9. Supplementary Material

9.1. Embedding-Based Visualization of Predicted Distortions

In this subsection, we provide further explanation of the training dataset with examples of synthetic noise that is added. The training dataset incorporates a broad spectrum of distortion classes designed to reflect realistic and challenging image degradations. These include different types of color, compression, noise, denoising, brightness and illumination, downsampling, sharpening, contrast, and geometric artefacts. Each distortion is applied at five levels of severity to enable fine-grained assessment of model robustness across diverse visual conditions. The distortion were only applied to TROI region only (the black area of TROI in Figure 5). Figure 7 illustrates three TROI maps for orgninal image and nine different synthetic artifacts that are applied on TROI 30% for a frame of video of "Netflix_WindAndNature"[1].

9.2. Dense Patch Matching via Nearest-Neighbor Cosine Similarity

To better understand how our model captures local quality degradations between a reference view and a processed (novel) view, we compute dense patch correspondences using nearest-neighbor cosine similarity. This produces a per-patch *mismatch heatmap* that highlights regions where the processed view deviates most strongly from the reference.

Method

1. **Patch embeddings.** Both images are divided into $g \times g$ non-overlapping patches and encoded using the NOVA backbone, yielding normalized patch embeddings $\hat{\mathbf{a}}_i, \hat{\mathbf{b}}_j \in \mathbb{R}^C$.

2. **Similarity matrix.** For every patch pair we compute

$$S_{ij} = \hat{\mathbf{a}}_i^\top \hat{\mathbf{b}}_j.$$

3. **Best matches.** For each patch in the processed image B,

$$s_j^{B\to A} = \max_i S_{ij},$$

which records the best-matching similarity to a reference patch.

4. **Mismatch score.** After min–max normalization, mismatch is defined as the inverted similarity:

$$m_j^B = 1 - \tilde{s}_j^{B\to A}.$$

If the match is not *reciprocal* (i.e., the best match is not mutual), a multiplicative penalty factor $\beta > 1$ is applied.

5. **Global normalization.** To ensure comparability across both directions, mismatch values from reference and processed images are pooled and globally min–max normalized:

$$m_j^{B,\text{norm}} = \frac{m_j^B - m_{\min}}{m_{\max} - m_{\min} + \varepsilon}.$$

6. **Heatmap construction.** The normalized values $\{m_j^{B,\text{norm}}\}$ are reshaped to the $g \times g$ patch grid and visualized as a heatmap. High intensity values correspond to poorly matched or inconsistent regions in the processed view.

This mismatch heatmap provides a dense, spatially localized measure of how well the processed view preserves the structure and content of the reference. Distortions such as blur, noise, ghosting, or rendering artifacts manifest as elevated mismatch responses. Importantly, because the measure relies on feature correspondences rather than pixel alignment, it is well-suited for NVS scenarios where geometric misalignment is inevitable.

In practice, these heatmaps allow us to visualize and quantify quality differences at a fine-grained level, complementing global quality scores. We include examples for both synthetic distortions (where ground-truth changes are known) and novel view synthesis outputs (where artifacts are scene- and viewpoint-dependent).

To illustrate the effectiveness of our approach, Figures 8 and 9 show qualitative examples for both synthetic distortions and novel view synthesis (NVS) distortions. In the synthetic setting, we include distorted images, unaligned references, TROI masks, and heatmaps from both the pretrained DINOv2 and our NOVA model. While DINOv2 produces diffuse and noisy responses, NOVA yields sharper, more localized mismatch patterns that align closely with the TROI regions, effectively capturing artifacts such as blur, noise, and compression.

In the NVS setting, distorted views are compared with nearby unaligned references, and NOVA highlights structural inconsistencies, ghosting, and texture artifacts characteristic of NVS pipelines. These responses concentrate in perceptually degraded areas, aligning well with human intuition.

Together, the results demonstrate that NOVA not only improves over DINOv2 in localizing synthetic distortions, but also generalizes to realistic NVS artifacts, providing dense, perceptually meaningful mismatch maps for non-aligned reference quality assessment.

10. NVS NAR-IQA Benchmark

To complement Figure 5 in the main paper, we present additional scene-wise scores for each IQA model in Ta-

[1] https://media.xiph.org/video/derf/

!

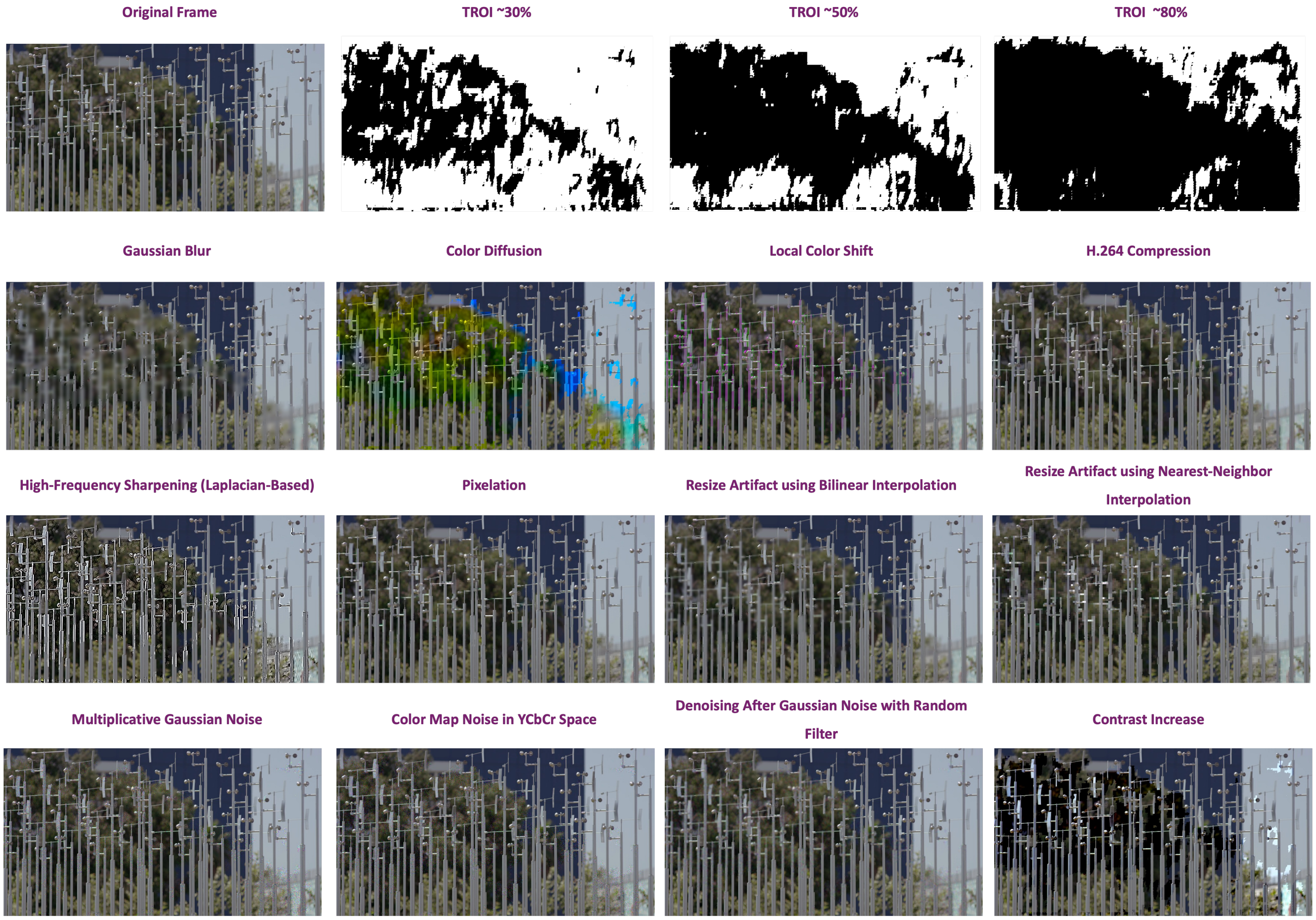
Original Frame
TROI ~30%
TROI ~50%
TROI ~80%
Gaussian Blur
Color Diffusion
Local Color Shift
H.264 Compression
High-Frequency Sharpening (Laplacian-Based)
Pixelation
Resize Artifact using Bilinear Interpolation
Resize Artifact using Nearest-Neighbor Interpolation
Multiplicative Gaussian Noise
Color Map Noise in YCbCr Space
Denoising After Gaussian Noise with Random Filter
Contrast Increase

Figure 7. Examples of distortion classes applied to the training dataset. Each row illustrates a different class of synthetic distortion (e.g., blur, color, compression, noise) applied selectively within Temporal Regions of Interest (TROIs) of varying coverage (30%, 50%, 80%) to simulate localized artifacts commonly observed in NVS.

ble 3. Additionally, we evaluated CrossScore [36] using five randomly selected reference frames before and after the aligned ground-truth frame, repeating inference 10 times. The mean score was 59.22 (std. 0.51), whereas our method NOVA achieved 80.19 with a single non-aligned reference in the benchmark. As CrossScore is trained to emulate SSIM, it does not surpass SSIM; however, incorporating multiple nearby references yields measurable improvements and indicates a promising direction for future extensions of NAR-NVSQA. Finally, we emphasize that our model was trained on a large collection of synthetic distortions derived from a video dataset, which does not overlap with our NVS NAR-IQA test dataset.

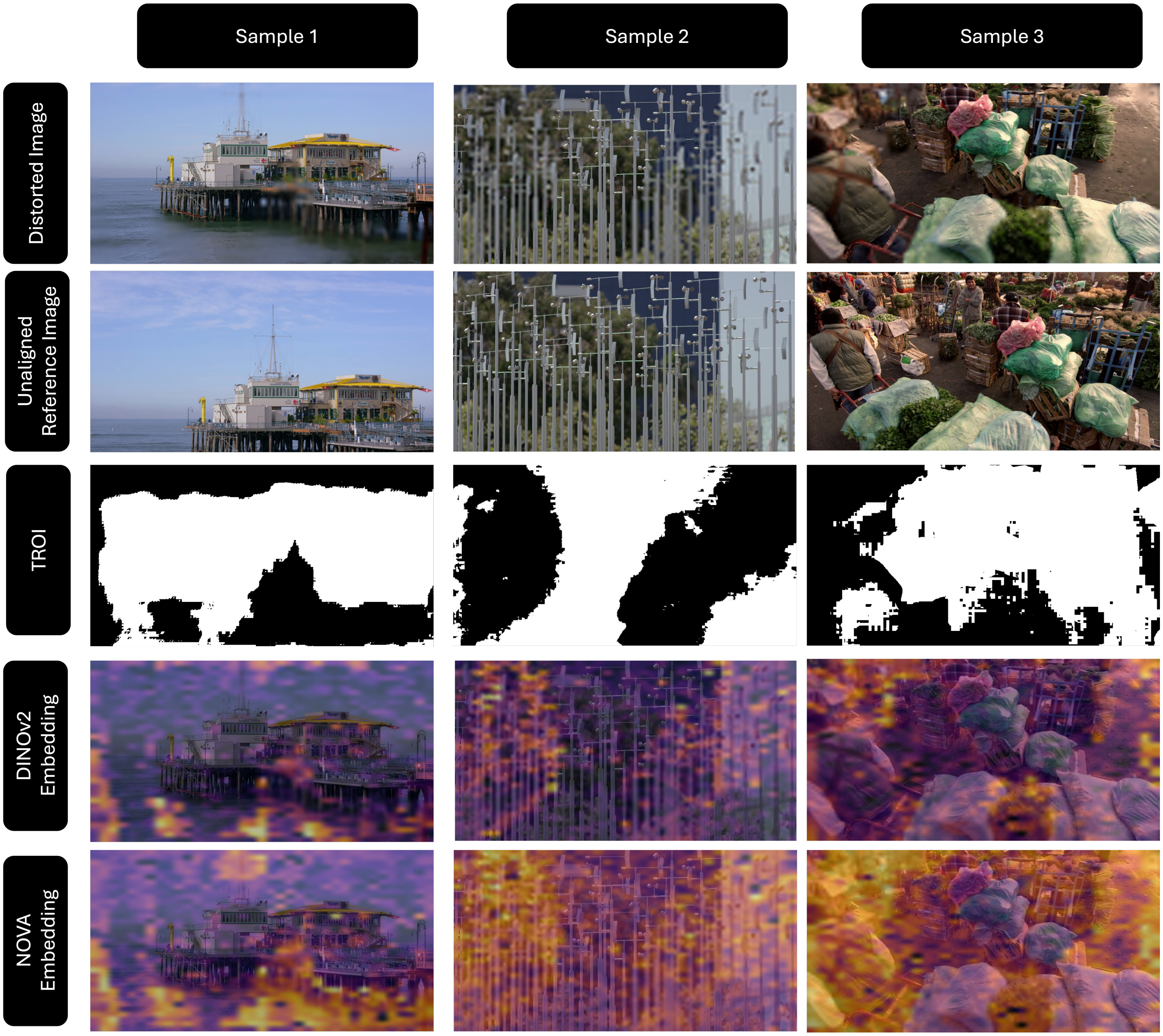

Figure 8. Visualization of embedding responses for distorted images relative to unaligned reference images. Each column shows a sample with: (1) the distorted image, (2) the corresponding unaligned reference image, (3) the Temporal Region of Interest (TROI) mask indicating motion-sensitive areas where distortions are most perceptually relevant, (4) heatmaps derived from pretrained DINOv2 embeddings, and (5) heatmaps derived from our proposed NOVA embeddings. Compared to DINOv2, NOVA produces sharper and more localized mismatch responses, especially within TROI regions, highlighting distortions such as structural artifacts and content degradation more faithfully.

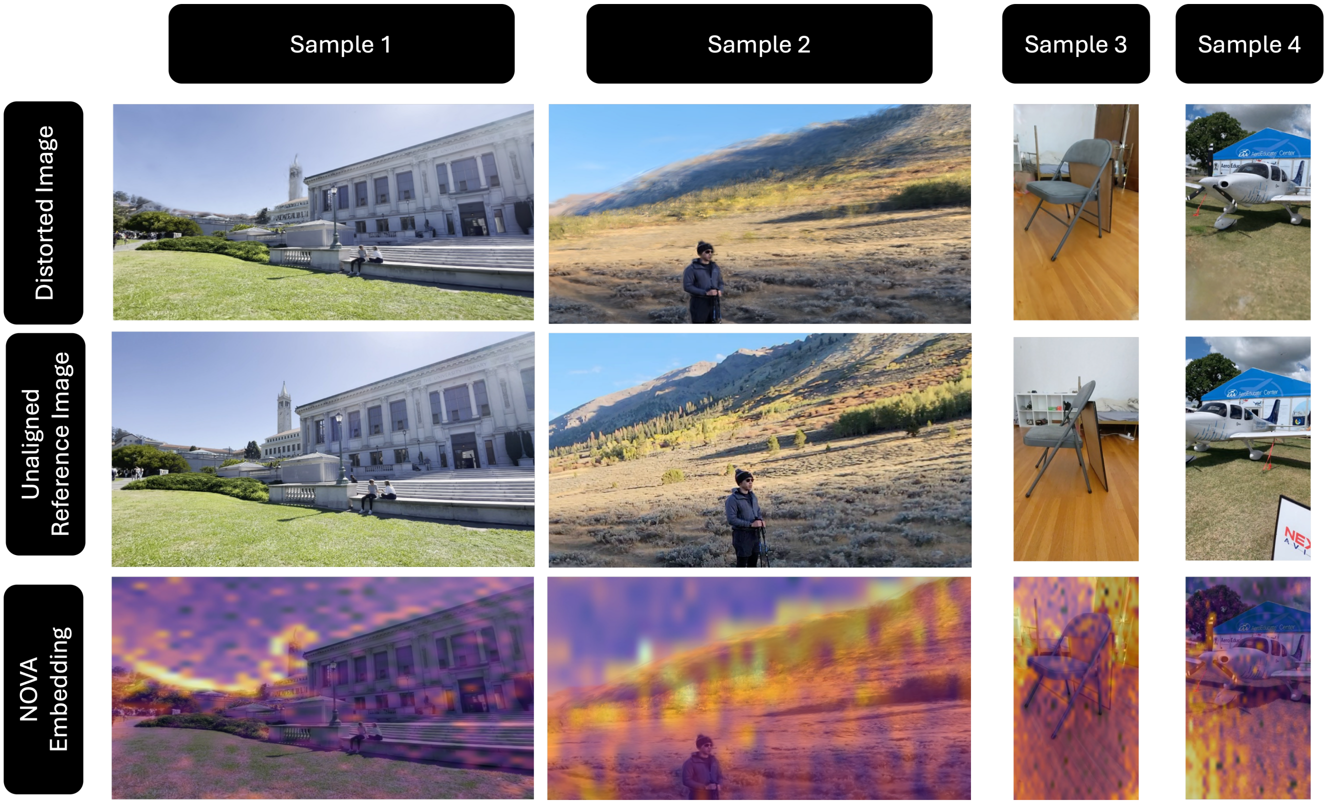

Figure 9. Examples from our proposed NVS NAR-IQA benchmark. Each column shows a distorted novel view (top), an unaligned reference image (middle), and the mismatch heatmap derived from our NOVA embeddings (bottom). The heatmaps reveal how distortions in synthesized views; including geometry errors, ghosting, and structural inconsistencies; are highlighted relative to nearby unaligned references. These examples demonstrate that NOVA can capture perceptually relevant degradations in realistic NVS outputs.

Metrics	Aligned	Non-Aligned	Egypt	Giannini-Hall	aspen	campanile	desolation	dozer	floating-tree	kitchen	library	person	plane	poster	redwoods2	sculpture	storefront	stump	vegetation	Total
Full Reference IQA																				
PSNR	✓		81.0	63.5	50.0	77.0	59.3	50.8	67.2	65.6	47.8	72.7	57.4	50.7	45.2	84.6	22.0	81.7	33.3	60.5
		✓	49.2	49.2	18.8	54.1	37.3	20.3	35.9	50.0	40.3	39.4	45.6	52.2	24.2	52.3	33.9	53.3	33.3	41.0
SSIM [37]	✓		81.0	65.1	57.8	77.0	61.0	55.9	65.6	68.8	37.3	69.7	61.8	38.8	41.9	90.8	32.2	75.0	37.5	60.7
		✓	69.8	47.6	31.2	45.9	47.5	30.5	39.1	57.8	50.7	50.0	44.1	61.2	29.0	80.0	44.1	36.7	37.5	47.8
LPIPS [44]	✓		71.4	71.4	65.6	80.3	74.6	55.9	78.1	85.9	67.2	72.7	79.4	50.7	45.2	86.2	40.7	83.3	37.5	68.7
(Alex)		✓	38.1	65.1	64.1	57.4	62.7	54.2	53.1	50.0	52.2	45.5	66.2	55.2	64.5	41.5	49.2	53.3	83.3	55.2
ST-LPIPS [9]	✓		74.6	74.6	82.8	80.3	62.7	59.3	89.1	76.6	67.2	57.6	75.0	56.7	48.4	86.2	44.1	80.0	83.3	70.1
(Alex)		✓	49.2	50.8	62.5	52.5	54.2	55.9	45.3	31.2	49.3	47.0	58.8	47.8	54.8	40.0	47.5	41.7	83.3	50.0
LPIPS [44]	✓		88.9	73.0	62.5	78.7	67.8	59.3	81.2	90.6	67.2	75.8	72.1	46.3	40.3	87.7	37.3	88.3	37.5	69.2
(VGG)		✓	58.7	63.5	64.1	68.9	67.8	57.6	57.8	73.4	73.1	60.6	54.4	70.1	56.5	43.1	40.7	65.0	58.3	61.0
ST-LPIPS [9]	✓		66.7	73.0	82.8	72.1	71.2	64.4	82.8	78.1	55.2	65.2	76.5	56.7	58.1	81.5	52.5	81.7	75.0	70.0
(VGG)		✓	47.6	49.2	59.4	44.3	52.5	66.1	54.7	45.3	41.8	45.5	63.2	56.7	64.5	43.1	54.2	45.0	83.3	52.8
LPIPS [44]	✓		82.5	73.0	75.0	80.3	74.6	62.7	81.2	85.9	67.2	81.8	79.4	53.7	41.9	86.2	44.1	86.7	41.7	71.7
(Squeeze)		✓	46.0	66.7	57.8	59.0	59.3	50.8	46.9	65.6	61.2	45.5	64.7	70.1	64.5	46.2	52.5	65.0	75.0	58.1
DISTS [5]	✓		71.4	71.4	75.0	75.4	76.3	81.4	75.0	84.4	76.1	75.8	73.5	80.6	62.9	87.7	57.6	88.3	45.8	75.2
		✓	60.3	71.4	76.6	65.6	55.9	74.6	70.3	67.2	77.6	75.8	73.5	83.6	66.1	76.9	67.8	73.3	66.7	71.1
FLIP [2]	✓		84.1	68.3	50.0	57.4	57.6	47.5	65.6	57.8	49.3	60.6	57.4	55.2	53.2	84.6	32.2	80.0	33.3	59.5
		✓	46.0	54.0	31.2	36.1	49.2	27.1	50.0	59.4	53.7	42.4	42.6	58.2	33.9	61.5	33.9	43.3	50.0	45.5
DeepDC [47]	✓		71.4	79.4	78.1	78.7	71.2	79.7	84.4	82.8	67.2	80.3	77.9	82.1	71.0	87.7	64.4	86.7	75.0	77.7
		✓	54.0	71.4	73.4	68.9	72.9	83.1	79.7	70.3	61.2	80.3	67.6	77.6	75.8	75.4	72.9	71.7	66.7	72.1
ZS-IQA [7]	✓		88.9	71.4	62.5	80.3	74.6	59.3	79.7	93.8	73.1	77.3	64.7	40.3	50.0	87.7	37.3	86.7	33.3	69.7
(L2, Dinov1)		✓	42.9	57.1	51.6	67.2	44.1	32.2	43.8	48.4	62.7	56.1	50.0	49.3	59.7	43.1	44.1	41.7	58.3	50.0
ZS-IQA [7]	✓		88.9	71.4	60.9	80.3	69.5	61.0	81.2	93.8	70.1	74.2	63.2	41.8	43.5	89.2	39.0	90.0	29.2	69.0
(Cos, Dinov1)		✓	55.6	63.5	56.2	67.2	62.7	45.8	57.8	65.6	67.2	63.6	58.8	55.2	56.5	40.0	47.5	56.7	66.7	57.8
DreamSim [6]	✓		63.5	68.3	43.8	68.9	55.9	45.8	64.1	65.6	53.7	48.5	57.4	34.3	37.1	76.9	59.3	70.0	54.2	56.9
		✓	20.6	20.6	35.9	19.7	30.5	42.4	28.1	10.9	20.9	10.6	26.5	52.2	43.6	13.9	44.1	15.0	37.5	27.3
No Reference IQA																				
ARNIQA [1]			38.1	44.4	57.8	41.0	55.9	42.4	60.9	67.2	49.2	51.5	55.9	64.2	71.0	46.2	69.5	58.3	62.5	55.1
CONTRIQUE [16]			42.9	69.8	60.9	37.7	52.5	52.5	60.9	59.4	47.8	50.0	61.8	52.2	64.5	66.1	47.5	80.0	41.7	55.8
LAR-IQA [11]			42.9	39.7	43.8	22.9	49.1	44.1	62.5	50.0	49.2	39.4	55.9	77.6	77.4	35.4	84.8	51.7	54.2	51.8
MANIQA [42]			34.9	55.6	64.1	32.8	64.4	54.2	60.9	65.6	43.3	57.6	69.1	61.2	61.3	41.5	62.7	60.0	83.3	57.2
HyperIQA [27]			38.1	42.9	71.9	32.8	61.0	61.0	54.7	64.1	53.7	56.1	63.2	68.7	70.9	40.0	67.8	61.7	66.7	57.0
GraphIQA [28]			55.6	71.4	78.1	36.1	72.9	71.2	68.8	78.1	56.7	72.7	75.0	73.1	80.6	78.5	76.3	78.3	91.7	54.5
TRIQA [29]			58.3	58.7	60.5	58.1	57.1	61.0	59.4	59.4	59.7	58.6	58.6	58.6	58.6	58.6	58.6	58.6	58.6	58.3
AGAIQA [34]			55.6	71.4	78.1	36.1	72.9	71.2	68.8	78.1	56.7	72.7	75.0	73.1	80.6	78.5	76.3	78.3	91.7	70.7
NVS-SQA [23]			33.3	66.7	75.0	21.3	69.5	62.7	67.2	48.4	53.7	48.5	64.7	53.7	74.2	29.2	54.2	73.3	75.0	56.3
Non-Aligned Reference IQA																				
CVRKD [43]	✓		60.3	60.3	50.0	70.5	47.5	50.8	43.8	32.8	52.2	59.1	44.1	29.9	46.8	47.7	42.4	50.0	50.0	49.2
		✓	65.1	58.7	54.7	60.7	55.9	44.1	51.6	43.8	53.7	60.6	36.8	46.3	46.8	44.6	55.9	48.3	37.5	51.3
CrossScore [36]	✓		76.2	44.4	51.6	29.5	55.9	61.0	67.2	70.3	52.2	34.8	67.6	76.1	72.6	33.8	59.3	48.3	62.5	56.5
		✓	71.4	41.3	48.4	26.2	59.3	55.9	67.2	60.9	62.7	51.5	70.6	65.7	62.9	38.5	62.7	45.0	79.2	56.3
NOVA	✓		84.1	85.7	79.7	65.6	76.3	79.7	90.6	89.1	83.6	78.8	82.3	94.0	71.0	93.8	64.4	85.0	70.8	**81.45**
(ours)		✓	82.5	84.1	85.9	65.6	69.5	76.3	90.6	90.6	83.6	80.3	77.9	95.5	67.7	87.7	66.1	78.3	70.8	**80.19**

Table 3. Quantitative comparison (in % accuracy) of IQA models on the NVS NAR-IQA benchmark across 17 diverse scenes. We evaluate full-reference, non-aligned reference, and no-reference metrics under both aligned and non-aligned reference conditions.