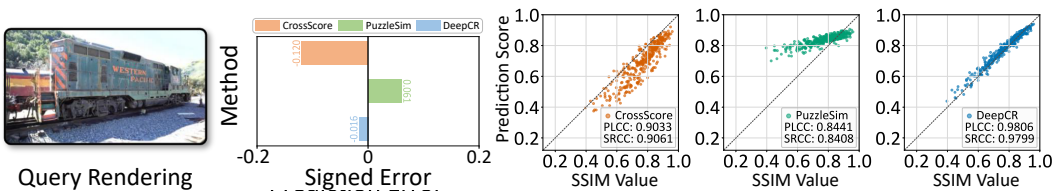
1 INTRODUCTION

Image quality assessment (IQA) supports a broad range of computer vision tasks by providing objective criteria for algorithm benchmarking, failure diagnosis, and system optimization (Wang, 2011; Chen et al., 2024). This capability is also particularly important for 3D Gaussian Splatting (3DGS) (Kerbl et al., 2023), where sparse observations, occlusions, and limited viewpoint coverage can introduce blur, floating primitives, and structural artifacts (Xiong et al., 2025). Reliable quality estimates are essential to compare reconstruction methods objectively. Beyond global benchmarking, view-specific assessment exposes failures and identifies where additional observations or optimization are required.

IQA methods can be categorized into full-reference (FR-IQA), no-reference (NR-IQA), and cross-reference (CR-IQA) methods according to the reference information available during assessment. FR-IQA methods such as PSNR (Huynh-Thu & Ghanbari, 2008), SSIM (Wang et al., 2004), and LPIPS (Zhang et al., 2018) compare a query with an aligned ground-truth image. Although these metrics provide reliable fidelity measurements, they cannot evaluate freely chosen or unseen viewpoints without corresponding observations. NR-IQA methods assess the query alone and therefore remain applicable in such settings (Mittal et al., 2012; 2013; Venkatanath et al., 2015; Ying et al., 2020), but lack the scene-specific context needed to reliably identify and localize reconstruction artifacts. CR-IQA addresses this limitation by leveraging non-aligned observations from the same scene, making it well suited for evaluating 3DGS renderings without aligned ground truth.

Recent CR-IQA methods demonstrate the value of such evidence. CrossScore (Wang et al., 2024) predicts dense quality maps from unregistered views, PuzzleSim (Hermann et al., 2025) localizes unsupported content using cross-view patch statistics, NOVA (Ghildyal et al., 2026) predicts similarity from a nearby reference, and PR-IQA (Choi et al., 2026) combines geometric overlap with reference-conditioned quality completion. Despite this progress, effectively exploiting cross-view evidence remains challenging. First, reference utility depends on geometric and visual relevance to the query, as distant viewpoints may provide limited evidence about its content. Second, quality

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Query rendering and predicted errors. (b) Prediction agreement over the scene.

Figure 1: Image- and scene-level quality assessment. (a) A query rendering together with the signed errors of CrossScore, PuzzleSim, and DeepCR, computed as $\hat{s} - s_{\mathrm{SSIM}}$; values closer to zero indicate higher accuracy. (b) Predictions of the three methods versus ground-truth SSIM over the scene. DeepCR exhibits the strongest correlation with ground-truth SSIM.

assessment requires balancing sensitivity to local artifacts with higher-level scene semantics that facilitate cross-view correspondence. Finally, dense prediction must recover spatial quality variations from the coarse representations used for cross-view aggregation.

To address these challenges, we propose DeepCR, a pose-aware method that assesses the structural quality of 3DGS renderings through deep cross-reference reasoning. For each query, normalized camera-pose distances prioritize nearby references as scene-specific evidence. Multi-level features from the query and selected references combine local details with higher-level scene context. Cross-attention incorporates reference information into the query representation while preserving its spatial layout, and a PixelShuffle-based decoder predicts a full-resolution quality map from the coarse patch grid. During training, pixel-level SSIM maps computed from aligned ground-truth images supervise the main and auxiliary predictions; inference requires only the query and selected non-aligned references. Spatial averaging yields an image-level score. Figure 1 compares query-level prediction errors and scene-level agreement with ground-truth SSIM.

Our main contributions are summarized as follows:

- We introduce DeepCR for dense CR-IQA of 3DGS renderings, using query-adaptive pose normalization to prioritize relevant non-aligned references.
- We integrate features across encoder depths for cross-view reasoning, combining local distortion cues with higher-level scene context.
- We use a PixelShuffle-based decoder with auxiliary supervision to predict quality at the query resolution from coarse patch representations.
- Experiments support improved image- and pixel-level structural quality prediction, stronger within-scene agreement with human annotations, and useful active view selection.

2 Related Work

2.1 Image Quality Assessment Metrics

IQA methods differ in the reference information available during evaluation. FR-IQA metrics compare a distorted image with aligned ground truth. PSNR (Huynh-Thu & Ghanbari, 2008) measures pixel fidelity, SSIM (Wang et al., 2004) compares local structure, and LPIPS (Zhang et al., 2018) evaluates deep-feature differences. Although effective when aligned ground truth is available, FR-IQA metrics cannot assess unseen viewpoints for which no corresponding observation exists. NR-IQA methods instead estimate quality from the query alone. BRISQUE (Mittal et al., 2012) and NIQE (Mittal et al., 2013) characterize deviations from natural-scene statistics, PIQE (Venkatanath et al., 2015) detects locally distorted regions, and PaQ-2-PiQ (Ying et al., 2020) aggregates patch-level estimates. NR-IQA methods (Qu et al., 2024) for novel view synthesis further model spatial and angular characteristics. GSOQA (Wan et al., 2026) instead assesses 3DGS content directly from spatial and photometric attributes of Gaussian primitives. Without scene context, however, NR-IQA methods struggle to distinguish reconstruction artifacts from valid scene content or localize such artifacts reliably.

CR-IQA addresses this limitation by assessing a query against non-aligned observations of the same scene. CrossScore (Wang et al., 2024) applies cross-attention over multiple references to predict maps supervised by pixel-level SSIM. PuzzleSim (Hermann et al., 2025) matches scene-specific patch statistics to localize unsupported content, NOVA (Ghildyal et al., 2026) predicts image-level similarity from a nearby reference, and PR-IQA (Choi et al., 2026) completes quality estimated in geometrically overlapping regions through reference-conditioned cross-attention. Effective assessment depends on reference relevance, the balance between local and semantic features, and spatial decoding. DeepCR addresses these factors through normalized pose-based selection, multi-level cross-view reasoning, and full-resolution quality prediction.

2.2 Radiance Fields and Active View Selection

Under limited observation budgets, quality-guided active view selection identifies observations that best improve reconstruction. NeRF (Mildenhall et al., 2020) represents scenes as a continuous radiance field, whereas 3DGS (Kerbl et al., 2023) employs explicit anisotropic Gaussians for real-time rendering. Sparse observations leave regions poorly constrained and cause viewpoint-dependent errors (Xiong et al., 2025), motivating the active selection of informative views. ActiveNeRF (Pan et al., 2022) selects views according to the expected reduction in predictive uncertainty. FisherRF (Jiang et al., 2024) and GauSS-MI (Xie et al., 2025) estimate information gain through Fisher and mutual information, respectively. POp-GS (Wilson et al., 2025) formulates view selection through optimal experimental design. Learned approaches predict image- and depth-based uncertainty (Shen et al., 2026) or construct hierarchical uncertainty spaces (Li et al., 2026).

Unlike uncertainty- and information-gain-based approaches, Active View Selector (Wang et al., 2025) applies the CrossScore framework to select low-quality candidate renderings, linking view assessment to reconstruction optimization. We further apply DeepCR to quality-guided active view selection, demonstrating that its predictions can direct view acquisition toward improved 3DGS reconstruction.

3 Method

3.1 Problem Formulation and Overview

Given a rendered query image I_q and a set of real observations $\mathcal{I} = \{I_i\}_{i=1}^{N}$ captured from other viewpoints in the same scene, DeepCR predicts spatial quality without requiring an aligned ground-truth image at inference. Each image is associated with a camera pose $P_i = (\mathbf{R}_i, \mathbf{c}_i)$, where $\mathbf{R}_i \in \mathrm{SO}(3)$ is its orientation and $\mathbf{c}_i \in \mathbb{R}^3$ is its camera center. The predicted full-resolution quality map $\hat{\mathbf{M}}_q \in [0,1]^{H\times W}$ approximates the local SSIM between I_q and its unavailable ground-truth counterpart. Spatial aggregation gives the image-level quality score

$$\hat{s}_q = \frac{1}{HW}\sum_{x=1}^{W}\sum_{y=1}^{H}\hat{\mathbf{M}}_q(y,x). \tag{1}$$

Figure 2 presents the overall framework. DeepCR first retrieves a subset $\mathcal{R}_q \subset \mathcal{I}$ of geometrically relevant reference views. A frozen DINOv2 encoder (Oquab et al., 2024) then extracts multi-level semantic patch features from the query and references. After multi-level integration, cross-view reasoning aggregates reference evidence into the query representation. A PixelShuffle-based pixel decoder finally recovers the full-resolution quality map, while an auxiliary branch provides additional spatial supervision during training.

3.2 Pose-Aware Reference Selection

The real observations of a scene provide useful reference evidence, but their relevance to a query varies with viewpoints. We prioritize candidate references by camera pose, where each pose comprises a camera-center position and a viewing orientation. For query q and candidate i, we compute translational and rotational distances

$$d_t(q,i) = \|\mathbf{c}_q - \mathbf{c}_i\|_2, \qquad d_R(q,i) = \arccos\left(\mathrm{clip}_{[-1,1]}\left(\frac{\mathrm{tr}(\mathbf{R}_i^\top \mathbf{R}_q) - 1}{2}\right)\right), \tag{2}$$

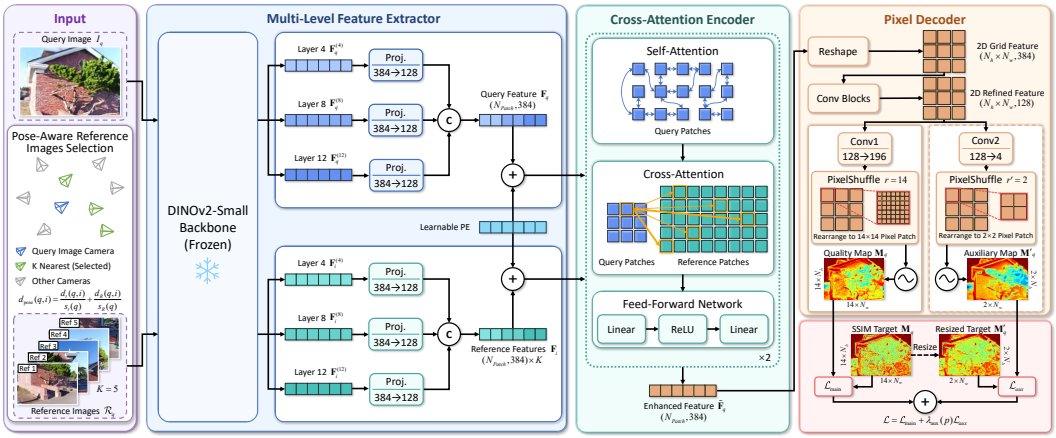


Figure 2: Overview of DeepCR. The selector retrieves the K references with the smallest normalized pose scores. A frozen DINOv2 backbone extracts multi-level features, and cross-view reasoning incorporates reference evidence into the query representation. A PixelShuffle-based decoder predicts a full-resolution quality map. The dashed auxiliary branch supplies additional supervision during training; the main branch provides the final prediction without an aligned ground-truth image at inference.

where d_R is the geodesic angle on $\mathrm{SO}(3)$ and clipping prevents numerical errors. Let $\mathcal{V}_q$ contain candidates with available images, recorded poses, and finite distances. We compute query-specific scales from all valid candidates,

$$s_t(q) = \mathrm{median}_{i\in\mathcal{V}_q}\, d_t(q,i), \qquad s_R(q) = \mathrm{median}_{i\in\mathcal{V}_q}\, d_R(q,i). \tag{3}$$

We rank candidates with the normalized score

$$d_{\mathrm{pose}}(q,i) = \frac{d_t(q,i)}{s_t(q)} + \frac{d_R(q,i)}{s_R(q)}, \tag{4}$$

and select the K smallest-scoring candidates as $\mathcal{R}_q$. Pose proximity serves as a geometric relevance prior for subsequent feature-level reasoning. Appendix B.2 describes degenerate-scale handling.

3.3 Multi-Level Semantic Reasoning

With geometrically relevant references selected, DeepCR constructs a cross-view representation for quality prediction. Features from any single encoder depth provide limited abstraction: shallow features preserve appearance and structure but are viewpoint-sensitive, whereas deep features encode viewpoint-stable scene semantics but may obscure local distortions. We therefore extract multi-level patch features from a frozen DINOv2 encoder to balance local artifact sensitivity with stable cross-view correspondence. Let $\mathbf{F}_q^{(l)} \in \mathbb{R}^{N_{\mathrm{patch}}\times D_l}$ denote the query features at depth l, where $N_{\mathrm{patch}} = N_{\mathrm{patch}}^h \times N_{\mathrm{patch}}^w$ is the number of spatial patch tokens and D_l is the channel dimension. We use $\mathcal{L} = \{4, 8, 12\}$ to span shallow, intermediate, and deep DINOv2 representations, and a layer-specific linear projection ϕ_l maps the features to a common channel dimension. We concatenate the projected features to obtain

$$\mathbf{F}_q = \mathrm{Concat}_{l\in\mathcal{L}}\left[\phi_l\left(\mathbf{F}_q^{(l)}\right)\right]. \tag{5}$$

The same shared operations produce $\mathbf{F}_i$ for every $I_i \in \mathcal{R}_q$. This integration combines complementary semantics from different encoder depths while preserving the common spatial patch layout. Learnable spatial positional embeddings are subsequently added to both query and reference features.

All reference tokens are concatenated into a memory $\mathbf{F}_{\mathcal{R}}$. Self-attention first refines the query features by modeling dependencies within the query image, after which query-to-reference cross-attention retrieves relevant evidence from $\mathbf{F}_{\mathcal{R}}$. The resulting enhanced query feature $\widetilde{\mathbf{F}}_q$ retains the spatial organization of the query while incorporating complementary scene evidence from the selected references, providing a structured input for quality map decoding.

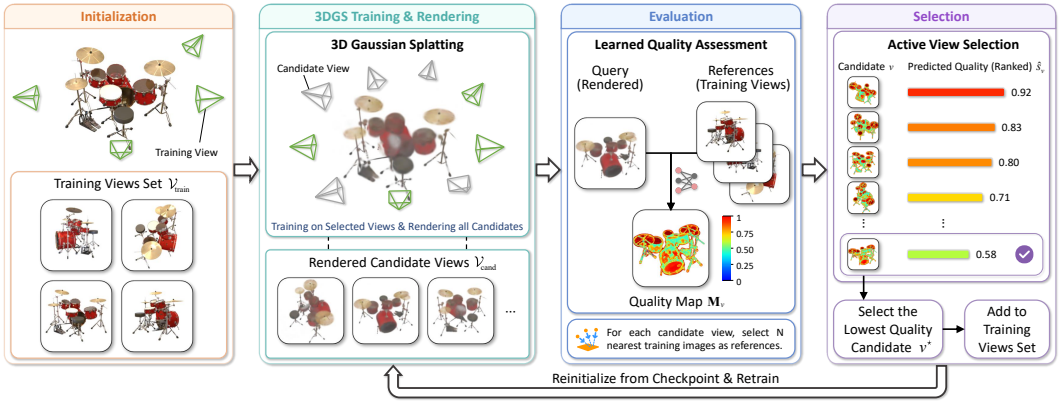


Figure 3: Quality-guided active view selection with DeepCR. Starting from an initial training view set, 3DGS renders candidate views. DeepCR evaluates each candidate using nearby training views as non-aligned references, selects the lowest-quality candidate, and adds its real observation to the training set for the next reconstruction round.

3.4 Full-Resolution Quality Decoding

Although cross-view reasoning enriches the query tokens, its output remains a coarse patch grid. We reshape the tokens into a two-dimensional feature map and refine them with convolutional layers. Each branch uses an output convolution followed by PixelShuffle (Shi et al., 2016). The main branch generates r^2 channels and applies a factor of r to produce $\hat{\mathbf{M}}_q$ at the preprocessed query resolution $H \times W$, which we call full resolution. The auxiliary branch uses $(r')^2$ channels and a smaller factor r' to produce $\hat{\mathbf{M}}'_q$. Sigmoid activations constrain both outputs to $[0, 1]$.

The auxiliary branch supplies lower-resolution supervision to the coarse patch representation during training, encouraging spatial quality learning alongside the main objective. The main branch alone provides the final quality map and image-level score.

The ground-truth SSIM map $\mathbf{M}_q$ is computed between the rendered query and its aligned ground-truth image, with values clipped to $[0, 1]$ throughout training and evaluation. Its spatial mean defines the ground-truth image-level score. The aligned image supplies targets only and is unavailable at inference. We optimize the main prediction using the mean absolute error (MAE)

$$\mathcal{L}_{\text{main}} = \frac{1}{HW} \sum_{x=1}^{W} \sum_{y=1}^{H} \left| \hat{\mathbf{M}}_q(y, x) - \mathbf{M}_q(y, x) \right|. \tag{6}$$

For the auxiliary branch, we resize the target map to its output resolution $H' \times W'$ and define

$$\mathcal{L}_{\text{aux}} = \frac{1}{H'W'} \sum_{x=1}^{W'} \sum_{y=1}^{H'} \left| \hat{\mathbf{M}}'_q(y, x) - \mathbf{M}'_q(y, x) \right|, \tag{7}$$

where $\mathbf{M}'_q = \text{Resize}(\mathbf{M}_q, H', W')$. The total objective combines both predictions,

$$\begin{aligned} \mathcal{L} &= \mathcal{L}_{\text{main}} + \lambda_{\text{aux}}(p)\mathcal{L}_{\text{aux}}, \\ \lambda_{\text{aux}}(p) &= \lambda_0 \frac{1 + \cos(\pi p)}{2}. \end{aligned} \tag{8}$$

where $p \in [0, 1]$ denotes normalized training progress and $\lambda_0 = 0.3$. The cosine schedule emphasizes low-resolution supervision early and gradually shifts optimization to the full-resolution prediction. Once trained, DeepCR reliably predicts quality maps and corresponding image-level scores.

3.5 Quality-Guided Active View Selection

Under limited observation budgets, active view selection prioritizes the least reliably reconstructed viewpoints. DeepCR provides this signal through the image-level score in Equation 1. As illustrated

in Figure 3, we begin with a 3DGS model optimized using the current training-view set $\mathcal{V}_{\text{train}}$. The model renders images at a pool of candidate camera poses $\mathcal{V}_{\text{cand}}$. For each candidate v, DeepCR treats its rendering as the query, selects non-aligned references from $\mathcal{V}_{\text{train}}$ using Equation 4, and spatially aggregates the predicted map $\hat{\mathbf{M}}_v$ into $\hat{s}_v$. A low score indicates that the reconstruction is unreliable at the corresponding viewpoint. We therefore acquire the candidate with the lowest predicted quality,

$$v^\star = \arg\min_{v \in \mathcal{V}_{\text{cand}}} \hat{s}_v, \tag{9}$$

and add its corresponding real observation to the training set. The selected pose is removed from the candidate pool, and 3DGS is optimized again with the expanded set $\mathcal{V}_{\text{train}} \cup \{v^\star\}$. Repeating this evaluate-select-update cycle allocates the observation budget to views with visible reconstruction deficiencies. By connecting view quality assessment with iterative reconstruction improvement, this quality-guided strategy achieves strong active view selection performance.

4 Experiments

4.1 Experimental Setup

Datasets and data generation. We follow the open-source CrossScore pipeline (Wang et al., 2024) to construct IQA benchmarks from the MapFree Relocalisation (MFR) (Arnold et al., 2022), RealEstate10K (RE10K) (Zhou et al., 2018), and Tanks and Temples (T&T) (Knapitsch et al., 2017) datasets. For each scene, 3DGS renderings generated at different optimization stages form query images with varying reconstruction artifacts. The real image at the query pose is used only to compute the supervision and evaluation SSIM maps and is explicitly excluded from the reference pool; observations from other viewpoints form the non-aligned reference pool. MFR contains 460 scenes, which we randomly split by scene into 80% for training, 10% for validation, and 10% for testing. To evaluate cross-dataset generalization, we additionally test on 100 RE10K scenes and all 21 T&T scenes without using either dataset for training.

Compared methods and metrics. We compare FR-IQA baselines (PSNR (Huynh-Thu & Ghanbari, 2008), SSIM (Wang et al., 2004), and LPIPS (Zhang et al., 2018)), zero-shot natural-image NR-IQA baselines (NIQE (Mittal et al., 2013), PIQE (Venkatanath et al., 2015), BRISQUE (Mittal et al., 2012), and PaQ-2-PiQ (Ying et al., 2020)), and CR-IQA baselines (PR-IQA (Choi et al., 2026), CrossScore (Wang et al., 2024), and PuzzleSim (Hermann et al., 2025)). Trainable CR-IQA baselines are retrained on the same MFR split with their original objectives and official configurations. All CR-IQA methods follow the same aligned-image exclusion rule. All methods use the same queries and input resolution, with spatial predictions evaluated at a common resolution. Map predictions are spatially averaged, and PLCC (Pearson, 1895) and SRCC (Spearman, 1904) are computed against SSIM over each benchmark. For methods whose score direction is reversed, we retain signed correlations in the dataset columns and report average magnitude separately.

Implementation details. DeepCR uses a frozen DINOv2-S/14 backbone, extracting multi-level features from layers 4, 8, and 12. Images are cropped to 518×518, with $K = 5$ references by default. Each layer is projected to 128 dimensions, yielding a 384-dimensional concatenated representation. The main and auxiliary pixel-decoder branches use PixelShuffle factors of 14 and 2, respectively. We train for 5 epochs on two NVIDIA RTX 5090D GPUs with a per-GPU batch size of 24, using AdamW (Loshchilov & Hutter, 2019) at a learning rate of 5×10^{-4}. Further details are in the Appendix.

4.2 Comparison with State-of-the-Art Methods

Quantitative comparison. Table 1 compares image-level predictions against ground-truth SSIM, with FR-IQA methods serving as full-reference baselines. NR-IQA methods show weak, dataset-dependent correlations, whereas CR-IQA generally benefits from non-aligned scene observations. DeepCR achieves the best NR-IQA/CR-IQA results on every benchmark, outperforming all other methods by at least 8 and 11 percentage points in average $|\text{PLCC}|$ and $|\text{SRCC}|$, respectively, while maintaining strong performance on the unseen RE10K and T&T datasets. Table S2 and Figure S3

Table 1: Quantitative experimental results of image quality assessment. PLCC and SRCC are computed against ground-truth SSIM scores over all test queries in each benchmark. Dataset-specific columns retain signed correlations, while Average Magnitude reports the mean absolute correlation across benchmarks. Among NR-IQA and CR-IQA methods, the best results are in bold and the second best are underlined.

IQA Type	Method	MFR		RE10K		T&T		Average Magnitude	
		PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	$\|PLCC\| \uparrow$	$\|SRCC\| \uparrow$
FR-IQA	PSNR†	0.7983	0.8412	0.8111	0.9452	0.8969	0.9187	0.8354	0.9017
	SSIM†	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	LPIPS†	-0.8752	-0.8728	-0.8945	-0.9674	-0.8834	-0.9215	0.8844	0.9206
NR-IQA	NIQE	-0.0167	-0.0852	0.1077	-0.1328	-0.0267	-0.3509	0.0504	0.1896
	PIQE	-0.0941	-0.0284	0.0888	0.1269	-0.3871	-0.4696	0.1900	0.2083
	BRISQUE	-0.1316	-0.1103	-0.0362	0.0665	-0.3536	-0.3983	0.1738	0.1917
	PaQ-2-PiQ	-0.0123	-0.0349	0.3588	0.3048	0.5668	0.6151	0.3126	0.3183
CR-IQA	PR-IQA	0.1728	0.1345	0.5176	0.4634	0.3134	0.2913	0.3346	0.2964
	CrossScore	0.7792	0.7221	0.8720	0.8526	0.8519	0.8448	0.8344	0.8065
	PuzzleSim	0.7006	0.6375	0.8566	0.8742	0.6843	0.7325	0.7472	0.7481
	DeepCR	**0.8956**	**0.8830**	**0.9170**	**0.9326**	**0.9449**	**0.9557**	**0.9192**	**0.9238**

† These metrics require an aligned ground-truth image and are reported only as full-reference baselines.

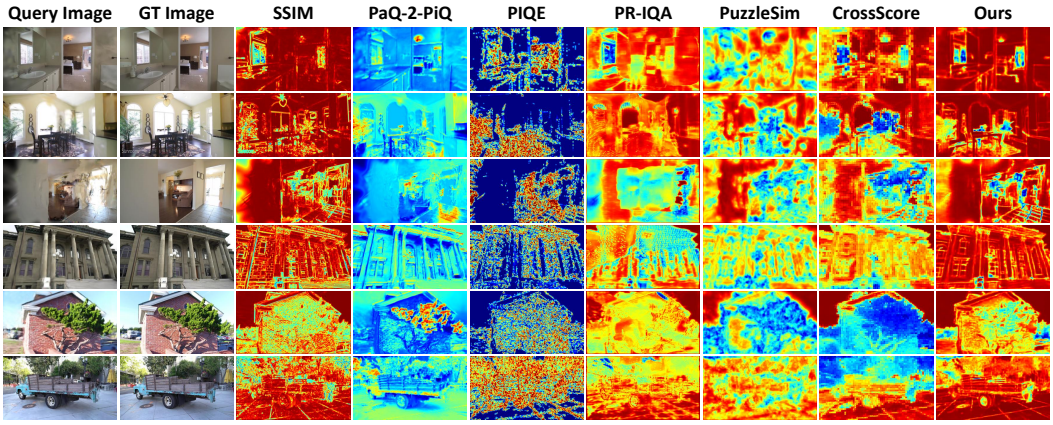


Figure 4: Qualitative comparison of quality map prediction. Each example shows a rendered query, its aligned ground-truth image, the target SSIM map, and maps predicted by representative NR-IQA and CR-IQA methods. DeepCR more closely follows the target's spatial quality variations and fine-grained structures.

further show the highest mean and smallest spread of scene-wise correlations, supporting consistency across evaluated scenes.

Qualitative comparison. Figure 4 presents spatial quality predictions alongside the target SSIM maps, which highlight localized errors in poorly reconstructed structures and appearance details. Existing methods show weak responses to these errors or produce spatially inconsistent maps. In contrast, DeepCR more accurately captures fine-grained quality variations and target structures, consistent with its stronger image-level correlations.

4.3 Ablation Studies and Further Evaluation

Component ablation. Table 2 examines reference selection, feature hierarchy, and spatial decoding within DeepCR. The baseline uses random references, only the final DINOv2 layer, and a token-wise MLP head. Removing any component lowers both correlations across all benchmarks,

Table 2: Ablation on architectural components. The best results are in bold.

Model variants	MFR		RE10K		T&T	
	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
Baseline	0.7993	0.7352	0.9002	0.8626	0.8750	0.8796
w/o pose selection	0.8728	0.8467	0.9156	0.9100	0.8896	0.8930
w/o multi-level features	0.8015	0.7514	0.8676	0.8645	0.9040	0.9143
w/o pixel decoder	0.8915	0.8763	0.9128	0.9209	0.9348	0.9482
Full model	**0.8956**	**0.8830**	**0.9170**	**0.9326**	**0.9449**	**0.9557**

Table 3: Dense-map pixel MAE against SSIM targets. Lower is better.

Method	MFR	RE10K	T&T
PR-IQA	0.3181	0.2614	0.4265
PuzzleSim	0.2182	0.2122	0.2698
CrossScore	0.1459	0.1160	0.1465
DeepCR	**0.1306**	**0.0823**	**0.1119**

Table 4: Within-scene correlation with human artifact annotations. Higher is better.

Method	wi-PLCC	wi-SRCC
PR-IQA	-0.0044	0.0160
PuzzleSim	0.1888	0.2190
CrossScore	0.2986	0.2553
DeepCR	**0.4787**	**0.5030**

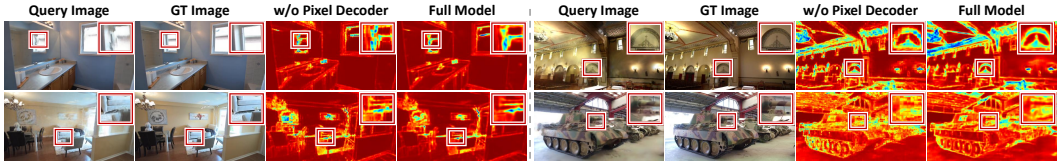


Figure 5: Effect of the pixel decoder. It reduces patch-wise discontinuities and recovers finer spatial quality variations, yielding a more detailed and spatially coherent prediction.

supporting the roles of nearby reference evidence, features across encoder depths, and learned spatial decoding. The complete model achieves the best results among these variants. Figure 5 illustrates reduced patch-wise discontinuities and finer spatial structure with the pixel decoder.

Dense-map quantitative evaluation. We compute pixel-wise MAE between each predicted map and its SSIM target, then average per-image errors within each dataset. Table 3 shows the lowest error for DeepCR on all benchmarks, including unseen RE10K and T&T. Its improvement over CrossScore, which also predicts SSIM maps, supports more accurate spatial prediction and complements the image-level correlations in Table 1.

Human-perception validation. We evaluate on annotated-3DGS-artifacts from PuzzleSim (Hermann et al., 2025), with 12 scenes at 60%, 70%, and 80% retained-view ratios. Binary judgments are converted to quality-oriented maps, so larger values indicate better perceived quality. We spatially average predicted and annotation maps, subtract their respective scene-wise means, and correlate the pooled residuals (wi-PLCC and wi-SRCC; Appendix C.2). DeepCR achieves the highest correlations among the compared methods (Table 4), supporting image-level agreement with human-perceived quality variations within scenes on this benchmark.

4.4 Active View Selection for 3DGS

Experimental protocol. We evaluate DeepCR for active view selection on eight Blender scenes (Mildenhall et al., 2020) and nine Mip-NeRF360 scenes (Barron et al., 2022). Following FisherRF (Jiang et al., 2024), every reconstruction begins with four training views and runs for 30,000 iterations. We add one view at iterations 400, 900, 1,500, 2,200, 3,000, 3,900, 4,900, 6,000, 7,200, 8,500, 9,900, 11,400, 13,000, 14,700, 16,500, and 18,400, yielding 20 training views in total. All methods share the same initial views, candidate pool, acquisition budget, and evaluation views. DeepCR is used without fine-tuning, and we compare against FisherRF, Auto3R (Shen et al., 2026), GauSS-MI (Xie et al., 2025), and the NR-IQA and CR-IQA baselines. At each step, IQA methods

Table 5: Quantitative results for the active view selection task. DeepCR-guided selection consistently yields high-quality reconstructions. The best results are in bold.

Method	Blender			Mip-NeRF360		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
FisherRF	29.2610	0.9411	0.0510	20.2784	0.5848	0.3308
Auto3R	28.5737	0.9378	0.0548	20.5953	0.6169	0.3133
GauSS-MI	27.0183	0.9185	0.0715	19.4086	0.5546	0.3593
NIQE	26.7398	0.9246	0.0669	18.5945	0.5596	0.3522
PIQE	27.6205	0.9279	0.0625	18.5956	0.5472	0.3594
BRISQUE	26.4139	0.9131	0.0761	18.5487	0.5321	0.3725
PaQ-2-PiQ	28.0692	0.9322	0.0586	19.1574	0.5578	0.3506
PR-IQA	27.9251	0.9348	0.0576	19.4168	0.5837	0.3340
PuzzleSim	28.7897	0.9410	0.0514	20.8223	0.5993	0.3189
CrossScore	28.3448	0.9365	0.0554	20.9099	0.6223	0.3067
DeepCR	**29.3183**	**0.9441**	**0.0496**	**20.9916**	**0.6274**	**0.3002**

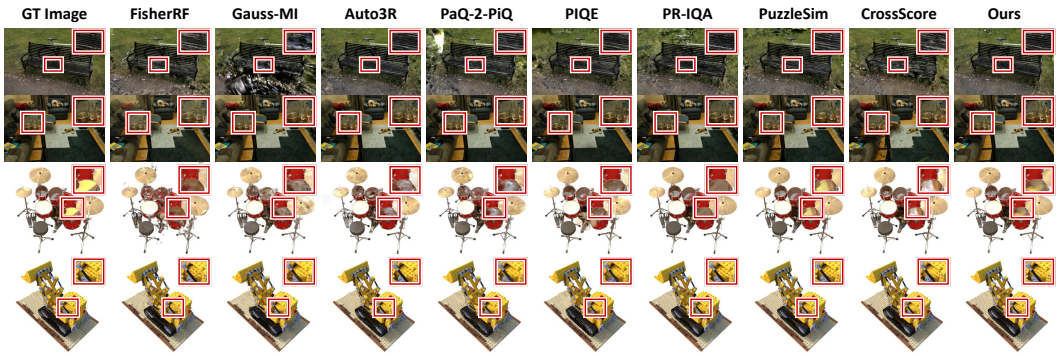


Figure 6: Qualitative comparison for active view selection. Each column shows a final 3DGS rendering using views acquired by the corresponding method. The highlighted regions illustrate differences in reconstructed appearance and structure.

rank rendered candidates and acquire the lowest-quality view; final reconstructions are evaluated by PSNR, SSIM, and LPIPS.

Results. Table 5 shows that DeepCR achieves the best PSNR, SSIM, and LPIPS on both benchmarks, indicating effective view acquisition across synthetic and real-world scenes. Figure 6 further supports the effectiveness of DeepCR for quality-guided active view selection. Although some strategies perform well in individual scenes, DeepCR-guided selection consistently yields high-quality reconstructions across all scenes, with fewer appearance and structural errors and closer agreement with the ground-truth images.

5 Conclusion

We presented DeepCR, a CR-IQA method for assessing 3DGS renderings without aligned ground truth at inference. Pose-aware selection identifies geometrically relevant references to reduce ambiguous evidence, multi-level reasoning integrates local distortion cues with stable scene semantics, and the pixel decoder recovers fine-grained quality variations from coarse patch representations. These experiments support improved structural quality prediction across the evaluated datasets, within-scene image-level agreement with human annotations, and useful guidance for active view selection.

AI USE STATEMENT

We used generative AI tools to assist with designing the experimental methodology, providing feedback on experiments, implementing parts of the method, drafting portions of the manuscript, editing paragraphs for clarity, and revising the organization of the paper. The tools were not used to generate datasets, fabricate data or results, or make final scientific decisions. The authors reviewed and verified all AI-assisted code, text, and methodological suggestions through implementation checks, experiments, and comparison with the underlying records. The authors take full responsibility for the final content, claims, results, and artifacts of this work.

ETHICS STATEMENT

This work studies image quality assessment for 3D Gaussian Splatting renderings using publicly available datasets, including aggregate human artifact annotations, and standard reconstruction pipelines. It does not involve new human-subject recruitment or intervention, collection of personally identifiable information, or release of new sensitive data. We use the existing human annotations only in aggregate and do not attempt to identify individual subjects. We do not foresee direct harmful applications, and we expect the method to support failure diagnosis and more reliable benchmarking of view synthesis systems.

REPRODUCIBILITY STATEMENT

The datasets used in this work (MapFree Relocalisation, RealEstate10K, Tanks and Temples, annotated-3DGS-artifacts, Blender, and Mip-NeRF360) are publicly available. Sections 3–4 and the Appendix describe the method, data processing, training settings, and evaluation protocols. The anonymized supplementary code includes the DeepCR implementation, configuration files, training and evaluation scripts, preprocessing utilities, an AVS integration, and a processed example scene.

REFERENCES

Eduardo Arnold, Jamie Wynn, Sara Vicente, Guillermo Garcia-Hernando, Aron Monszpart, Victor A. Prisacariu, Daniyar Turmukhambetov, and Eric Brachmann. Map-free visual relocalization: Metric pose relative to a single image. In *European Conference on Computer Vision*, pp. 690–708. Springer, 2022. doi: 10.1007/978-3-031-19769-7_40.

Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-NeRF 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5470–5479, 2022. doi: 10.1109/CVPR52688.2022.00539.

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. doi: 10.1109/TIP.2024.3378466.

Inseong Choi, Siwoo Lee, Seung-Hun Nam, and Soohwan Song. PR-IQA: Partial-reference image quality assessment for diffusion-based novel view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Abhijay Ghildyal, Rajesh Sureddi, Nabajeet Barman, Saman Zadtootaghaj, and Alan C. Bovik. Non-aligned reference image quality assessment for novel view synthesis. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 6350–6359, 2026. doi: 10.1109/WACV61042.2026.00614.

Nicolai Hermann, Jorge Condor, and Piotr Didyk. Puzzle similarity: A perceptually-guided cross-reference metric for artifact detection in 3d scene reconstructions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 28881–28891, 2025. doi: 10.1109/ICCV51701.2025.02682.

Quan Huynh-Thu and Mohammed Ghanbari. Scope of validity of PSNR in image/video quality assessment. *Electronics Letters*, 44(13):800–801, 2008. doi: 10.1049/el:20080522.

Wen Jiang, Boshu Lei, and Kostas Daniilidis. FisherRF: Active view selection and mapping with radiance fields using fisher information. In *European Conference on Computer Vision*, pp. 422–440. Springer, 2024. doi: 10.1007/978-3-031-72624-8_24.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):139:1–139:14, 2023. doi: 10.1145/3592433.

Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics*, 36(4):78:1–78:13, 2017. doi: 10.1145/3072959.3073599.

Yan Li, Yingzhao Li, and Gim Hee Lee. Active3D: Active high-fidelity 3D reconstruction via multi-level uncertainty quantification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 6557–6565, 2026. doi: 10.1609/aaai.v40i8.37585.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

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 *European Conference on Computer Vision*, pp. 405–421. Springer, 2020. doi: 10.1007/978-3-030-58452-8_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. doi: 10.1109/TIP.2012.2214050.

Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013. doi: 10.1109/LSP.2012.2227726.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaa El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Xuran Pan, Zihang Lai, Shuran Song, and Gao Huang. ActiveNeRF: Learning where to see with uncertainty estimation. In *European Conference on Computer Vision*, pp. 230–246. Springer, 2022. doi: 10.1007/978-3-031-19827-4_14.

Karl Pearson. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58:240–242, 1895. doi: 10.1098/rspl.1895.0041.

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. doi: 10.1109/TVCG.2024.3372037.

Chentao Shen, Sizhe Zheng, Bingqian Wu, Yaohua Feng, Yuanchen Fei, Mingyu Mei, Hanwen Jiang, and Xiangru Huang. Auto3R: Automated 3D reconstruction and scanning via data-driven uncertainty quantification. In *European Conference on Computer Vision*, 2026.

Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P. Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1874–1883, 2016. doi: 10.1109/CVPR.2016.207.

Charles Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904. doi: 10.2307/1412159.

N. Venkatanath, D. Praneeth, Maruthi Chandrasekhar Bh, Sumohana S. Channappayya, and Swarup S. Medasani. Blind image quality evaluation using perception based features. In *2015 Twenty First National Conference on Communications*, pp. 1–6, 2015. doi: 10.1109/NCC.2015.7084843.

Zhaolin Wan, Yining Diao, Jingqi Xu, Hao Wang, Zhiyang Li, Xiaopeng Fan, Wangmeng Zuo, and Debin Zhao. Perceptual quality assessment of 3D gaussian splatting: A subjective dataset and prediction metric. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 9657–9665, 2026. doi: 10.1609/aaai.v40i12.37928.

Zhou Wang. Applications of objective image quality assessment methods. *IEEE Signal Processing Magazine*, 28(6):137–142, 2011. doi: 10.1109/MSP.2011.942295.

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. doi: 10.1109/TIP.2003.819861.

Zirui Wang, Wenjing Bian, and Victor Adrian Prisacariu. CrossScore: Towards multi-view image evaluation and scoring. In *European Conference on Computer Vision*, pp. 492–510. Springer, 2024. doi: 10.1007/978-3-031-72673-6_27.

Zirui Wang, Yash Bhalgat, Ruining Li, and Victor Adrian Prisacariu. Active view selector: Fast and accurate active view selection with cross reference image quality assessment. *arXiv preprint arXiv:2506.19844*, 2025.

Joey Wilson, Marcelino Almeida, Sachit Mahajan, Martin Labrie, Maani Ghaffari, Omid Ghasemalizadeh, Min Sun, Cheng-Hao Kuo, and Arnab Sen. POp-GS: Next best view in 3D-gaussian splatting with p-optimality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3646–3655, 2025. doi: 10.1109/CVPR52734.2025.00345.

Yuhan Xie, Yixi Cai, Yinqiang Zhang, Lei Yang, and Jia Pan. GauSS-MI: Gaussian splatting shannon mutual information for active 3d reconstruction. In *Robotics: Science and Systems*, 2025. doi: 10.15607/RSS.2025.XXI.030.

Haolin Xiong, Sairisheek Muttukuru, Hanyuan Xiao, Rishi Upadhyay, Pradyumna Chari, Yajie Zhao, and Achuta Kadambi. SparseGS: Sparse view synthesis using 3D gaussian splatting. In *2025 International Conference on 3D Vision (3DV)*, pp. 1032–1041, 2025. doi: 10.1109/3DV66043.2025.00100.

Zhenqiang Ying, Haoran Niu, Praful Gupta, Dhruv Mahajan, Deepti Ghadiyaram, and Alan C. Bovik. From patches to pictures (PaQ-2-PiQ): Mapping the perceptual space of picture quality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3575–3585, 2020. doi: 10.1109/CVPR42600.2020.00363.

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/CVF Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018. doi: 10.1109/CVPR.2018.00068.

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *ACM Transactions on Graphics*, 37(4):65:1–65:12, 2018. doi: 10.1145/3197517.3201323.

APPENDIX

This appendix provides additional details and analyses supporting the main paper. It describes the datasets, benchmark generation, implementation, evaluation metrics, and compared methods, followed by scene-wise IQA results and additional experiments on reference-count sensitivity and inference runtime.

A DATASETS

This section describes the source datasets and splits used to construct the IQA benchmarks, the benchmark-generation procedure, and the datasets used for AVS evaluation.

A.1 IQA SOURCE DATASETS AND SPLITS

We derive our image quality assessment (IQA) benchmarks from three multi-view datasets with complementary scene distributions: MapFree Relocalisation (MFR) (Arnold et al., 2022), RealEstate10K (RE10K) (Zhou et al., 2018), and Tanks and Temples (T&T) (Knapitsch et al., 2017). MFR serves as the source domain for training, validation, and in-domain testing, while RE10K and T&T serve as two distinct target domains for cross-domain evaluation (Table S1).

MapFree Relocalisation. MFR contains 655 small places of interest captured worldwide, with substantial variation in viewpoint, illumination, and visual overlap. These properties provide diverse cross-view evidence and make reference selection nontrivial. We use the 460 scenes from its official training split as our source domain and randomly divide them at the scene level into 80% training, 10% validation, and 10% in-domain test subsets, with no scene overlap.

RealEstate10K. RE10K contains approximately 10 million posed frames from about 80,000 real-estate video clips, covering diverse indoor and outdoor residential environments. Its video trajectories and natural capture variation differ from MFR's place-centric imagery. We randomly sample 100 scenes and use them exclusively to evaluate cross-domain generalization without adaptation or fine-tuning.

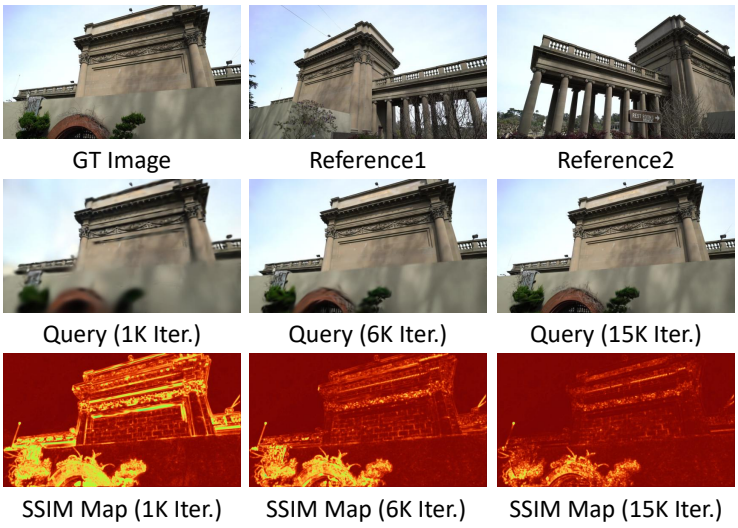


Figure S1: A representative sample from the IQA benchmark. The top row shows the aligned ground-truth image and two non-aligned reference observations from the same scene. The middle row shows query renderings at 1,000, 6,000, and 15,000 3DGS optimization iterations, and the bottom row shows the corresponding SSIM maps.

Table S1: IQA source datasets and their roles in our evaluation. Only MFR contributes training samples; the selected RE10K subset and all T&T scenes are reserved for cross-dataset evaluation.

Dataset	Used in this work	Experimental role	Representative characteristics
MFR	460 scenes	Train/val/in-domain test	Small places of interest; large viewpoint and condition changes; variable visual overlap
RE10K	100 scenes	Cross-domain test	Real-estate video trajectories; diverse indoor/outdoor residential content and natural capture variation
T&T	21 scenes	Cross-domain test	7 training, 8 intermediate, and 6 advanced scenes; reconstruction at scales ranging from individual objects to buildings

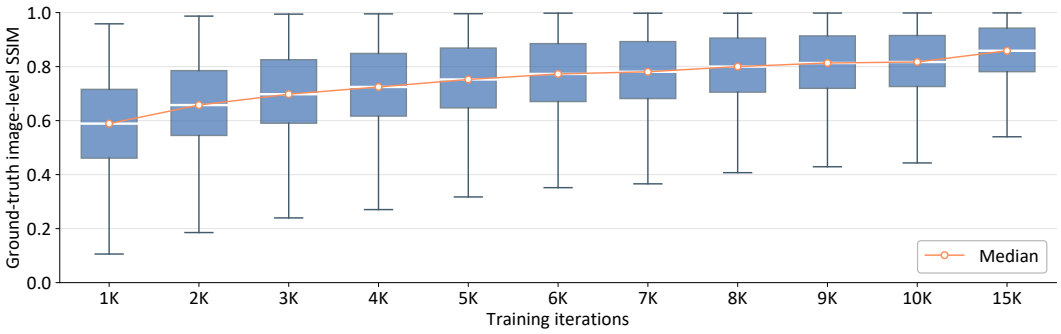


Figure S2: Distribution of ground-truth image-level SSIM across the 11 saved 3DGS optimization iterations, pooling query renderings from MFR, RE10K, and T&T. Boxes span the interquartile range, white bars indicate medians, and whiskers extend to the most extreme observations within $1.5\times$ the interquartile range; the orange curve connects the medians.

Tanks and Temples. T&T comprises 21 scenes organized into three official groups: seven training, eight intermediate, and six advanced scenes. The intermediate group includes sculptures, large vehicles, and house-scale buildings captured with outside-looking-in trajectories, while the advanced group contains large indoor and outdoor scenes with complex geometry and camera trajectories. We use all 21 scenes exclusively for cross-domain testing, covering reconstruction at scales ranging from individual objects to buildings under realistic conditions.

A.2 IQA Benchmark Generation

We build on the open-source CrossScore data-generation pipeline (Wang et al., 2024) to convert the selected multi-view scenes into IQA benchmarks. As illustrated in Figure S1, each benchmark instance consists of a rendered query, an aligned real image and its target SSIM map, non-aligned reference observations, and the associated camera poses.

Queries with controlled quality variation. For each scene, we optimize a 3DGS model (Kerbl et al., 2023) and save query renderings at iterations 1,000, 2,000, 3,000, 4,000, 5,000, 6,000, 7,000, 8,000, 9,000, 10,000, and 15,000. A rendering and its camera pose constitute one query. Renderings from earlier iterations generally exhibit stronger structural errors, floating primitives, blur, and incomplete appearance, whereas those from later iterations tend to more closely match the aligned real image. Sampling these 11 iterations creates a broad quality range while preserving the same scene content and camera distribution.

As shown in Figure S2, the pooled SSIM distribution shifts toward higher values as optimization proceeds, while adjacent iterations retain substantial overlap. The benchmark therefore covers a continuous range of rendering quality rather than a set of sharply separated quality levels.

Targets and reference pools. The real image captured at the query pose is spatially aligned with the rendering and is used only to derive the supervision and evaluation targets. We compute the ground-truth SSIM map from this image pair using the definition in Section 3.4; its spatial average defines the ground-truth image-level quality score. The same target definition is used for training and evaluation. All other real observations in the same scene form the non-aligned reference pool. The aligned image is excluded from this pool. At inference, NR-IQA methods receive only the query, while CR-IQA methods receive the query and non-aligned observations; neither setting accesses the aligned target image. This separation ensures that the benchmark evaluates quality prediction from the information available at deployment rather than direct full-reference comparison.

The resulting benchmarks contain 37,521 MFR queries, 20,779 RE10K queries, and 10,384 T&T queries. Together with their corresponding ground-truth SSIM maps, non-aligned reference observations, and camera metadata, the processed data occupy approximately 1.6 TB.

A.3 Datasets for Active View Selection

We evaluate active view selection (AVS) on the eight Blender (NeRF Synthetic) scenes (Mildenhall et al., 2020) and all nine real-world Mip-NeRF360 scenes (Barron et al., 2022). These benchmarks are distinct from the IQA datasets because AVS evaluates whether an IQA signal can guide sequential observation acquisition and improve the final 3DGS reconstruction under a fixed view budget.

Blender (NeRF Synthetic). The Blender dataset contains eight object-centric scenes rendered synthetically: *Chair*, *Drums*, *Ficus*, *Hotdog*, *Lego*, *Materials*, *Mic*, and *Ship*. The scenes provide exact camera parameters and clean ground-truth images under controlled rendering conditions. They cover diverse reconstruction challenges, including thin structures, fine geometric details, specular highlights, and heterogeneous materials, and therefore provide a controlled test of whether different selection criteria allocate the same acquisition budget to informative viewpoints.

Mip-NeRF360. Mip-NeRF360 contains nine real-world unbounded scenes with 360-degree captures. The four indoor scenes are *Bonsai*, *Counter*, *Kitchen*, and *Room*; the five outdoor scenes are *Bicycle*, *Flowers*, *Garden*, *Stump*, and *Treehill*. These scenes contain cluttered backgrounds, large depth ranges, fine foliage, reflective or view-dependent appearance, and complex geometry. They complement the controlled Blender benchmark by testing AVS under real imaging conditions and unbounded scene extent.

B Implementation Details

This section presents the implementation pipeline, including input preprocessing, the DeepCR architecture, the training configuration and computing environment, and the AVS and 3DGS settings.

B.1 Data Processing

During training and validation, we resize each image with bilinear interpolation so that its shorter side is 540 pixels while preserving the aspect ratio, and then randomly crop a 518×518 region. The query and its target SSIM map use identical crop coordinates to maintain pixel alignment, whereas each reference is cropped independently. During benchmark evaluation, we resize the shorter side to 518 pixels while preserving the aspect ratio and do not apply cropping. For AVS, candidate renderings and references are directly resized to 518×518. All RGB inputs are normalized using the ImageNet mean and standard deviation.

B.2 Architecture

Feature fusion and positional encoding. DeepCR extracts frozen DINOv2-S/14 features (Oquab et al., 2024) from layers 4, 8, and 12. A layer-specific linear projection maps each 384-dimensional feature to 128 dimensions. Their concatenation gives $\mathbf{F}_q$ as defined in the main paper; a learned affine layer $\psi : \mathbb{R}^{384} \to \mathbb{R}^{384}$ then fuses its channels at each patch location. These projections and ψ are shared across query and reference images. The attention inputs are

$$\mathbf{X}_0 = \psi(\mathbf{F}_q) + \mathbf{E}_q, \qquad \mathbf{Z} = \mathrm{Concat}_{i \in \mathcal{R}_q} \left[\psi(\mathbf{F}_i) + \mathbf{E}_i \right], \tag{10}$$

where reference concatenation is along the token dimension. The positional embeddings $\mathbf{E}_q$ and $\mathbf{E}_i$ come from the same learned two-dimensional grid, bilinearly interpolated to the patch grid and repeated for each reference. They encode within-image positions; no separate view-identity embedding is added. For 518×518 inputs, $\mathbf{X}_0$ has 37×37 tokens and $\mathbf{Z}$ has $K \times 37 \times 37$ tokens.

Cross-view attention. Two Transformer decoder layers update the query tokens while retaining the reference memory $\mathbf{Z}$. Each layer uses eight attention heads, model dimension 384, a $384 \to 768 \to 384$ feed-forward network (FFN) with ReLU, and dropout of 0.1. With layer index $j \in \{0, 1\}$, the residual updates use post-normalization:

$$\mathbf{U}_j = \mathrm{LN}_{j,1}(\mathbf{X}_j + \mathrm{Drop}(\mathrm{SA}_j(\mathbf{X}_j))), \tag{11}$$

$$\mathbf{V}_j = \mathrm{LN}_{j,2}(\mathbf{U}_j + \mathrm{Drop}(\mathrm{CA}_j(\mathbf{U}_j, \mathbf{Z}))), \tag{12}$$

$$\mathbf{X}_{j+1} = \mathrm{LN}_{j,3}(\mathbf{V}_j + \mathrm{Drop}(\mathrm{FFN}_j(\mathbf{V}_j))). \tag{13}$$

Self-attention (SA) derives queries, keys, and values from the query tokens. In each cross-attention (CA) head, queries come from $\mathbf{U}_j$, while keys and values come from $\mathbf{Z}$:

$$\mathrm{Attn}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{48}} + \mathbf{B}\right)\mathbf{V}. \tag{14}$$

Here $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ are learned affine projections with head dimension 48. Head outputs are concatenated and linearly projected to 384 dimensions. Attention-weight dropout and FFN hidden-layer dropout also use probability 0.1 during training. The mask $\mathbf{B}$ is $-\infty$ for tokens of padded reference views and zero elsewhere. There is no causal or geometric attention mask: poses are used for reference selection, and cross-view associations are learned from image features. The output $\mathbf{X}_2$ retains one token per query patch, allowing it to be reshaped to the query grid for decoding.

Spatial decoding. The pixel decoder applies two 3×3 convolutional blocks with 128 hidden channels and LeakyReLU activations. Its main branch predicts 196 channels and applies PixelShuffle with factor 14, mapping a 37×37 grid to a one-channel 518×518 quality map. The auxiliary branch predicts four channels and applies PixelShuffle with factor 2, producing a 74×74 map. A sigmoid constrains both outputs to $[0, 1]$. Full resolution refers to the preprocessed query dimensions, rather than the native source-image dimensions. Auxiliary predictions provide training supervision; only the main map determines reported quality scores. The supplied forward pass computes both outputs.

Pose-selection edge cases. Normalization uses all valid candidates before ranking. A valid candidate has an available image and pose and finite translation and rotation distances; padded `empty_image` entries are excluded. For each distance dimension, a median greater than 10^{-8} is used directly. Otherwise, the scale is the smallest strictly positive valid distance; if none exists, that dimension contributes zero to every score. If the query pose is missing or no valid candidate remains, selection falls back to random sampling without replacement, or all candidates when the pool has at most K entries. IQA batches pad short reference sets and mask their tokens; AVS uses up to K valid references. The aligned ground-truth image is excluded from the candidate pool before selection. Pose proximity is a relevance prior and does not guarantee visual overlap under occlusion or diversity among selected views.

B.3 Training Configuration and Environment

We train DeepCR for five epochs on two NVIDIA GeForce RTX 5090 D GPUs, each with 32 GB of memory, using a per-GPU batch size of 24. We optimize the trainable parameters with AdamW (Loshchilov & Hutter, 2019) using a fixed learning rate of 5×10^{-4} throughout training. We use 16-bit mixed precision, clip the gradient norm at 1.0, and set the global random seed to 1. The experiments run on Ubuntu 22.04.5 LTS with an Intel Xeon Silver 4314 CPU, two NVIDIA GeForce RTX 5090 D GPUs, CUDA 12.8, cuDNN 9.13.0, Python 3.10, and PyTorch 2.1.2.

The main prediction is supervised by the mean absolute error between the predicted full-resolution quality map and the target SSIM map. The auxiliary target is obtained by bilinearly resizing the same SSIM map to 74×74. Its loss is also a pixel-level mean absolute error. The initial auxiliary loss weight is $\lambda_0 = 0.3$ and is decayed according to the cosine schedule described in the main paper,

so that optimization gradually shifts from coarse spatial guidance to the full-resolution objective. Unless stated otherwise, the pose-aware selector uses $K = 5$ references with the query-adaptive normalization described in Section 3.2.

B.4 AVS AND 3DGS CONFIGURATION

Following FisherRF (Jiang et al., 2024), every reconstruction begins with four training views and runs for 30,000 iterations. We add one view at iterations 400, 900, 1,500, 2,200, 3,000, 3,900, 4,900, 6,000, 7,200, 8,500, 9,900, 11,400, 13,000, 14,700, 16,500, and 18,400, yielding 20 training views in total. All methods share the same initial views, candidate pool, acquisition budget, and evaluation views. DeepCR is used without fine-tuning, and we compare against FisherRF, Auto3R (Shen et al., 2026), GauSS-MI (Xie et al., 2025), and the NR-IQA and CR-IQA baselines.

At each acquisition step, the current 3DGS model renders every candidate view, and each view-selection method ranks the candidates according to its own criterion. IQA-based methods acquire the candidate with the lowest predicted quality, accounting for each method's score direction. For DeepCR, each candidate rendering serves as the query, while the current real training observations form the reference pool; the pose-aware module selects up to $K = 5$ references. After a method selects a candidate, the corresponding real image is added to the training set, the selected pose is removed from the candidate pool, and 3DGS is optimized with the expanded training set. No aligned candidate image is available before selection.

The held-out evaluation views remain fixed across acquisition rounds and are disjoint from both the training views and candidate pool. We set the global random seed to 0 for all AVS experiments, covering Python, NumPy, and PyTorch. We perform reconstruction with the official 3DGS codebase (Kerbl et al., 2023) as integrated into the FisherRF implementation. All methods use identical 3DGS optimization settings and procedures, including the acquisition schedule and total iteration budget.

C EVALUATION METRICS

This section describes the metrics used in the IQA and AVS experiments. PSNR, SSIM, and LPIPS evaluate the fidelity of rendered images against aligned ground truth and quantify the quality of the final 3DGS reconstructions in AVS, while PLCC and SRCC measure the agreement between predicted image-level quality scores and ground-truth SSIM scores.

C.1 RECONSTRUCTION QUALITY METRICS

For a rendered RGB image I and its aligned ground truth G, both represented in $[0,1]^{H\times W\times 3}$, the mean squared error is

$$\mathrm{MSE}(I,G) = \frac{1}{3HW}\sum_{c=1}^{3}\sum_{y=1}^{H}\sum_{x=1}^{W}\big(I_{yxc} - G_{yxc}\big)^2. \tag{15}$$

Peak signal-to-noise ratio (PSNR) (Huynh-Thu & Ghanbari, 2008) is then

$$\mathrm{PSNR}(I,G) = -10\log_{10}\mathrm{MSE}(I,G). \tag{16}$$

Our PyTorch implementation computes MSE directly over all RGB pixels without border cropping or luminance conversion. PSNR is reported in decibels, with higher values indicating better pixel-level fidelity.

Structural similarity (SSIM) (Wang et al., 2004) evaluates local luminance, contrast, and structure. For corresponding local windows x and y from I and G,

$$\mathrm{SSIM}(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}. \tag{17}$$

We implement SSIM in PyTorch using an 11×11 Gaussian window with standard deviation 1.5, $C_1 = 0.01^2$, and $C_2 = 0.03^2$. Grouped convolution processes the three RGB channels independently with a padding of five pixels. Channel averaging yields a spatial map, processed according

to the convention in Section 3.4. Spatial averaging of the resulting map gives the image-level SSIM score used for evaluation. Higher SSIM indicates stronger structural agreement, making it suitable both as the IQA target and as a reconstruction metric.

Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al., 2018) measures distance in a learned feature space:

$$\begin{aligned} \mathrm{LPIPS}(I, G) &= \sum_l \frac{1}{H_l W_l} \sum_{h,w} \Big\| \mathbf{w}_l \odot \Delta_l(h, w) \Big\|_2^2, \\ \Delta_l(h, w) &= \hat{\phi}_l(I)_{hw} - \hat{\phi}_l(G)_{hw}, \end{aligned} \tag{18}$$

where $\hat{\phi}_l$ denotes channel-normalized features and $\mathbf{w}_l$ denotes learned channel weights. For the LPIPS baseline in the IQA correlation experiments, we use the official PyTorch `lpips` implementation with an AlexNet backbone. For AVS reconstruction evaluation, we use the vendored `lpipsPyTorch` implementation of LPIPS v0.1 with a VGG16 backbone, matching the FisherRF evaluation code. Lower LPIPS indicates greater perceptual similarity. As a learned perceptual metric, LPIPS complements the pixel-level and structural measurements provided by PSNR and SSIM.

For AVS, we evaluate the final reconstruction of each scene on its fixed set of held-out views and average the per-view metric values to obtain a scene-level result. We then report the mean over the eight Blender scenes or nine Mip-NeRF360 scenes. The same evaluation procedure is applied to every view-selection method.

C.2 Correlation Metrics for IQA

Let $\{\hat{s}_i\}_{i=1}^n$ be the predicted image-level quality scores and $\{s_i\}_{i=1}^n$ the ground-truth image-level SSIM scores for the n evaluated queries in a benchmark. For a method that predicts a quality map $\hat{\mathbf{M}}_i \in \mathbb{R}^{H \times W}$, we obtain its predicted image-level score by spatial averaging:

$$\hat{s}_i = \frac{1}{HW} \sum_{y=1}^{H} \sum_{x=1}^{W} \hat{\mathbf{M}}_i(y, x). \tag{19}$$

For DeepCR, $\hat{\mathbf{M}}_i$ is trained to approximate the SSIM map defined above, so $\hat{s}_i$ represents a predicted image-level SSIM score rather than a second computation of ground-truth SSIM. The Pearson linear correlation coefficient (PLCC) (Pearson, 1895) is

$$\mathrm{PLCC} = \frac{\sum_{i=1}^n (\hat{s}_i - \bar{\hat{s}})(s_i - \bar{s})}{\sqrt{\sum_{i=1}^n (\hat{s}_i - \bar{\hat{s}})^2} \sqrt{\sum_{i=1}^n (s_i - \bar{s})^2}}, \tag{20}$$

where $\bar{\hat{s}}$ and $\bar{s}$ are sample means. PLCC measures linear agreement with the target quality scale.

The Spearman rank-order correlation coefficient (SRCC) (Spearman, 1904) applies Pearson correlation to the ranks:

$$\mathrm{SRCC} = \mathrm{PLCC}\big(\mathrm{rank}(\hat{\mathbf{s}}), \mathrm{rank}(\mathbf{s})\big). \tag{21}$$

SRCC measures whether a method preserves the ordering of rendering quality and is invariant to strictly increasing transformations of the score scale; a strictly decreasing transformation reverses the sign of the correlation. We compute PLCC and SRCC with `scipy.stats.pearsonr` and `scipy.stats.spearmanr`, respectively, using SciPy 1.11.4 without nonlinear score fitting. SciPy assigns average ranks to tied observations.

For each of MFR, RE10K, and T&T, we concatenate the image-level predictions and targets of all evaluated queries and compute each correlation coefficient once. The reported values are therefore global benchmark-level correlations rather than averages of per-scene correlations. The average magnitude in the main paper is

$$\mathrm{AvgAbs}(r) = \frac{1}{3} \left(|r_{\mathrm{MFR}}| + |r_{\mathrm{RE10K}}| + |r_{\mathrm{T\&T}}| \right), \tag{22}$$

where r is PLCC or SRCC. A larger average magnitude indicates stronger overall correlation across the three IQA benchmarks.

Dense-map errors. For each query, we average absolute pixel errors between its predicted map and SSIM target at a common spatial resolution, then average these per-image errors within each benchmark. MAE measures agreement with the target values and therefore depends on score calibration as well as spatial structure. Comparisons with methods trained for different quality targets should be interpreted in this light; DeepCR and CrossScore both learn SSIM-map prediction.

Within-scene human-annotation agreement. On annotated-3DGS-artifacts, we evaluate 12 scenes at retained-view ratios of $60/70/80\%$. Binary artifact judgments are aggregated and converted to quality-oriented annotation maps before spatial averaging. Let p_i and h_i denote the mean predicted quality and mean human-annotation value for image i, and let $\mathcal{J}_s$ index images from scene s. We center each score separately within its scene:

$$\widetilde{p}_i = p_i - \frac{1}{|\mathcal{J}_{s(i)}|} \sum_{k \in \mathcal{J}_{s(i)}} p_k, \qquad \widetilde{h}_i = h_i - \frac{1}{|\mathcal{J}_{s(i)}|} \sum_{k \in \mathcal{J}_{s(i)}} h_k. \tag{23}$$

We concatenate these residuals across scenes and compute

$$\text{wi-PLCC} = \text{Pearson}(\widetilde{\mathbf{p}}, \widetilde{\mathbf{h}}), \qquad \text{wi-SRCC} = \text{Spearman}(\widetilde{\mathbf{p}}, \widetilde{\mathbf{h}}). \tag{24}$$

Thus, ranks are computed after centering; the reported coefficients are not averages of per-scene correlations. Centering removes differences in scene-wise mean scores, assessing agreement with image-level quality variations within scenes. This protocol does not quantify pixel-level agreement with human artifact locations.

D Compared Methods

This section introduces the full-reference, no-reference, and cross-reference IQA baselines, followed by the methods used for AVS comparison. Trainable CR-IQA baselines are retrained on the same MFR training split using their original objectives and official configurations. All methods use the same input resolution, and spatial predictions are evaluated at the same output resolution.

D.1 Full-Reference IQA Methods

Full-reference methods compute quality using the aligned ground-truth image corresponding to each rendered query; this image is unavailable to DeepCR during inference. PSNR compares pixel-level reconstruction errors, SSIM measures local structural agreement, and LPIPS measures distance in learned perceptual feature spaces. Since the ground-truth target of the IQA benchmark is image-level SSIM, SSIM has a correlation of one by construction. PSNR and LPIPS show how other aligned fidelity measures agree with this target. LPIPS is a distance, so better reconstruction quality corresponds to a lower score and typically produces a negative correlation with SSIM.

D.2 No-Reference IQA Methods

BRISQUE (Mittal et al., 2012) and NIQE (Mittal et al., 2013) quantify deviations from natural-scene statistics without using a reference image. PIQE (Venkatanath et al., 2015) estimates perceptual quality from locally distorted regions. PaQ-2-PiQ (Ying et al., 2020) predicts patch-level quality and aggregates local estimates into an image-level score. These methods are applicable to arbitrary query viewpoints, but they cannot use scene-specific evidence to distinguish genuine content from reconstruction artifacts. We do not include GSOQA (Wan et al., 2026) in the quantitative comparison because it operates directly on native 3D Gaussian primitives and predicts reconstruction-level perceptual quality, whereas our IQA benchmark evaluates individual rendered images from image-space inputs.

D.3 Cross-Reference IQA Methods

CrossScore (Wang et al., 2024) encodes the query and multiple non-aligned reference views with DINOv2, associates their features through cross-attention, and regresses a dense quality map supervised by pixel-level SSIM. PuzzleSim (Hermann et al., 2025) builds a scene-specific distribution of

Table S2: Scene-wise IQA correlations, reported as mean ± standard deviation across scenes. Correlations are computed independently within each scene against ground-truth SSIM. The best results are shown in **bold**.

IQA Type	Method	MFR		RE10K		T&T	
		PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
NR-IQA	NIQE	-0.141 ± 0.259	-0.331 ± 0.264	-0.147 ± 0.330	-0.261 ± 0.327	-0.130 ± 0.256	-0.512 ± 0.166
	PIQE	-0.480 ± 0.157	-0.357 ± 0.210	-0.213 ± 0.312	-0.143 ± 0.316	-0.401 ± 0.117	-0.510 ± 0.148
	BRISQUE	-0.445 ± 0.202	-0.414 ± 0.200	-0.212 ± 0.326	-0.153 ± 0.335	-0.325 ± 0.164	-0.383 ± 0.204
	PaQ-2-PiQ	0.392 ± 0.161	0.343 ± 0.160	0.567 ± 0.197	0.605 ± 0.253	0.637 ± 0.099	0.691 ± 0.113
CR-IQA	PR-IQA	0.084 ± 0.158	0.082 ± 0.162	0.347 ± 0.320	0.302 ± 0.298	0.344 ± 0.206	0.353 ± 0.211
	CrossScore	0.804 ± 0.126	0.783 ± 0.129	0.921 ± 0.104	0.877 ± 0.110	0.907 ± 0.062	0.896 ± 0.071
	PuzzleSim	0.720 ± 0.141	0.700 ± 0.158	0.898 ± 0.101	0.883 ± 0.113	0.754 ± 0.161	0.803 ± 0.097
	DeepCR	$\mathbf{0.876} \pm 0.071$	$\mathbf{0.871} \pm 0.075$	$\mathbf{0.948} \pm 0.082$	$\mathbf{0.926} \pm 0.093$	$\mathbf{0.954} \pm 0.043$	$\mathbf{0.958} \pm 0.033$

perceptual patches from the reference views. For each query patch, it finds the most similar reference patch and combines the resulting layer-wise similarity maps to localize regions unsupported by the observed scene. PR-IQA (Choi et al., 2026) first derives a partial quality map from geometrically overlapping query-reference regions and then uses reference-conditioned cross-attention to infer quality in the remaining regions, producing a dense quality map. These methods share DeepCR's inference setting: they access non-aligned observations from the same scene but not the aligned ground-truth image. We exclude NOVA (Ghildyal et al., 2026) because its public release does not provide the training code or training-pair construction procedure required to retrain it on the same MFR split.

D.4 Active View Selection Methods

FisherRF (Jiang et al., 2024) selects views according to information gain approximated through Fisher information. GauSS-MI (Xie et al., 2025) uses mutual information to quantify the expected utility of candidate observations. Auto3R (Shen et al., 2026) provides a learned reconstruction signal for active acquisition. We additionally convert every NR-IQA and CR-IQA baseline into a view selector by evaluating each candidate rendering and acquiring the candidate with the lowest predicted quality, accounting for each method's score direction. All selectors receive the same initial observations, candidate poses, acquisition budget, and held-out evaluation views.

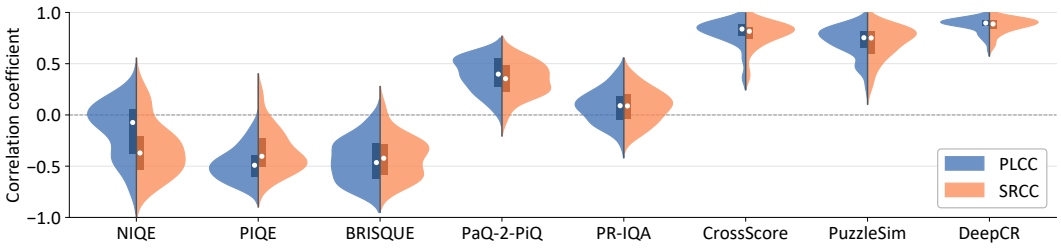
E Scene-Wise IQA Results

This section complements the global benchmark-level correlations in the main paper with scene-wise statistics. For each scene, we compute PLCC and SRCC over its query images and then summarize the resulting correlations across scenes. These statistics characterize variation across scene content and reconstruction conditions, reducing reliance on pooled benchmark-level correlations.

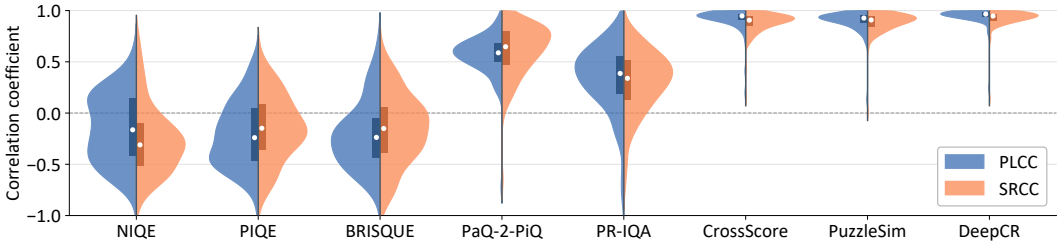
E.1 Scene-Wise Statistical Summary

Table S2 reports the mean and standard deviation of the scene-wise correlations. All 47 MFR test scenes, 100 RE10K scenes, and 21 T&T scenes yield valid PLCC and SRCC values and are included. The correlations retain their signs because the score direction differs across IQA methods; for example, distortion scores may decrease as quality increases. These scene-wise means give equal weight to every scene and therefore should not be interpreted as recomputations of the global correlations reported in the main paper, which pool all query images before computing each coefficient.

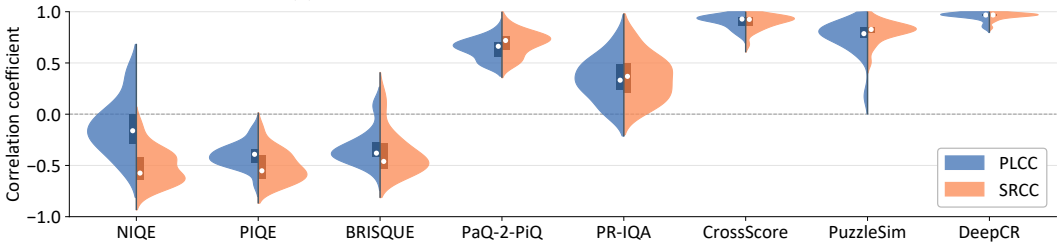
DeepCR achieves the highest mean scene-wise PLCC and SRCC on all three datasets. Compared with the strongest competing CR-IQA result for each metric, it improves PLCC and SRCC by approximately 7 and 9 percentage points on MFR, 3 and 4 percentage points on RE10K, and 5 and 6 percentage points on T&T, respectively. DeepCR also has the smallest standard deviation in every dataset–metric combination. These are descriptive statistics across scenes, rather than confidence intervals or estimates of variation across training seeds.



(a) Scene-wise correlation distributions on MFR



(b) Scene-wise correlation distributions on RE10K



(c) Scene-wise correlation distributions on T&T

Figure S3: Distributions of scene-wise PLCC (blue, left half) and SRCC (orange, right half) on the three IQA benchmarks. Each violin contains one correlation coefficient per scene, with white dots denoting medians and thick bars denoting interquartile ranges. Higher positive correlations indicate stronger agreement with ground-truth SSIM for methods whose scores increase with quality.

E.2 Distribution Analysis

Figure S3 presents the distributions underlying Table S2. DeepCR has high medians and comparatively narrow spreads across all three datasets, supporting consistency across the evaluated scenes. CrossScore and PuzzleSim also produce predominantly positive correlations, with greater scene-to-scene variation, particularly for PuzzleSim on T&T. The zero-shot natural-image NR-IQA baselines generally show lower correlation magnitudes; their performance also reflects the mismatch between their original domain and 3DGS renderings.

F Additional Experimental Results

F.1 Reference-Count Sensitivity

We additionally evaluate the sensitivity to the number of references selected by the pose-aware strategy. Table S3 reports image-level PLCC and SRCC for $K \in \{1, 3, 5, 7, 9\}$. Performance remains relatively stable across this range, while the best setting varies by dataset and metric. Smaller reference sets are competitive on MFR and RE10K, whereas $K = 9$ performs best on T&T. The non-monotonic trend suggests that additional views can provide useful evidence but may also introduce redundant or weakly related content. The default $K = 5$ therefore offers a practical balance between prediction quality and the cost of feature extraction and cross-view reasoning.

Table S3: Sensitivity to the number of selected references K.

Number	MFR		RE10K		T&T	
	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
$K = 1$	0.8895	0.8727	**0.9182**	0.9326	0.9275	0.9388
$K = 3$	**0.8959**	0.8826	0.9177	0.9335	0.9415	0.9528
$K = 5$	0.8956	**0.8830**	0.9170	0.9326	0.9449	0.9557
$K = 7$	0.8954	0.8829	0.9168	0.9321	0.9467	0.9573
$K = 9$	0.8935	0.8812	0.9168	**0.9343**	**0.9543**	**0.9662**

Table S4: Inference runtime normalized by DeepCR's total inference runtime on the same benchmark. Lower values indicate shorter inference runtime.

Category	Method	Relative runtime ↓
NR-IQA	PIQE	0.106×
	BRISQUE	0.174×
	PaQ-2-PiQ	0.335×
	NIQE	0.797×
CR-IQA	**DeepCR**	**1.000**×
	CrossScore	1.008×
	PR-IQA	2.886×
	PuzzleSim	2.911×

F.2 Inference Runtime

This subsection compares the inference runtime of the IQA methods. We measure total inference runtime on the same benchmark under identical hardware, input resolution, batch size, warm-up, and timing settings, handle model loading and disk I/O consistently across methods, and normalize each result by DeepCR's inference runtime. Because every method processes the same set of queries, these ratios are equivalent to the relative mean inference runtime per query. Table S4 reports the normalized runtime of each method.

Among the evaluated CR-IQA methods, DeepCR achieves the highest IQA correlations, with inference runtime comparable to CrossScore and substantially lower than PR-IQA and PuzzleSim. Specifically, PR-IQA and PuzzleSim take approximately $2.89\times$ and $2.91\times$ as long as DeepCR at inference, respectively. The NR-IQA methods are generally faster because they process only the query image, but their lower computational cost is accompanied by substantially weaker correlations with ground-truth quality. Thus, DeepCR combines the strongest IQA performance among the evaluated CR-IQA methods with inference runtime comparable to CrossScore.