gQIR: Generative Quanta Image Reconstruction

Supplementary Material

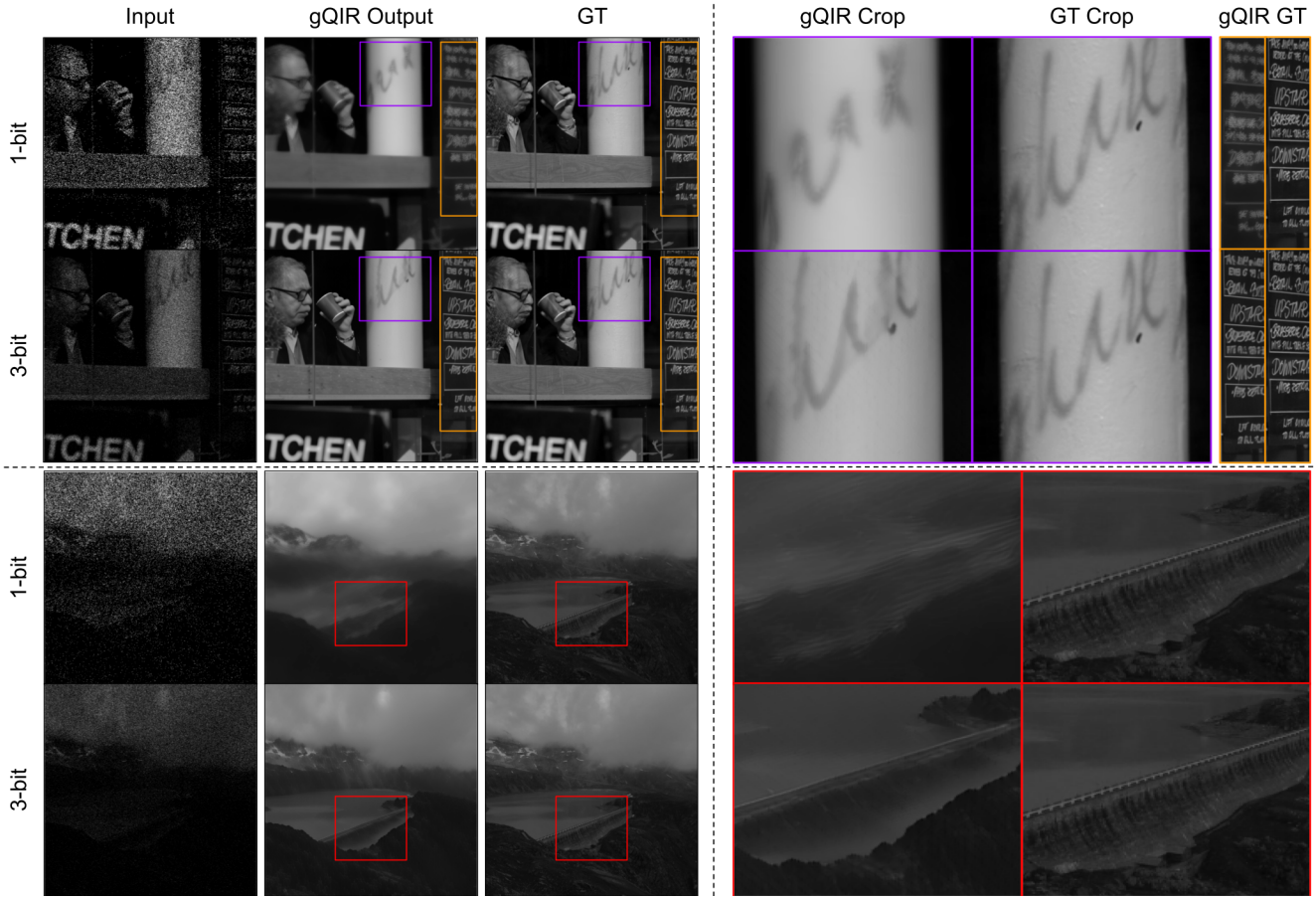

Figure 1. **Extreme 1-bit Quanta Reconstruction.** Lower bit-depth leads to higher downstream hallucinations, detrimental to any image restoration task, thereby motivating the need for 3-bit inputs or *nano-bursts*.

1. Extreme 1-bit Quanta Reconstruction

In Fig. 1, we relax the constraint or the requirement of *nano-bursts* altogether and demonstrate an extreme version of our key paradigm shift in quanta reconstruction by using just a single binary quanta frame for reconstruction.

It is worth noting that hallucinations increase significantly, as visible in the cropped regions in Fig. 1. This led us to the use of *nano-bursts* or 7 binary frame averages to slightly reduce noise or increase the input SNR throughout the main paper. This additionally leads to a dowstream increase in fidelity, as observed quantitatively in Tab. 1. Relying solely on a generative prior in this extreme regime of single binary frame input could prove harmful for true-to-scene reconstruction.

We consider quantifying this or providing a confidence interval for the generated output's trueness as future work.

Table 1. **1-bit vs 3-bit Monochrome gQIR Fidelity.** We relax the constraint of also needing *nano-bursts* by directly utilizing single-bit binary frames. However, this comes at the cost of more hallucinations leading to lower overall fidelity. We evaluated both variants on the same 334-image test set used in the main paper table for single frame quanta reconstruction.

gQIR	PSNR ↑	SSIM ↑	LPIPS ↓
1-bit	25.03	0.76	0.43
3-bit	27.28	0.84	0.32

2. Different SPAD Simulators

Realistic FPS-preserving simulation. When 7 i.i.d binary frames are sampled from 1 GT, motion-blur is completely eliminated. In a realistic setting, the reconstruction algo-

Table 2. **Quanta Video Datasets.** Legend: Sim = Simulated from conventional CMOS sensor, Syn = Synthetically rendered, Min. Res. = Minimum Resolution of each sample. Note that we resize all our datasets to 512^2 due to VRAM limitations.

			Dataset Characteristics				Dataset Complexity		
Dataset	**Videos**	**Sim/Real/Syn**	fps	Min. Res.	Color	GT	Non-Rigid	Text	Faces
QBP [29]	15	Real (Indoor)	∼100k	512×512	×	×	×	×	×
Ma et al. [21]	137	Real	∼100k	512×256	×	×	×	✓	✓
QUIVER [5]	280	Sim (Outdoor)	2000	512×1024	✓	✓	×	✓	✓
bit2bit [19]	7	Real	∼100k	512×512	×	×	×	✓	✓
VisionSim [13]	50	Syn	100	800×800	✓	✓	×	×	×
Ours	390+44135+50	Sim+Real+Syn	24 - 100k+	512×512	✓	✓	✓	✓	✓

rithm/ISP would only have access to a photon-cube and a 3-bit 11 frame burst can only be realized by averaging 77 binary frames into 11 non-overlapping windows of 7 temporal frames each. This results in scene-fps preserving and much more challenging inputs due to motion-blur, as shown in Fig. 2. gQIR was trained to work with a realistic simulator but it also enjoys the SNR boost of QUIVER-like simulations thus setting a new benchmark on the I2-2000fps dataset, as shown in Tab. 3.

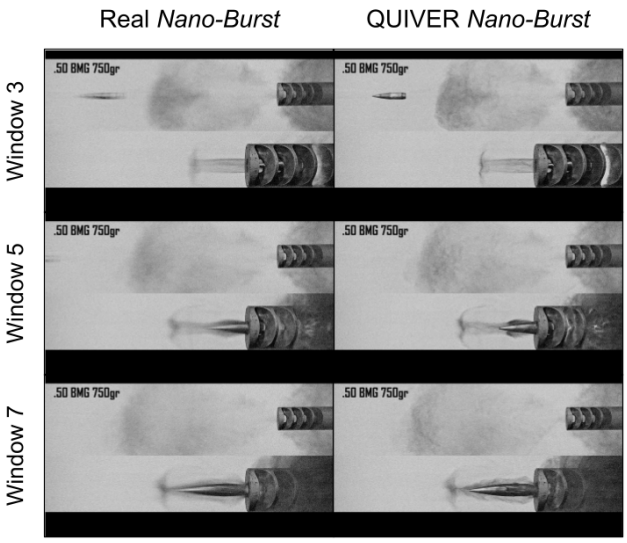

Figure 2. **SPAD Simulators - Realistic vs. QUIVER.** Realistic nano-bursts have motion-blur in each window making the reconstruction task significantly more challenging but closer to real-world operation settings.

3. Training Datasets

We provide comprehensive categorization and statistics for image and vieo datasets used for training our method, while comparing with popularly used quanta datasets in Tab. 2.

4. Stage 2 - Training Stability & Implementation Practices

It is essential for any adversarially trained framework to demonstrate training curves that oscillate around the nash equilibrium [10] instead of converging/collapsing. This provides essential cues for gauging mode or diversity collapse.

4.1. GAN Loss Stabilized At Nash Equilibrium

We visualize the sequential training phases of Stage 2 ($\mathcal{G}_{\text{LoRA}}$): adversarial-only, reconstruction-only, and joint training in Fig. 3.. Clearly, the discriminator and generator are oscillating around or have reached an equilibrium state. See top row in Fig. 3

4.2. Small Initial Gradients Verification

See Fig. 4 for grad norm curves ensuring stability of the GAN training or no vanishing/exploding gradient conditions. This is guaranteed by the diffusion prior weight initialization.

4.3. Architecture Details.

$\mathcal{G}_{lora}$'s Specifications. The LoRA [11] rank r and α_r are set to 256 adding a total of 0.26B parameters to the 0.8B U-Net backbone. We add LoRA modules to the following layers of the latent diffusion U-Net backbone, similar to [18]:

`[to_k, to_q, to_v, to_out.0, conv, conv1, conv2, conv_shortcut, conv_out, proj_in, proj_out, ff.net.2, ff.net.0.proj]`

ConvNext Pyramid Discriminator Modification. We modify HYPIR's [18] Multilevel ConvNext-Large [20] backbone Discriminator's topmost level to accomodate different (smaller) input resolutions dynamically instead of fixed $\geq 512^2$ px. We achieve this by adding a 1×1 bias-free convolutional projection layer followed by spectral norm and LeakyReIU activation, similar to the rest of the pyramid levels. The projection layer dynamically sets the output

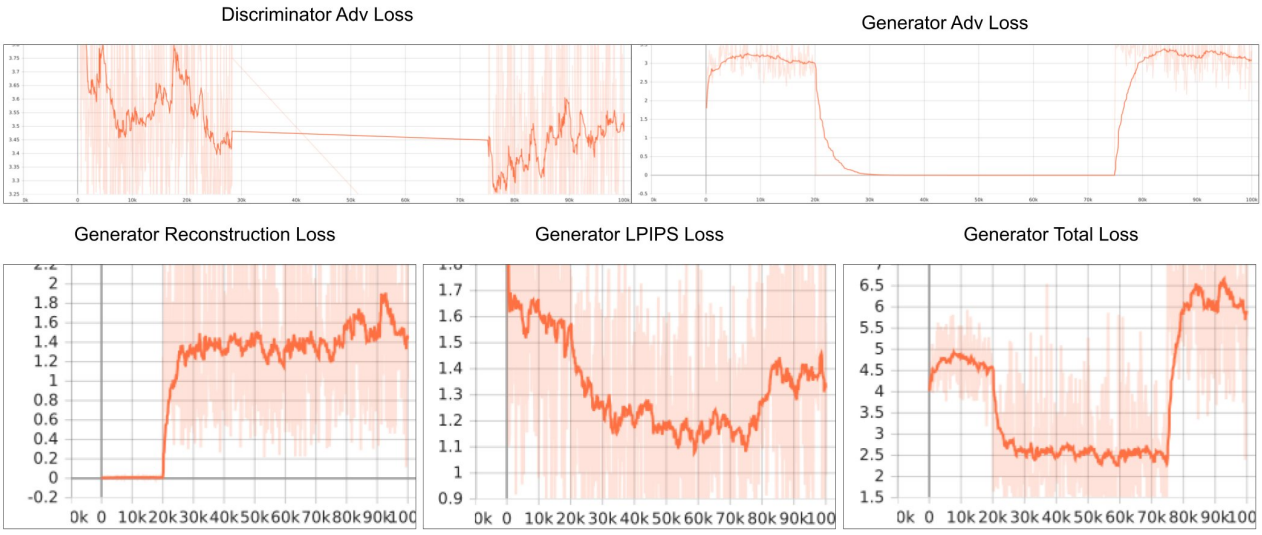

Figure 3. **Stage 2 Training Loss Curves.** We train Stage 2 ($\mathcal{G}_{lora}$) in a hybrid fashion. First, purely adversarial, then purely reconstruction to reduce hallucinations/increase fidelity for the reconstruction task and then combined training to regain any lost generative prowess.

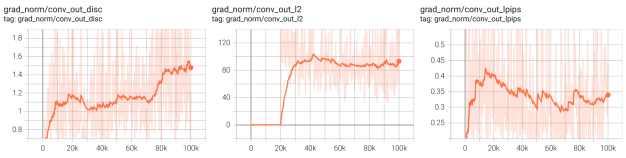

Figure 4. Stable grad norms due to small initial gradients guaranteed by diffusion prior initialization of $\mathcal{G}_{lora}$. Shown are the grad norms of the discriminator, generator and lpips models on the VRAM during training.

projected channels to be: $\max(96, c_{in}/2)$, where c_{in} is the features channels coming from ConvNext-Large.

5. Stage 3 - Burst Extension

5.1. Architecture Details.

FusionVIT. For Stage 3, FusionViT's uses patches of size 2×4 along the temporal volume with 8 attention heads per MHSA block and a total of 3 temporal MHSA blocks that feed the volume into 2 spatial MHSA blocks in window sizes of 8×8 for Swin [16]-like attention computation. This only amounts to 13.23M parameters for the FusionViT. RAFT [25] runs for 20 iterations per flow calculation and then downsamples $\mathcal{O}$ by a factor of $256/64 = 4$ to warp the $4 \times 64 \times 64$ latents.

6. More Qualitative Results

6.1. Single Image Reconstructions

Monochrome. See Fig. 22, Fig. 23, Fig. 24 and Fig. 25 for more comparisons, similar to main paper Fig 3.
Color. See Fig. 26, Fig. 27, Fig. 28 and Fig. 29 for more comparisons on our test set (similar to main paper Fig 3).
Facial Reconstruction. Including FFHQ-dataset [14] with 70,000 high quality faces helps improve facial reconstruction for our method. Notice the detail in hair, which is hard to reconstruct and is a sub research field within facial reconstruction community [28].

6.2. Comparing With A Color-Burst Baseline

Color-QBP [22] proposed color filter arrays (CFAs) over monochrome SwissSPAD 512 and demonstrated simulated results of color burst reconstructions using their proposed RGBW filter. Our real color SPAD prototypes use the conventional Bayer pattern, instead of the RGBW pattern proposed by [22]. Hence, to maintain fairness, we compare against Color-QBP [22] with our Color Burst-gQIR on true prototype Bayer pattern color, 3-bit 11 frame realistic burst, in Fig. 6, using a .

6.3. Real Color SPAD Capture Reconstructions

Our method is sensor-agnostic and light-level agnostic within a certain range (as shown in Fig. 14). This in combination with the generative prior leveraging capabilities, leads to successful, **finetuning-free** transfer to real world settings. We show more qualitative results *without* processing the input (to account for dark count rates and hot pixels)

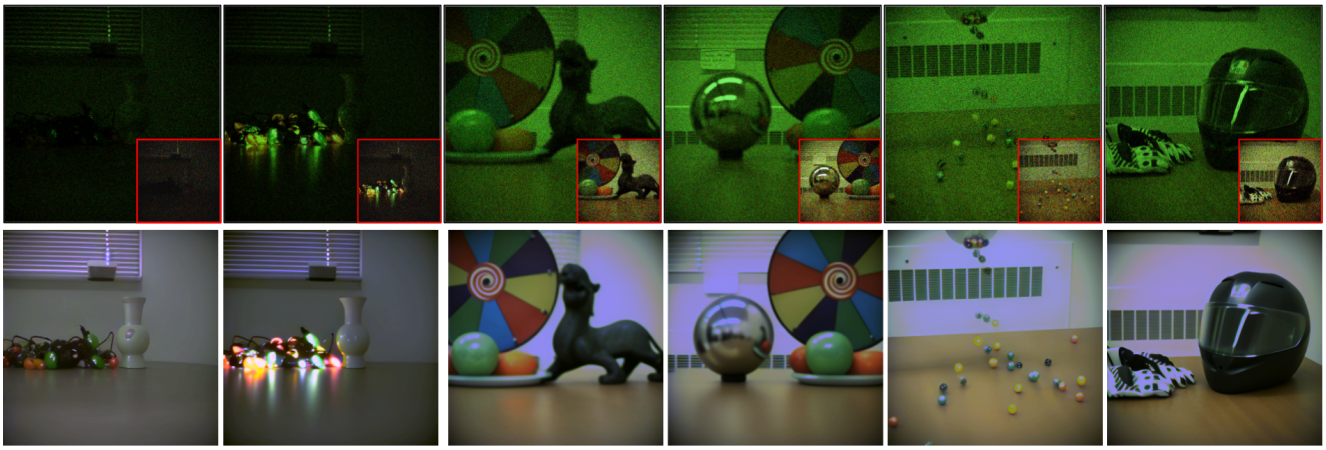

Figure 5. **More Qualitative Real World Testing** We turn on/off the Christmas lights to verify if the lighting + vase physics is implicitly understood and the reconstructions stay plausible. The third column is color-alignment/testing scene while the fourth and last column shows reconstruction on non-lambertian surfaces (zoom in at the metal sphere and motorcycle helmet's visor)

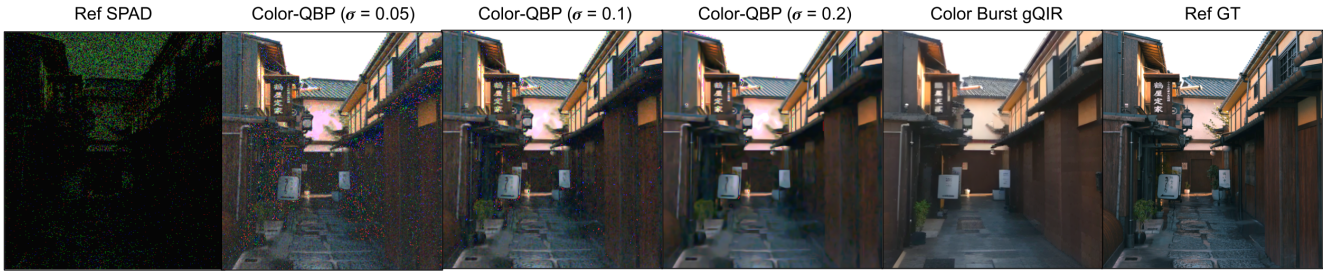

Figure 6. **Color Burst Reconstructions - Color-QBP [22] vs Ours.** As σ for BM3D [8] is increased, the noise reduces for Color-QBP at the cost of blurring. Our method fails to reconstruct text due to the prior's text-drawing inability but excels in creating a perceptually pleasing output, without tweaking any hyper-parameters. We used the realistic burst simulations on the static JapanAlley Blender scene.

in Fig. 5. Automatic white balancing with gray world assumption is applied.

7. More Quantitative Results

7.1. I2-2000fps Full Benchmark

See Tab. 3 for our method's performance on slightly out-of-domain PPP (3.25 instead of train-time 3.5), QUIVER [5]-style simulations of the I2-2000fps test set. We include all the baselines, designed or adapted for the quanta reconstruction task so far. Our method sets a new top score.

7.2. Burst Reconstruction - Scene by Scene Metrics

Our method suffers from 'content-drift' due to the inherent reliance on generative prior to retain perceptualness; leading to a slightly poorer E^* error compared to learning based methods as shown in Tab. 4. However, our method achieves superior fidelity, both visually and quantitatively.

Table 3. **Burst Reconstruction Fidelity on I2-2000fps.** Despite being trained on a slightly different photon-per-pixel level (3.25 instead of 3.5), gQIR operates out-of-domain while reconstructing superiorly. gQIR sets a new benchmark high score.

Method	PSNR ↑	SSIM ↑
Transform Denoise [2]	21.317	0.718
QBP [29]	21.548	0.703
Student-Teacher [7]	18.720	0.401
RVRT [17]	19.412	0.354
EMVD [23]	20.019	0.587
FloRNN [15]	21.034	0.679
MemDeblur [12]	19.488	0.387
Spk2ImgNet [27]	20.395	0.564
QUIVER [5]	26.214	0.790
QuDi [6]	28.641	0.811
(Ours) Burst-gQIR (3.25 PPP)	**30.811**	**0.868**

Table 4. **Burst Reconstruction: Test Sequence-wise Fidelity and Video Coherence.** We slide the burst window to reconstruct the entire video except the first 5 and trailing 5 frames to compute the temporal stability or warping metric ($E^* = 10^3.E_{warp}$).

		QBP [29]			QUIVER [5]			**Burst SD2.1-gQIR**		
Video	GT fps	PSNR↑	SSIM↑	E^*↓	PSNR↑	SSIM↑	E^*↓	PSNR↑	SSIM↑	E^*↓
XVFI boat	1000	10.978	0.491	**0.7**	20.811	0.639	0.9	**27.594**	**0.716**	1.7
XVFI train	1000	13.049	0.249	**0.3**	**25.191**	0.862	1.8	24.050	**0.707**	3.4
I2-2k test-14	2000	13.286	0.532	**0.6**	26.092	**0.857**	1.5	**30.221**	0.830	1.8
I2-2k test-20	2000	18.631	0.707	1.9	23.338	0.917	2.8	**32.089**	**0.945**	**1.2**
I2-2k test-21	2000	18.255	0.485	**0.7**	23.495	0.867	1.1	**32.089**	**0.892**	1.2
I2-2k test-31	2000	14.001	0.470	**0.7**	27.318	**0.858**	1.5	**30.458**	0.845	1.8
XD jetengine	2000	14.706	0.162	**0.2**	20.588	**0.861**	1.7	**30.121**	0.853	2.5
XD tank	12,000	8.870	0.479	2.3	22.391	0.853	**2.2**	**26.116**	**0.864**	3.0
XD explosion	50,000	19.783	0.161	**0.4**	17.456	0.573	2.1	**28.976**	**0.880**	4.4
XD padlock	80,000	11.278	0.546	**0.5**	24.826	0.945	1.6	**34.685**	**0.964**	1.4
XD moreguns	100,000	9.262	0.696	2.4	15.219	0.718	1.9	**31.758**	**0.915**	**1.1**
Cumulative:		13.380	0.448	**1.0**	22.429	0.814	1.7	**29.832**	**0.856**	2.1

8. More Ablations

8.1. Running RAFT Directly on 3-bit SPADs

Directly running RAFT [25] on noisy but still easier demosaiced toy inputs still leads to incorrect flow estimates $\mathcal{O}$ as shown in Fig. 7. This led us to the recursive design of using

Figure 7. We try to solve the toy problem of aligning and merging (much easier) demosaiced RGB inputs. However, notice the highly erroneous optical flow measure that leads to ghosting artifacts on naively aligning and merging the 11 frame input burst.

our Stage 2 first to completely reconstruct every frame in the burst for a cleaner optical flow estimate ($\mathcal{O}$).

8.2. Effect of Boundary Convolution in FusionViT.

Due to the ViT's unpojection of patches, a grid-boundary artifact persists through the pipeline which the decoder reconstructs as is, resulting in a patch-y reconstruction, as shown in Fig. 8. We fix this issue by introducing a critical Group-

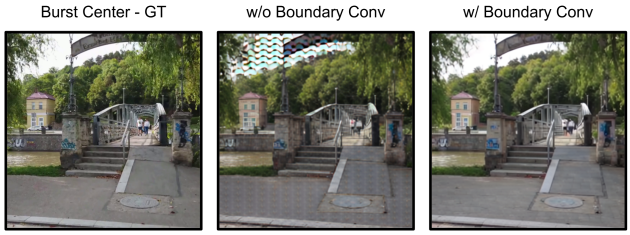

Figure 8. **FusionViT - Boundary Convolution Ablation.** Without the boundary convolution, the fused tokens forming the clean merged latent representation, do not smooth out along the boundary of each patch resulting in the shown grid-like artifact in the latent code which the VAE decoded similarly.

Norm [26] and 3×3 2D convolutional layer.

8.3. Latent Fidelity Visualization. Stage 1 vs 3

We qualitatively compare Stage 1: qVAE and Stage 3: qVAE+FusionViT in Fig. 10. Clearly, adding the fusionViT and the extra context from the burst frames leads to a higher alignment with the GT latent.

8.4. Spatial vs Temporal Fidelity - S1 vs. S2 vs. S3

In Fig. 19 and Fig. 9, we also show the visual fidelity difference between all 3 stages (using QUIVER sim) and s2

Figure 9. **Stage 2 (gQIR) vs Stage 3 (Burst-gQIR) On Realistic Burst of Nano-Bursts.** Motion-blur ridden nano-bursts result in Stage 2 to replicate motion blur in the output due to lack of any temporal context.

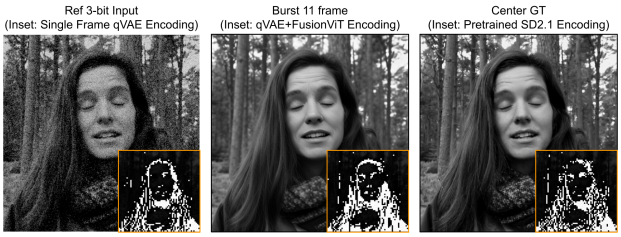

Figure 10. Insets show the encoded latents just by the qVAE in the first column, the qVAE + FusionViT in the second and the frozen SD2.1 VAE for the GT in the last column.

vs s3 (using realistic sim) respectively. Often, Stage 2 deviates from the ground truth or introduces blur due to the motion-blur present in the realistic nano-burst. Whenever it deviates due to the inherent generativeness of the network, it amplifies overall perceptualness at the cost of scene fidelity.

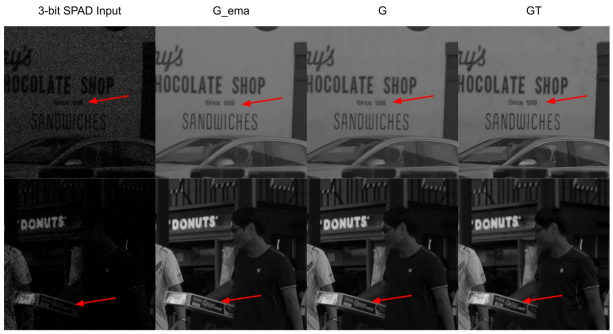

Figure 11. **Text Limitation of SD2.1 T2I prior.** The qVAE fails to reconstruct text in ambiguous regions that can not be further corrected by the LoRA UNet due to the prior's inability to draw meaningful text [3, 4]. Zoom in for better detail.

9. Text Problem With SD 2.1

It is a common and widely known issue with T2I priors to draw coherent and meaningful text [3, 4]. We are afflicted by the same problem due to our selection of Stable Diffusion 2.1 [24]. However, since we adapt the prior to a reconstruction task, this issue is observed only in regions made ambiguous due to shot noise, as shown in Fig. 11.

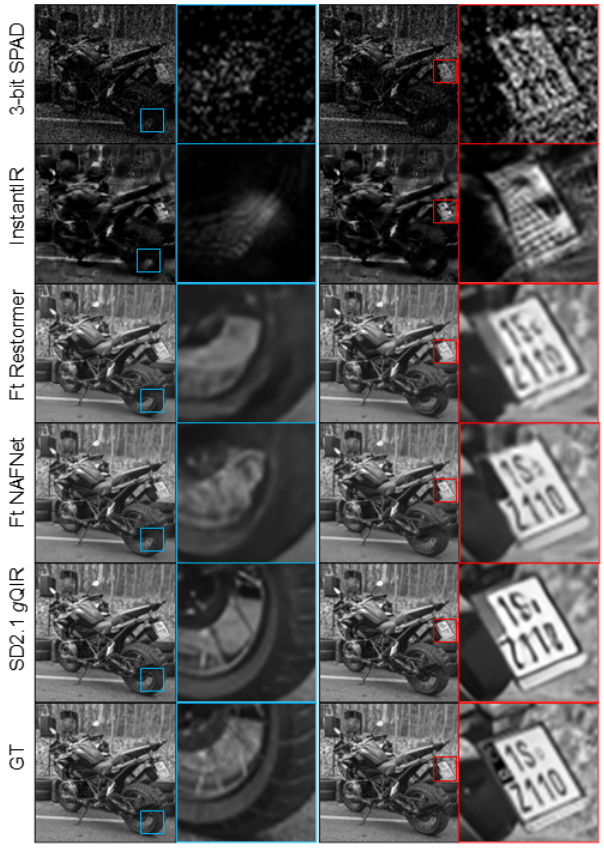

Figure 12. **Prior's Ability in Ambiguous Regions.** The prior provides, even better than GT sometimes, when the 'to-reconstruct' region is supported by the learned natural image manifold but significantly suffers when text is present, owing to the lack of text-focused pre-training of the chosen (SD2.1 [24]) prior. This can happen simultaneously in the same scene as shown.

The generative prior excels in these ambiguous regions, as shown in Fig. 12 but completely breaks down when text

is present due to the T2I prior's limited representational power. Instead of fully understanding the text, it generates some artifacts with text-like structure.

Table 5. **Ablation - qVAE Replacement.** Comparison of SD2.1 and SD3.5 qVAEs after quanta alignment training. Increasing latent capacity $4\times$ yields a pronounced perceptual gain, consistent with visually sharper text and high-frequency content.

Model	# Params	PSNR ↑	MUSIQ ↑
SD 2.1 qVAE	83.65 M	26.320	31.649
SD 3.5 qVAE	83.82 M	**26.443**	**36.845**

Scaling Latent Dimensionality of qVAEs. Stable Diffusion 3 ([9]) increases the representation power of the VAE by increasing the number of latent channels by $4\times$. We fix the text problem for our use case (as shown in main paper figure 7) by aligning SD3.5's VAE (Stage 1). In Fig. 18, we demonstrate the higher recovery of high-frequency details due to the $4\times$ larger latent. Tab. 5 shows quantitative consistency.

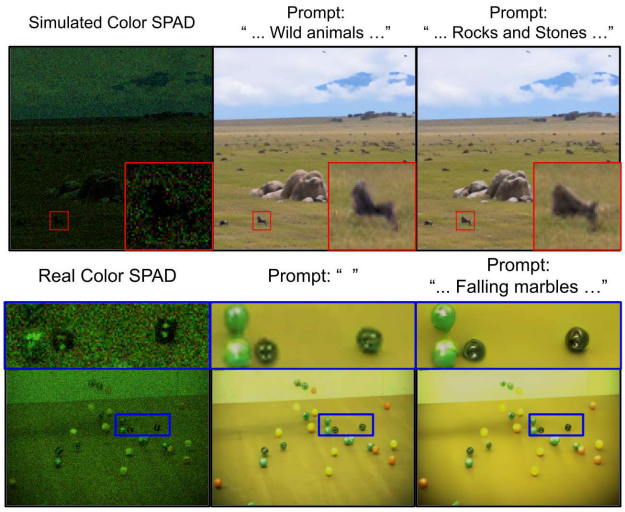

Figure 13. **Prompts for semantic guidance.** Reconstruction in ambiguous regions can be guided during inference for free (null-prompt trainings), owing to the rich text-image cross-attention features in our base prior: Stable Diffusion 2.1 [24]. Our method generalizes to real color SPADs as well.

10. Prompts for Semantic Guidance

Photon-sparse regions are inherently ambiguous, often causing the generative prior to hallucinate details. Incorporating additional modalities can help guide and constrain the reconstruction. For instance, the rich semantic and textual understanding acquired during the prior's large-scale pretraining is preserved due to the addition of only LoRA parameters to the base latent diffusion U-Net. This allows the pretrained embeddings to be leveraged without additional textual training or prompt engineering, for free.

We present results for a simulated SPAD color scene (Row 1) and a real-world SPAD color acquisition captured by our system (Row 2) in Fig. 13. Aligning the scene's ambiguous regions with a true content description can significantly boost reconstruction fidelity or enable novel creative applications (e.g., Column 3).

The guiding signal exploits the basic fact that the task is inherently ill-posed or probabilistic in nature thereby allowing multiple plausible outcomes. This property also makes our framework future-proof and extensible to additional control modalities.

11. Generalizability

We show more qualitative results across two challenging axes that our model was never trained on: i) varying light levels (≤ 3.5 PPP) and ii) Deforming scenes.

11.1. Extreme Low Light

We vary the light levels all the way from extreme low light to brighter than training photons-per-pixel (PPP) level in Fig. 14. Our (stage 2) method can generalize to low-light regimes as well, barring the brightness of the output due to constant PPP training. One can train our method on varying PPP levels to fix this issue.

11.2. Deforming Scenes

We test the scene-rigidity constraint that most prior work assumes due to the inherent limitation of pre-trained or traditional optical flow estimators in their pipelines.

As a result of the generative prior's extensive pretraining, which encodes strong physical inductive biases, our method can operate in deforming conditions without the need for long temporal context. We show realistically simulated burst outputs using stage 3 on deformable scenes in Fig. 15. Additionally, we show an extremely challenging glass shattering video originally captured at 375,000 fps, reconstructed by sliding the burst window across in Fig. 16

12. A Small Note on Latency.

All network latencies are measured at an input resolution of 512^2 and averaged over 10 iterations. See Tab. 6.

Table 6. **Stage-wise Latencies.** Our propsed FusionViT is extremely lightweight at $\sim$ 13M parameters and adds minimal latency (in the order of 10s of milliseconds) to the iterative processing of 11 single bit frames using S2 (11 * 0.26 = 2.86s).

Network	Recon Speed (sec/iter)	Effective fps
(S1) qVAE	0.147s	6.825 fps
(S2) gQIR	0.269s	3.724 fps
(S3) Burst-gQIR	2.890s	0.346 fps

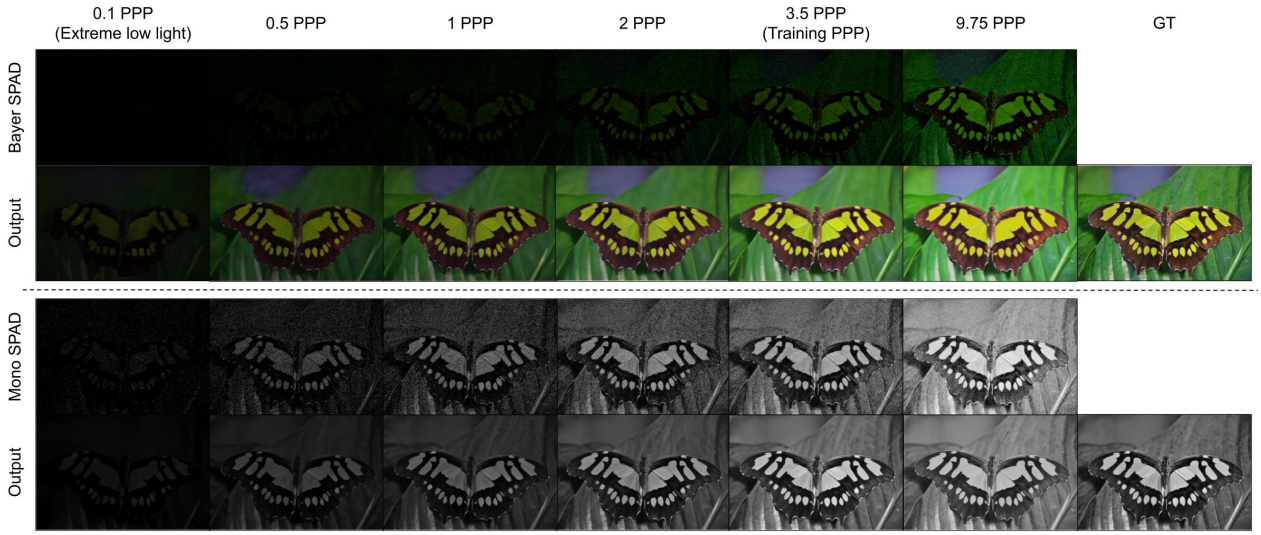

Figure 14. **Varying Light Levels.** Our method generalizes well to non-training photons-per-pixel regimes as well. However, since it is not trained on the shown PPPs, it fails to correct for the output brightness. Focus on the butterfly's antenna to see increasing output fidelity with the ground truth.

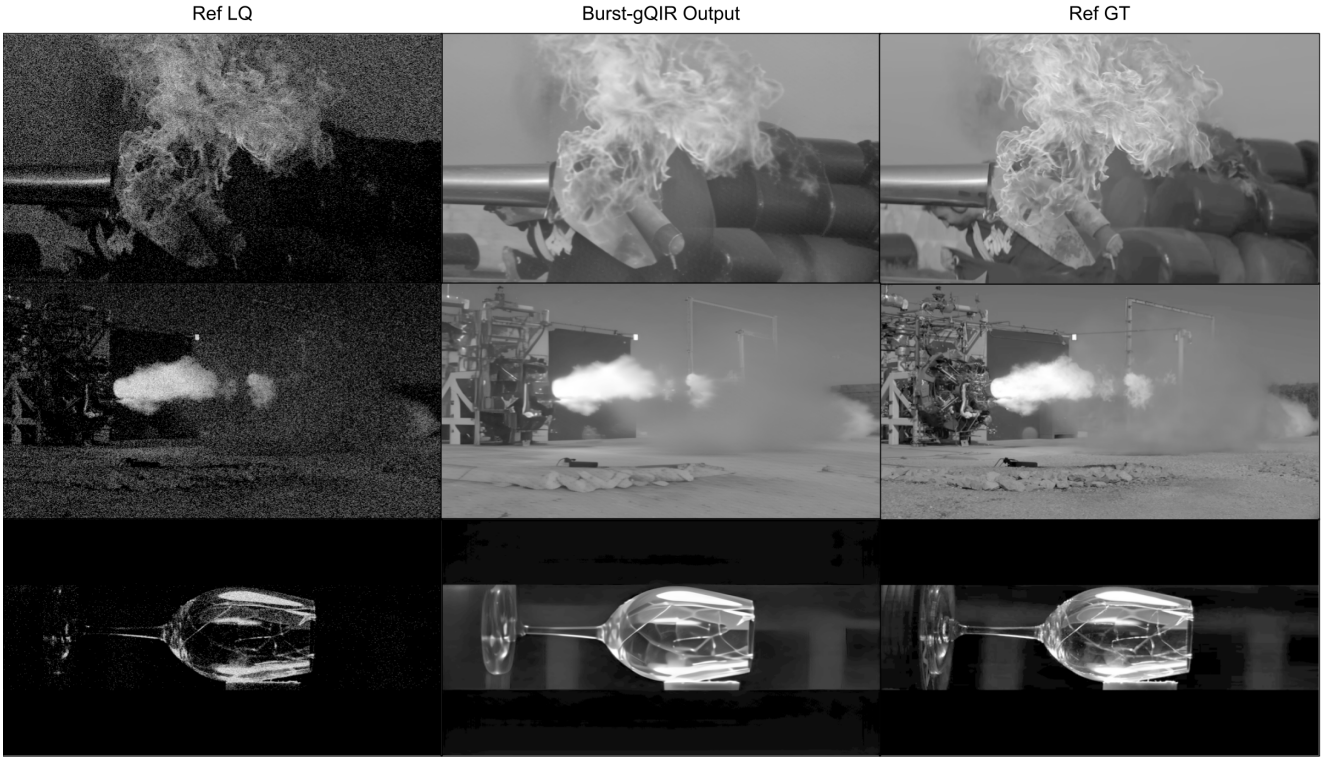

Figure 15. **Burst-gQIR Generalizes to Deforming Scenes.** This emergent behavior is due to the inherent understanding of physics learned during the prior's (large) internet-scale pre-training.

13. Discussion & Open Questions

The entire quest to design a Quanta ISP begins from the failure of pre-trained RGB denoising methods, as shown

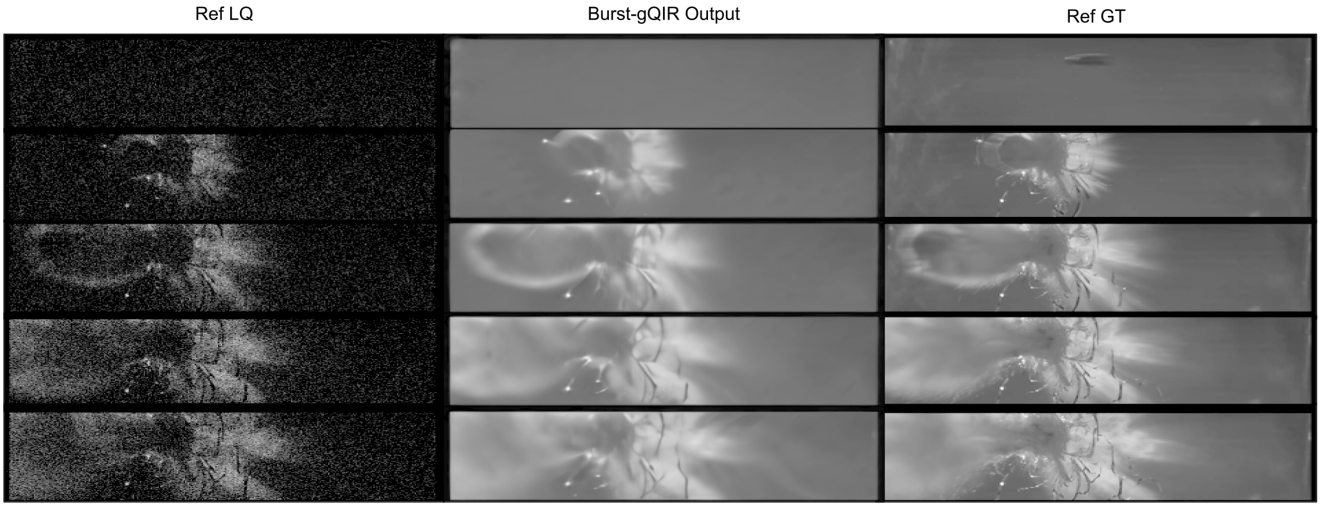

Figure 16. **Realistic Deformable Video Reconstruction using Burst-gQIR at 375,000fps.** Due to the averaging for each nano-burst frame and the extremely fast motion of the bullet, it completely fades into the scene and is un-recoverable.

in Fig. 17. However, quanta literature has only seen burst denoising/imaging pipelines. We present gQIR that brings a two-front paradigm shift in quanta imaging. First, it enables single quanta frame reconstructions instead of relying on long temporal bursts. Second, it utilizes an internet-scale generative prior to enable reconstruction for the first time, serving as a bridge between generative techniques and quanta sensors.

gQIR achieves the highest fidelity and perceptual scores in photon-starved regimes, compared to adapted as well as prior works, to the best of our knowledge. However, gQIR comes with its own limitations and poses several open-ended questions:

1. Is it possible to quantify hallucination levels or provide a confidence interval for true-to-scene reconstruction?
2. Can a +prompt training aid fidelity? Also, could text-based reasoning solve occlusion-disocclusion for video reconstruction?

References

[1] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6228–6237, 2018. 13, 14

[2] Stanley H. Chan, Omar A. Elgendy, and Xiran Wang. Images from bits: Non-iterative image reconstruction for quanta image sensors. *Sensors*, 16(11), 2016. 4

[3] Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. Textdiffuser: Diffusion models as text painters. *arXiv preprint arXiv:2305.10855*, 2023. 6

[4] Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. Textdiffuser-2: Unleashing the power of language models for text rendering. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part V*, page 386–402, Berlin, Heidelberg, 2024. Springer-Verlag. 6

[5] Prateek Chennuri, Yiheng Chi, Enze Jiang, GM Dilshan Godaliyadda, Abhiram Gnanasambandam, Hamid R Sheikh, Istvan Gyongy, and Stanley H Chan. Quanta video restoration. In *European Conference on Computer Vision*, pages 152–171. Springer, 2024. 2, 4, 5

[6] Prateek Chennuri, Dongdong Fu, and Stanley H. Chan. Quanta diffusion, 2025. 4

[7] Yiheng Chi, Abhiram Gnanasambandam, Vladlen Koltun, and Stanley H. Chan. Dynamic low-light imaging with quanta image sensors. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI*, page 122–138, Berlin, Heidelberg, 2020. Springer-Verlag. 4

[8] Kostadin Dabov, Alessandro Foi, Vladimir Katkovnik, and Karen O. Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Transactions on Image Processing*, 16:2080–2095, 2007. 4

[9] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*. JMLR.org, 2024. 7

[10] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, page 2672–2680, Cambridge, MA, USA, 2014. MIT Press. 2

[11] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *In-*

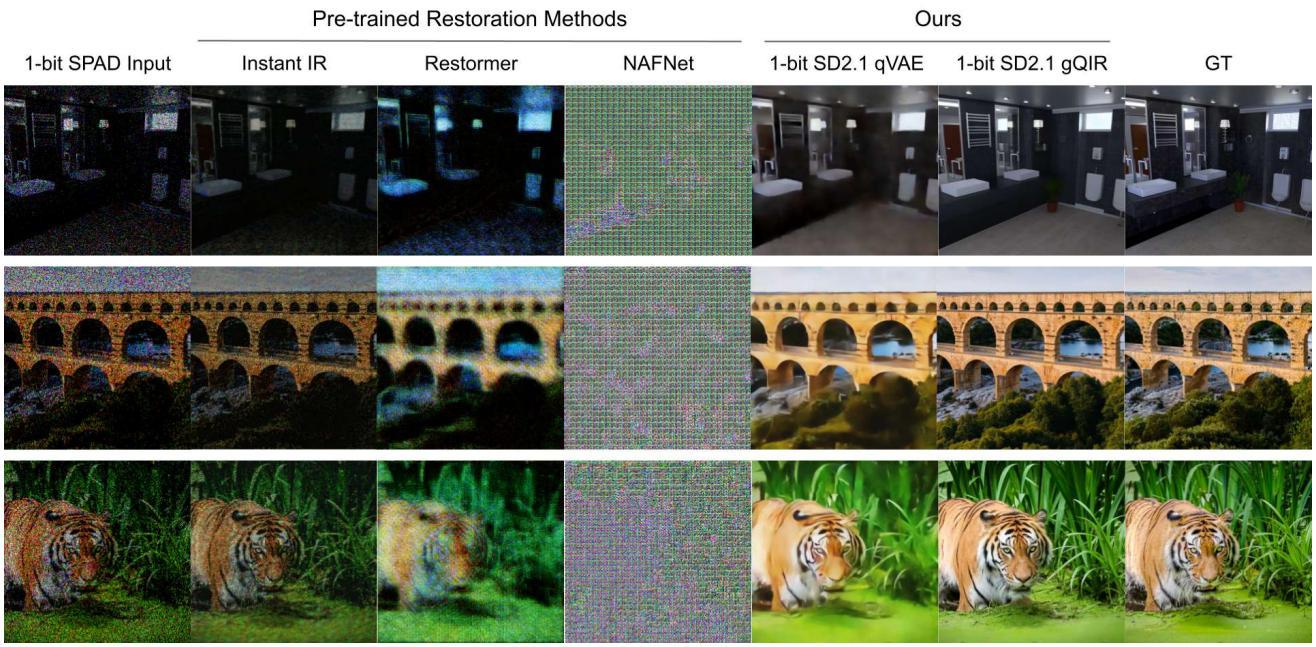

Figure 17. **Conventional Pretraining Failure on Quanta Sensing & Extreme 1-bit Single Frame Quanta Reconstruction.** The extreme regime of quanta sensing is highly unsuitable for conventional sensor trained models. We show the most extreme case of single, 1-bit quanta frame reconstruction. Our adaptation of the generative prior results in highly photorealistic results. However, the hallucinations also increase significantly compared to 3-bit nano-bursts. For ex - The generative prior decides to close the tiger's mouth while the true reality shows otherwise highlighting the increasing ill-posedness of 1-bit quanta reconstruction.

ternational Conference on Learning Representations, 2022. 2

[12] Bo Ji and Angela Yao. Multi-Scale Memory-Based Video Deblurring . In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1909–1918, Los Alamitos, CA, USA, 2022. IEEE Computer Society. 4

[13] Sacha Jungerman, Max Leblang, Shantanu Gupta, and Kaustubh Sadekar. visionsim. https://github.com/WISION-Lab/visionsim, 2025. 2

[14] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4396–4405, 2018. 3

[15] Junyi Li, Xiaohe Wu, Zhenxing Niu, and Wangmeng Zuo. Unidirectional video denoising by mimicking backward recurrent modules with look-ahead forward ones. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVIII*, page 592–609, Berlin, Heidelberg, 2022. Springer-Verlag. 4

[16] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. *arXiv preprint arXiv:2108.10257*, 2021. 3

[17] Jingyun Liang, Yuchen Fan, Xiaoyu Xiang, Rakesh Ranjan, Eddy Ilg, Simon Green, Jiezhang Cao, Kai Zhang, Radu Timofte, and Luc Van Gool. Recurrent video restoration transformer with guided deformable attention. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2022. Curran Associates Inc. 4

[18] Xinqi Lin, Fanghua Yu, Jinfan Hu, Zhiyuan You, Wu Shi, Jimmy S. Ren, Jinjin Gu, and Chao Dong. Harnessing diffusion-yielded score priors for image restoration, 2025. 2

[19] Yehe Liu, Alexander Krull, Hector Basevi, Ales Leonardis, and Michael W. Jenkins. bit2bit: 1-bit quanta video reconstruction via self-supervised photon prediction. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 2

[20] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. *CoRR*, abs/2201.03545, 2022. 2

[21] Sizhuo Ma, Paul Mos, E Charbon, and Mohit Gupta. Burst vision using single-photon cameras. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023. 2

[22] Sizhuo Ma, Varun Sundar, Paul Mos, Claudio Brushini, Edoardo Charbon, and Mohit Gupta. "seeing photons in color". *"ACM Transactions on Graphics (TOG)"*, 2023. 3, 4

[23] Matteo Maggioni, Yibin Huang, Cheng Li, Shuai Xiao, Zhongqian Fu, and Fenglong Song. Efficient multi-stage video denoising with recurrent spatio-temporal fusion. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3465–3474, 2021. 4

[24] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and BjÃ¶rn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the*

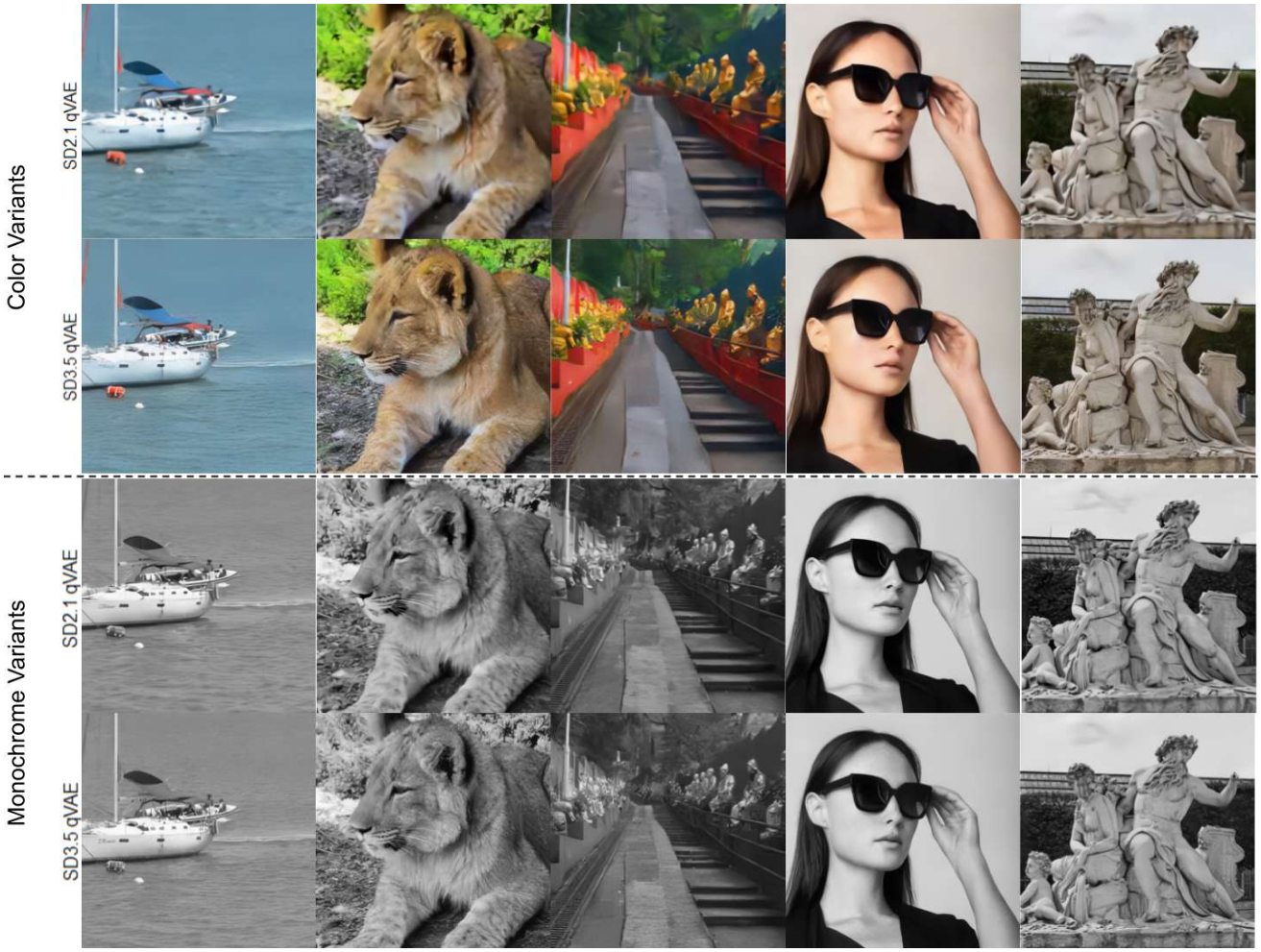

Figure 18. **SD2.1 vs SD3.5 qVAE Comparison.** Increasing latent dimensionality ($4\times$) recovers more high-frequency details in the qVAE's alignment phase at the mere cost of 0.2% increase in VAE parameters.

IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2022. 6, 7

[25] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II*, page 402–419, Berlin, Heidelberg, 2020. Springer-Verlag. 3, 5

[26] Yuxin Wu and Kaiming He. Group normalization, 2018. 5

[27] Jing Zhao, Ruiqin Xiong, Hangfan Liu, Jian Zhang, and Tiejun Huang. Spk2imgnet: Learning to reconstruct dynamic scene from continuous spike stream. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11991–12000, 2021. 4

[28] Yuxiao Zhou, Menglei Chai, Alessandro Pepe, Markus Gross, and Thabo Beeler. Groomgen: A high-quality generative hair model using hierarchical latent representations. *ACM Trans. Graph.*, 42(6), 2023. 3

[29] Sizhuo "Ma, Shantanu Gupta, Arin C. Ulku, Claudio Brushini, Edoardo Charbon, and Mohit" Gupta. "quanta burst photography". *"ACM Transactions on Graphics (TOG)"*, "39"("4"), "2020". 2, 4, 5

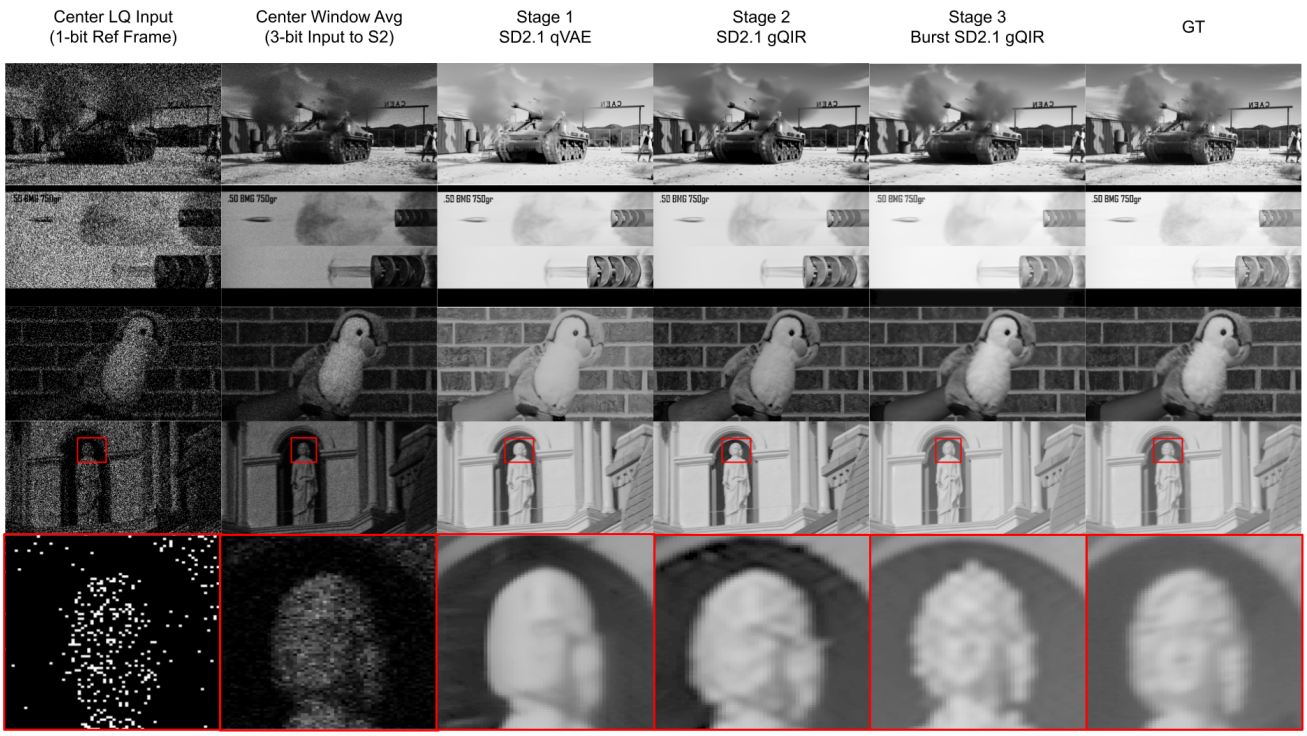

Figure 19. **Ablation - Perceptual Fidelity Difference between Stage 1, 2 and 3.** We use QUIVER simulations for these reconstructions.

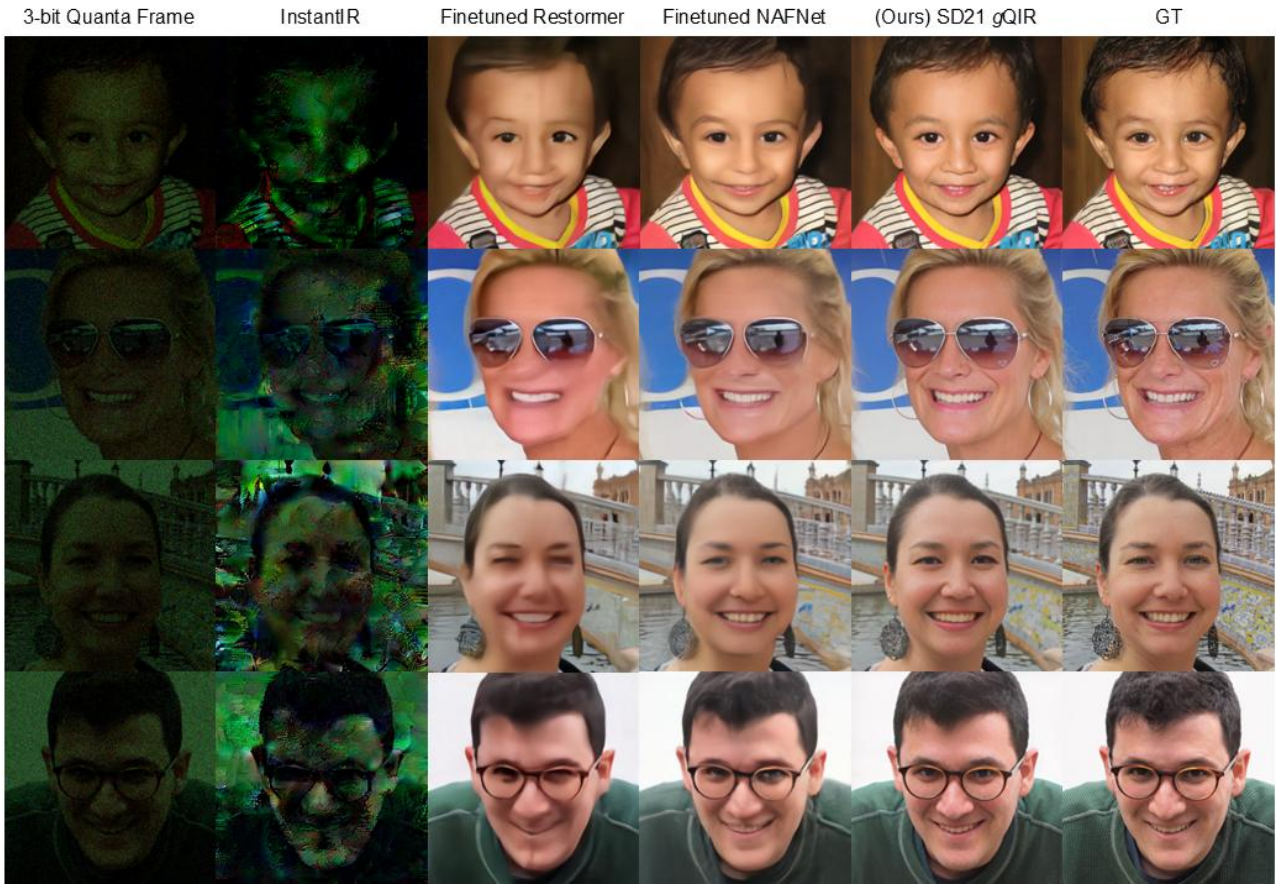

Figure 20. **Color Facial Reconstruction.** We notice the perception-distortion trade-off [1] with conventional baselines keeping the distortion low by oversmoothing. This is detrimental to the facial reconstruction task. Our method is also limited by the fact that it does not preserve the identity of the subject completely. However, notice the background in row 3, showing our method's superior attention-to-detail in completely lost regions due to shot-noise.

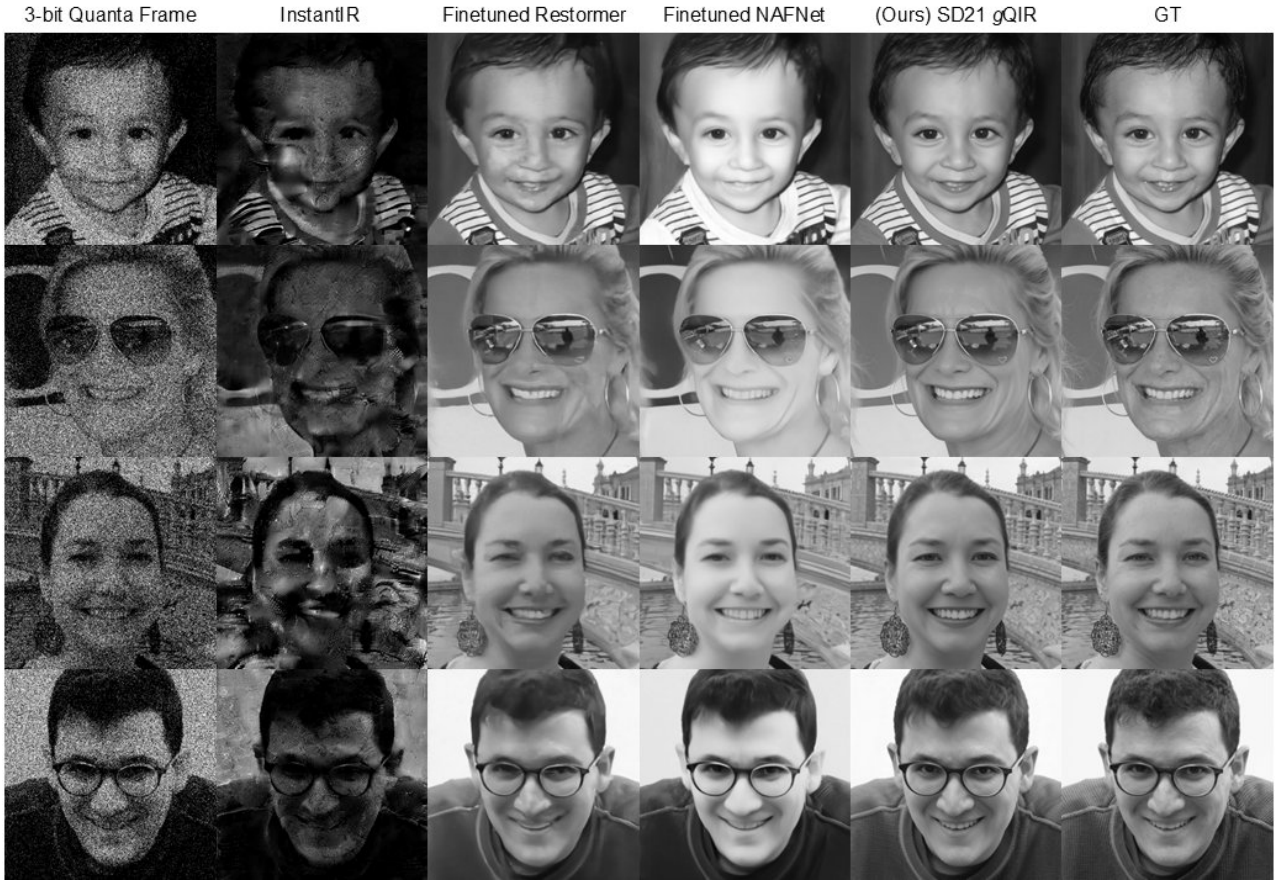

Figure 21. **Monochrome Facial Reconstruction.** We notice the perception-distortion trade-off [1] with conventional baselines keeping the distortion low by oversmoothing. This is detrimental to the facial reconstruction task. Our method is also limited by the fact that it does not preserve the identity of the subject completely.

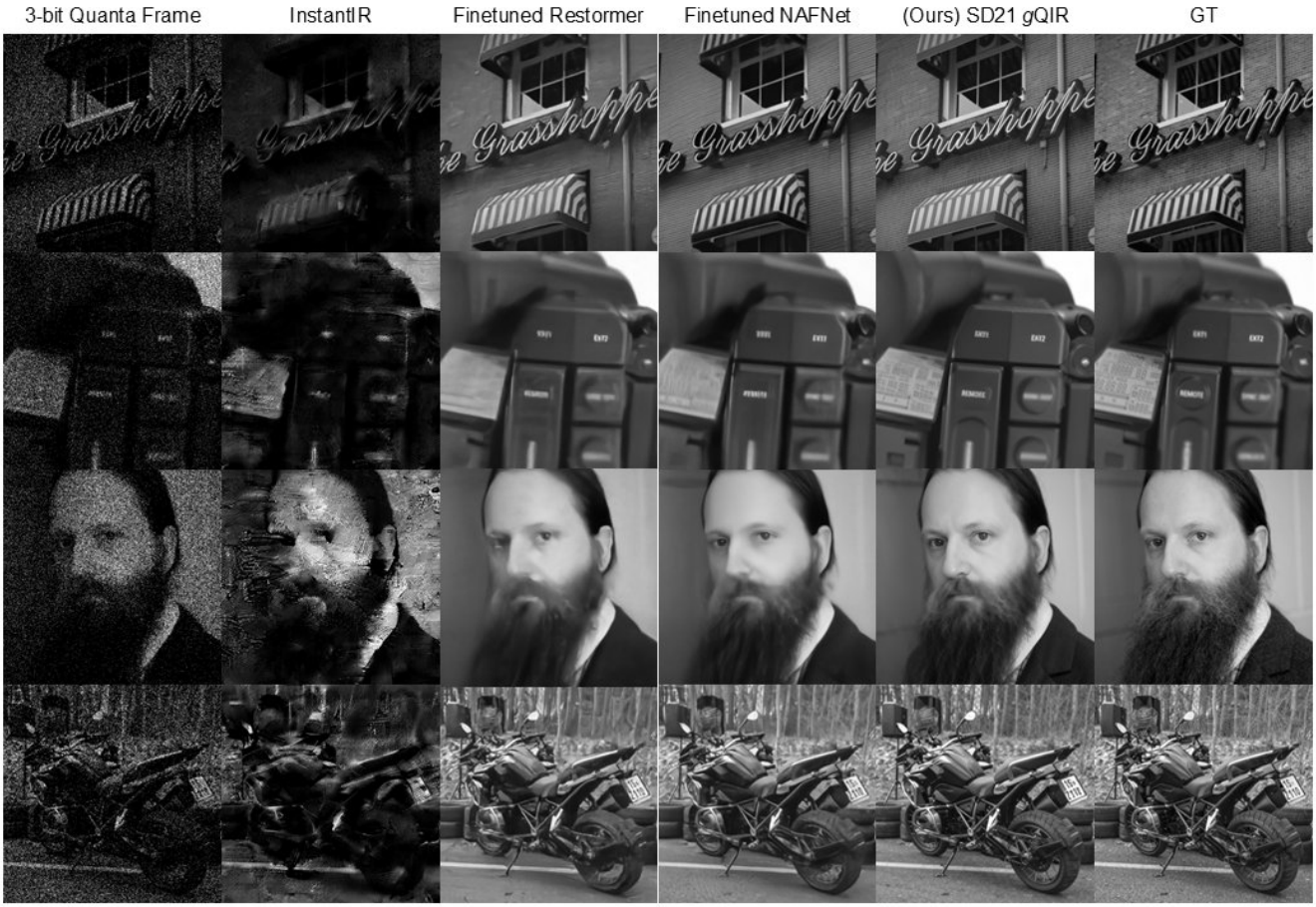

Figure 22. **Mono Single Frame Reconstruction Comparison (1/4).**

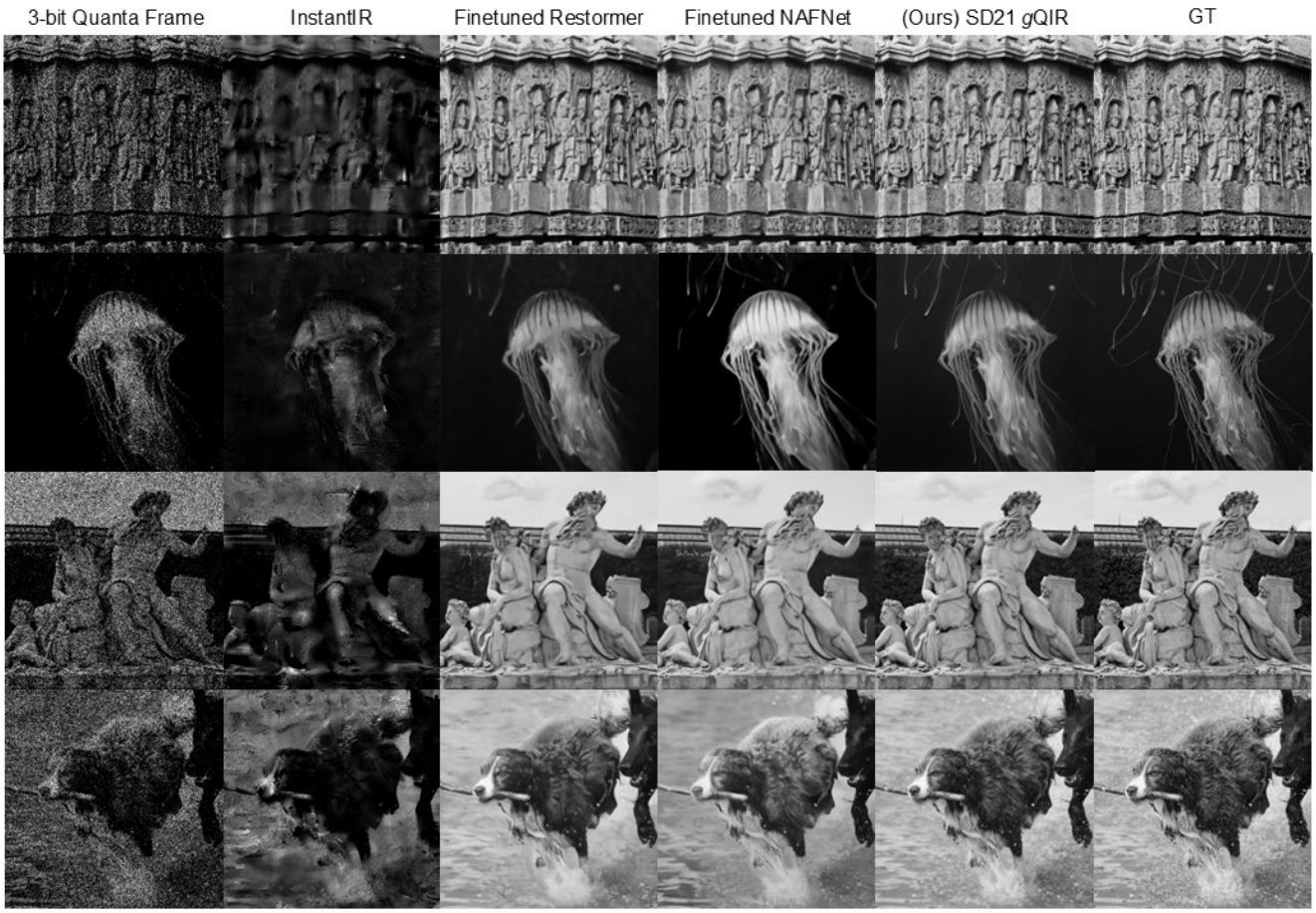

Figure 23. **Mono Single Frame Reconstruction Comparison (2/4).**

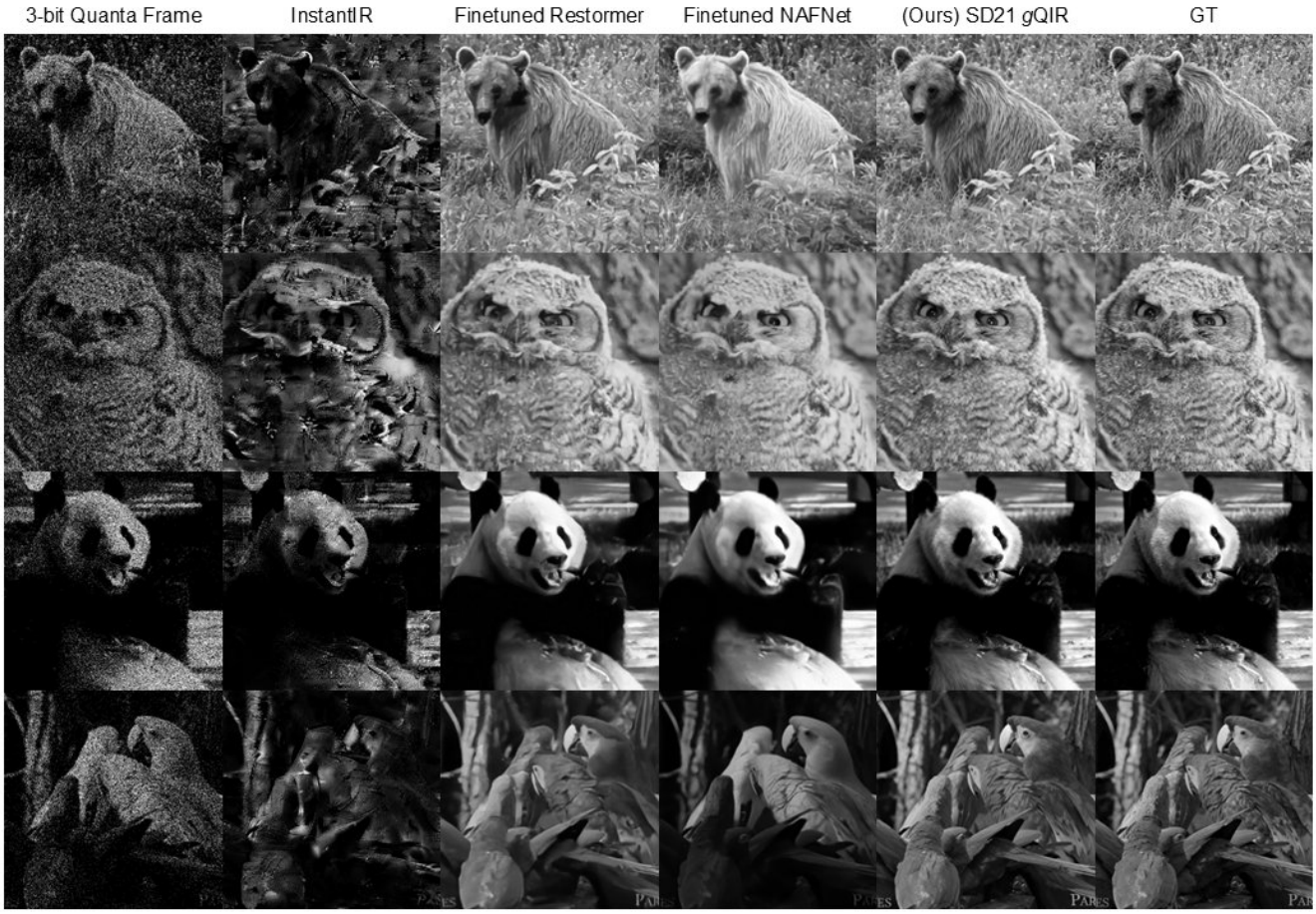

Figure 24. **Mono Single Frame Reconstruction Comparison (3/4).**

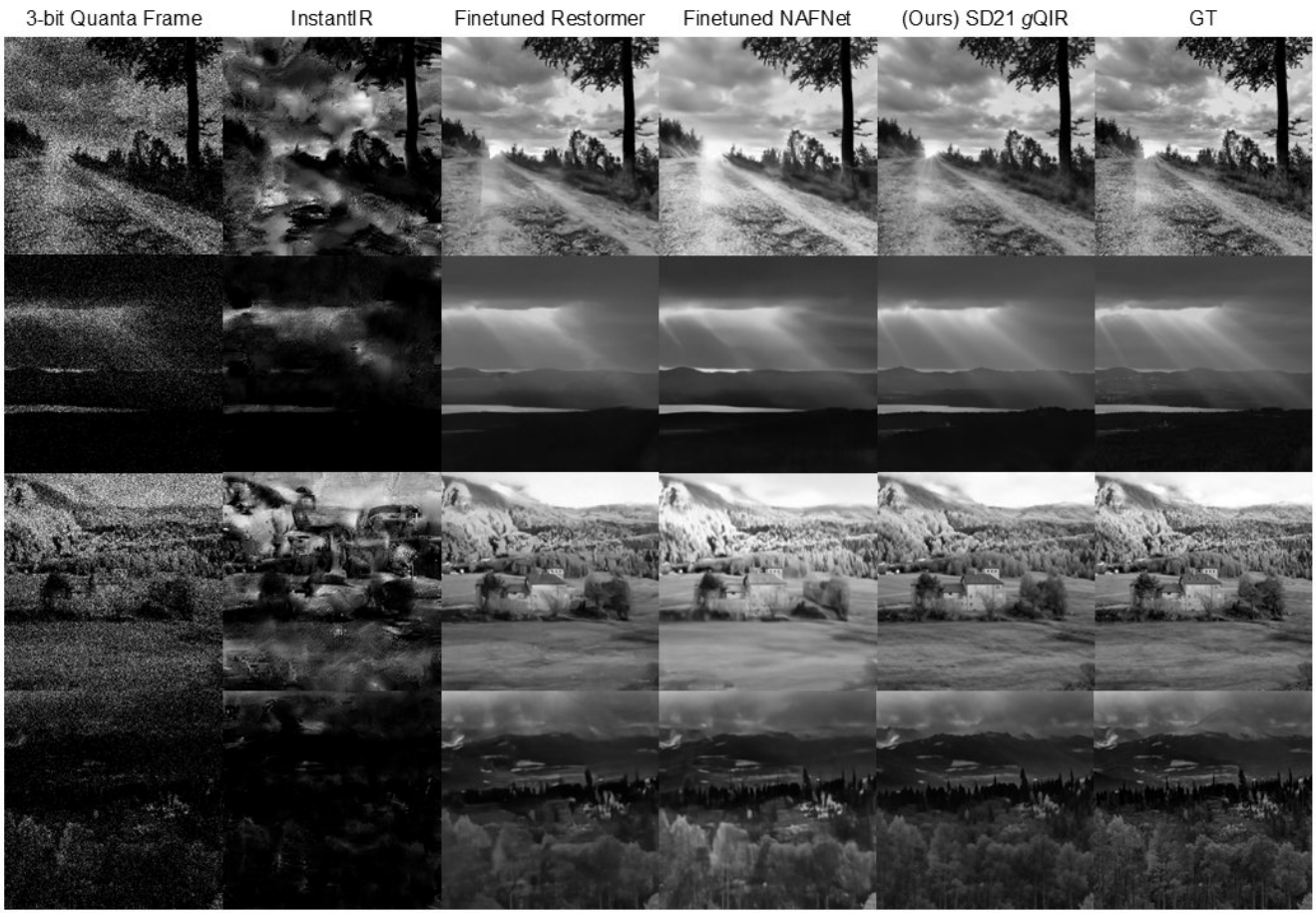

Figure 25. **Mono Single Frame Reconstruction Comparison (4/4).**

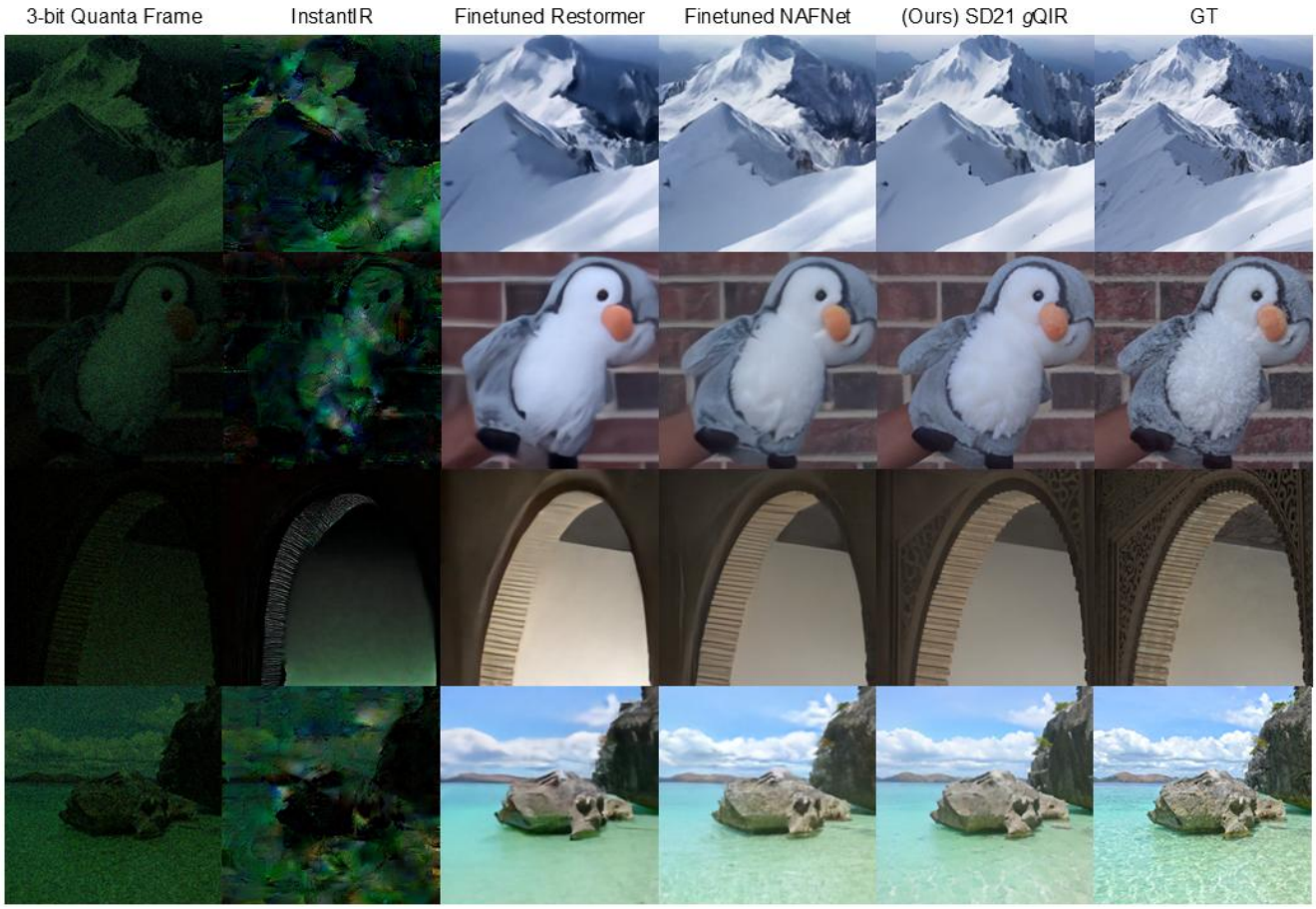

Figure 26. **Color Single Frame Reconstruction Comparison (1/4).**

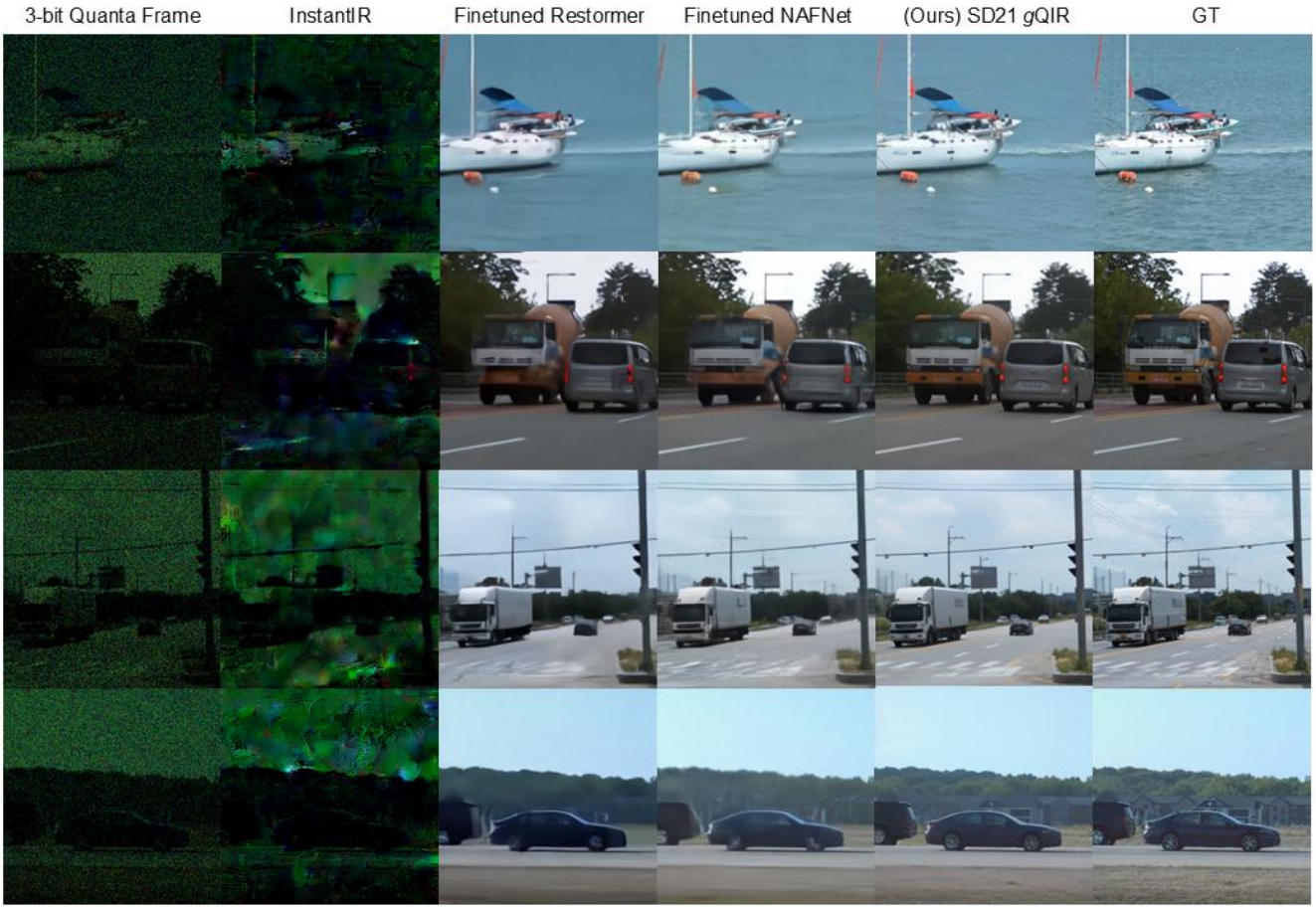

Figure 27. **Color Single Frame Reconstruction Comparison (2/4).**

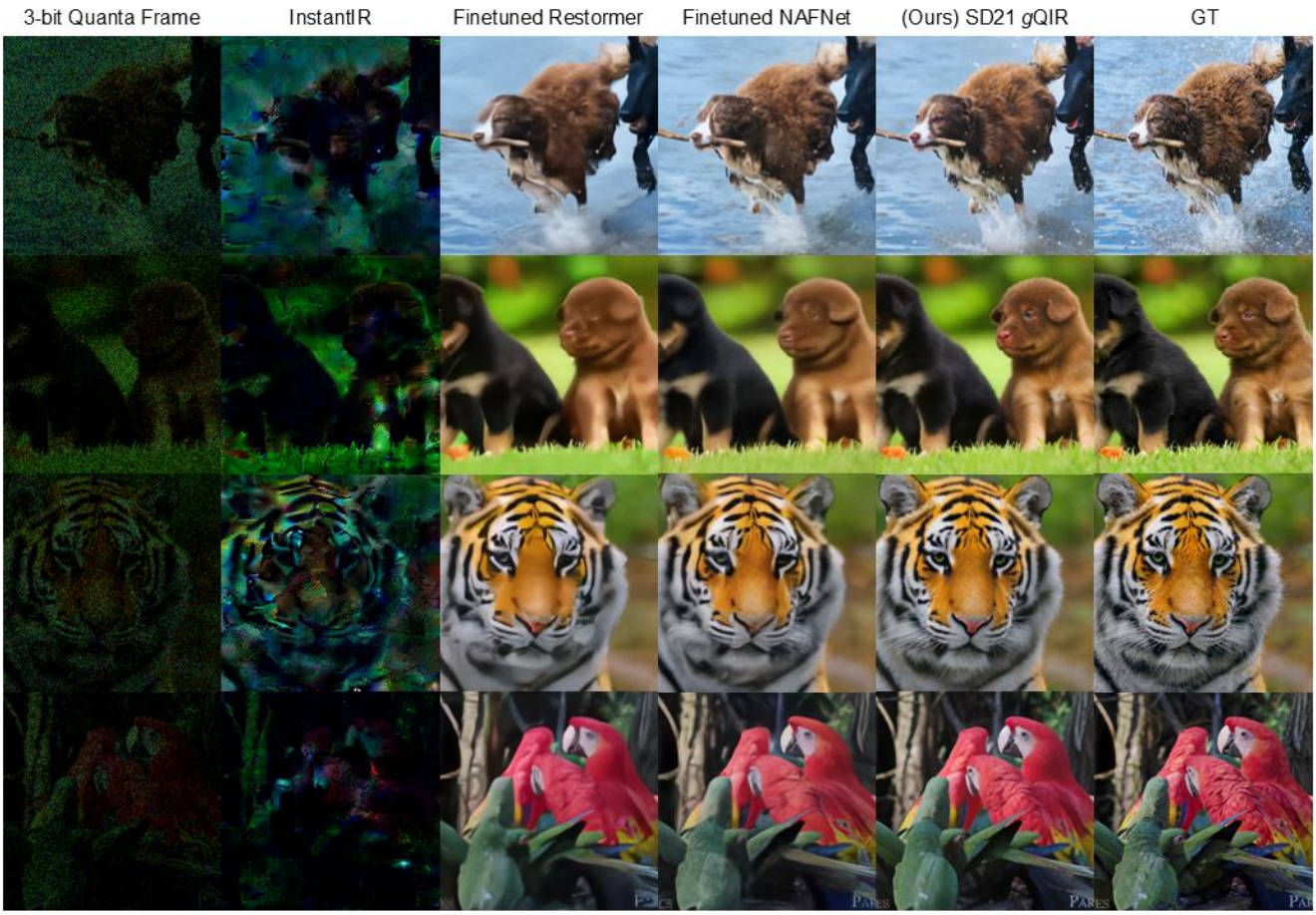

Figure 28. **Color Single Frame Reconstruction Comparison (3/4).**

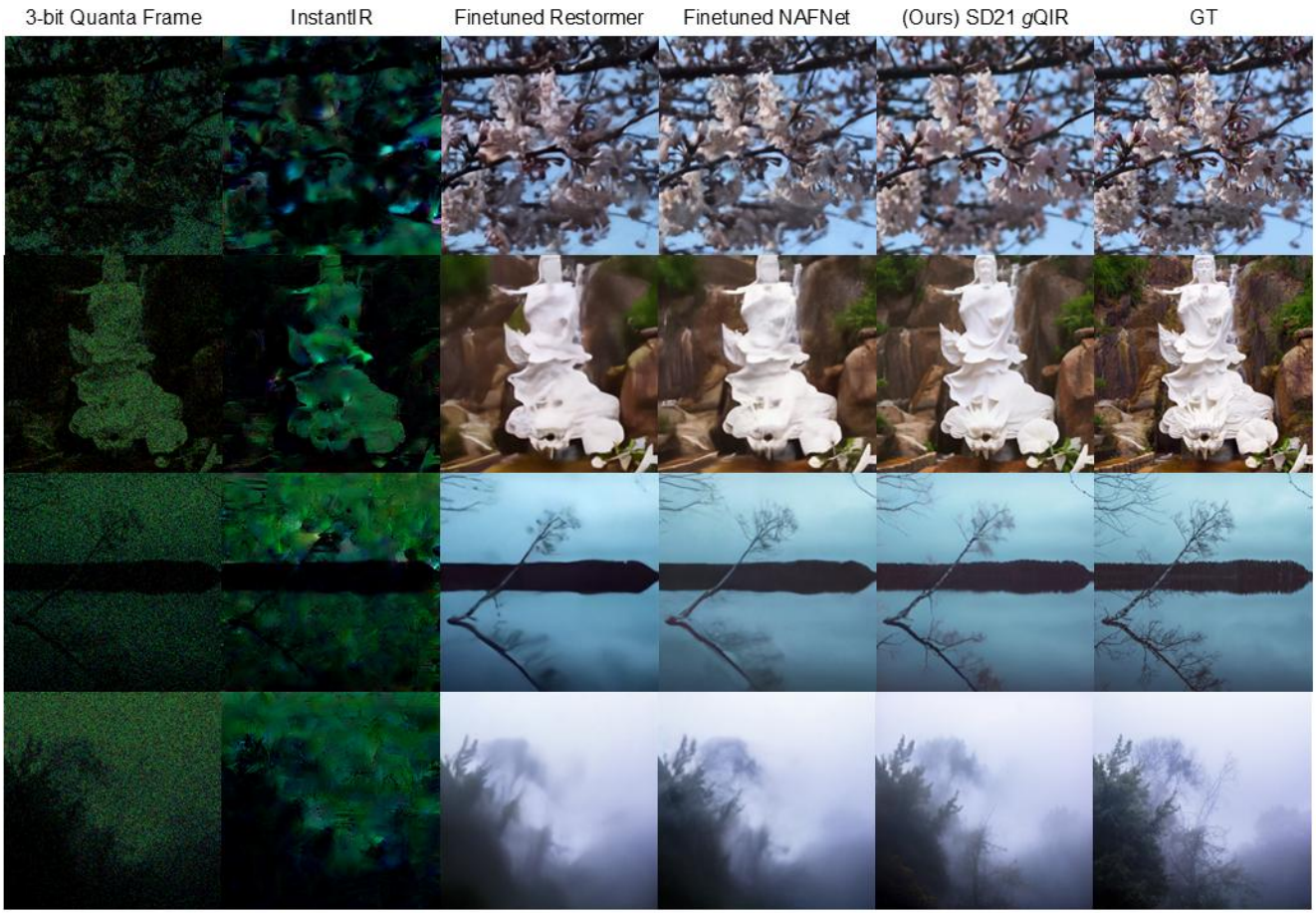

Figure 29. **Color Single Frame Reconstruction Comparison (4/4).**