Geometry-Aware Representation Denoising for Robust Multi-view 3D Reconstruction

Jin Hyeon Kim[1*] **Jaeeun Lee**[1*] **Claire Kim**[1] **Kyoungjin Oh**[1]
Paul Hyunbin Cho[1] **Jaewon Min**[1] **Yeji Choi**[1]
Jihye Park[2] **Hyunhee Park**[2] **Minkyu Park**[2] **Seungryong Kim**[1†]

[1]KAIST AI [2]Samsung Electronics

https://github.com/cvlab-kaist/GARD

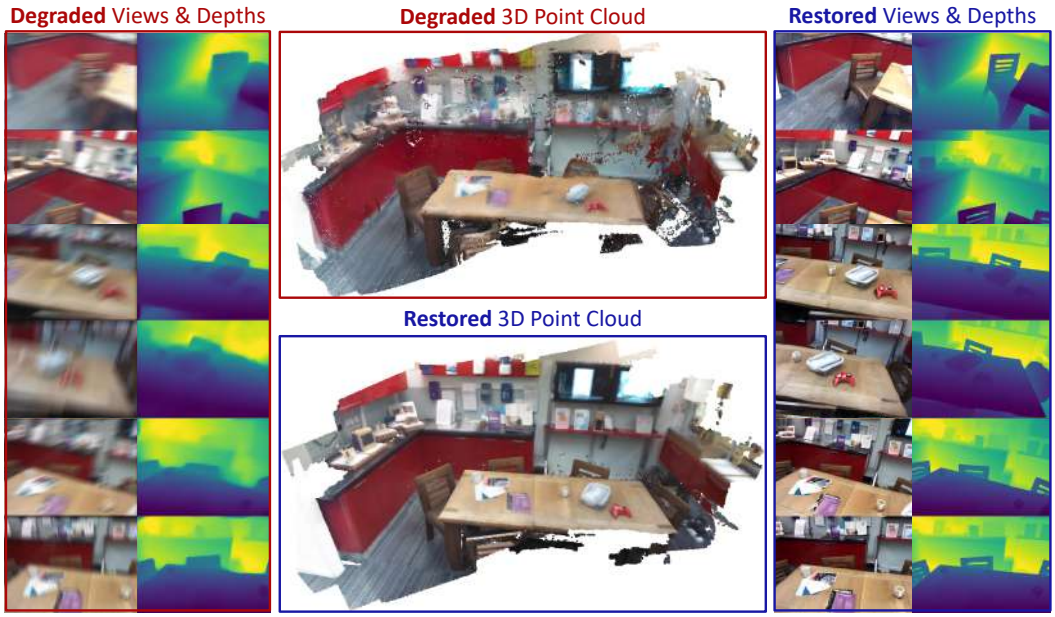


Figure 1: **Geometry-Aware Representation Denoising (GARD) framework.** Given degraded multi-view input images, our approach performs denoising on geometry-aware representations, thereby enabling simultaneous recovery of accurate 3D scene geometry and high-quality multi-view imagery.

Abstract

Multi-view 3D reconstruction has achieved remarkable progress with the advent of feed-forward 3D reconstruction models. However, these models are typically trained and evaluated under ideal, degradation-free imaging conditions, whereas real-world observations often contain degradations that differ significantly from such settings. Improving robustness for multi-view 3D reconstruction under degraded conditions therefore remains an important challenge. We present **Geometry-Aware Representation Denoising (GARD)**, a novel framework that performs diffusion-based multi-view restoration directly in the feature space of a feed-forward 3D reconstruction model. This design exploits the *geometry-aware feature representations* of the 3D reconstructor to effectively recover accurate scene geometry. Furthermore, by employing an additional RGB image decoder, the refined

*: Equal contribution
†: Corresponding authors

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

representations can also be used to restore high-quality RGB images, thereby enabling the simultaneous recovery of 3D scene geometry and high-quality imagery. Comprehensive experiments on the Depth Anything 3 (DA3) benchmark demonstrate the effectiveness of the proposed GARD framework.

1 Introduction

Reconstructing the 3D structure of a scene from 2D observations, commonly formulated as multi-view 3D reconstruction [15, 50, 12, 49, 7], is a foundational problem in computer vision. It unlocks a wide range of real-world applications, including autonomous navigation [3, 58, 73], robotics [22, 54, 32], and augmented and virtual reality [43, 65, 40]. Recent feed-forward reconstruction models [29, 60, 64, 62, 27] have significantly advanced this task by replacing traditional multi-stage 3D reconstruction pipelines [15, 50, 12] with end-to-end architectures that directly infer scene geometry from multi-view inputs. Built on transformer architectures [59, 9], their attention mechanism encode cross-view information to learn *geometry-aware* representations, enabling accurate and scalable reconstruction under ideal imaging conditions.

However, real-world multi-view observations often deviate from this ideal setting. In practice, captured images and video sequences frequently suffer from degradations such as motion blur induced by camera motion [37, 46, 34, 31, 14]. These effects obscure fine textures and structural cues essential for reliable feature extraction and cross-view matching. As a result, the learned representations become less discriminative, disrupting geometric consistency across views. Since feed-forward models [29, 60, 64, 62, 27] directly infer scene geometry from these features in a single forward pass, they lack mechanisms to explicitly correct such errors, allowing them to propagate through the network and accumulate in the final reconstruction. Improving robustness to such imperfect inputs therefore remains a key challenge for achieving reliable and consistent performance in multi-view 3D reconstruction.

A key design question is where restoration should be performed in the 3D reconstruction pipeline. A straightforward approach would be to adopt a *restore-then-reconstruct* paradigm (Fig. 2 (a)), where degraded inputs are first restored in image space using existing image restoration models [70, 6, 5, 28, 68, 35, 69] before being passed to the feed-forward reconstructor. However, current image restoration models are predominantly designed for single-view restoration and thus fail to leverage multi-view information and cannot enforce cross-view geometry consistency during restoration. While a recent multi-view restoration approach [35] partially addresses this issue, it operates in a heavily compressed VAE-based latent space [23], where information bottlenecks hinder the preservation of fine-grained details and geometric fidelity. Consequently, existing approaches remain suboptimal for multi-view image restoration and 3D reconstruction.

On the other hand, recent advances in Representation Autoencoders (RAEs) [72, 56, 24] further highlight the inherent shortcomings of conventional latent spaces. In particular, the compressed representations learned by VAEs [23, 66] introduce information bottlenecks that hinder the preservation of fine-grained details and structural fidelity, which are essential for accurate multi-view 3D reconstruction. RAEs address this limitation by adopting high dimensional, semantically rich latent representations that better retain both global structure and local details, while maintaining a dedicated decoder for reconstructing high fidelity images. Motivated by this insight, we move beyond conventional image space [70, 5, 6, 28, 68] and VAE-based formulations [35] and instead exploit the geometry-aware feature space inherently encoded by feed-forward reconstruction models as a more suitable domain for denoising, while enabling image recovery through an auxiliary decoder.

To this end, we propose **Geometry-Aware Representation Denoising (GARD)**, a novel framework that learns a diffusion-based multi-view restoration denoiser model operating directly within the *geometry-aware feature space* of a feed-forward reconstructor (Fig. 2 (b)). By conducting denoising in this feature space, the proposed approach exploits high-dimensional representations inherently structured for scene geometry estimation, as well as the cross-view consistency encoded by feed-forward reconstruction models. This design preserves geometric fidelity while mitigating the information bottlenecks associated with VAE-based latent spaces and the inconsistencies introduced by image-space restoration. Furthermore, we adopt a dedicated image decoder [17] to reconstruct high-quality RGB images from the refined representations, thereby enabling the joint recovery of high-quality imagery and accurate 3D scene geometry within a unified framework.

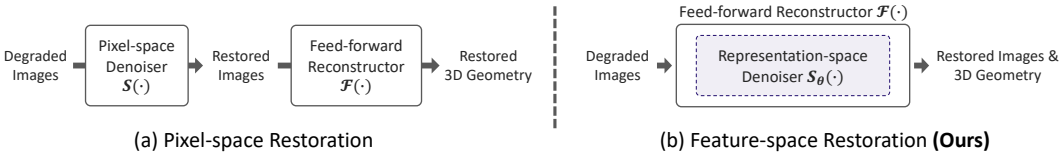


Figure 2: **Comparison of restoration denoising spaces.** (a) A restore-then-reconstruct pipeline first performs pixel-space restoration prior to 3D reconstruction. However, performing restoration in a single-view setting [70, 6, 5] or within a heavily compressed VAE-based latent space [35, 23] fails to preserve cross-view consistency and fine-grained geometric details, which often results in suboptimal geometric reconstruction. (b) In contrast, our **Geometry-Aware Representation Denoising (GARD)** operates on geometry-aware latent representations within a feed-forward reconstruction model, enabling the joint recovery of restored images and consistent 3D geometry across views.

We validate our approach through extensive experiments on the Depth Anything 3 benchmark [29], where controlled degradations are introduced to establish a rigorous evaluation protocol for restoration and reconstruction under motion blur degradation. To enable a fair comparison, we train the multi-view diffusion restoration model in both the geometry-aware feature space of the feed-forward reconstructor [29] and a conventional VAE-based latent space [23], thereby isolating the impact of representation choice. We further compare our approach against restore-then-reconstruct pipelines for image restoration and scene geometry recovery, as well as dedicated image restoration methods evaluated using standard image quality metrics. Experimental results demonstrate that operating in the geometry-aware feature space yields improved geometric fidelity and visual quality, resulting in strong performance across pose estimation, 3D reconstruction, and image restoration benchmarks.

2 Related Work

Robust multi-view 3D reconstruction. The advent of feed-forward 3D reconstruction models [62, 27, 60, 64, 29] has substantially advanced the field of multi-view 3D reconstruction. These models enable direct inference of scene geometry from multi-view images in a single forward pass, effectively replacing conventional multi-stage optimization pipelines [15, 50, 51, 12, 13, 8, 2, 55, 53]. Nevertheless, their performance degrades significantly in real-world settings [37, 46, 34, 31, 14], where observations often contain noise such as distractors and degradations, since these models are primarily trained under ideal conditions with clean inputs. Prior works [14, 42] have addressed robust multi-view 3D reconstruction in the presence of distractors. For example, RobustVGGT [14] introduces an outlier rejection mechanism to eliminate irrelevant distractor views, while VGTW [42] learns a dedicated distractor prediction head to identify and suppress distractor objects during reconstruction. In contrast, our approach is orthogonal to these methods, as we focus on degradations rather than distractors. In particular, camera motion blur is one of the most common and practically important degradations, arising frequently from handheld capture and dynamic imaging conditions. Such blur severely distorts fine textures, edges, and structural details, leading to unreliable geometric correspondence estimation and significantly hindering accurate 3D reconstruction.

Multi-view image restoration. Image restoration [18, 4, 70, 68, 74, 38] aims to recover clean images from degraded observations, including deblurring [28, 70, 5], denoising [70, 71, 69], and super-resolution [30, 10, 36, 21]. Early CNN-based methods [74, 71] were later surpassed by transformer-based approaches [70, 5] such as Restormer [70] and Hi-Diff [5], which better capture long-range dependencies. InstructIR [6] further enables unified restoration through language conditioning. However, these methods operate on single images and cannot leverage multi-view complementary information for effective restoration. Although video restoration models such as VRT [28] and FMA-Net [68] process multi-frame inputs, they are predominantly trained on temporally adjacent video sequences and rely heavily on temporal coherence, limiting their applicability to multi-view scenarios with substantial viewpoint variation. Furthermore, while SIR-Diff [35] introduces a sparse multi-view diffusion-based restoration framework, it operates in compressed VAE-based latent spaces [23] which may discard fine-grained visual structures. These limitations motivate restoration within geometry-aware representation spaces that preserve both cross-view consistency and detailed scene structure.

Representation space learning. Large-scale pretraining has made semantically rich visual representations a core component of modern vision systems [41, 16, 45, 57, 26]. Pretrained encoders produce high-dimensional feature spaces that generalize effectively across diverse downstream tasks [72, 17, 56, 24]. In generative modeling, latent diffusion models (LDMs) [48, 11, 44, 20, 21, 36] improve efficiency by operating in compact VAE-based latent spaces [23]. However, since these latents are optimized mainly for reconstruction, they often lack rich semantic structure and geometric consistency, creating information bottlenecks for downstream reasoning. Representation Autoencoders (RAEs) [72, 56, 24] address this by replacing the VAE encoder with a frozen pretrained representation network, yielding richer latent spaces for diffusion. Similarly, feed-forward 3D reconstruction models [29, 60, 27, 62, 64] learn strong geometric multi-view feature representations through transformer-based attention mechanisms, enabling representation spaces optimized for cross-view reasoning and scene geometry inference. Motivated by these advances, our approach performs diffusion-based feature restoration directly in such geometry-aware representation spaces, leveraging the already optimized geometric representations to better preserve and restore geometric feature representations during denoising while avoiding the limitations of VAE-based formulations.

3 Method

We propose a novel framework that performs denoising directly in the geometry-aware feature space of a frozen feed-forward 3D reconstruction model [29]. Our framework learns a multi-view latent diffusion model [61, 72] to restore the degraded intermediate feature representations produced by the feed-forward reconstructor. This design enables simultaneous recovery of clean RGB images and accurate 3D scene geometry in a single forward pass with dedicated decoders [29, 17], without retraining the underlying backbone.

3.1 Task Formulation and Motivation

Given a set of V degraded multi-view images $\mathbf{I}_{\text{deg}} \in \mathbb{R}^{V\times H\times W\times 3}$, where H and W denote the image height and width respectively, our objective is to recover both the restored images $\mathbf{I}_{\text{res}} \in \mathbb{R}^{V\times H\times W\times 3}$ and the underlying 3D scene geometry $\mathbf{G} = \{\mathbf{G}_{\text{depth}}, \mathbf{G}_{\text{pose}}\}$. Here, $\mathbf{G}_{\text{depth}} \in \mathbb{R}^{V\times H\times W\times 1}$ denotes the per-view depth maps, and $\mathbf{G}_{\text{pose}} \in \mathbb{R}^{V\times 9}$ represents the corresponding camera pose parameters, consisting of translation and rotation quaternions.

Limitations of pixel-space denoising. As illustrated in Fig. 2 (a), a straightforward solution is to adopt a restore-then-reconstruct pipeline, in which degraded inputs are first processed by a restoration denoiser $\mathcal{S}(\cdot)$ to obtain restored images, $\mathbf{I}_{\text{res}} = \mathcal{S}(\mathbf{I}_{\text{deg}})$, which are subsequently fed into the feed-forward reconstructor $\mathcal{F}(\cdot)$ for geometry estimation, i.e., $\mathbf{G} = \mathcal{F}(\mathbf{I}_{\text{res}})$. In practice, $\mathcal{S}(\cdot)$ can be instantiated using existing image restoration models [70, 5, 6, 69, 28, 68]. Despite its conceptual simplicity, this pipeline inherits two fundamental limitations. First, the majority of restoration methods operate on single-view inputs, consequently, applying $\mathcal{S}(\cdot)$ independently to each view fails to leverage multi-view complementary information for effective restoration and cannot enforce cross-view geometric consistency due to its fundamental architecture design. This often leads to view-dependent artifacts and inconsistencies that propagate to the feed-forward 3D reconstructor, resulting in poor geometry estimation performance.

Second, existing multi-view restoration methods remain constrained by both their modeling assumptions and underlying representation spaces. While video restoration models [28, 68] can exploit complementary information across multiple views, these models are predominantly trained on temporally adjacent video frames and rely heavily on short-range temporal coherence, limiting their generalization to sparse multi-view scenarios. Furthermore, although a recent multi-view restoration approach [35] explicitly models sparse cross-view interactions, it performs restoration in a heavily compressed VAE latent space [23], thereby introducing an information bottleneck that removes fine-grained spatial structures and high-frequency details. Such information is critical for establishing accurate cross-view correspondences, and its absence can substantially degrade downstream 3D reconstruction performance. Collectively, these limitations hinder the effective exploitation of complementary multi-view information and restrict the preservation of geometric consistency and fine-grained scene structure during restoration.

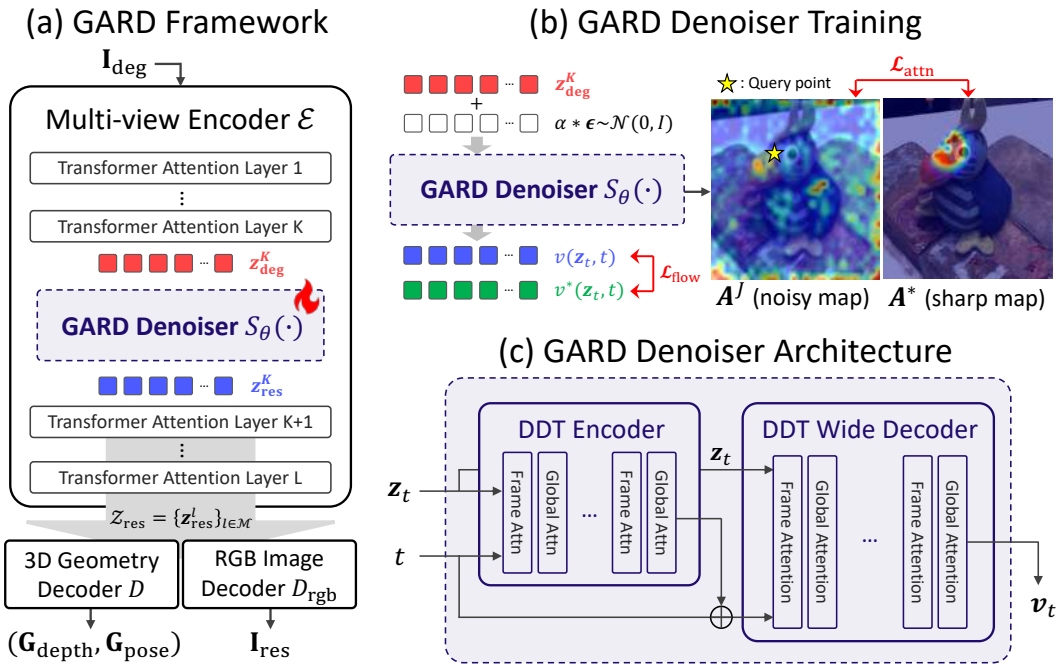


Figure 3: **Overview of the GARD framework.** (a) The GARD denoiser $\mathcal{S}_\theta(\cdot)$ is learned within the representation space of a frozen multi-view encoder [29] to restore degraded intermediate representations $\mathbf{z}^K_{\text{deg}}$ into restored representations $\mathbf{z}^K_{\text{res}}$ before they are propagated through the remaining encoder layers. The restored representations $\mathcal{Z}_{\text{res}}$ are then decoded by their respective decoders to produce geometry predictions and restored RGB images. (b) The GARD denoiser is optimized using an interpolated flow matching loss together with an attention alignment loss, which jointly learns the mapping from degraded to clean feature representations while preserving geometric consistency through explicit alignment of attention maps. (c) The GARD denoiser adopts a multi-view latent diffusion architecture [72], comprising a DDT encoder [61] and a DDT wide decoder, with global attention layers inserted to enable multi-view modeling, thereby facilitating global context aggregation and reconstruction of high-dimensional multi-view representations.

Representation space perspective. Recent advances in Representation Autoencoders (RAEs) [72, 56, 24] further highlight the limitations of conventional compressed latent spaces such as VAEs, demonstrating that richer, high-dimensional representations are crucial for preserving structural and semantic information. This observation suggests that restoration performance is fundamentally tied to the expressiveness of the underlying representation. In our case, the intermediate feature space $\mathbf{z}$ of the feed-forward 3D reconstruction model $\mathcal{F}(\cdot)$ is inherently geometry-aware due to cross-view interactions encoded by the multi-view transformer encoder $\mathcal{E}(\cdot)$, making it a more suitable domain for multi-view denoising. These observations further motivate restoration mechanisms that operate directly in geometry-aware representation spaces, enabling the joint preservation of visual fidelity and multi-view consistency.

3.2 GARD: Geometry-Aware Representation Denoising

Overview. To address the limitations of pixel-based and VAE-based latent-space restoration [35, 23], we propose **Geometry-Aware Representation Denoising (GARD)** to perform denoising directly in the geometry-aware feature space of a feed-forward reconstructor $\mathcal{F}(\cdot)$. As illustrated in Fig. 3 (a), the restoration denoiser $\mathcal{S}_\theta(\cdot)$, which we refer to as the GARD denoiser, operates directly within the feature representation space of a pretrained feed-forward reconstruction model $\mathcal{F}(\cdot)$.

Specifically, the reconstructor $\mathcal{F}(\cdot)$ comprises a multi-view encoder $\mathcal{E}(\cdot)$ and a geometry decoder $\mathcal{D}(\cdot)$. The encoder $\mathcal{E}(\cdot)$ maps the input images to a latent feature representation $\mathbf{z} = \mathcal{E}(\mathbf{I})$, where $\mathbf{z} \in \mathbb{R}^{V\times N\times C}$, with N denoting the number of tokens and C the feature dimensionality. The geometry decoder $\mathcal{D}(\cdot)$ subsequently predicts the scene geometry from the latent representation, i.e.,

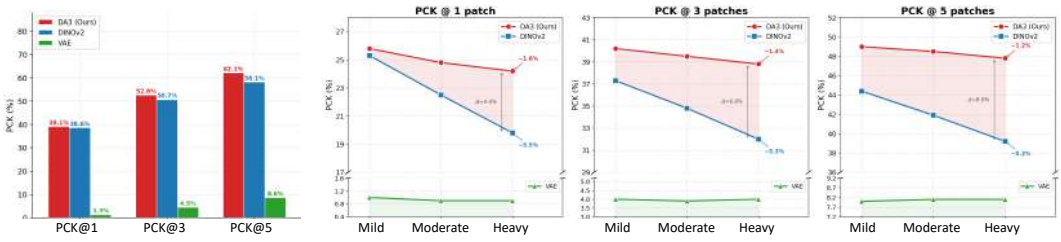


Figure 4: **Geometry-aware feature analysis conducted on ETH3D [52].** We evaluate the PCK accuracy of three feature cost volumes [23, 41, 29] under two experimental settings to validate the effectiveness of our proposed denoising space. (a) PCK performance on high-quality (HQ) input images. (b) PCK performance under progressively increasing levels of degradation (mild, moderate, and heavy), demonstrating robustness to input corruption. The correspondence visualizations for each cost volume are provided in Fig. 9 of the supplementary material.

$\mathbf{G} = \mathcal{D}(\mathbf{z})$. Specifically, given degraded inputs $\mathbf{I}_{\text{deg}}$, they are first encoded by an L-layer multi-view encoder $\mathcal{E}(\cdot)$ to produce layer-wise latent representations $\{\mathbf{z}_{\text{deg}}^l\}_{l=1}^L = \mathcal{E}(\mathbf{I}_{\text{deg}})$. The GARD denoiser $\mathcal{S}_\theta(\cdot)$, implemented as a multi-view latent diffusion model, composed of transformer attention layers is inserted at the K-th layer of $\mathcal{E}$ to refine the degraded feature representation at layer K, yielding $\mathbf{z}_{\text{res}}^K = \mathcal{S}_\theta(\mathbf{z}_{\text{deg}}^K)$. This refined representation is subsequently propagated through the remaining encoder layers, producing restored features $\{\mathbf{z}_{\text{res}}^l\}_{l=K}^L$. From these restored features, four feature levels $\mathcal{Z}_{\text{res}} = \{\mathbf{z}_{\text{res}}^l\}_{l\in\mathcal{M}}$, with $|\mathcal{M}| = 4$, are selected and provided as input to two task-specific decoders: a geometry decoder $\mathcal{D}(\cdot)$ for 3D geometry estimation and an RGB decoder $\mathcal{D}_{\text{rgb}}(\cdot)$ for image restoration. Note that the geometry decoder $\mathcal{D}(\cdot)$ is part of the feed-forward reconstructor, while the RGB image decoder $\mathcal{D}_{\text{rgb}}(\cdot)$ is adapted from [17] and fine-tuned for our framework. The final outputs are given by $\mathbf{G} = \mathcal{D}(\mathcal{Z}_{\text{res}})$ and $\mathbf{I}_{\text{res}} = \mathcal{D}_{\text{rgb}}(\mathcal{Z}_{\text{res}})$. By learning the GARD denoiser within the multi-view encoder and performing denoising directly in the geometry-aware feature space, the proposed framework enables simultaneous image restoration and 3D scene reconstruction in a single forward pass, without requiring separate restoration and reconstruction stages or retraining the underlying backbone.

Geometry-aware feature analysis. We first investigate the geometric encoding capability of feature representations produced by the multi-view encoder $\mathcal{E}(\cdot)$ to evaluate their effectiveness for representation-level denoising. Specifically, we adopt Depth Anything 3 (DA3) as our representative multi-view encoder and compare its feature representations with those of VAE [23] and DINOv2 [41]. To this end, we measure keypoint correspondence accuracy using PCK by constructing feature-based cost volumes from the three different representations. Through two experiments, we demonstrate that DA3 encodes more geometry-aware representations than the alternatives. (1) Under clean high-quality (HQ) multi-view inputs, DA3 achieves the highest PCK across all three evaluation thresholds. (2) Under progressively increasing degradation levels (mild, moderate, and heavy), DA3 features exhibit superior robustness to input corruption. As shown in Fig. 4, the feature representations of DA3 consistently produce higher PCK scores and maintain stronger robustness across degradation levels compared to VAE- and DINOv2-based features. These results indicate that the multi-view encoder more effectively preserves geometric structure, making it more suitable for representation-level denoising and downstream 3D reconstruction tasks. Refer to Fig. 9 in the supplementary material for qualitative visualizations of feature cost-volume correspondences for each representation.

Representation denoiser model. The GARD denoiser $\mathcal{S}_\theta(\cdot)$ is a multi-view diffusion architecture built on the DiT^{DH} design from RAE [72], which is well-suited for denoising in high-dimensional feature spaces. As illustrated in Fig. 3 (c), it adopts an encoder and a wide decoder structure augmented with interleaved global attention layers to enable multi-view representation learning. Specifically, the model interleaves frame-level attention and global cross-view attention to aggregate contextual information both within and across views. Frame-level attention captures local spatial structure within each view, while global attention enables the model to exploit cross-view correspondences and enforce geometric consistency across multi-view feature representations.

Table 1: **Quantitative pose estimation results.** We report AUC5↑ and AUC30↑ for camera pose estimation on the DA3 benchmark [29], comparing single-view and multi-view restoration approaches. The best result is highlighted in **bold** and the second best is underlined.

Model	HiRoom		ETH3D		DTU		7Scenes		ScanNet++	
	AUC5↑	AUC30↑	AUC5↑	AUC30↑	AUC5↑	AUC30↑	AUC5↑	AUC30↑	AUC5↑	AUC30↑
Input Conditions										
HQ Input	87.20	96.65	53.45	84.68	92.44	98.70	42.47	86.91	82.66	92.95
LQ Input	4.10	32.90	16.72	61.38	20.83	66.43	7.55	51.39	34.55	71.02
Single-View Restoration										
Restormer [70]	3.69	26.68	16.51	57.68	21.08	67.80	24.94	75.12	39.20	76.12
HI-Diff [5]	1.37	16.02	15.51	52.56	10.54	51.13	7.30	52.72	33.91	71.61
InstructIR [6]	3.89	27.68	14.71	53.80	54.80	85.91	18.42	74.22	32.84	68.23
MoCE-IR [69]	2.88	28.60	21.73	63.25	45.85	84.61	16.38	65.20	38.64	72.99
Multi-View Restoration										
VRT [28]	3.67	30.17	14.90	58.98	19.83	67.01	11.11	48.02	36.55	72.67
FMA-Net [68]	2.10	18.29	10.78	53.81	7.09	40.62	7.55	38.23	24.66	55.40
VAE_{MVD}	2.84	28.70	7.88	35.20	8.18	60.50	13.90	76.50	31.20	75.00
GARD (Ours)	**12.00**	**67.22**	**35.75**	**74.68**	**62.24**	**92.37**	**35.55**	**84.73**	**56.44**	**87.45**

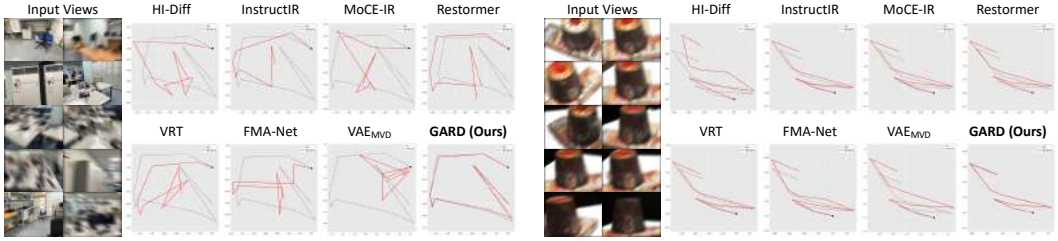


Figure 5: **Qualitative results for camera trajectory prediction.** We visualize the top-down camera trajectories for degraded multi-view inputs. Compared to the baselines, the proposed GARD produces more accurate camera pose trajectories. The black dot indicates the starting camera point.

Interpolated flow matching loss. The GARD denoiser $\mathcal{S}_\theta(\cdot)$ is trained to perform restoration on the K-th layer feature representation $\mathbf{z}^K_{\text{deg}}$ within the multi-view encoder $\mathcal{E}$. Unlike prior works that define flow trajectories from Gaussian noise to the clean representations [72, 10], we use the degraded latent itself as the source distribution, since $\mathbf{z}^K_{\text{deg}}$ already retains meaningful structural and geometric information beneficial for restoration. To improve robustness, we introduce a noise-perturbed source representation $\tilde{\mathbf{z}}^K_{\text{deg}} = \mathbf{z}^K_{\text{deg}} + \alpha\boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$ and $\alpha \in [0, 1]$ controls the perturbation magnitude. The resulting optimization objective follows:

$$\mathcal{L}_{\text{flow}} = \mathbb{E}_{t, \mathbf{z}^K_{\text{deg}}, \mathbf{z}^K_{\text{clean}}} \left[\| v(\mathbf{z}_t, t) - v^*(\mathbf{z}_t, t) \|_2^2 \right]. \tag{1}$$

where $\mathbf{z}_t = (1-t)\tilde{\mathbf{z}}^K_{\text{deg}} + t\mathbf{z}^K_{\text{clean}}$, $t \sim \mathcal{U}(0, 1)$ with the predicted velocity field as $v(\mathbf{z}_t, t) = \mathbf{v}_t = \mathcal{S}_\theta(\mathbf{z}_t, t)$ and ground-truth velocity field as $v^*(\mathbf{z}_t, t) = \mathbf{v}^*_t = \mathbf{z}^K_{\text{clean}} - \tilde{\mathbf{z}}^K_{\text{deg}}$. The restored representation is obtained by integrating the learned velocity field with an ODE solver, after which it is propagated through the remaining encoder layers and decoded for both geometry estimation and image restoration. An illustration is provided in Fig. 3 (b).

Attention alignment loss. While the flow-matching objective enforces feature-level alignment, it does not explicitly encourage learning of cross-view correspondences. We therefore regularize the attention maps within the GARD denoiser $\mathcal{S}_\theta(\cdot)$ to align with geometrically consistent correspondence maps following prior attention alignment works [25, 39, 19]. Specifically, we encourage the global attention weights for the J-th layer of the GARD denoiser to focus on geometrically corresponding regions rather than spurious artifacts. Let $\mathbf{A}^J \in \mathbb{R}^{V \times N \times N}$ denote the global attention map of the J-th layer within $\mathcal{S}_\theta(\cdot)$, and $\mathbf{A}^* \in \mathbb{R}^{V \times N \times N}$ denote the target correspondence maps obtained from the point cloud of clean multi-view inputs. The alignment loss is defined using cross-entropy as follows:

$$\mathcal{L}_{\text{attn}} = -\mathbb{E}\left[\mathbf{A}^* \log \mathbf{A}^J \right]. \tag{2}$$

Table 2: **Quantitative 3D reconstruction results.** We report Overall↓ and F-Score↑ for 3D reconstruction on the DA3 benchmark [29], comparing single-view and multi-view restoration approaches. The best result is highlighted in **bold** and the second best is underlined.

Model	HiRoom		ETH3D		DTU	7Scenes		ScanNet++	
	Overall↓	F-score↑	Overall↓	F-score↑	Overall↓	Overall↓	F-score↑	Overall↓	F-score↑
Input Conditions									
HQ Input	0.069	84.05	0.812	60.81	2.475	0.159	45.15	0.265	50.25
LQ Input	1.634	11.74	1.564	37.50	6.611	0.363	18.40	0.335	24.13
Single-View Restoration									
Restormer [70]	0.842	11.21	2.116	33.97	7.272	0.388	27.92	0.326	30.45
HI-Diff [5]	1.778	8.07	3.256	35.10	7.758	0.392	25.92	0.334	25.83
InstructIR [6]	0.992	12.41	2.263	33.71	5.563	0.311	29.80	0.366	26.06
MoCE-IR [69]	1.275	10.62	1.539	37.15	6.120	0.291	26.31	0.324	25.97
Multi-View Restoration									
VRT [28]	1.289	9.45	1.493	35.14	7.570	0.538	19.53	0.319	26.72
FMA-Net [68]	1.988	9.61	1.865	35.65	7.415	1.368	13.37	0.389	19.66
VAE_{MVD}	0.750	11.26	2.046	25.64	7.745	0.259	28.16	0.343	28.38
GARD (Ours)	**0.293**	**18.25**	**1.136**	**45.79**	**4.760**	**0.190**	**36.08**	**0.277**	**35.77**

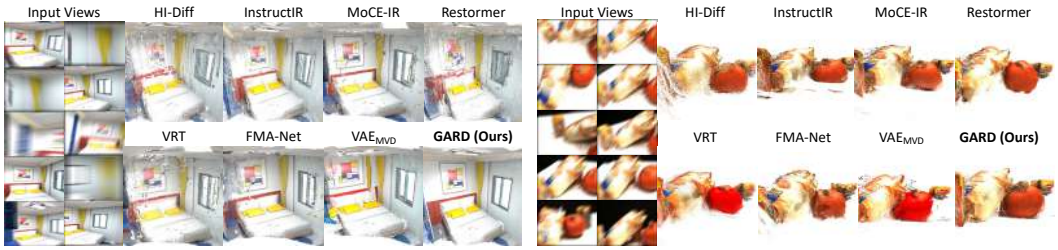


Figure 6: **Qualitative 3D reconstruction results.** We visualize the reconstructed 3D point clouds from degraded multi-view inputs. Compared with baseline approaches, the proposed GARD produces more accurate and geometrically consistent 3D reconstructions.

and is optimized jointly with the flow-matching objective, yielding the total loss $\mathcal{L} = \mathcal{L}_{\text{flow}} + \lambda_{\text{attn}}\mathcal{L}_{\text{attn}}$, where λ_{attn} denotes the attention alignment loss coefficient. This supervision promotes sharper and more coherent attention patterns, thereby improving both reconstruction accuracy and structural fidelity. Further details and visualizations of the attention target correspondence maps and the effects of attention alignment are provided in Fig. 11 and Fig. 12 of the supplementary material.

4 Experiments

4.1 Experimental Settings

Baselines. We compare against restore-then-reconstruct pipelines (Fig. 2), which apply image restoration before feed-forward 3D reconstruction. For single-view restoration, we use Restormer [70], HI-Diff [5], InstructIR [6], and MoCE-IR [69], which together cover conventional, diffusion-based, and all-in-one restoration models. For multi-view restoration, we construct VAE_{MVD} (Multi-View Denoiser). It shares the GARD denoiser architecture but operates in a VAE latent space [23] instead of the geometry-aware feature space. We also include the video restoration models VRT [28] and FMA-Net [68], since they can process multi-view inputs. Results for the multi-view restoration model SIR-Diff [35] are provided in the supplementary material.

Implementation details. We use Depth Anything 3 (DA3) [29] as our primary feed-forward reconstructor, initialized from the DA3-GIANT-1.1 checkpoint. Its multi-view encoder has $L = 40$ layers with hidden dimension $C = 1536$. The GARD denoiser is applied at layer $K = 18$, and the geometry and RGB decoders take features from layers $M = \{20, 28, 34, 40\}$. The denoiser adopts DiT^{DH} from RAE [72] with added global attention layers, and uses 8 encoder and 6 decoder layers. For training, we set the interpolated flow matching noise coefficient to $\alpha = 0.3$, the attention alignment layer to $J = 9$, and $\lambda_{\text{attn}} = 1.0$. Target correspondence maps $\mathbf{A}^*$ are computed from point clouds reconstructed from the clean multi-view images. All evaluations use $V = 10$ input views. Datasets, evaluation metrics, and further training details are provided in the supplementary material.

Table 3: **Quantitative restoration results.** We report PSNR↑ and LPIPS↓ for image restoration on the DA3 benchmark [29], comparing single-view and multi-view restoration approaches. The best result is highlighted in **bold** and the second best is underlined.

Model	HiRoom		ETH3D		DTU		7Scenes		ScanNet++	
	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓
Single-View Restoration										
Restormer [70]	17.49	0.544	20.97	0.672	17.73	0.588	21.30	0.428	21.50	0.415
HI-Diff [5]	17.35	0.551	20.45	0.586	17.39	0.591	19.82	0.428	20.68	0.413
InstructIR [6]	17.51	0.549	20.93	0.666	20.38	0.558	20.93	0.440	21.15	0.402
MoCE-IR [69]	17.69	0.560	21.00	0.611	20.38	0.562	20.60	0.491	21.19	0.431
Multi-View Restoration										
VRT [28]	17.47	0.562	20.82	0.606	17.61	0.606	19.79	0.510	20.83	0.437
FMA-Net [68]	17.14	0.551	20.65	**0.571**	17.13	0.572	18.84	0.534	19.98	0.455
VAE_{MVD}	19.76	0.493	21.37	0.638	20.54	0.434	21.74	0.404	21.19	0.379
GARD (Ours)	**21.89**	**0.362**	**21.88**	0.635	**21.25**	**0.418**	**22.67**	**0.249**	**22.19**	**0.345**

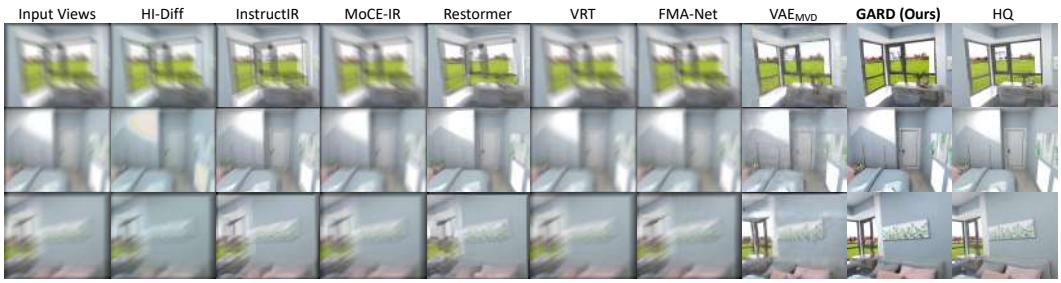


Figure 7: **Qualitative image restoration results.** We visualize restored RGB images from degraded multi-view inputs. Compared with baseline approaches, the proposed GARD effectively recovers high-fidelity multi-view images while preserving fine-grained details.

4.2 Experimental Results

Pose estimation. Tab. 1 and Fig. 5 report camera pose estimation results (AUC↑) on the DA3 benchmarks [29] under severe motion blur. Applying a feed-forward reconstructor directly to degraded inputs substantially degrades pose accuracy, as such models are predominantly trained and evaluated on clean data. Single- and multi-view restoration baselines also fall short of GARD. They are limited by single-view processing or restoration in heavily compressed VAE latent spaces. In contrast, GARD denoises directly in the reconstructor's geometry-aware feature space. This preserves structural fidelity and cross-view consistency and yields more accurate and stable poses. Additional visualizations are in Fig. 13 of the supplementary material.

3D reconstruction. Tab. 2 and Fig. 6 report 3D reconstruction results (Overall↓, F-score↑) under the same setting. Consistent with the pose results, GARD performs best on all datasets. Single-view restoration models may improve perceptual quality, but they cannot exploit complementary multi-view information, which leads to incomplete or misaligned geometry. VAE-based latent restoration suffers from information loss that hinders recovery of fine structures. By operating in a geometry-aware feature space, GARD preserves both global structural coherence and fine local detail. Since reconstruction quality also depends on detailed geometry such as depth, we provide depth evaluation in Tab. 6 and Fig. 16. Additional reconstruction visualizations are in Fig. 14 of the supplementary material.

Image restoration. Tab. 3 and Fig. 7 report image restoration quality (PSNR↑, LPIPS↓) under the same setting. Single-view models cannot exploit multi-view information, so their restoration capability under severe degradation is limited. Video restoration models such as VRT [28] and FMA-Net [68] use multiple frames, but they are designed for temporally adjacent sequences and struggle with long-range inputs. Among the baselines, only the trained VAE_{MVD} yields noticeable gains. GARD achieves the best overall PSNR and LPIPS, showing that restoration in a high-dimensional geometry-aware feature space better preserves visual fidelity and structural consistency. Additional comparisons are in Fig. 15 of the supplementary material.

Model	Interp. Flow	Alignment	AUC30↑		
			ETH3D	DTU	ScanNet++
A	×	×	67.30	87.21	84.12
B	×	✓	66.42	85.49	84.90
C	✓	×	73.85	89.99	85.90
D (Full)	✓	✓	**74.68**	**92.37**	**87.45**

(a) **Pose estimation accuracy.** We report the area under the curve (AUC30)↑ for pose evaluation.

Model	Interp. Flow	Alignment	F-score↑ / Overall↓		
			ETH3D	DTU	ScanNet++
A	×	×	39.91	5.43	31.52
B	×	✓	38.44	5.40	30.63
C	✓	×	44.65	4.92	32.40
D (Full)	✓	✓	**45.79**	**4.76**	**35.77**

(b) **3D reconstruction accuracy.** We report Overall↓ for DTU, while F-score↑ is reported for the remaining benchmarks.

Table 4: **Ablation on GARD training components.** We conduct an ablation study on the training components of GARD across three representative benchmarks spanning outdoor, object-centric, and indoor scenes: ETH3D [52], DTU [1], and ScanNet++ [67], respectively.

# Views	AUC30↑				
	HiRoom	ETH3D	DTU	7Scenes	ScanNet++
4 views	63.61	35.80	84.25	71.62	60.12
10 views	**67.22**	74.68	92.37	84.73	87.45
30 views	-	82.35	94.29	85.22	94.61
50 views	-	**82.62**	**98.00**	**86.85**	**95.30**

(a) **Pose estimation accuracy.** We report the area under the curve (AUC30)↑ for pose evaluation.

# Views	F-score↑ / Overall↓				
	HiRoom	ETH3D	DTU	7Scenes	ScanNet++
4 views	8.65	10.56	6.95	12.00	10.10
10 views	**18.25**	45.79	4.76	36.08	35.77
30 views	-	58.82	3.58	40.47	57.62
50 views	-	**65.46**	**2.03**	**45.92**	**65.72**

(b) **3D reconstruction accuracy.** We report Overall↓ for DTU, while F-score↑ is reported for the remaining benchmarks.

Table 5: **Ablation study on the number of input views.** Increasing the number of views improves both camera pose estimation accuracy and 3D reconstruction quality, indicating that richer cross-view information substantially benefits geometric reconstruction.

4.3 Ablation Experiments

GARD denoiser training components. Tab. 4 ablates the two training components of the GARD denoiser: the interpolated flow matching loss and the attention alignment loss. Models A–D cover the possible combinations, with Model D being the full configuration. Attention alignment does not consistently help under standard flow matching, but yields consistent gains when combined with interpolated flow matching. Standard flow matching starts denoising from pure Gaussian noise, which carries no structural prior, so correspondences are hard to learn even with attention supervision. Interpolated flow matching instead starts from a mixture of Gaussian noise and the degraded low-quality (LQ) input. The partial structure in this starting point lets the denoiser exploit the attention alignment signal to learn the LQ-to-high-quality (HQ) mapping, improving restoration. Fig. 12 of the supplementary material visualizes the effect of attention alignment.

Number of input views. Tab. 5 ablates the number of input views used for multi-view restoration. Adding views consistently improves both pose estimation and 3D reconstruction on the DA3 benchmarks, confirming that multiple views provide complementary geometric and visual information for restoration. Since each HiRoom scene in DA3 contains at most 20 views, we report results only up to 10 input views.

5 Conclusion

We presented **GARD**, a Geometry-Aware Restoration Denoising framework that restores degraded multi-view inputs in the feature space of feed-forward 3D reconstruction models. GARD refines degraded geometry-aware features, which are then decoded by the 3D geometry decoder and the RGB image decoder. This representation-level denoising jointly recovers accurate scene geometry and high-quality multi-view images. Extensive experiments show that GARD consistently outperforms existing single-view and multi-view restoration methods, highlighting the effectiveness of restoration in geometry-aware feature space.

Limitation and future directions. As a diffusion-based method, GARD requires iterative denoising, which limits its use in latency-sensitive settings. Future work will explore more efficient denoiser designs, as well as multi-layer denoising strategies to reduce error accumulation and further improve reconstruction quality.

Acknowledgments and Disclosure of Funding

This research was supported by Institute of Information communications Technology Planning Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2019-II190075, RS-2024-00509279, RS-2025-II212068, RS2023-00227592, RS-2025-02214479, RS-2024-00457882, RS-2025-25441838, RS2025-02217259, RS-2026-25519202, RS-2024-00436680) and the Culture, Sports, and Tourism RD Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism (RS-2024-00345025, RS-2024-00333068), the InnoCORE program of the Ministry of Science and ICT (N10250156) and National Research Foundation of Korea (RS-2024-00346597).

References

[1] Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl. Large-scale data for multiple-view stereopsis. *International Journal of Computer Vision*, 120(2):153–168, 2016.

[2] Honggyu An, Jinhyeon Kim, Seonghoon Park, Jaewoo Jung, Jisang Han, Sunghwan Hong, and Seungryong Kim. Cross-view completion models are zero-shot correspondence estimators. *arXiv preprint arXiv:2412.09072*, 2024.

[3] Andrea Ceccarelli and Francesco Secci. Rgb cameras failures and their effects in autonomous driving applications. *IEEE Transactions on Dependable and Secure Computing*, 20(4):2731–2745, 2022.

[4] Xiang Chen, Hao Li, Jiangxin Dong, Jinshan Pan, Xin Li, Xin He, Naiwei Chen, Shengyuan Li, Fengning Liu, Haoyi Lv, Haowei Peng, Yilian Zhong, Yuxiang Chen, Shibo Yin, Yushun Fang, Xilei Zhu, Yahui Wang, Chen Lu, Kaibin Chen, Xu Zhang, Xuhui Cao, Jiaqi Ma, Ziqi Wang, Shengkai Hu, Yuning Cui, Huan Zhang, Shi Chen, Bin Ren, Lefei Zhang, Guanglu Dong, Qiyao Zhao, Tianheng Zheng, Chunlei Li, Lichao Mou, Chao Ren, Wangzhi Xing, Xin Lu, Enxuan Gu, Jingxi Zhang, Diqi Chen, Qiaosi Yi, Bingcai Wei, Mingyu Liu, Pengyu Wang, Ce Liu, Miaoxin Guan, Boyu Chen, Hongyu Li, Jian Zhu, Xinrui Luo, Ziyang He, Jiayu Wang, Yichen Xiang, Huayi Qi, Haoyu Bian, Yiran Li, and Sunlichen Zhou. Lovif 2026 challenge on real-world all-in-one image restoration: Methods and results, 2026.

[5] Zheng Chen, Yulun Zhang, Liu Ding, Xia Bin, Jinjin Gu, Linghe Kong, and Xin Yuan. Hierarchical integration diffusion model for realistic image deblurring. In *NeurIPS*, 2023.

[6] Marcos V Conde, Gregor Geigle, and Radu Timofte. Instructir: High-quality image restoration following human instructions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

[7] Andrei Cramariuc, Lukas Bernreiter, Florian Tschopp, Marius Fehr, Victor Reijgwart, Juan Nieto, Roland Siegwart, and Cesar Cadena. maplab 2.0–a modular and multi-modal mapping framework. *IEEE Robotics and Automation Letters*, 8(2):520–527, 2022.

[8] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[10] Zheng-Peng Duan, Jiawei Zhang, Xin Jin, Ziheng Zhang, Zheng Xiong, Dongqing Zou, Jimmy Ren, Chun-Le Guo, and Chongyi Li. Dit4sr: Taming diffusion transformer for real-world image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

[11] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

[12] Yasutaka Furukawa, Brian Curless, Steven M Seitz, and Richard Szeliski. Towards internet-scale multi-view stereo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 1434–1441. IEEE, 2010.

[13] Yasutaka Furukawa and Carlos Hernández. *Multi-View Stereo: A Tutorial*. 01 2015.

[14] Jisang Han, Sunghwan Hong, Jaewoo Jung, Wooseok Jang, Honggyu An, Qianqian Wang, Seungryong Kim, and Chen Feng. Emergent outlier view rejection in visual geometry grounded transformers. *arXiv preprint arXiv:2512.04012*, 2025.

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

[16] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[17] Wooseok Jang, Seonghu Jeon, Jisang Han, Jinhyeok Choi, Minkyung Kwon, Seungryong Kim, Saining Xie, and Sainan Liu. Repurposing geometric foundation models for multi-view diffusion. *arXiv preprint arXiv:2603.22275*, 2026.

[18] Junjun Jiang, Zengyuan Zuo, Gang Wu, Kui Jiang, and Xianming Liu. A survey on all-in-one image restoration: Taxonomy, evaluation and future trends. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(12):11892–11911, December 2025.

[19] Siyoon Jin, Seongchan Kim, Dahyun Chung, Jaeho Lee, Hyunwook Choi, Jisu Nam, Jiyoung Kim, and Seungryong Kim. Matrix: Mask track alignment for interaction-aware video generation, 2025.

[20] Maciej Kilian, Varun Jampani, and Luke Zettlemoyer. Computational tradeoffs in image synthesis: Diffusion, masked-token, and next-token prediction. *arXiv preprint arXiv:2405.13218*, 2024.

[21] Jin Hyeon Kim, Paul Hyunbin Cho, Claire Kim, Jaewon Min, Jaeeun Lee, Jihye Park, Yeji Choi, and Seungryong Kim. Unified diffusion transformer for high-fidelity text-aware image restoration, 2025.

[22] Michael D Kim and Jun Ueda. Real-time image de-blurring and image processing for a robotic vision system. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1899–1904. IEEE, 2015.

[23] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[24] Amandeep Kumar and Vishal M Patel. Learning on the manifold: Unlocking standard diffusion transformers with representation encoders. *arXiv preprint arXiv:2602.10099*, 2026.

[25] Minkyung Kwon, Jinhyeok Choi, Jiho Park, Seonghu Jeon, Jinhyuk Jang, Junyoung Seo, Min-Seop Kwak, Jin-Hwa Kim, and Seungryong Kim. Cameo: Correspondence-attention alignment for multi-view diffusion models. *arXiv preprint arXiv:2512.03045*, 2025.

[26] Gyuseong Lee, Wooseok Jang, Jin Hyeon Kim, Jaewoo Jung, and Seungryong Kim. Domain generalization using large pretrained models with mixture-of-adapters. *arXiv preprint arXiv:2310.11031*, 2023.

[27] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European conference on computer vision*, pages 71–91. Springer, 2024.

[28] Jingyun Liang, Jiezhang Cao, Yuchen Fan, Kai Zhang, Rakesh Ranjan, Yawei Li, Radu Timofte, and Luc Van Gool. Vrt: A video restoration transformer. *IEEE Transactions on Image Processing*, 33:2171–2182, 2024.

[29] Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

[30] Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Wanli Ouyang, Yu Qiao, and Chao Dong. Diffbir: Towards blind image restoration with generative diffusion prior, 2024.

[31] Zhanwen Liu, Yuhang Li, Yang Wang, Bolin Gao, Yisheng An, and Xiangmo Zhao. Boosting visual recognition in real-world degradations via unsupervised feature enhancement module with deep channel prior. *IEEE Transactions on Intelligent Vehicles*, 2024.

[32] Chen-Chou Lo and Patrick Vandewalle. Depth estimation from monocular images and sparse radar using deep ordinal regression network. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 3343–3347, 2021.

[33] Li Ma, Xiaoyu Li, Jing Liao, Qi Zhang, Xuan Wang, Jue Wang, and Pedro V Sander. Deblur-nerf: Neural radiance fields from blurry images. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 12851–12860. IEEE, 2022.

[34] Alexandra Malyugina, Yini Li, Joanne Lin, and Nantheera Anantrasirichai. Unsupervised methods for video quality improvement: a survey of restoration and enhancement techniques. *arXiv preprint arXiv:2507.08375*, 2025.

[35] Yucheng Mao, Boyang Wang, Nilesh Kulkarni, and Jeong Joon Park. Sir-diff: Sparse image sets restoration with multi-view diffusion model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21620–21630, 2025.

[36] Jaewon Min, Jin Hyeon Kim, Paul Hyunbin Cho, Jaeeun Lee, Jihye Park, Minkyu Park, Sangpil Kim, Hyunhee Park, and Seungryong Kim. Text-aware image restoration with diffusion models. *arXiv preprint arXiv:2506.09993*, 2025.

[37] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3883–3891, 2017.

[38] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring, 2018.

[39] Jisu Nam, Soowon Son, Dahyun Chung, Jiyoung Kim, Siyoon Jin, Junhwa Hur, and Seungryong Kim. Emergent temporal correspondences from video diffusion transformers, 2025.

[40] Bunyo Okumura, Masayuki Kanbara, and Naokazu Yokoya. Augmented reality based on estimation of defocusing and motion blurring from captured images. In *2006 IEEE/ACM International Symposium on Mixed and Augmented Reality*, pages 219–225. IEEE, 2006.

[41] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision.

[42] Tianbo Pan, Xingyi Yang, Shizun Wang, and Xinchao Wang. Visual geometry transformer in the wild: Distractor-free 3d reconstruction.

[43] Youngmin Park, Vincent Lepetit, and Woontack Woo. Handling motion-blur in 3d tracking and rendering for augmented reality. *IEEE transactions on visualization and computer graphics*, 18(9):1449–1459, 2011.

[44] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.

[45] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[46] Jaesung Rim, Haeyun Lee, Jucheol Won, and Sunghyun Cho. Real-world blur dataset for learning and benchmarking deblurring algorithms. In *European conference on computer vision*, pages 184–201. Springer, 2020.

[47] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10912–10922, 2021.

[48] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[49] Antoni Rosinol, Marcus Abate, Yun Chang, and Luca Carlone. Kimera: an open-source library for real-time metric-semantic localization and mapping. In *2020 IEEE international conference on robotics and automation (ICRA)*, pages 1689–1696. IEEE, 2020.

[50] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016.

[51] Johannes L. Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[52] Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3260–3269, 2017.

[53] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8922–8931, June 2021.

[54] Bin Tan, Changjiang Sun, Xiage Qin, Hanat Adai, Zelin Fu, Tianxiang Zhou, Han Zhang, Yinghao Xu, Xing Zhu, Yujun Shen, and Nan Xue. Masked depth modeling for spatial perception. *arXiv preprint arXiv:[2601.17895]*, 2026.

[55] Engin Tola, Vincent Lepetit, and Pascal Fua. A fast local descriptor for dense matching. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[56] Shengbang Tong, Boyang Zheng, Ziteng Wang, Bingda Tang, Nanye Ma, Ellis Brown, Jihan Yang, Rob Fergus, Yann LeCun, and Saining Xie. Scaling text-to-image diffusion transformers with representation autoencoders. *arXiv preprint arXiv:2601.16208*, 2026.

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

[58] Jorge Vargas, Suleiman Alsweiss, Onur Toker, Rahul Razdan, and Joshua Santos. An overview of autonomous vehicles sensors and their vulnerability to weather conditions. *Sensors*, 21(16):5397, 2021.

[59] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[60] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025.

[61] Shuai Wang, Zhi Tian, Weilin Huang, and Limin Wang. Ddt: Decoupled diffusion transformer. *arXiv preprint arXiv:2504.05741*, 2025.

[62] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20697–20709, 2024.

[63] Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4909–4916. IEEE, 2020.

[64] Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-Equivariant Visual Geometry Learning. *arXiv preprint arXiv:2507.13347*, 2025.

[65] Matthias M Wloka and Robert C Zeleznik. Interactive real-time motion blur. *The Visual Computer*, 12(6):283–295, 1996.

[66] Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15703–15712, 2025.

[67] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12–22, 2023.

[68] Geunhyuk Youk, Jihyong Oh, and Munchurl Kim. Fma-net: Flow-guided dynamic filtering and iterative feature refinement with multi-attention for joint video super-resolution and deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 44–55, June 2024.

[69] Eduard Zamfir, Zongwei Wu, Nancy Mehta, Yuedong Tan, Danda Pani Paudel, Yulun Zhang, and Radu Timofte. Complexity experts are task-discriminative learners for any image restoration, 2024.

[70] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5728–5739, 2022.

[71] Kai Zhang, Wangmeng Zuo, and Lei Zhang. Ffdnet: Toward a fast and flexible solution for CNN based image denoising. *IEEE Transactions on Image Processing*, 2018.

[72] Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. *arXiv preprint arXiv:2510.11690*, 2025.

[73] Chenming Zhu, Tai Wang, Wenwei Zhang, Jiangmiao Pang, and Xihui Liu. Llava-3d: A simple yet effective pathway to empowering lmms with 3d-awareness. *arXiv preprint arXiv:2409.18125*, 2024.

[74] Wangmeng Zuo, Kai Zhang, and Lei Zhang. *Convolutional Neural Networks for Image Denoising and Restoration*, pages 93–123. Springer International Publishing, Cham, 2018.

Appendix

A Extended Experimental Results

A.1 Feature Similarity Analysis

To further validate the effectiveness of the proposed GARD framework, we investigate the feature similarity across the transformer layers of the feed-forward reconstructor. Since the GARD denoiser is trained to map low-quality (LQ) degraded feature representations to their corresponding high-quality (HQ) clean representations, we measure the similarity between the restored representations and the clean representations to assess the extent to which the degraded feature representations are effectively recovered. Specifically, the feature similarity at transformer layer l is computed as follows:

$$\mathrm{Sim}^l = \frac{\mathbf{z}^l_{\text{res}} \cdot \mathbf{z}^l_{\text{clean}}}{\left\|\mathbf{z}^l_{\text{res}}\right\| \left\|\mathbf{z}^l_{\text{clean}}\right\|}, \tag{3}$$

where $\mathbf{z}^l_{\text{res}}$ and $\mathbf{z}^l_{\text{clean}}$ are the restored and clean multi-view representations at layer l, respectively. Fig. 8 visualizes this similarity for two independent scenes. In the baseline setting, without using the GARD denoiser, the similarity to the clean representation progressively decreases as the representations propagate through deeper layers, indicating that severe degradations cause a gradual loss of geometric information in the encoded features. In contrast, applying GARD at an early layer initially improves the similarity between restored and clean representations. These results provide a rationale for the improved reconstruction and pose estimation performance, demonstrating that GARD effectively maintains high-fidelity representations throughout the pipeline.

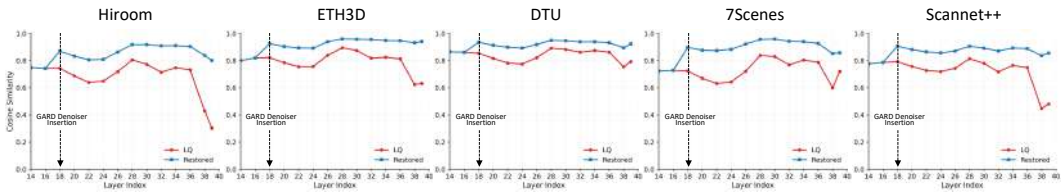


Figure 8: **Feature similarity analysis across layers.** We evaluate the cosine similarity between the restored feature representations and the corresponding clean HQ representations across the multi-view encoder layers of the feed-forward 3D reconstruction model [29]. The GARD denoiser is applied at layer $K = 18$. Across all DA3 benchmark [29] datasets, the feature similarity of the degraded LQ representations (red) progressively decreases in deeper layers due to the accumulation of degradation-induced feature distortion. In contrast, the restored representations produced by the GARD denoiser (blue) recover feature representations that remain substantially closer to the clean HQ representations throughout the deeper layers. This demonstrates that the GARD denoiser effectively restores geometry-aware features and suppresses degradation propagation, resulting in substantially improved consistency with the clean HQ representations. Please zoom in for clearer visualization.

A.2 Feature Cost Volume Visualization

Fig. 9 further presents qualitative correspondence attention visualizations of the three feature cost volumes [23, 41, 29] analyzed in Fig. 4 of the main paper. Specifically, for each red query point in the reference view, we visualize the attention region under (a) clean high-quality (HQ) multi-view inputs and (b) progressively increasing degradation levels of degraded multi-view inputs (mild, moderate, and heavy). Both quantitative and qualitative results reveal a clear performance hierarchy of DA3 $>$ DINOv2 $\gg$ VAE in terms of correspondence accuracy and robustness. Comparing the three feature cost volume representations through attention correspondence, DA3 features produce the sharpest and most geometrically consistent correspondences across views under both clean and degraded conditions, whereas DINOv2 achieves reasonably accurate matching with less precise and less stable response regions. In contrast, VAE features exhibit highly scattered and ambiguous correspondences under both clean and degraded conditions, reflecting limited geometry-aware representation quality caused by heavy latent compression. As the degradation severity increases, DA3 maintains strong localization and geometric consistency, whereas DINOv2 gradually deteriorates and VAE fails to preserve reliable correspondences.

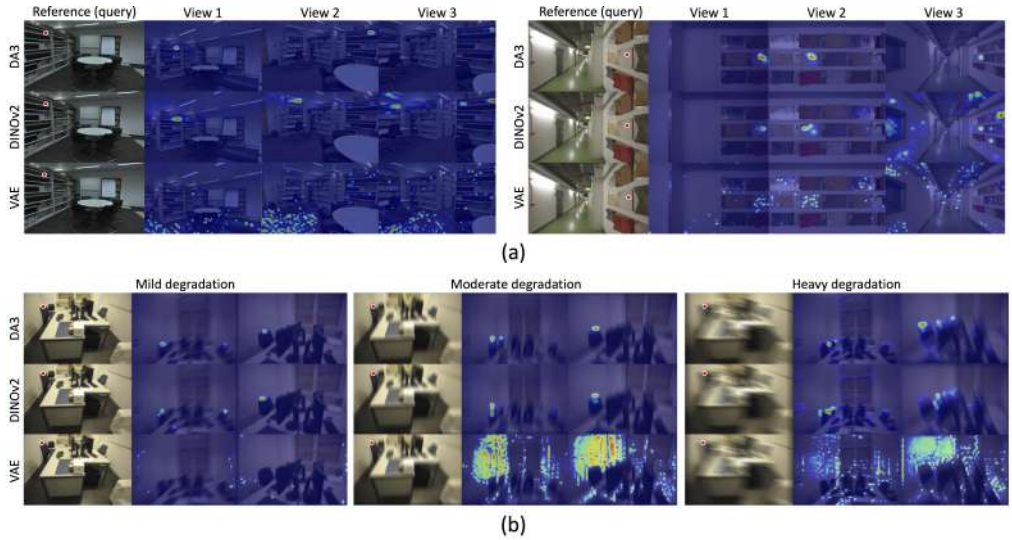


Figure 9: **Correspondence visualization of feature cost volumes.** Cross-view correspondence visualization of feature cost volumes constructed from VAE [23], DINOv2 [41], and DA3 [29] feature cost volumes. Please zoom in for clearer visualization.

A.3 Multi-View Depth Estimation

Tab. 6 further presents the multi-view depth estimation evaluation. We report AbsRel↓ and δ_1↑, which measure the absolute relative depth error and the percentage of predicted depths within a threshold ratio of 1.25 from the ground-truth depth, respectively. GARD demonstrates superior depth estimation performance on the DA3 [29] benchmark datasets, outperforming all baseline methods on four datasets while exhibiting only a marginal performance gap on DTU. These results underscore the effectiveness of restoration in a geometry-aware representation space for enhancing downstream multi-view depth estimation. By restoring degraded observations directly in this space, our method preserves cross-view geometric consistency and fine-grained structural details that are essential for accurate depth prediction. In contrast, single-view restoration methods are inherently unable to enforce geometric consistency across views, while VAE-based multi-view baselines are constrained by compression bottlenecks that limit the preservation of detailed geometric structures. As a result, baseline methods are more vulnerable to severe degradations, often producing blurred structures and visible artifacts, whereas our method yields sharper and more geometrically coherent depth maps across views, as further illustrated in the qualitative depth evaluation results in Fig. 16.

Table 6: **Quantitative depth estimation results.** We report AbsRel↓ and δ_1 ↑ across the DA3 benchmarks [29]. The best result is highlighted in **bold** and the second best is underlined.

Model	HiRoom		ETH3D		DTU		7Scenes		ScanNet++	
	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑
Input Conditions										
HQ Input	0.012	99.4	0.021	99.7	0.055	95.2	0.062	93.7	0.035	97.5
LQ Input	0.191	78.6	0.087	90.4	0.061	94.5	0.137	81.9	0.092	89.5
Single-View Restoration										
Restormer [70]	0.210	74.1	0.140	82.2	0.061	94.6	0.117	85.0	0.097	89.1
HI-Diff [5]	0.239	70.2	0.105	86.3	0.065	93.8	0.131	82.2	0.097	89.2
InstructIR [6]	0.227	71.7	0.142	81.3	0.056	95.0	0.124	83.3	0.128	83.4
MoCE-IR [69]	0.198	76.8	0.096	88.3	**0.055**	95.6	0.125	84.5	0.092	89.6
Multi-View Restoration										
VRT [28]	0.193	77.9	0.087	90.2	0.060	94.7	0.144	80.5	0.092	89.8
FMA-Net [68]	0.240	70.8	0.101	88.0	0.064	94.1	0.153	76.8	0.121	85.5
VAE_{MVD}	0.188	80.2	0.069	95.4	0.057	95.9	0.093	89.0	0.068	93.2
GARD (Ours)	**0.048**	**97.2**	**0.036**	**98.4**	0.057	**96.1**	**0.071**	**92.7**	**0.040**	**96.7**

A.4 Multi-View Restoration

We additionally compare our method with SIR-Diff [35], a multi-view diffusion-based restoration model that leverages Spatial-3D ResNet blocks and 3D self-attention Transformers for effective restoration of degraded multi-view inputs. As the official pretrained checkpoints and training code were not publicly available, we contacted the authors for clarification and implemented the method based on the available unofficial training code together with the details provided in the paper. While we carefully followed the described training protocol to ensure a fair comparison, some components of the unofficial implementation were incomplete, which made full reproduction of the reported results challenging. Nevertheless, we made our best effort to faithfully follow the method as described. The quantitative comparisons are presented in Tab. 7, and the qualitative results are provided in Fig. 10. While SIR-Diff can produce reasonable restorations under mild degradation settings, its performance degrades noticeably under stronger corruption levels, where it often introduces artifacts and structural inconsistencies. As a result, its effectiveness for downstream geometric tasks becomes limited in challenging scenarios.

Model	AUC30↑				
	HiRoom	ETH3D	DTU	7Scenes	ScanNet++
SIR-Diff [35]	9.73	20.94	16.59	2.79	30.02
GARD (Ours)	**67.22**	**74.68**	**92.37**	**84.73**	**87.45**

(a) **Pose estimation accuracy.** We report the area under the curve (AUC30)↑ for pose evaluation.

Model	F-score↑ / Overall↓				
	HiRoom	ETH3D	DTU	7Scenes	ScanNet++
SIR-Diff [35]	5.12	11.72	8.101	0.00	12.74
GARD (Ours)	**18.25**	**45.79**	**4.760**	**36.08**	**35.77**

(b) **3D reconstruction accuracy.** We report Overall↓ for DTU, while F-score↑ is reported for the remaining benchmarks.

Table 7: **Quantitative evaluation of SIR-Diff [35].** We report the pose estimation and 3D reconstruction results on the DA3 benchmark [29], comparing SIR-Diff [35] with our GARD framework.

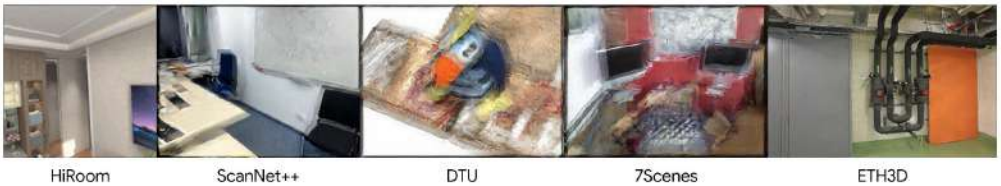


Figure 10: **Multi-view image restoration results of SIR-Diff [35].** We show qualitative results on the DA3 benchmark [29] based on our reproduction of SIR-Diff.

A.5 Multi-View Encoder Backbone

We further evaluate the effectiveness of the GARD denoiser across encoder backbones of varying capacity by training it on a different multi-view encoder. Specifically, we train the GARD denoiser using representations from DA3-BASE. In contrast to DA3-GIANT, which comprises 40 transformer layers and 1.10B parameters, DA3-BASE consists of 12 layers and 0.11B parameters. Following the same design as DA3-GIANT, we apply the GARD denoiser within the DA3-BASE feature space by inserting it at encoder layer $K = 4$, denoted as $\mathbf{GARD_{BASE}}$. We report the pose estimation, 3D reconstruction, and depth estimation results in Tab. 8, Tab. 9, and Tab. 10, respectively.

B Implementation Details

B.1 Evaluation Metrics

We follow the DA3 [29] evaluation protocol for pose estimation and 3D reconstruction. For pose estimation, we report the area under the curve (AUC) of Relative Rotation Accuracy (RRA) and Relative Translation Accuracy (RTA) at thresholds of 5 and 30, reflecting strict and relaxed evaluation criteria. For 3D reconstruction, we report accuracy and completeness, and compute the Chamfer Distance (CD) as their average. We additionally report the F1-score under a distance threshold to assess overall reconstruction quality. For image restoration, we use PSNR and LPIPS to evaluate pixel-level fidelity and perceptual quality, respectively. PSNR measures reconstruction accuracy against the ground-truth image, while LPIPS assesses perceptual similarity in a deep feature space. Together, they evaluate both low-level fidelity and perceptual quality of the restored images. Higher PSNR and lower LPIPS indicate better restoration performance.

Table 8: **Quantitative pose estimation results.** We report AUC5↑ and AUC30↑ for camera pose estimation on the DA3 benchmark [29] using the DA3-BASE [29] checkpoint as the feed-forward 3D reconstruction model. The best result is highlighted in **bold** and the second best is underlined.

Model	HiRoom		ETH3D		DTU		7Scenes		ScanNet++	
	AUC5↑	AUC30↑	AUC5↑	AUC30↑	AUC5↑	AUC30↑	AUC5↑	AUC30↑	AUC5↑	AUC30↑
Input Conditions										
HQ Input	35.21	84.11	23.79	62.61	53.53	91.21	34.47	83.02	37.68	69.81
LQ Input	0.56	12.14	1.01	11.44	5.97	58.78	3.04	28.26	0.40	10.49
Single-View Restoration										
Restormer [70]	0.45	12.78	2.18	17.25	4.44	56.56	4.88	48.20	5.20	37.84
HI-Diff [5]	0.19	05.68	1.13	19.33	4.84	50.57	1.96	25.55	3.02	26.56
InstructIR [6]	0.66	11.62	1.41	16.58	22.76	76.01	3.61	39.58	2.75	25.80
MoCE-IR [69]	0.91	17.15	2.02	22.36	16.92	67.94	4.12	35.75	4.24	27.72
Multi-View Restoration										
VRT [28]	0.62	12.79	1.65	16.70	6.38	57.11	1.33	24.21	2.40	23.30
FMA-Net [68]	0.04	06.92	0.80	15.35	2.52	36.87	0.25	12.96	1.00	19.15
VAE_{MVD}	7.88	**58.91**	0.24	7.05	11.11	60.42	11.49	51.48	1.37	20.60
$\text{GARD}_{\text{BASE}}$ (Ours)	**8.14**	58.60	**6.50**	**32.55**	**24.06**	**76.12**	**23.11**	**68.56**	**17.22**	**55.69**

Table 9: **Quantitative 3D reconstruction results.** We report Overall↓ and F-Score↑ for 3D reconstruction on the DA3 benchmark [29] using the DA3-BASE [29] checkpoint as the feed-forward 3D reconstruction model. The best result is highlighted in **bold** and the second best is underlined.

Model	HiRoom		ETH3D		DTU	7Scenes		ScanNet++	
	Overall↓	F-score↑	Overall↓	F-score↑	Overall↓	Overall↓	F-score↑	Overall↓	F-score↑
Input Conditions									
HQ Input	0.349	22.29	1.220	37.92	5.843	0.186	31.89	0.340	26.09
LQ Input	0.946	7.58	3.337	11.85	8.429	0.529	12.56	0.476	13.12
Single-View Restoration									
Restormer [70]	0.889	7.67	3.212	14.67	9.064	0.755	13.24	0.489	12.16
HI-Diff [5]	1.173	4.58	3.737	8.02	8.826	0.551	12.95	0.612	8.33
InstructIR [6]	1.183	6.23	3.653	7.29	7.960	0.501	14.19	0.562	9.03
MoCE-IR [69]	0.870	7.81	2.909	13.95	7.942	0.504	12.17	0.554	9.45
Multi-View Restoration									
VRT [28]	0.882	7.87	3.226	10.68	8.770	0.750	6.98	0.558	8.96
FMA-Net [68]	1.319	4.45	3.623	11.92	8.653	1.171	4.53	0.813	5.80
VAE_{MVD}	**0.389**	11.63	5.518	4.63	7.571	0.314	20.15	0.636	6.21
$\text{GARD}_{\text{BASE}}$ (Ours)	0.512	**12.69**	**2.047**	**20.38**	**6.304**	**0.232**	**23.41**	**0.347**	**21.60**

B.2 RGB Image Decoder

Our GARD framework jointly restores 3D scene geometry and high-quality RGB images by propagating the restored representations to their corresponding decoders. The underlying feed-forward 3D reconstruction model already incorporates a 3D geometry decoder $\mathcal{D}$ for scene geometry prediction. To reconstruct restored RGB images, we adopt the ViT-based RGB decoder $\mathcal{D}_{\text{rgb}}$ introduced in GLD [17]. As the original RGB decoder was designed exclusively for the DA3-BASE backbone, we further adapt and fine-tune it for the DA3-GIANT backbone, which serves as the primary feed-forward 3D reconstruction model in our framework. Concretely, we initialize the RGB decoder using the released DA3-BASE checkpoint and introduce an additional DA3 linear projection adapter to bridge the feature dimensional discrepancy between the DA3-BASE and DA3-GIANT representations. The restored multi-scale features extracted from the four selected feature levels $\mathcal{M} = \{20, 28, 34, 40\}$ are concatenated and subsequently fed into the RGB decoder to reconstruct the restored RGB images. We refer readers to the original GLD work [17] for additional architectural and implementation details.

B.3 Training Details

Training dataset. We train our model on two photo-realistic synthetic datasets: Hypersim [47] and TartanAir [63], which together cover diverse 3D environments. Hypersim provides high-fidelity indoor scenes with realistic lighting and materials, making it well-suited for learning appearance-aware geometric representations. In contrast, TartanAir includes both indoor and outdoor environments

Table 10: **Quantitative depth estimation results.** We report AbsRel↓ and δ_1 ↑ on the DA3 benchmark [29] using the DA3-BASE [29] checkpoint as the feed-forward 3D reconstruction model. The best result is highlighted in **bold** and the second best is underlined.

Model	HiRoom		ETH3D		DTU		7Scenes		ScanNet++	
	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑
Input Conditions										
HQ Input	0.039	98.28	0.035	98.89	0.051	97.13	0.068	93.50	0.044	97.08
LQ Input	0.195	77.75	0.124	84.75	0.057	95.66	0.144	79.19	0.123	85.07
Single-View Restoration										
Restormer [70]	0.202	76.56	0.147	80.91	0.057	95.71	0.119	84.47	0.112	86.36
HI-Diff [5]	0.237	71.45	0.122	84.88	0.063	94.36	0.143	80.08	0.131	84.02
InstructIR [6]	0.215	74.66	0.142	81.43	0.054	95.94	0.123	83.69	0.143	82.27
MoCE-IR [69]	0.186	79.12	0.124	85.07	**0.053**	**96.16**	0.132	81.52	0.122	85.51
Multi-View Restoration										
VRT [28]	0.195	77.72	0.122	85.24	0.056	95.93	0.147	78.14	0.125	84.70
FMA-Net [68]	0.251	69.54	0.137	82.12	0.063	95.13	0.178	73.26	0.159	79.07
VAE_{MVD}	0.099	92.80	0.248	63.74	0.054	95.85	0.106	87.43	0.160	77.57
$\text{GARD}_{\text{BASE}}$ (Ours)	**0.079**	**95.04**	**0.063**	**95.54**	0.060	95.23	**0.081**	**91.60**	**0.059**	**95.76**

with dynamic camera trajectories, emphasizing temporal consistency and large viewpoint changes for multi-view geometry learning. By combining them, we leverage complementary strengths in visual realism and motion diversity, enabling more robust and generalizable geometric representations. Degraded images are generated by applying motion blur kernels using the codebase[1]. During training, we resize the input frames so that the longest side is 504 pixels and randomly select up to 4 views per iteration to train the model in various multi-view settings.

Training hyperparameters. We train the GARD denoiser for 10 epochs with a global batch size of 8 on NVIDIA H100 GPUs, using the AdamW optimizer with a learning rate of 2×10^{-4}, $(\beta_1, \beta_2) = (0.9, 0.95)$, and no weight decay. All parameters are updated in full precision (fp32). We apply exponential moving average (EMA) with a decay rate of 0.9995 to stabilize training. The flow-matching objective uses a linear path with velocity prediction and logit-normal time sampling. Gradients are clipped to a maximum norm of 1.0.

Attention alignment. Fig. 11 visualizes the target correspondence map $\mathbf{A}^*$ under different temperature values T. To construct $\mathbf{A}^*$, we first estimate depth and camera pose from the clean high-quality (HQ) input views and unproject each pixel into a shared 3D world coordinate system. The resulting point cloud is spatially pooled to the ViT patch resolution (patch size 14) to align with the spatial token grid of the GARD denoiser $\mathcal{S}_\theta(\cdot)$. We then compute pairwise negative ℓ_2 distances across all VN patch tokens and apply a temperature-scaled softmax to obtain $\mathbf{A}^*$. Lower temperatures produce sharper and more spatially localized correspondence distributions, providing a stronger supervisory signal for the cross-entropy-based attention alignment loss $\mathcal{L}_{\text{attn}}$, whereas higher temperatures yield softer distributions across neighboring patches. For training the GARD denoiser, we used $T = 0.01$ to construct the target correspondence map.

Fig. 12 presents visualizations of the GARD denoiser attention maps before and after applying the attention alignment loss. Specifically, we first train the GARD denoiser using only the flow-matching objective and visualize the global attention maps from the wide decoder layers within the denoiser. While the 11th and 13th decoder layers exhibit relatively accurate attention toward the green query point, the 9th decoder layer fails to establish reliable geometric correspondences. Motivated by both this observation and the significance of wide decoder layers demonstrated in RAE [72], we explicitly regularize the global attention weights of the J-th decoder layer ($J = 9$) to encourage attention toward geometrically corresponding regions rather than spurious artifacts. As reported in Tab. 4, although attention alignment does not consistently improve performance under standard flow matching, combining it with interpolated flow matching yields additional gains. We attribute this improvement to the initialization from an LQ structural prior rather than pure noise, which enables more effective learning of geometrically consistent correspondences.

[1] https://github.com/LeviBorodenko/motionblur

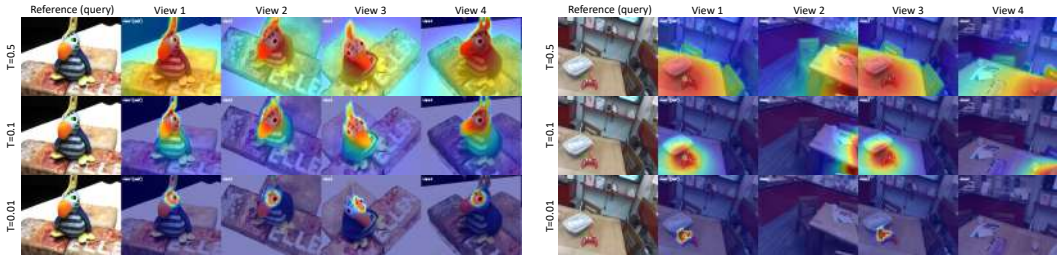
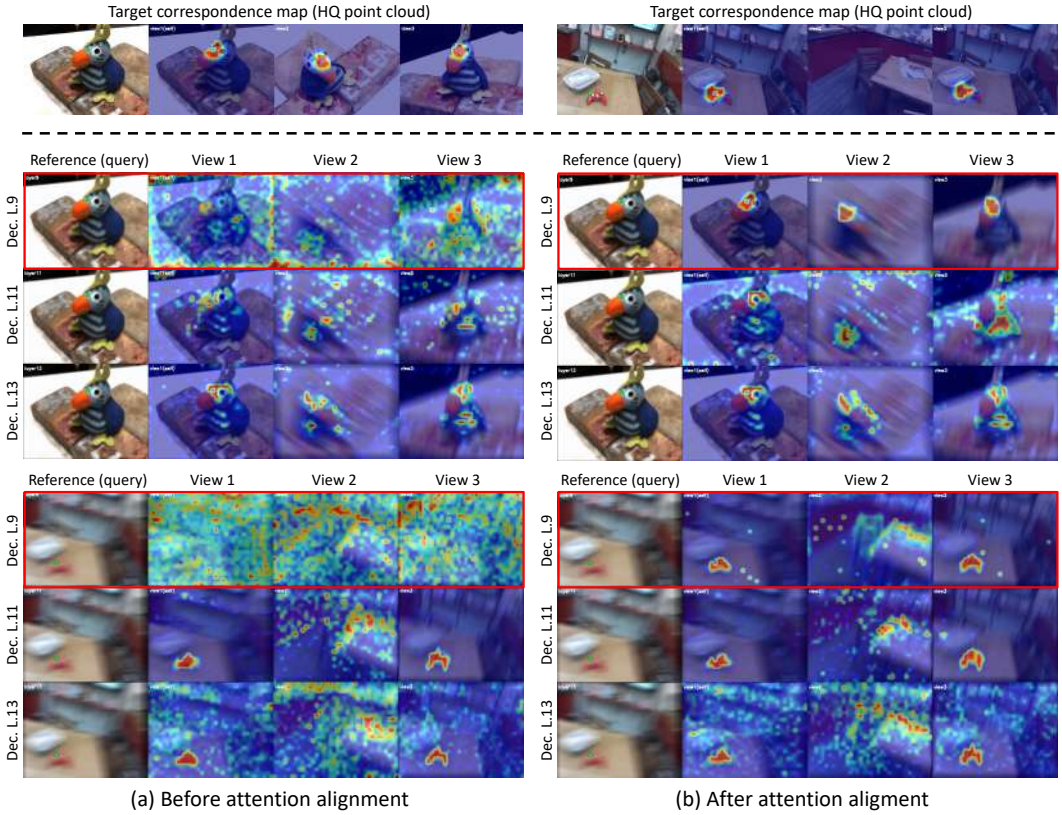


Figure 11: **Visualization of target correspondence maps** We visualize the effect of attention alignment training which augments the learning of the attention of the GARD denoiser.



(a) Before attention alignment (b) After attention aligment

Figure 12: **Visualization of attention alignment** We visualize the effect of attention alignment training which augments the learning of the attention of the GARD denoiser.

View selection strategy. To construct multi-view training samples, we adopt a view selection strategy based on the ground-truth camera poses provided in each dataset. Given a target number of views V, we retrieve neighboring frames using an expansion ratio that defines a larger candidate window to select the required number of views. For Hypersim, we use a *near-camera* strategy since adjacent views often exhibit relatively large viewpoint changes. For every anchor frame, we first precompute the ranking between frame pairs based on their geometric overlap, and then select additional candidates to form the multi-view input. We use an expansion ratio of 2, which encourages stronger geometric consistency across selected views. In contrast, TartanAir consists of long video sequences with strong temporal continuity, where consecutive frames are often highly redundant. To increase viewpoint diversity during training, we employ a *near-random* strategy with an expansion ratio of 4. Using this strategy, we first collect a larger set of nearby frames around the anchor frame and then randomly sample the remaining views from this candidate window. For training, we set the maximum number of training input views to $V = 4$.

B.4 Inference Details

During inference, degraded multi-view images are first processed using the DA3-GIANT-1.1 multi-view encoder [29]. We extract the intermediate geometry-aware representation from the $K = 18$-th encoder layer and use it as input to the proposed GARD denoiser. The denoiser restores the degraded representation by modeling the flow field in the latent feature space. To perform denoising, we solve the corresponding probability flow ordinary differential equation (ODE) using a first-order Euler solver with 50 discretized sampling steps. The restored representation is subsequently propagated through the remaining layers of the multi-view encoder to obtain refined multi-scale features. From these features, four feature levels $\mathcal{Z}_{\text{res}} = \mathbf{z}^{l}_{\text{res}l\in\mathcal{M}}$ are selected, where $\mathcal{M} = \{20, 28, 34, 40\}$ for DA3-GIANT-1.1 and $\mathcal{M} = \{6, 8, 10, 12\}$ for DA3-BASE. These features are provided to the geometry decoder and RGB decoder [17] for downstream 3D reconstruction and image restoration, respectively. Overall, the proposed GARD framework jointly reconstructs high-quality multi-view images and accurate 3D geometry in a single forward pass, eliminating the need for separate restoration and reconstruction stages.

C Additional Qualitative Results

We present additional qualitative evaluation results for pose estimation, 3D reconstruction, image restoration and depth estimation in Fig. 13, Fig. 14, Fig. 15, and Fig. 16 respectively.

Real-world evaluation. We further provide qualitative evaluation results on the real-world multi-view deblurring benchmark [33]. Since clean ground-truth images are unavailable for these real-world scenarios, quantitative evaluation is not feasible. Following the established evaluation protocol, we therefore qualitatively assess the quality of the restored depth maps in Fig. 17. Despite being trained solely on realistic synthetic degradations, GARD demonstrates reliable generalization to real-world degradation scenarios.

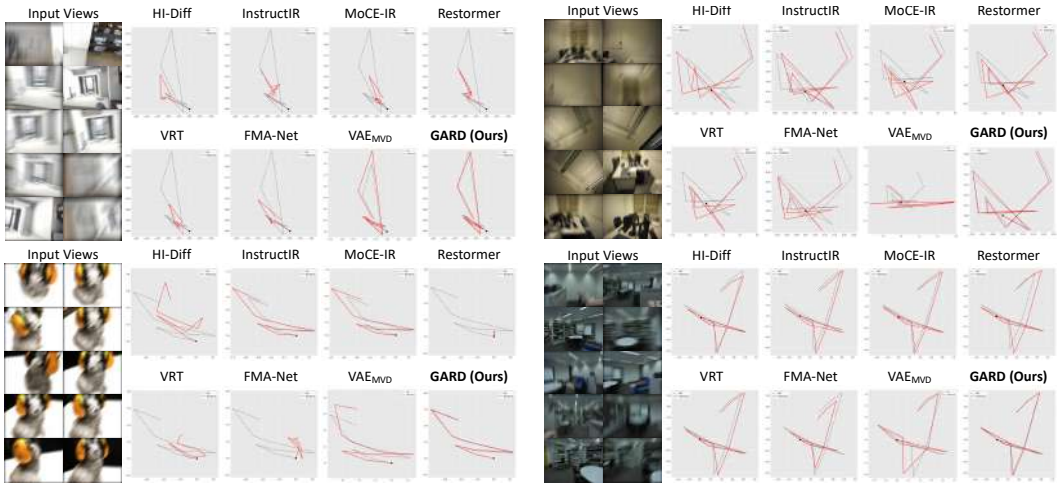


Figure 13: **Qualitative camera pose estimation on the DA3 benchmark [29].** We visualize the top-down camera trajectory results for ten input views. The black dot indicates the starting camera point. Please zoom in for clearer visualization.

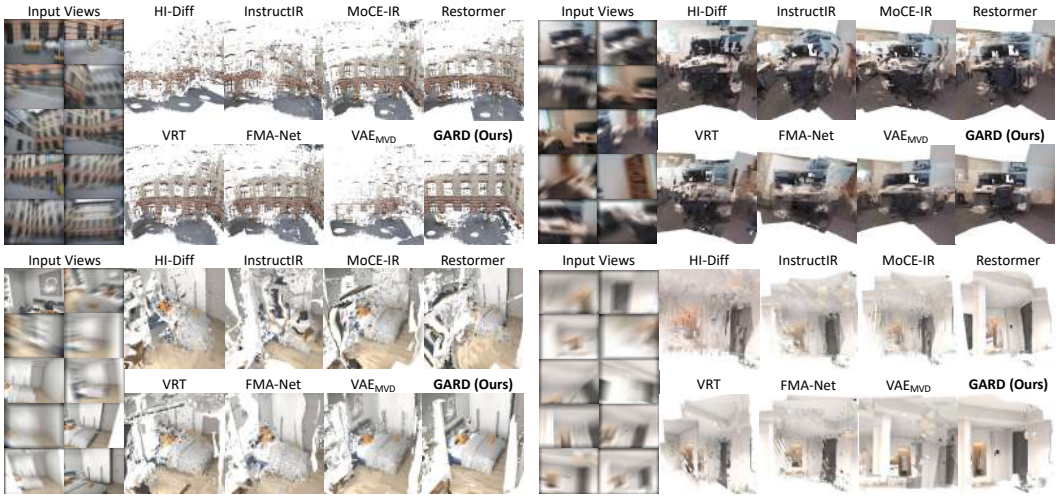


Figure 14: **Qualitative 3D reconstruction results on the DA3 benchmark [29].** We visualize the 3D reconstruction point cloud results for ten input views. Please zoom in for clearer visualization.

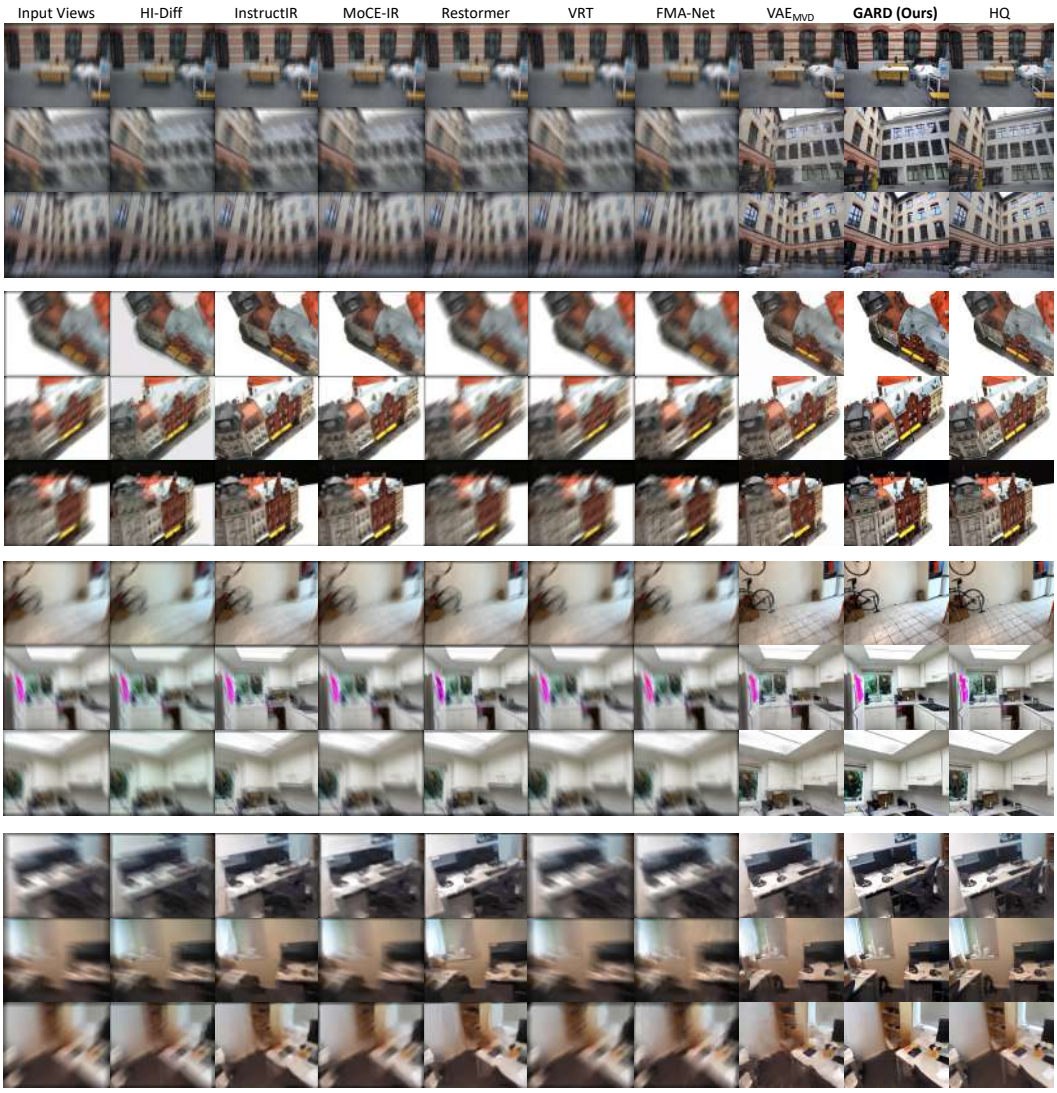


Figure 15: **Qualitative image restoration results on the DA3 benchmark [29].** We visualize three selected views out of ten input views for each dataset. Please zoom in for clearer visualization.

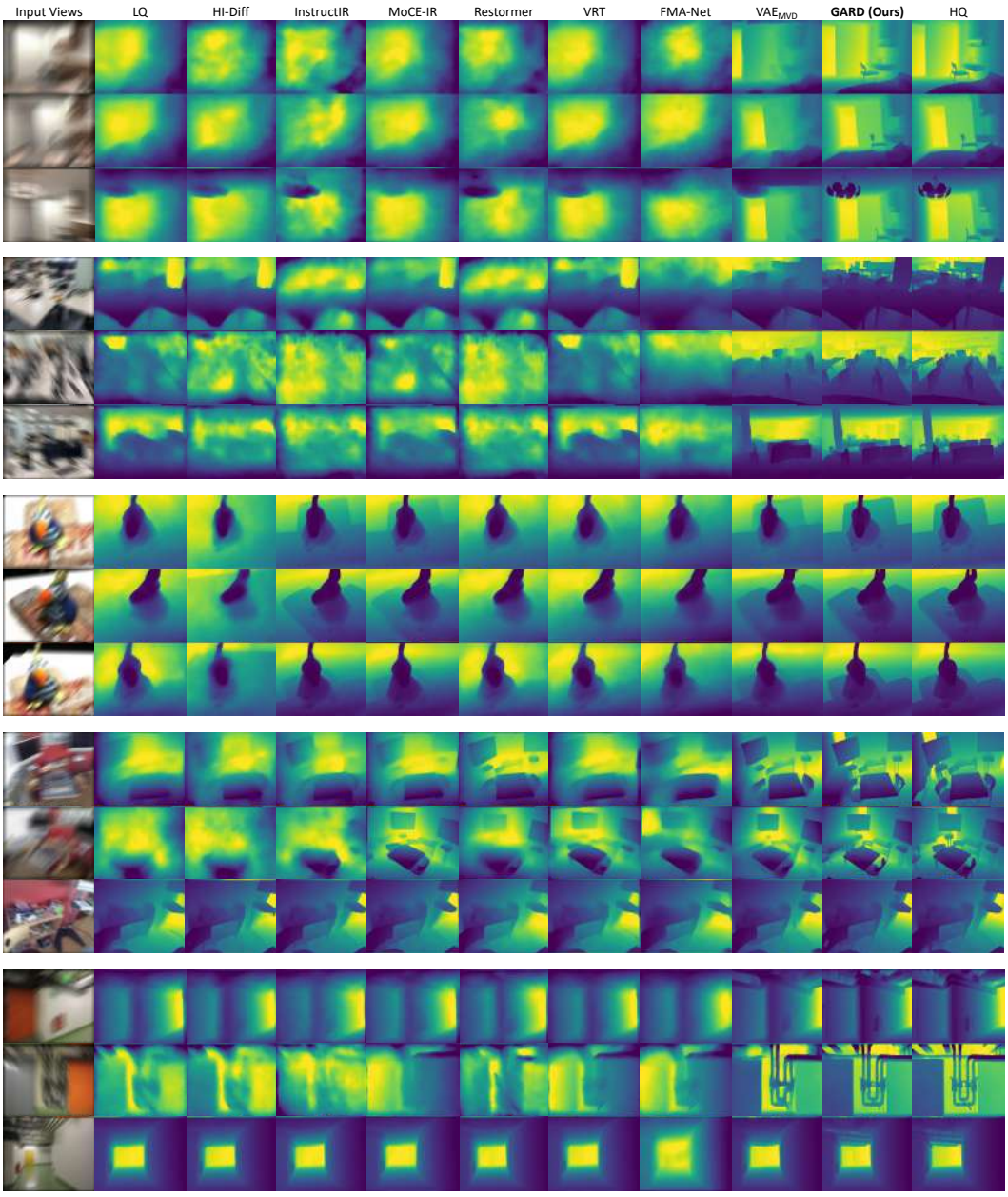


Figure 16: **Qualitative depth estimation results on the DA3 benchmark [29].** We visualize three selected views out of ten input views for each dataset. Please zoom in for clearer visualization.

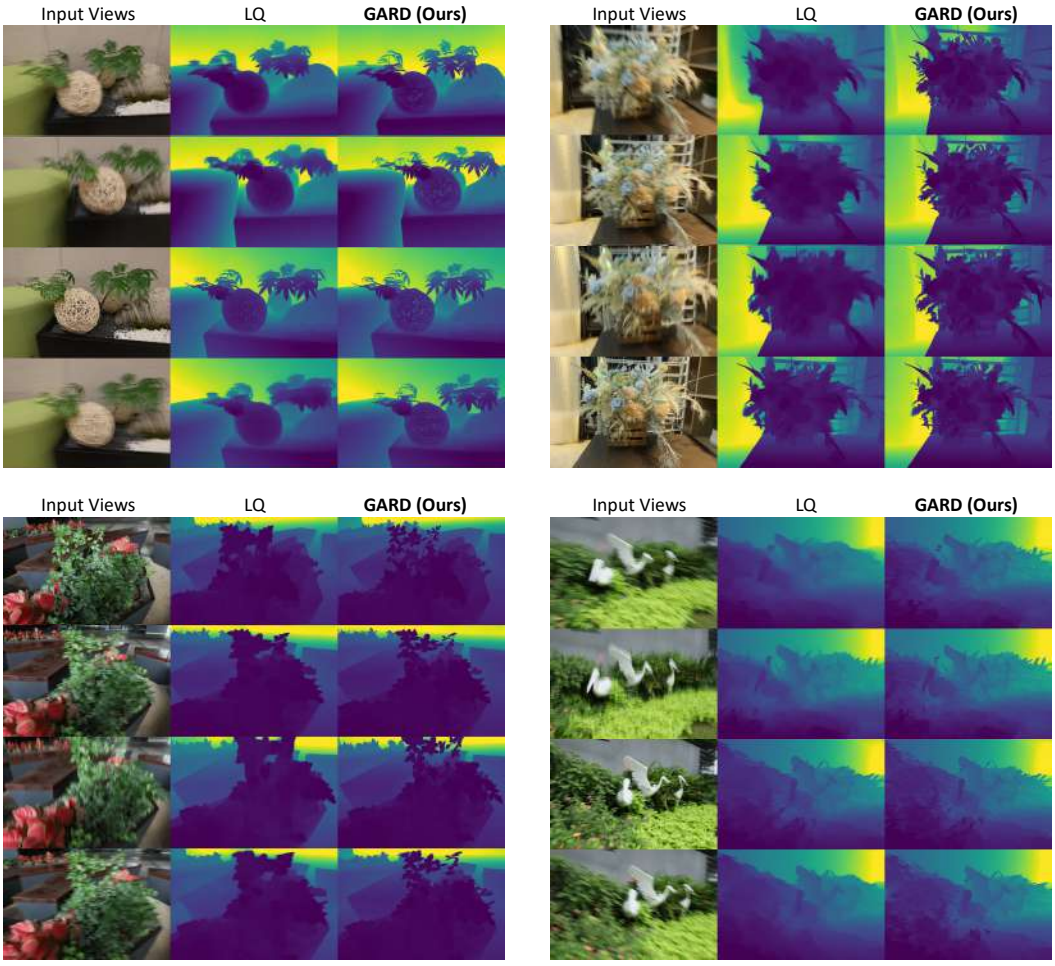


Figure 17: **Qualitative depth estimation results on the Deblur-Nerf benchmark [33].** We visualize four selected views out of ten input views for each dataset. Please zoom in for clearer visualization.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction reflects the paper's contribution.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper contains a limitations section.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We provide equations and figures for assumptions and proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We explain full details of our experiments in the main paper and in the supplementary material.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We will release our code upon acceptance.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide the settings and details in the experimental settings section.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We instead provide quantitative evaluation accompanied by ablation studies in the paper.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide the details in the implementation details.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: All the authors review the code of ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss this in the main paper.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: Our paper doesn't have such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We appropriately cite the original papers.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We explain the details and the motivation of our proposed model.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [No]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [No]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [No]

Justification: We do not use LLMs.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.