PR-IQA: Partial-Reference Image Quality Assessment for Diffusion-Based Novel View Synthesis

Supplementary Material

This supplementary material complements the main paper by providing comprehensive implementation details, extended experimental results, and in-depth ablation studies. Sections 1 and 2 establish the foundation for reproducibility by detailing the network architecture, loss functions, dataset generation protocols, and training configurations. We expand our experimental analysis in Section 3 to cover alternative FR-IQA targets (PSNR, LPIPS) and validate the reliability of image-level view selection. Furthermore, Sections 4 and 5 present systematic ablation studies concerning IQA design choices (e.g., reference count, fusion strategy, geometric robustness) and 3DGS parameters (e.g., guidance metric, masking threshold, soft vs. binary masking), respectively. Section 6 provides extensive qualitative visualizations for both quality map estimation and 3D reconstruction results. Finally, Section 7 discusses the limitations of the proposed method and outlines potential future directions.

1. Method Details

1.1. Architecture Details

As illustrated in Fig. 1, our architecture adopts a U-Net-like [14] encoder-decoder design, leveraging DINOv2 [12] as the feature backbone. The network utilizes GELU [8] as the activation function throughout all layers. Detailed specifications, including resolution, channel dimensions, and the number of blocks for each level, are summarized in Table 1.

The encoder is structured into four stages with $[2, 3, 3, 4]$ encoding blocks and $[1, 2, 4, 8]$ attention heads, respectively. The encoders for the query and reference branches share weights, whereas the encoder for the partial branch remains independent. The channel dimensions scale progressively as $[48, 96, 192, 384]$ from Level 1 to Level 4.

To effectively integrate information across branches, we employ a ConvFuse operation at each encoding stage. Specifically, the feature maps from the query and partial branches are concatenated along the channel dimension and then projected back to the original channel size via a convolutional layer. The resulting fused features serve as the input for the subsequent stage of the query branch, while the partial branch retains its original, unfused features for its own propagation.

The decoder consists of three stages containing $[3, 3, 2]$ decoding blocks and $[4, 2, 1]$ attention heads. Corresponding to the encoder levels, the decoder maintains channel widths of 192, 96, and 96, respectively, following skip connection fusion and 1×1 channel reduction.

The resulting model comprises approximately 60M trainable parameters. In terms of resource consumption, it is highly efficient, requiring approximately 2 GB of GPU memory for single-image inference and 6 GB for training with a batch size of 1.

1.2. Loss Functions

To ensure robust performance, our training objective combines three complementary loss terms: the pixel-wise $\mathcal{L}_1$ loss for local accuracy, the Jensen-Shannon Divergence (JSD) [5] loss for global distributional alignment, and the Pearson Linear Correlation Coefficient (PLCC) [3] loss for ranking consistency.

Distribution Alignment (JSD Loss). We employ the JSD loss to align the global distribution of predicted quality scores with the GT, thereby preventing mode collapse where the network predicts overly uniform values. We first flatten the quality maps $\hat{Q}$ and Q into vectors $\hat{\mathbf{p}}, \mathbf{g} \in [0, 1]^N$, where $N = H \times W$. Since $\hat{\mathbf{p}}$ and $\mathbf{g}$ are bounded, we first apply a logit transformation to map them into an unbounded space suitable for softmax normalization:

$$\tilde{p}_i = \log\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right), \quad \tilde{g}_i = \log\left(\frac{g_i}{1-g_i}\right). \tag{1}$$

Next, we convert these logits into probability distributions P and G using a temperature-scaled softmax function:

$$P_i = \frac{\exp(\tilde{p}_i/\tau)}{\sum_j \exp(\tilde{p}_j/\tau)}, \quad G_i = \frac{\exp(\tilde{g}_i/\tau)}{\sum_j \exp(\tilde{g}_j/\tau)}. \tag{2}$$

where τ is the temperature parameter, empirically set to 0.2. The symmetric JSD loss is then defined as the average Kullback-Leibler (KL) divergence from the mixture distribution $M = (P + G)/2$:

$$\mathcal{L}_{\text{JSD}} = \frac{1}{2}\mathcal{D}_{\text{KL}}(P \parallel M) + \frac{1}{2}\mathcal{D}_{\text{KL}}(G \parallel M). \tag{3}$$

This prevents mode collapse by penalizing uniform predictions: when the network predicts similar quality values everywhere, P becomes nearly uniform, resulting in a large JSD loss against the typically non-uniform ground truth (GT) G.

Ranking Consistency (Pearson Loss). To strictly enforce the relative ranking of quality, we utilize the PLCC

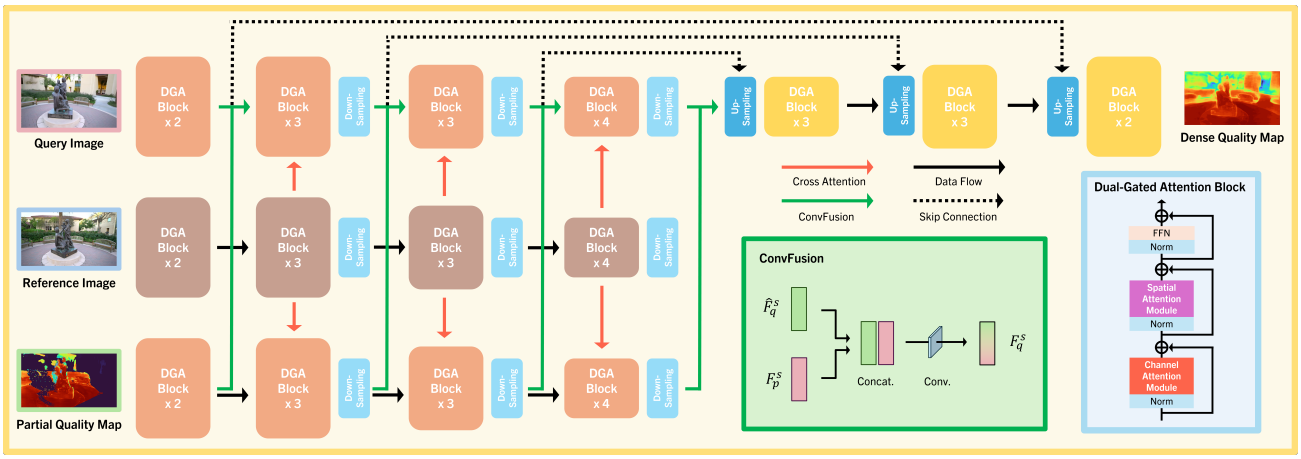


Figure 1. Detailed architecture of the proposed model. The network employs an encoder–decoder design featuring cross- and self-attention modules, query fusion, and mask-aware pixel-shuffle downsampling. Key specifications, including stage-wise block counts, attention heads, and the status of component sharing (frozen vs. trainable), are explicitly annotated.

Table 1. Detailed architecture specifications of the proposed PR-IQA network. We report the spatial resolution, channel dimensions, number of attention heads, and block counts for each stage of the encoder (Enc0- Enc3) and decoder (Dec3- Dec1).

Level	Resolution	Channels	Heads	Blocks	Output Channels
Input	224 x 224	4	-	-	-
Enc0	224 x 224	48	1	2	48
Enc1	112 x 112	96	2	3	96
Enc2	56 x 56	192	4	3	192
Enc3	28 x 28	384	8	4	384
Dec3	56 x 56	192	4	3	192
Dec2	112 x 112	96	2	3	96
Dec1	224 x 224	96	1	2	96
Output	224 x 224	1	-	-	1

loss. Let $\hat{\mathbf{y}}$ and $\mathbf{y}$ denote the flattened predicted and GT quality maps, respectively. We first center these vectors by subtracting their means $(\mu_{\hat{y}}, \mu_y)$. The correlation coefficient r is computed as:

$$r = \frac{\sum(\hat{y}_i - \mu_{\hat{y}})(y_i - \mu_y)}{\sqrt{\sum(\hat{y}_i - \mu_{\hat{y}})^2}\sqrt{\sum(y_i - \mu_y)^2}}. \tag{4}$$

The Pearson loss is defined as $\mathcal{L}_{\text{PLCC}} = 1 - r$. This term complements the pixel-wise $\mathcal{L}_1$ loss by focusing on linear trends and the relative ordering of salient regions, crucial for accurate quality assessment and downstream tasks, rather than solely minimizing absolute pixel errors.

2. Experimental Details

2.1. Training Data Generation

Frame Sampling. We utilize the Map-free Visual Relocalization (MFR) dataset [1] as our primary source. For each scene, we uniformly sample 200 frames along the camera trajectory, explicitly including the start and end frames.

Table 2. List of evaluation scenes. We enumerate the specific scenes and sequence IDs selected from the Mip-NeRF 360, Tanks and Temples, and RealEstate10K datasets used for our experimental benchmarks.

Dataset	Scene					
Mip-NeRF 360	Bonsai	Counter	Garden	Kitchen	Room	Treehill
Tanks and Temples	Barn	Caterpillar	Family	Horse	Ignatius	Truck
RealEstate10K	87f03b8928fc286e		d932fa3862974507		9ea61697c238be3d	
	7bab7b21dbaf38ab		2e7ffcba51990c93		f48829b917629fe0	

This uniform sampling strategy serves two purposes: it reduces the computational overhead for the Video Diffusion Model (VDM) [21] and prevents redundancy by mitigating negligible pose changes between adjacent frames.

View Synthesis and Distortion. Following the ViewCrafter protocol [21], we organize the sampled frames into sliding windows of size 25. Within each window, two anchor images are used to synthesize novel views. It is a known characteristic of VDMs that generation fidelity degrades as the target viewpoint deviates further from the conditioning camera poses. We explicitly leverage this property to induce a diverse spectrum of realistic artifacts and geometric distortions in the generated images. This strategy enriches our training distribution with challenging samples, thereby enhancing the model's robustness to reconstruction errors.

Reference Selection and Annotation. To ensure sufficient baseline separation and avoid trivial correlations from high-overlap pairs, we systematically select reference frames relative to the query. For a given query frame I_q, we identify four reference candidates $\{I_r\}$ at relative indices of ± 10 and ± 20 within the sampled sequence. For each resulting query-reference pair (I_q, I_r), we generate pseudo-

Table 3. Quantitative comparisons of predicted quality maps against GT quality maps (PLCC↑, SRCC↑), targeting PSNR and LPIPS. Red, orange, and yellow cells denote the 1st, 2nd, and 3rd best methods per column (excluding FR settings †), while gray cells indicate identity cases where the IQA prediction matches the GT quality map.

IQA Type	IQA Method	Mip-NeRF 360				Tanks and Temples				RealEstate10K			
		PSNR		LPIPS		PSNR		LPIPS		PSNR		LPIPS	
		PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
FR-IQA	PSNR†	1.000	1.000	0.434	0.384	1.000	1.000	0.478	0.459	1.000	1.000	0.370	0.347
	SSIM†	0.517	0.487	0.565	0.554	0.486	0.487	0.598	0.595	0.392	0.386	0.452	0.460
	LPIPS†	0.434	0.384	1.000	1.000	0.478	0.459	1.000	1.000	0.370	0.347	1.000	1.000
	DINOv2†	0.407	0.338	0.557	0.472	0.396	0.361	0.582	0.581	0.248	0.241	0.489	0.516
NR-IQA	PAL4VST	0.016	0.016	0.024	0.021	0.004	0.004	0.004	0.004	0.013	0.012	0.078	0.074
	PaQ-2-PiQ	-0.179	-0.181	-0.047	-0.047	-0.136	-0.095	0.007	0.053	-0.126	-0.134	0.029	0.030
	PIQE	-0.110	-0.114	0.031	0.035	0.223	0.242	0.194	0.208	0.227	0.235	0.047	0.062
CR-IQA	MEt3R*	0.056	0.055	0.057	0.042	0.106	0.120	0.181	0.196	0.125	0.117	0.363	0.352
	CrossScore	0.082	0.081	0.224	0.238	0.206	0.182	0.312	0.304	0.195	0.149	0.169	0.161
	PuzzleSim	0.179	0.172	0.286	0.264	0.250	0.259	0.456	0.433	0.208	0.200	0.458	0.447
	$\text{Ours}_{\text{partial}}$*	0.161	0.184	0.189	0.173	0.131	0.134	0.225	0.256	0.070	0.150	0.208	0.298
	$\text{Ours}_{\text{DINOv2}}$	0.259	0.227	0.280	0.229	0.273	0.258	0.401	0.384	0.206	0.215	0.304	0.333
	$\text{Ours}_{\text{SSIM}}$	0.338	0.345	0.235	0.229	0.340	0.334	0.340	0.334	0.284	0.244	0.171	0.175

† Metrics require a same-pose GT image. * Metrics are computed only over the valid overlapping region.

ground-truth supervision by applying the procedure described in the Partial Map Generation section (Sect. 3.3 of the main manuscript). This involves estimating global point clouds via dense stereo matching, performing z-buffered reprojection to align views, and finally computing the partial quality map $\hat{Q}$.

Data Structure. Consequently, training samples are formed as tuples $(I_q, I_r, \hat{Q}, Q^*)$, where Q^* represents the GT quality map. This structure enables a systematic evaluation of robustness in the CR-IQA setting.

2.2. Evaluation Data Generation

Dataset Selection. We conduct our evaluation across three standard benchmarks: Mip-NeRF 360, Tanks and Temples, and RealEstate10K. The specific scenes selected for these experiments are listed in Table 2 (with RealEstate10K sequences indexed as Scenes 1–6). To ensure consistency, we employ the identical set of scenes for both the standalone IQA performance assessment and the downstream IQA-guided 3DGS experiments.

Query Image Synthesis. To generate the synthesized query images used for evaluation, we adopt a standardized pipeline. We utilize the sequence endpoints (i.e., the first and last frames) as the reference views. For the intermediate target frames, we employ DUSt3R [17] to estimate dense point clouds by matching the endpoints with the current frame. These point clouds are subsequently rendered into the target viewpoint and processed by the VDM to refine the details, producing the final query images.

2.3. Model Training

All input images are resized to a resolution of 294×518. The model is trained using the AdamW optimizer [11] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, starting with an initial learning rate of 1×10^{-4}. We employ a Cosine Annealing with Warm Restarts schedule [11], where the learning rate decays to 1×10^{-6} with a restart period of 135,000 iterations. The entire training process spans 270,000 iterations (approximately 20 hours) on four NVIDIA RTX 3090 GPUs, utilizing a total batch size of 12 (3 frames per GPU).

2.4. Baseline Details

We compare our method against a comprehensive set of baselines across three categories: Full-Reference (FR), No-Reference (NR), and Cross-Reference (CR) IQA.

- **FR-IQA:** We utilize PSNR and SSIM [18] as representative metrics for measuring pixel-wise reconstruction error and structural similarity, respectively. Additionally, LPIPS [23] is employed to assess perceptual similarity based on deep feature distances extracted from pre-trained networks.
- **NR-IQA:** PAL4VST [22] is a segmentation-based model trained on pixel-level artifact masks. PaQ-2-PiQ [20] uses a ResNet-based [7] architecture to jointly learn local (patch-level) and global (image-level) quality. PIQE [15] is a training-free method that quantifies distortions, such as blur and noise, by analyzing the statistical properties of spatially active blocks.
- **CR-IQA:** MEt3R [2] evaluates multi-view consistency by using dense stereo to project DINO [4] and FeatUp [6] features into a shared 3D space, followed by cosine similarity computation. CrossScore [19] utilizes a DI-

Table 4. Image selection evaluation. We report the correlation (PLCC, SRCC) between per-image quality scores and ground-truth quality scalars derived from DINOv2 feature similarity and SSIM across three datasets. $\text{Ours}_{\text{DINOv2}}$ demonstrates strong alignment with feature-based quality, achieving the highest performance on Tanks and Temples and RealEstate10K, and competitive results on Mip-NeRF 360.

		Mip-NeRF 360				Tanks and Temples				RealEstate10K			
IQA Type	IQA Method	PLCC (DINOv2)	SRCC (DINOv2)	PLCC (SSIM)	SRCC (SSIM)	PLCC (DINOv2)	SRCC (DINOv2)	PLCC (SSIM)	SRCC (SSIM)	PLCC (DINOv2)	SRCC (DINOv2)	PLCC (SSIM)	SRCC (SSIM)
NR-IQA	PaQ-2-PiQ	0.002	0.012	-0.014	-0.009	0.113	0.112	-0.166	-0.179	0.022	0.012	0.032	0.042
	PIQE	0.047	0.044	0.348	0.347	0.075	0.075	0.329	0.385	-0.126	-0.128	-0.133	-0.118
CR-IQA	CrossScore	-0.090	-0.104	0.366	0.366	0.188	0.188	-0.097	-0.126	-0.095	-0.074	-0.035	-0.026
	PuzzleSim	0.607	0.518	0.516	0.494	0.616	0.612	0.539	0.399	0.772	0.727	0.827	0.747
	$\text{Ours}_{\text{SSIM}}$	0.186	0.164	0.629	0.620	0.278	0.287	0.457	0.511	0.595	0.600	0.666	0.684
	$\text{Ours}_{\text{DINOv2}}$	0.597	0.547	0.571	0.541	0.627	0.619	0.590	0.557	0.790	0.802	0.746	0.783

NOv2 [12] encoder with a cross-attention module to compare the query against multiple references, predicting a patch-level map approximating SSIM. PuzzleSim [9] operates in the feature space of a pre-trained network, producing a similarity map based on patch statistics learned from scene training views.

For all learning-based baselines, we use the publicly available pre-trained models without additional fine-tuning.

3. More Experimental Results

3.1. Evaluation on Alternative FR-IQA Targets

Although our Partial-Reference (PR-IQA) framework is trained to optimize DINOv2-SIM and SSIM maps, we extend our evaluation to alternative FR-IQA targets, specifically PSNR and LPIPS, to assess the generalization capability of our predicted quality maps. Table 3 summarizes the Pearson Linear Correlation Coefficient (PLCC) and Spearman Rank Correlation Coefficient (SRCC) between our predicted maps Q and the GT quality maps $Q*$ derived from these unseen metrics.

Evaluation on PSNR. As shown in Table 3, our method demonstrates robust generalization to the PSNR target, which measures pixel-level fidelity. Specifically for the PSNR target, $\text{Ours}_{\text{SSIM}}$ achieves state-of-the-art performance, ranking first across all datasets. $\text{Ours}_{\text{DINOv2}}$ also shows competitive correlations, generally outperforming other baselines. In stark contrast, NR-IQA baselines (PAL4VST, PaQ-2-PiQ, PIQE) exhibit extremely low or even negative correlations. This suggests that traditional natural-image quality predictors fail to capture the specific rendering artifacts inherent in novel view synthesis. While some CR-IQA methods like PuzzleSim show moderate success, our method proves significantly more effective at approximating the pixel-wise accuracy required for PSNR prediction.

Evaluation on LPIPS. For the LPIPS target, the CR-IQA baseline PuzzleSim generally ranks first. This performance is likely attributable to architectural bias: PuzzleSim relies on VGG features, which structurally align with the VGG backbone used in LPIPS. Despite this advantage, $\text{Ours}_{\text{DINOv2}}$ achieves highly competitive results, consistently ranking second on Mip-NeRF 360 and Tanks and Temples. This indicates that our method effectively captures perceptual quality variations even without relying on the same feature backbone as the target metric. Other CR-IQA methods (MEt3R, CrossScore) show lower correlations, and NR-IQA methods again fail to provide meaningful estimates.

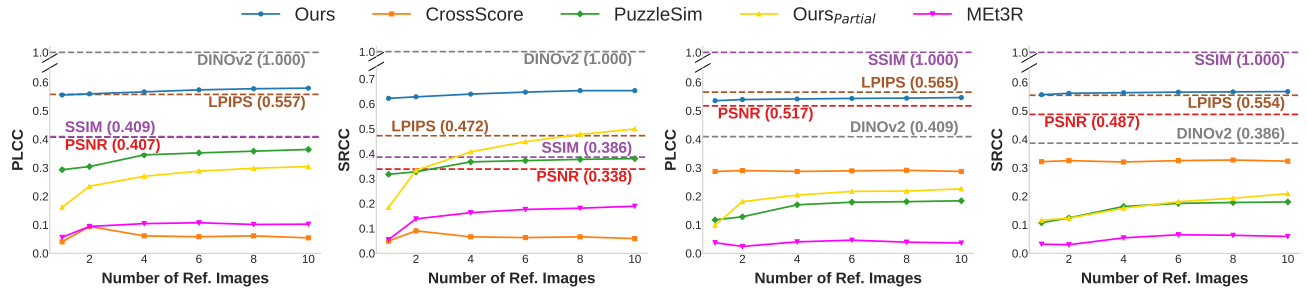
Our approach demonstrates superior generalization compared to existing methods. $\text{Ours}_{\text{SSIM}}$ and $\text{Ours}_{\text{DINOv2}}$ effectively generalize PSNR and LPIPS targets respectively, significantly outperforming NR-IQA. Furthermore, compared to CR-IQA baselines, our strategy of learning quality completion from partial references proves to be a more robust solution for estimating diverse quality metrics.

3.2. Evaluation on Image Selection for 3DGS

We employ an image-level quality score to select the optimal pseudo-ground-truth candidate from the diffusion-generated pool (in Sect. 4 of the main manuscript). In this section, we quantitatively evaluate the reliability of various IQA metrics for this selection task.

To validate whether image-level scores effectively represent semantic quality, we analyze the correlation between the predicted scores and the GT DINOv2 feature similarity maps. Specifically, for each generated image, we compute the pixel-wise cosine similarity between its DINOv2 features and those of the corresponding real image at the same pose. This dense similarity map is then spatially averaged to derive a single scalar GT score. We verify the alignment by measuring the PLCC and SRCC correlations between this scalar and the scores predicted by different IQA methods (CR-IQA and NR-IQA) across all test frames.

Table 4 presents the correlation results on three benchmark datasets. $\text{Ours}_{\text{DINOv2}}$ demonstrates robust and consistent performance across all datasets. It generally achieves the highest correlations on Tanks and Temples and RealEstate10K, significantly outperforming baselines. On Mip-NeRF 360, it remains highly competitive, showing



(a) DINOv2 target evaluation (b) SSIM target evaluation

Figure 2. Impact of the number of reference views on IQA performance. We plot the PLCC and SRCC against the number of reference images used for evaluation. FR-IQA baselines are indicated by constant horizontal lines. The results are shown for (a) DINOv2 and (b) SSIM targets. In both scenarios, our model achieves the highest correlation among learned metrics (CrossScore, PuzzleSim) and demonstrates robustness even with a single reference image.

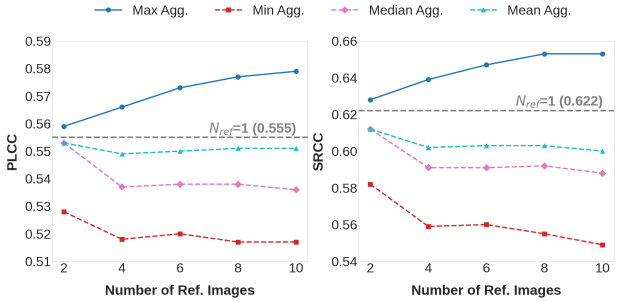


Figure 3. Impact of quality map fusion strategies on DINOv2 target evaluation. We evaluate the performance of four aggregation strategies (Max, Min, Median, and Mean) as a function of the number of reference images ($N_{\text{ref}} \in [2, 10]$).

results comparable to PuzzleSim. While PuzzleSim also exhibits strong correlations thanks to its VGG-based feature representation, our method proves to be more effective in scenarios requiring precise semantic alignment, such as RealEstate10K.

In stark contrast, NR-IQA methods (PaQ-2-PiQ, PIQE) exhibit weak or near-zero correlations across all datasets. This indicates that no-reference metrics, which focus on low-level perceptual artifacts, fail to capture the reference-relative semantic quality required for 3DGS. Similarly, CrossScore displays inconsistent behavior, yielding negative correlations on Mip-NeRF 360, suggesting that its matching-based mechanism does not reliably align with dense feature similarity.

3.3. Generalization to Unseen Generators

To evaluate cross-generator generalization, we applied PR-IQA directly to images synthesized by unseen generators (**GEN3C [13]** and **SEVA [24]**) without any retraining. As shown in Table 5, the evaluation demonstrates that our model maintains a stable correlation across various datasets and target metrics, indicating that the learned quality cues are not tied to the rendering characteristics of a specific generator. Unlike prior methods that exhibit significant performance fluctuations depending on the generator or evaluation metric, PR-IQA consistently yields competitive and superior results. This suggests that our partial-reference formulation effectively captures transferable perceptual correspondences rather than overfitting to generator-specific artifacts, thereby demonstrating robust generalization capabilities on previously unseen generative models.

4. More Ablation Studies on IQA

4.1. Impact of the Number of Reference Images

We conducted an ablation study to analyze the sensitivity of our PR-IQA framework to the number of available reference images N_{ref}. In this experiment, we varied N_{ref} from 1 to 10 by selecting reference views at regular intervals from the corresponding image sequence.

Fig. 2 illustrates the evolution of PLCC and SRCC scores for both DINOv2 and SSIM targets as the number of reference views increases. In contrast to CrossScore, where performance saturates, all other methods exhibit a steady gain in performance with additional reference views. Notably, our models (Ours$_{\text{DINOv2}}$ and Ours$_{\text{SSIM}}$) demonstrate high robustness even with a single reference view and continue to improve monotonically.

A significant finding is that our method achieves parity with, or even surpasses, established FR metrics without requiring GT supervision. Specifically, as shown in the DINOv2 target evaluation, our method begins to outperform the LPIPS baseline (orange dotted line) once $N_{\text{ref}} \geq 4$. This confirms that with sufficient cross-view context, our framework can predict quality maps with FR-level accuracy.

The $\hat{Q}$ variant (yellow solid line), which relies solely on geometrically overlapping regions, shows a steep performance increase as N_{ref} grows. This trend validates our design rationale: increasing the number of reference views expands the geometric coverage of the partial quality map $\hat{Q}$, thereby providing a richer guidance signal for the subse-

Table 5. Cross-generator generalization results on two unseen generators, GEN3C and SEVA, evaluated without any retraining. We report PLCC and SRCC on Mip-NeRF 360 and Tanks and Temples using DINOv2- and SSIM-based target quality scores.

IQA Method	Mip-NeRF 360 (GEN3C)				Tanks and Temples (GEN3C)				Mip-NeRF (SEVA)				Tanks and Temples (SEVA)			
	DINOv2		SSIM		DINOv2		SSIM		DINOv2		SSIM		DINOv2		SSIM	
	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
MEt3R*	0.251	0.279	0.070	0.061	0.254	0.231	0.121	0.121	0.168	0.227	0.086	0.098	0.142	0.144	0.120	0.141
CrossScore	0.076	0.092	0.229	0.220	0.345	0.365	0.530	0.520	-0.005	0.026	0.187	0.185	0.204	0.276	0.395	0.381
PuzzleSim	0.258	0.271	0.153	0.153	0.422	0.435	0.420	0.413	0.312	0.338	0.160	0.164	0.331	0.367	0.327	0.320
$\text{Ours}_{\text{partial}}$*	0.308	0.409	0.174	0.178	0.344	0.433	0.067	0.084	0.258	0.409	0.119	0.164	0.318	0.504	0.087	0.113
$\text{Ours}_{\text{DINOv2}}$	0.368	0.401	0.303	0.287	0.548	0.596	0.392	0.403	0.358	0.472	0.306	0.313	0.418	0.543	0.299	0.276
$\text{Ours}_{\text{SSIM}}$	0.113	0.136	0.340	0.341	0.328	0.340	0.558	0.553	0.085	0.143	0.431	0.420	0.211	0.294	0.547	0.521

Table 6. Ablation study on the contribution of loss components. We compare the full model with variants trained without the JSD loss (w/o $\mathcal{L}_{\text{JSD}}$) or without the PLCC loss (w/o $\mathcal{L}_{\text{PLCC}}$). All metrics are evaluated on the Mip-NeRF 360 and Tanks and Temples datasets using PLCC and SRCC for the target of DINOv2. Bold indicates the best performance.

Loss variants	Mip-NeRF 360		Tanks and Temples	
	PLCC	SRCC	PLCC	SRCC
w/o $\mathcal{L}_{\text{JSD}}$	-0.181	-0.202	-0.242	-0.274
w/o $\mathcal{L}_{\text{PLCC}}$	-0.119	-0.134	-0.147	-0.150
Full Model	**0.555**	**0.622**	**0.573**	**0.649**

quent quality completion network.

4.2. Quality Fusion Strategy

We investigate the optimal strategy for aggregating quality predictions when multiple reference images are available. As illustrated in Fig. 3, we evaluate four pixel-wise fusion operators, Max, Min, Median, and Mean, across varying reference counts ($N_{\text{ref}} = 1$ to 10) to determine the most effective aggregation method.

Given K reference images yielding predicted quality maps $\{Q_i\}_{i=1}^{K}$ for a query image I_q, the fused map values at pixel p are computed as follows:

$$\begin{aligned} Q_{\max}(p) &= \max_i \{Q_i(p)\}, \\ Q_{\min}(p) &= \min_i \{Q_i(p)\}, \\ Q_{\text{mean}}(p) &= \frac{1}{K}\sum_{i=1}^{K} Q_i(p), \\ Q_{\text{median}}(p) &= \text{median}_i \{Q_i(p)\}. \end{aligned} \tag{5}$$

The quantitative results demonstrate that the Max fusion strategy consistently outperforms all other aggregation methods. As shown in Fig. 3, the performance of Max fusion improves monotonically as the number of reference images increases, reaching peak correlations at $N_{\text{ref}} = 10$. This represents a substantial gain over the single-reference baseline.

In contrast, Min fusion exhibits the poorest performance, showing a degrading trend where accuracy drops significantly as more references are added. The Mean and Median strategies remain relatively stagnant and fail to consistently surpass the single-reference baseline.

The widening gap between Max fusion and other methods suggests that an optimistic aggregation strategy is crucial for robust cross-reference evaluation. By selecting the maximum quality score per pixel, the framework effectively isolates the best matching evidence from the available views. This approach allows the model to filter out low scores caused by occlusions, view-dependent artifacts, or poor geometric correspondences in specific reference frames, ensuring that the final quality map reflects the most reliable visual information.

4.3. Ablation Study on Loss Components

In this section, we evaluate the contribution of the individual loss terms defined in the training objective (Eq. (5) in the main manuscript). Our full objective function combines a pixel-wise reconstruction loss ($\mathcal{L}_1$) with two distribution-aware losses: the Jensen-Shannon Divergence loss ($\mathcal{L}_{\text{JSD}}$) and the Pearson Linear Correlation Coefficient loss ($\mathcal{L}_{\text{PLCC}}$). To isolate the impact of these auxiliary terms, we trained variants of our model by removing them one at a time.

Table 6 presents the performance comparison on the Mip-NeRF 360 and Tanks and Temples datasets. The results demonstrate that $\mathcal{L}_{\text{JSD}}$ and $\mathcal{L}_{\text{PLCC}}$ are not merely supplementary but are fundamental to the learning process.

As shown in the table, removing either the JSD loss ("w/o $\mathcal{L}_{\text{JSD}}$") or the PLCC loss ("w/o $\mathcal{L}_{\text{PLCC}}$") leads to a catastrophic performance drop, resulting in negative correlation values across all metrics and datasets. A negative correlation implies that the model's predictions are inversely related to the GT, indicating a complete failure to learn the correct quality ranking.

In contrast, the "Full Model" achieves strong positive correlations (e.g., PLCC > 0.55). This sharp contrast suggests that the pixel-wise loss alone is insufficient for this task. The combination of $\mathcal{L}_{\text{JSD}}$ (which aligns score dis-

Table 7. Geometric robustness analysis under point cloud filtering and camera pose perturbations. We evaluate the sensitivity of our PR-IQA framework to geometric input quality on the Mip-NeRF 360 and Tanks and Temples datasets. We analyze the impact of varying VGGT depth confidence filtering thresholds (No filtering, 20%, 50%) and introduce synthetic Gaussian noise to camera parameters (5% and 10% levels). Red, orange, and yellow cells denote the 1st, 2nd, and 3rd best results, respectively. The results demonstrate that our default configuration (20% filtering) yields optimal performance, and the method remains robust, consistently outperforming baselines (CrossScore, PuzzleSim) even under significant geometric noise.

Method	Type	Mip-NeRF 360				Tanks and Temples			
		DINOv2		SSIM		DINOv2		SSIM	
		PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
CrossScore	-	0.094	0.090	0.290	0.325	0.237	0.272	0.444	0.462
PuzzleSim	-	0.304	0.327	0.128	0.124	0.351	0.369	0.348	0.347
$\text{Ours}_{\text{DINOv2}}$	(20% Conf Filtering)	0.555	0.622	0.261	0.241	0.573	0.649	0.387	0.367
	+ 50% Conf Filtering	0.476	0.522	0.252	0.241	0.495	0.559	0.362	0.325
	+ No Filtering	0.495	0.555	0.261	0.251	0.517	0.584	0.352	0.310
	+ 5% Random Noise on Cam	0.460	0.498	0.248	0.240	0.480	0.511	0.358	0.319
	+ 10% Random Noise on Cam	0.447	0.477	0.244	0.236	0.464	0.479	0.353	0.315
$\text{Ours}_{\text{SSIM}}$	(20% Conf Filtering)	0.320	0.367	0.535	0.556	0.309	0.344	0.625	0.642
	+ 50% Conf Filtering	0.301	0.348	0.514	0.534	0.294	0.326	0.607	0.624
	+ No Filtering	0.304	0.349	0.520	0.542	0.301	0.334	0.609	0.626
	+ 5% Random Noise on Cam	0.312	0.360	0.505	0.524	0.293	0.320	0.610	0.624
	+ 10% Random Noise on Cam	0.312	0.360	0.504	0.523	0.292	0.318	0.609	0.624

Table 8. Comparison of FR-IQA metrics as guidance signals for Quality-Aware 3DGS training. We evaluate the 3DGS modeling quality (PSNR, SSIM, LPIPS) when guiding the optimization using different IQA targets (PSNR, SSIM, LPIPS, and DINOv2). The results demonstrate that DINOv2 feature similarity consistently outperforms traditional metrics, even surpassing methods that directly optimize for the target metric itself, thereby justifying its selection as our primary prediction target.

	Tanks and Temples			Mip-NeRF 360		
IQA method	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
PSNR	7.09	0.435	0.575	7.33	0.371	0.574
SSIM	14.11	0.525	0.482	15.71	0.515	0.466
LPIPS	13.21	0.524	0.480	14.72	0.502	0.472
DINOv2-SIM	**16.05**	**0.548**	**0.465**	**18.29**	**0.526**	**0.453**

tributions) and $\mathcal{L}_{\text{PLCC}}$ (which enforces linear relationship) provides the necessary constraints to stabilize training and guide the model toward perceptually meaningful quality predictions.

4.4. Geometric Robustness Analysis

In this section, we investigate the sensitivity of the PR-IQA framework to geometric imperfections, specifically focusing on point cloud quality and camera pose accuracy. As detailed in our methodology (Sect. 3.3 of the main manuscript), our approach generates a partial quality map by warping features from the reference image to the query view using VGGT [16]. This process relies on estimating 3D points via stereo correspondences and reprojecting them for feature alignment. To mitigate artifacts arising from unreliable correspondences, our default configuration filters out 3D points falling within the bottom 20% of confidence scores, utilizing only the remaining high-confidence points for warping. To evaluate the robustness of this design, we conducted experiments varying this filtering threshold and introducing synthetic noise to the estimated camera poses.

Table 7 summarizes the performance of our method under these varying geometric conditions. A broad analysis reveals that our proposed methods ($\text{Ours}_{\text{DINOv2}}$ and $\text{Ours}_{\text{SSIM}}$) consistently achieve significantly higher PLCC and SRCC correlations compared to baselines like CrossScore and PuzzleSim across both Mip-NeRF 360 and Tanks and Temples datasets. This empirically validates the effectiveness of our geometry-guided feature matching approach.

Impact of Point Cloud Filtering. We analyzed how the density and reliability of the geometric input affect performance by adjusting the VGGT depth confidence filter. As shown in Table 7, the default setting, removing the bottom 20% of low-confidence points, yields optimal performance. This threshold strikes a critical balance: it effectively eliminates high-variance noise (e.g., sky regions or inaccurate depths) while preserving sufficient scene context essential for matching. Conversely, performance degrades under the "No Filtering" setting due to the inclusion of geometric outliers, as well as under the stricter "+50% Conf Filtering" setting, where the excessive removal of points leads to a loss of valuable visual information.

Robustness to Camera Pose Noise. To evaluate resilience against inaccurate camera poses, a common chal-

Table 9. Ablation on the masking threshold τ for 3DGS training. We evaluate the impact of the pixel retention rate τ on reconstruction quality. We compare aggressive ($\tau = 30$), default ($\tau = 50$), and lenient ($\tau = 70$) filtering strategies on the Mip-NeRF 360 and Tanks and Temples datasets. The results show that $\tau = 50$ achieves the best performance across datasets, validating it as a robust heuristic that balances noise removal with data retention. Red, orange, and yellow cells denote the 1st, 2nd, and 3rd best methods per column. (excluding FR settings †)

		$\tau = 30$						$\tau = 50$						$\tau = 70$					
		Mip-NeRF 360			Tanks and Temples			Mip-NeRF 360			Tanks and Temples			Mip-NeRF 360			Tanks and Temples		
IQA-Guided	Method	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
w/o IQA	Vanilla 3DGS	16.078	0.461	0.415	15.298	0.509	0.406	16.078	0.461	0.415	15.298	0.509	0.406	16.078	0.461	0.415	15.298	0.509	0.406
	ViewCrafter	16.179	0.474	0.452	15.773	0.523	0.455	16.179	0.474	0.452	15.773	0.523	0.455	16.179	0.474	0.452	15.773	0.523	0.455
w/ FR-IQA	SSIM†	16.837	0.494	0.413	16.331	0.551	0.397	16.676	0.487	0.421	16.228	0.556	0.399	16.779	0.491	0.425	16.405	0.557	0.407
	DINOv2†	16.892	0.494	0.400	16.401	0.551	0.392	17.178	0.498	0.399	16.777	0.562	0.384	17.209	0.497	0.412	16.784	0.560	0.391
w/ NR-IQA	PaQ-2-PiQ	16.148	0.456	0.432	15.345	0.511	0.435	16.298	0.472	0.425	15.769	0.534	0.421	16.370	0.477	0.430	16.137	0.546	0.414
	PIQE	15.858	0.462	0.450	15.275	0.521	0.443	16.313	0.479	0.440	15.671	0.534	0.433	16.426	0.478	0.441	15.936	0.543	0.427
w/ CR-IQA	CrossScore	16.036	0.469	0.441	15.196	0.515	0.440	16.312	0.476	0.431	15.856	0.537	0.427	16.463	0.480	0.442	16.195	0.547	0.424
	PuzzleSim	16.239	0.473	0.411	15.645	0.527	0.414	16.349	0.482	0.423	15.937	0.541	0.406	16.469	0.479	0.421	16.104	0.546	0.406
	$\text{Ours}_{\text{SSIM}}$	16.319	0.474	0.437	15.914	0.540	0.410	16.371	0.485	0.427	16.143	0.548	0.407	16.512	0.482	0.437	16.240	0.551	0.416
	$\text{Ours}_{\text{DINOv2}}$	16.529	0.482	0.417	15.981	0.540	0.406	16.756	0.493	0.414	16.238	0.551	0.403	16.736	0.489	0.424	16.370	0.554	0.405

† Metrics require a same-pose GT image.

Table 10. Ablation study comparing binary masking and soft masking strategies. We evaluate the robustness of our framework by comparing the default binary masking approach against a continuous soft weighting strategy.

		Mip-NeRF 360			Tanks and Temples		
Mask Type	Method	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
Binary Mask	$\text{Ours}_{\text{SSIM}}$	16.37	0.485	0.427	16.14	0.548	0.407
	$\text{Ours}_{\text{DINOv2}}$	16.76	**0.493**	**0.414**	16.24	0.551	**0.403**
Soft Mask	$\text{Ours}_{\text{SSIM}}$	16.61	0.485	0.437	16.34	**0.553**	0.418
	$\text{Ours}_{\text{DINOv2}}$	**16.78**	0.488	0.426	**16.44**	**0.553**	0.405

lenge in real-world sparse-view reconstruction, we introduced Gaussian noise to both intrinsic and extrinsic parameters. We defined two noise levels:

- **5% Noise Level:** Perturbations included rotation by approximately 5°, translation by 5% of the original magnitude, focal length by 5%, and principal point shifts by 5% of image dimensions.
- **10% Noise Level:** These perturbations were doubled (e.g., approximately 10° rotation).

As expected, the performance exhibits a gradual decline as noise levels increase (see Table 7). However, a crucial finding is that even under significant perturbations (10% noise), our method maintains competitive scores that continue to surpass the baseline methods (CrossScore and PuzzleSim). This confirms that the PR-IQA framework is not only effective under ideal conditions but also practically robust to the geometric errors frequently encountered in sparse-view scenarios.

4.5. Low-Overlap Robustness Analysis

To examine robustness under limited visual correspondence, we re-evaluated the test set by regrouping image pairs according to their overlap ratio. As shown in Table 11, the proposed PR-IQA remains stable even as the overlap becomes progressively smaller, while competing methods tend to degrade more noticeably under the same condition. This result indicates that our method does not rely solely on directly shared regions between the generated and reference images, but instead learns transferable quality cues that remain meaningful when only partial correspondence is available.

Fig. 4 further illustrates this behavior in challenging low-overlap examples. Even when the common visible region is very limited, our method produces quality maps that better preserve the perceptually important structure and object-level consistency than direct full-reference targets. In particular, the propagated responses remain coherent beyond the overlapping area, supporting reliable quality estimation in non-overlapping regions. These observations confirm that PR-IQA effectively extends local reference evidence into the unseen area and remains robust even in near-zero-overlap cases.

4.6. False Positive Analysis in Non-Overlapping Regions

To further analyze reliability in unseen areas, we measured the *False Positive Rate* (FPR@Top-$X\%$) specifically within non-overlapping regions, where hallucinated content is most likely to appear. As summarized in Table 12, $\text{Ours}_{\text{DINOv2}}$ consistently achieves the lowest false positive rate across all evaluation thresholds. This indicates that the proposed PR-IQA is less prone to incorrectly assigning high-quality scores to regions that are not supported by reference evidence, demonstrating stronger conservativeness and robustness in ambiguous areas.

Fig. 5 provides qualitative examples of this behavior on hallucinated objects and structures. Compared with existing methods, our predictions suppress spuriously high responses in boxed non-overlapping regions while preserving meaningful quality patterns in the valid area. In contrast, competing approaches more often produce overly confident activations on unsupported content. These results confirm that PR-IQA avoids false positives on hallucinated content.

Table 11. Low-overlap evaluation. We evaluate robustness by grouping image pairs from the original dataset by overlap ratio.

IQA Method	25% (81)				20% (52)				10% (17)				5% (9)			
	DINOv2		SSIM		DINOv2		SSIM		DINOv2		SSIM		DINOv2		SSIM	
	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
CrossScore	0.137	0.118	0.153	0.139	0.189	0.162	0.106	0.074	0.223	0.194	0.141	0.079	0.250	0.217	0.131	0.080
PuzzleSim	0.178	0.211	0.013	0.004	0.120	0.173	0.041	0.030	0.041	0.081	0.024	0.006	-0.007	0.027	0.058	0.041
$\text{Ours}_{\text{DINOv2}}$	0.469	0.502	0.374	0.366	0.485	0.503	0.396	0.382	0.501	0.511	0.383	0.370	0.486	0.484	0.386	0.369
$\text{Ours}_{\text{SSIM}}$	0.278	0.295	0.463	0.482	0.331	0.311	0.462	0.477	0.384	0.345	0.434	0.414	0.365	0.300	0.409	0.418

Note. %: overlap ratio; (): # of images.

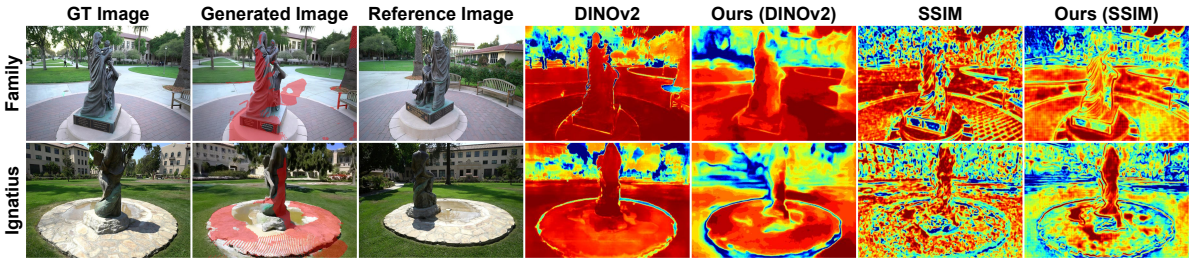


Figure 4. Low-overlap qualitative results. Red region in the generated image shows overlaps of 16% (Family) and 22% (Ignatius).

Table 12. FPR@Top-$X\%$ measures how often pixels in the top $X\%$ of scores within non-overlapping regions are falsely rated as high quality on Tanks and Temples.

Method	FPR@50%	FPR@40%	FPR@30%	FPR@20%	FPR@10%
CrossScore	0.380	0.300	0.240	0.183	0.105
PuzzleSim	0.328	0.273	0.222	0.162	0.093
$\text{Ours}_{\text{DINOv2}}$	**0.306**	**0.236**	**0.183**	**0.137**	**0.082**

5. More Ablation Studies on 3DGS

5.1. Effectiveness of DINOv2 Feature Similarity

We validate the rationale behind selecting DINOv2 feature similarity (i.e., DINOv2-SIM) as our primary optimization target by comparing its effectiveness against standard FR-IQA metrics: PSNR, SSIM, and LPIPS. To ensure a fair comparison, we integrated these metrics into the "Quality-Aware 3DGS Training" pipeline (described in Section 4 of the main manuscript) as alternative guidance signals. For consistency, all quality maps were normalized to the range $[0, 1]$ via min-max scaling, where higher values denote better quality.

As detailed in Table 8, the 3DGS reconstruction guided by DINOv2-based quality maps consistently yields superior performance across all evaluation metrics on both the Tanks and Temples and Mip-NeRF 360 datasets. A remarkable finding is that utilizing DINOv2 similarity as a training guide results in higher final PSNR and SSIM scores than using those specific metrics themselves as guidance targets.

This superiority stems from the inherent limitations of conventional metrics in the context of diffusion-based synthesis. Pixel-wise metrics like PSNR tend to unduly penalize regions that possess valid geometric structures but exhibit minor color shifts or lighting variations, thereby discarding potentially useful supervision signals. Similarly, SSIM and LPIPS often struggle to reliably distinguish between fine geometric details and artifacts in generated views. In contrast, our DINOv2-based approach prioritizes high-level semantic and geometric alignment. It effectively identifies and utilizes structurally consistent regions while remaining robust to benign photometric discrepancies, making it significantly more suitable for supervising 3D reconstruction from diffusion-generated imagery.

5.2. Impact of Masking Threshold τ

In this section, we provide a detailed ablation study to validate our choice of the masking threshold τ, which was set to a heuristic value of 50 in the main manuscript. In our framework, τ represents the retention rate, the percentage of pixels with the highest predicted quality scores that are used for 3DGS optimization. We determine a global quality threshold Q_{thresh} corresponding to the $(100 - \tau)$-th percentile of the score distribution; pixels exceeding this value are included in the training mask. Thus, a lower τ (e.g., $\tau = 30$) implies an aggressive filtering strategy that retains only the top 30% of pixels, whereas a higher τ (e.g., $\tau = 70$) is more lenient.

Table 9 presents the reconstruction performance across varying thresholds ($\tau \in \{30, 50, 70\}$). The aggressive strategy ($\tau = 30$) consistently yields the lowest performance across both datasets. This indicates that while removing low-quality regions is essential, discarding 70% of the gen-

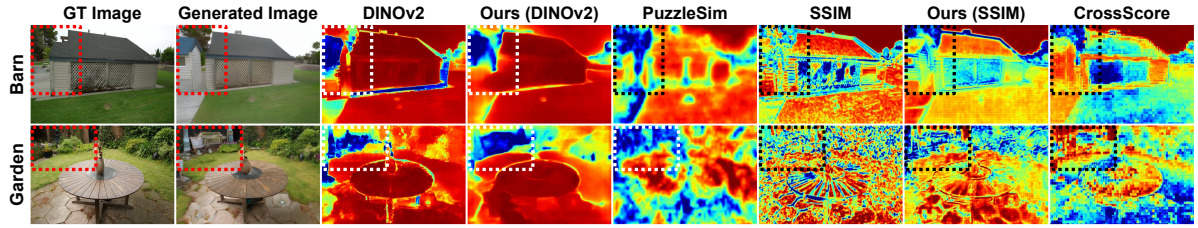


Figure 5. Quality estimation results on hallucinated non-overlapping regions (boxed) from the Barn and Garden scenes. The dashed boxes highlight unsupported areas that are visible in the generated image but not reliably matched to the reference view.

Table 13. Computational cost analysis. We report the averaged runtime (seconds) and memory usage (MB) for individual components of the PR-IQA pipeline and the 3DGS optimization process.

Method	Stage	Runtime (s)	Memory (MB)
PR-IQA	Feature Ext.	0.303	5509.250
	VGGT	0.207	2448.317
	Inference	0.510	530.950
3DGS	-	25.210	749.780

erated data eliminates too much valid supervision signal, thereby hindering the geometry convergence and degrading the final reconstruction quality.

Performance peaks between $\tau = 50$ and $\tau = 70$. For $\text{Ours}_{\text{DINOv2}}$, the heuristic $\tau = 50$ achieves the best PSNR (16.756) on the Mip-NeRF 360 dataset, outperforming both the stricter ($\tau = 30$) and looser ($\tau = 70$) settings. On the Tanks and Temples dataset, while $\tau = 70$ yields a marginal improvement, the performance at $\tau = 50$ remains highly competitive and robust.

This study confirms that $\tau = 50$ serves as an effective and robust heuristic across diverse scenes. It strikes a critical balance: it is strict enough to filter out significant artifacts and inconsistencies, yet lenient enough to preserve a sufficient density of high-confidence pseudo-ground-truth pixels for accurate 3D reconstruction.

5.3. Soft vs. Binary Masking Strategies

In our primary manuscript, we employ a binary masking strategy that strictly includes or excludes pixels based on a confidence threshold. In this section, we conduct an ablation study to evaluate an alternative "soft weighting" strategy. Instead of a hard binary selection (0 or 1), this approach utilizes the predicted continuous quality score directly as a pixel-wise loss weight (ranging from 0 to 1) during 3DGS optimization. This allows the influence of each pixel to be modulated gradually by its estimated quality.

Mathematical Formulation. Let $\mathcal{L}_{\text{base}}(p)$ denote the standard photometric loss (e.g., $\mathcal{L}_1$ or D-SSIM) for a pixel p during 3DGS training.

- **Binary Masking:** We define a binary mask $M(p)$ based on the quality threshold Q_τ derived from the percentile τ:

$$M(p) = \mathbf{1}(Q(p) \geq Q_\tau) = \begin{cases} 1 & \text{if } Q(p) \geq Q_\tau \\ 0 & \text{otherwise} \end{cases}. \quad (6)$$

The final loss function is given by:

$$\mathcal{L}_{\text{binary}} = \sum_{p \in \mathcal{P}} M(p) \cdot \mathcal{L}_{\text{base}}(p). \quad (7)$$

- **Soft Weighting:** We directly use the normalized predicted quality score $Q(p) \in [0, 1]$ as a weighting factor $W(p)$:

$$W(p) = Q(p). \quad (8)$$

The weighted loss function becomes:

$$\mathcal{L}_{\text{soft}} = \sum_{p \in \mathcal{P}} W(p) \cdot \mathcal{L}_{\text{base}}(p). \quad (9)$$

Table 10 compares the reconstruction performance of our method under both masking regimes. The quantitative results indicate that both strategies yield highly similar performance metrics across the Mip-NeRF 360 and Tanks and Temples datasets. For instance, while binary masking achieves a slightly better LPIPS score on Mip-NeRF 360, soft masking yields a marginally higher PSNR. Overall, the performance differences are negligible, suggesting that both approaches effectively guide the optimization process.

These findings demonstrate the inherent robustness of the PR-IQA framework. The fact that the optimization remains stable and high-performing under both hard-thresholding and continuous-weighting schemes confirms that our predicted quality maps provide reliable supervision signals regardless of the specific masking implementation. This flexibility suggests that practitioners can select either approach, prioritizing the interpretability of binary masks or the differentiability of soft weights, without compromising reconstruction quality.

5.4. Computational Analysis

In this section, we evaluate the computational efficiency of the proposed PR-IQA framework. Table 13 details the

runtime and memory usage for each stage of the pipeline: feature extraction, VGGT-based warping, and quality inference, measured on a single-image basis. A key advantage of our design is that the pipeline internally resizes all inputs to a fixed resolution, ensuring that these computational metrics remain invariant regardless of the original input image resolution.

To provide context for these costs, we compare them against the resource consumption of the standard 3DGS optimization process. This comparison was conducted on the 'Barn' scene from the Tanks and Temples dataset, initialized with 28,290 points.

As shown in Table 13, the total runtime for the PR-IQA pipeline is approximately 1.02 seconds per image (summing feature extraction, VGGT, and inference). In contrast, the 3DGS optimization for the corresponding scene requires 25.21 seconds. This indicates that the additional computational overhead introduced by our quality assessment module is negligible, making it a highly practical addition to the reconstruction pipeline without causing significant bottlenecks.

6. More Qualitative Results

6.1. More Qualitative Results for Quality Map

We provide extensive qualitative comparisons on scenes not featured in the main manuscript. Figs. 6, 7, and 8 illustrate results across the Mip-NeRF 360, Tanks and Temples, and RealEstate10K datasets, respectively. As shown in these figures, our PR-IQA generates quality maps that exhibit high fidelity to the GT DINOv2-SIM, accurately capturing fine-grained variations and sharp boundaries. In contrast, NR-IQA methods often struggle to provide meaningful estimates, while CR-IQA baselines tend to suffer from blocky artifacts, particularly in non-overlapping regions. Our method overcomes these limitations by effectively propagating quality information globally, resulting in smooth and accurate dense quality maps.

6.2. More Qualitative Results for SSIM Map

We provide extended qualitative comparisons for SSIM-based quality assessment. Figs. 9, 10, and 11 display results for the Mip-NeRF 360, Tanks and Temples, and RealEstate10K datasets, respectively. Notably, our $\text{Ours}_{\text{SSIM}}$ variant substantially outperforms CrossScore, despite both methods sharing the same SSIM target. This performance gap highlights the effectiveness of our reference-conditioned cross-attention and quality completion framework. Visually, $\text{Ours}_{\text{SSIM}}$ maintains consistent quality estimation across both textured and smooth regions, whereas baselines frequently exhibit noisy predictions. This validates that our framework adapts robustly to diverse quality metrics.

6.3. More Qualitative Results for 3DGS

We present additional visualization results highlighting the impact of our IQA-Guided 3DGS framework. Fig. 12 shows reconstructions from the Mip-NeRF 360, Tanks and Temples, and RealEstate10K datasets, respectively. These results illustrate how our quality-guided training effectively concentrates computational resources on high-quality regions, significantly improving the overall reconstruction quality.

7. Limitations and Discussion

While PR-IQA achieves state-of-the-art performance in CR-IQA and significantly enhances sparse-view 3DGS reconstruction, we acknowledge several limitations and outline avenues for future research.

First, PR-IQA is currently trained using pseudo-GT quality maps derived from FR metrics, specifically DINOv2 feature similarity or SSIM. While this proxy-supervision strategy is practical for our targeted downstream task and has proven effective for geometric reconstruction, it does not fully replace human perceptual quality assessment. The model's upper bound is inherently limited by the capability of the chosen FR metric to capture perceptual subtleties or domain-specific artifacts. Incorporating human annotations or learning from large-scale perceptual preference data [10] remains an exciting direction to align the quality predictions more closely with human visual perception.

Second, our experimental validation covers multiple standard benchmarks (Mip-NeRF 360, Tanks and Temples, RealEstate10K) and utilizes widely adopted backbones like ViewCrafter for view synthesis and standard 3DGS for reconstruction. However, the fields of generative AI and 3D vision are rapidly evolving, with new multi-view diffusion models and reconstruction primitives emerging frequently. A truly comprehensive evaluation across all recent architectures is beyond the scope of this work. Exploring the broader applicability of PR-IQA as a plug-and-play module for diverse generative pipelines and reconstruction methods is an interesting avenue for future research.

Finally, our framework relies on the generation of a partial quality map $\hat{Q}$, which is constructed using geometric correspondences (via VGGT and dense stereo). While our ablation studies demonstrate robustness to significant geometric noise, extreme scenarios, such as large textureless regions or severe lighting changes where stereo matching fails completely, could inevitably degrade the quality of the partial map. Future work could investigate end-to-end joint training strategies that simultaneously optimize for geometric alignment and quality estimation to mitigate this dependency.

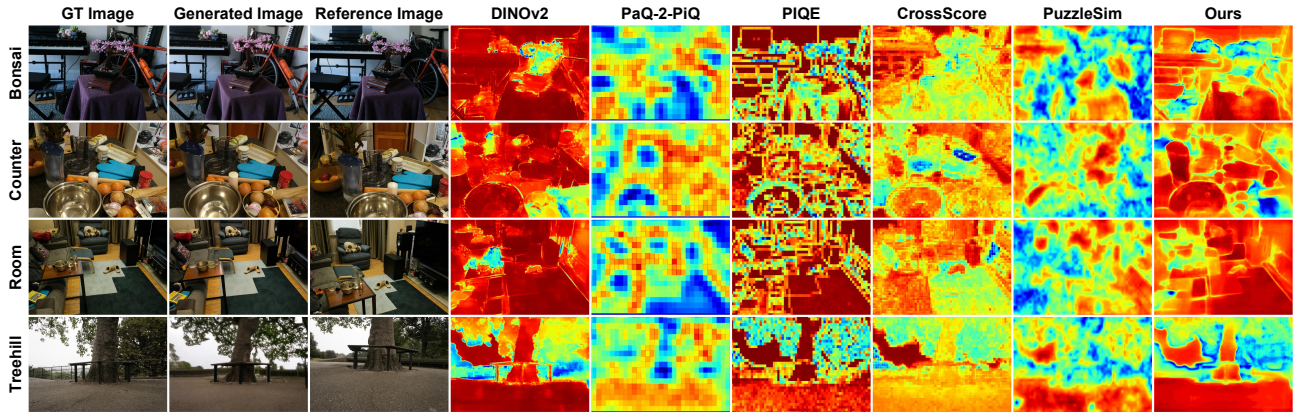


Figure 6. Additional quality map comparisons on Mip-NeRF 360 dataset (DINOv2-SIM target). Our method produces quality maps closely aligned with ground-truth DINOv2-SIM.

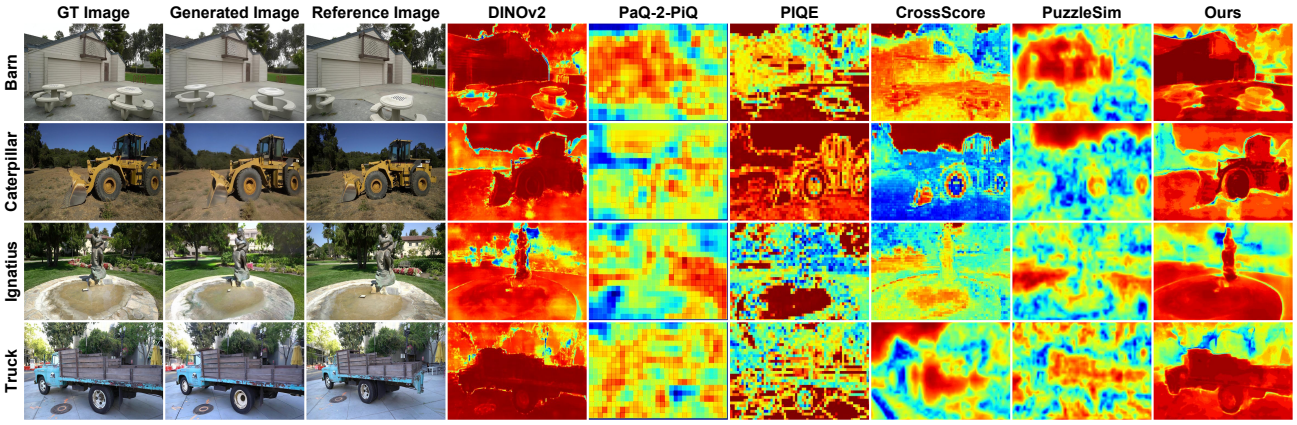


Figure 7. Additional quality map comparisons on Tanks and Temples dataset (DINOv2-SIM target). Our PR-IQA consistently estimates quality across complex outdoor scenes.

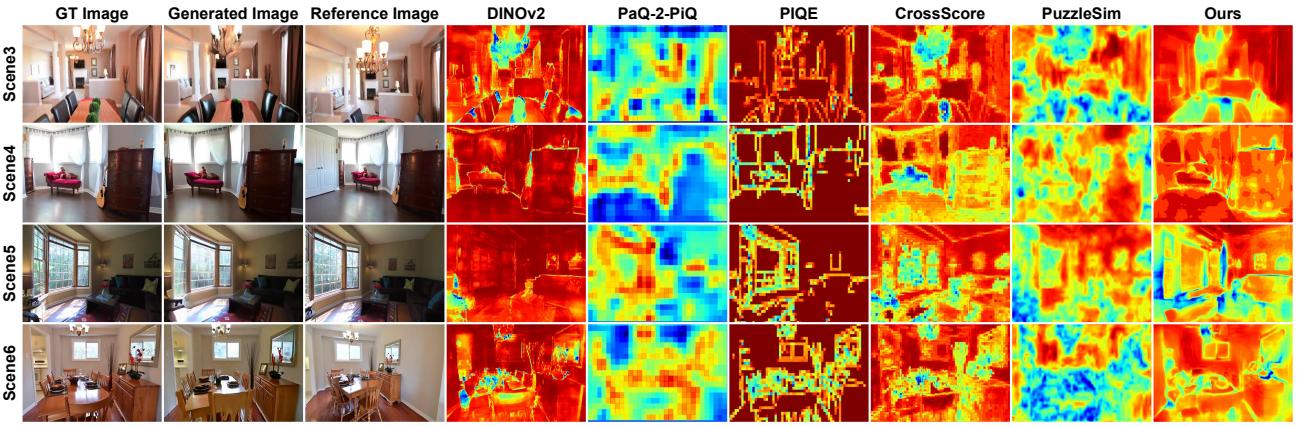


Figure 8. Additional quality map comparisons on RealEstate10K dataset (DINOv2-SIM target). Our method demonstrates robust performance on real estate scenes.

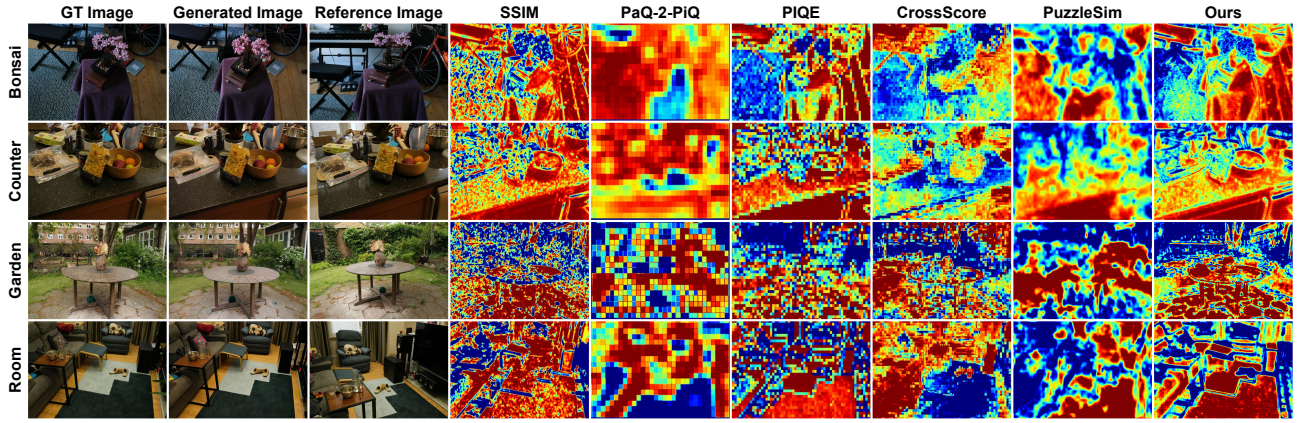


Figure 9. Additional quality map comparisons on Mip-NeRF 360 dataset (SSIM target). Ours$_{\text{SSIM}}$ variant effectively predicts SSIM maps, outperforming CrossScore across indoor scenes with various textures and structures.

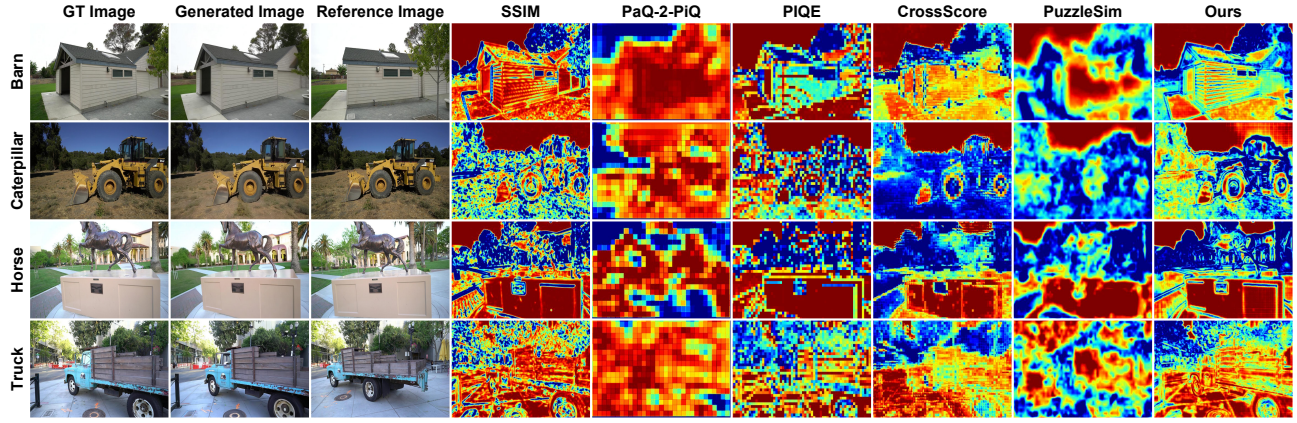


Figure 10. Additional quality map comparisons on Tanks and Temples dataset (SSIM target). Ours$_{\text{SSIM}}$ maintains consistent quality estimation in both textured and smooth regions, demonstrating superior performance over baseline methods in complex outdoor environments.

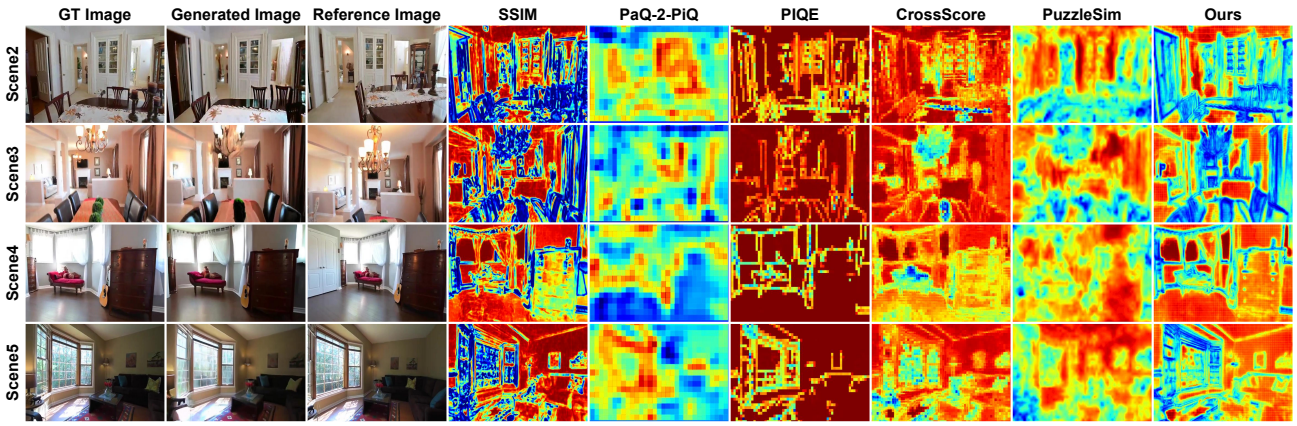


Figure 11. Additional quality map comparisons on RealEstate10K dataset (SSIM target). Our method accurately predicts SSIM maps, producing smooth and consistent results while baselines exhibit noisy or inconsistent predictions.

Figure 12. Qualitative comparison of 3DGS reconstruction quality. Our IQA-Guided 3DGS produces sharper geometry and more accurate textures compared to baselines by focusing computational resources on high-quality regions. Red boxes highlight representative areas where our method demonstrates superior reconstruction quality.

References

[1] Eduardo Arnold, Jamie Wynn, Sara Vicente, Guillermo Garcia-Hernando, Aron Monszpart, Victor Prisacariu, Daniyar Turmukhambetov, and Eric Brachmann. Map-free visual relocalization: Metric pose relative to a single image. In *ECCV*, pages 690–708, 2022. 2

[2] Mohammad Asim, Christopher Wewer, Thomas Wimmer, Bernt Schiele, and Jan Eric Lenssen. MET3R: Measuring multi-view consistency in generated images. In *CVPR*, pages 6034–6044, 2025. 3

[3] Weihan Cao, Yifan Zhang, Jianfei Gao, Anda Cheng, Ke Cheng, and Jian Cheng. PKD: general distillation framework for object detectors via pearson correlation coefficient. In *NeurIPS*, 2022. 1

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pages 9650–9660, 2021. 3

[5] Erik Englesson and Hossein Azizpour. Generalized jensen-shannon divergence loss for learning with noisy labels. In *NeurIPS*, pages 30284–30297, 2021. 1

[6] Stephanie Fu, Mark Hamilton, Laura Brandt, Axel Feldman, Zhoutong Zhang, and William T. Freeman. Featup: A model-agnostic framework for features at any resolution. In *ICLR*, 2024. 3

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3

[8] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus), 2016. 1

[9] Nicolai Hermann, Jorge Condor, and Piotr Didyk. Puzzle similarity: A perceptually-guided cross-reference metric for artifact detection in 3d scene reconstructions. In *ICCV*, pages 28881–28891, 2025. 4

[10] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. In *NeurIPS*, 2023. 11

[11] Ilya Loshchilov and Frank Hutter. SGDR: stochastic gradient descent with warm restarts. In *ICLR*, 2017. 3

[12] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin 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. 1, 4

[13] Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. Gen3C: 3d-informed world-consistent video generation with precise camera control. In *CVPR*, pages 6121–6132, 2025. 5

[14] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 1

[15] 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. 3

[16] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *CVPR*, pages 5294–5306, 2025. 7

[17] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3D vision made easy. In *CVPR*, pages 20697–20709, 2024. 3

[18] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE TIP*, 13(4):600–612, 2004. 3

[19] Zirui Wang, Wenjing Bian, and Victor Adrian Prisacariu. CrossScore: Towards multi-view image evaluation and scoring. In *ECCV*, pages 492–510, 2024. 3

[20] Zhenqiang Ying, Haoran Niu, Praful Gupta, Dhruv Mahajan, Deepti Ghadiyaram, and Alan Bovik. From patches to pictures (PaQ-2-PiQ): Mapping the perceptual space of picture quality. In *CVPR*, pages 3575–3585, 2020. 3

[21] Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. ViewCrafter: Taming video diffusion models for high-fidelity novel view synthesis. *IEEE TPAMI*, pages 1–18, 2025. 2

[22] Lingzhi Zhang, Zhengjie Xu, Connelly Barnes, Yuqian Zhou, Qing Liu, He Zhang, Sohrab Amirghodsi, Zhe Lin, Eli Shechtman, and Jianbo Shi. Perceptual artifacts localization for image synthesis tasks. In *ICCV*, pages 7579–7590, 2023. 3

[23] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 3

[24] Jensen Zhou, Hang Gao, Vikram Voleti, Aaryaman Vasishta, Chun-Han Yao, Mark Boss, Philip Torr, Christian Rupprecht, and Varun Jampani. Stable virtual camera: Generative view synthesis with diffusion models, 2025. 5