UG-DIFF: IMPROVING NEURAL RENDERING IN DISTRACTOR-CORRUPTED SCENES VIA UNCERTAINTY-GUIDED DIFFUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural rendering in real-world scenes is often compromised by transient distractors, leaving noticeable artifacts even after distractor suppression by existing distractor-free methods. Diffusion-based generative restoration can further improve rendering quality, but typically requires clean, distractor-free supervision, which is unavailable in distractor-corrupted scenes; directly applying frozen generative priors avoids this requirement but limits adaptation to scene-specific corruptions. We propose UG-Diff, an uncertainty-guided diffusion framework for robust neural rendering restoration. Our key insight is to employ uncertainty as a conditioning signal, enabling the model to simultaneously learn static structure restoration and transient distractor fitting under distractor-corrupted supervision. Specifically, we design an Uncertainty Mixing Strategy that modulates the diffusion input with spatially varying uncertainty, allowing the model to learn an uncertainty-conditioned graded restoration behavior directly from distractor-corrupted supervision: in low-uncertainty regions, it preserves more reliable static structures and performs restoration; whereas as uncertainty increases, it progressively strengthens distractor-dependent fitting. We further introduce Multi-view Condition Injection, which aggregates complementary cues from reference views to alleviate distractor-induced information loss and improve structural restoration accuracy. At inference, setting uncertainty to zero reduces the learned behavior to static structure restoration. Experiments on diverse distractor-corrupted datasets demonstrate that UG-Diff is applicable to multiple diffusion-based restoration methods and consistently outperforms existing baselines.

1 INTRODUCTION

Recently, neural rendering techniques represented by Neural Radiance Fields (NeRF) and 3D Gaussian Splatting (3DGS) have achieved remarkable performance in novel view synthesis and 3D reconstruction Du et al. (2023); Xu et al. (2024a; 2025); Jia et al. (2025). However, their performance relies heavily on multi-view geometric consistency Bao et al. (2025); He et al. (2026); in real-world scenes, transient distractors (e.g. pedestrians and vehicles) violate this assumption, making the geometric constraints ill-posed and causing rendering failures.

Several distractor-free neural rendering methods attempt to incorporate uncertainty estimation to suppress or filter transient distractors. NeRF on-the-go Ren et al. (2024) employs a dedicated uncertainty network with a robust loss to adaptively reduce distractor influence; WildGaussians Kulhanek et al. (2024) learns implicit uncertainty embeddings for individual Gaussians to improve 3DGS robustness in dynamic scenes; MU-GeNeRF Mu et al. (2026) further leverages uncertainty modeling to identify distractors and decouple cross-view structural inconsistencies, enabling adaptive modulation of supervision. However, since distractor-occluded regions are unobservable, these methods still suffer from noticeable artifacts and missing textures.

Recently, Diffusion-based generative restoration has emerged as an effective approach for improving neural rendering quality Yu et al. (2025); Wu et al. (2025b); Dong et al. (2026); Ji et al. (2026). Difix3D Wu et al. (2025a) utilizes a single-step diffusion model to restore 3DGS rendering artifacts in under-constrained regions, and GeoQuery Cao et al. (2026) introduces a geometry-aware

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

cross-attention mechanism to improve sparse-view reconstruction by integrating 3D geometric priors. However, these methods rely on clean, distractor-free supervision, which is rarely available in real-world scenes. RogSplat Kong et al. (2025) targets restoration in distractor-corrupted scenes, it relies on frozen generative priors that cannot sufficiently adapt to scene-specific corruptions. Moreover, its conditional guidance is derived from feature similarity between rendered and ground-truth images, which makes it difficult to accurately distinguish structural errors from transient distractors.

To achieve robust rendering restoration in distractor-corrupted scenes, we propose UG-Diff, an uncertainty-guided diffusion framework. As shown in Figure 1, we use uncertainty that characterizes distractors as a conditional signal, enabling the model to simultaneously learn static structure restoration and transient distractor fitting under distractor-corrupted supervision. Benefiting from this conditioning paradigm, training and inference are unified within a single framework: during training, the model learns corresponding restoration behaviors under varying uncertainty conditions; at inference, setting uncertainty to zero naturally reduces the learned behavior to static structure restoration. Specifically, we design an Uncertainty Mixing Strategy that modulates the diffusion input with spatially varying uncertainty, allowing the model to learn an uncertainty-conditioned graded restoration behavior directly from distractor-corrupted supervision: in low-uncertainty regions, it preserves more reliable static structures and performs restoration; whereas as uncertainty increases, it progressively strengthens distractor-dependent fitting. To alleviate distractor-induced information loss, we introduce Multi-view Condition Injection, which aggregates complementary cues from multiple reference views through cross-view feature interaction to improve the accuracy of structural restoration. Experiments demonstrate that UG-Diff is compatible with multiple diffusion-based restoration methods, effectively enhancing rendering quality in distractor-corrupted scenes and significantly outperforming existing baselines.

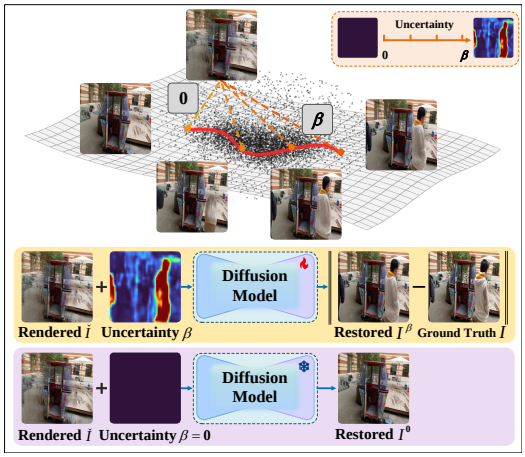

Figure 1: **Unified conditional paradigm of UG-Diff.** UG-Diff conditions on uncertainty β to guide the diffusion-based restoration process, enabling the model to jointly learn different restoration behaviors within the latent space, ranging from static structure restoration to distractor-dependent fitting; at inference, by setting $\beta = \mathbf{0}$, the model reduces to a static-structure restorer.

In summary, our main contributions are as follows: **1)** We propose UG-Diff, an uncertainty-guided diffusion framework that eliminates reliance on clean supervision, achieving robust rendering restoration in distractor-corrupted scenes. **2)** We introduce Uncertainty Mixing Strategy and Multi-view Condition Injection. The former modulates the diffusion input with uncertainty, enabling the model to learn robust restoration behavior from distractor-corrupted supervision, while the latter guides restoration by providing complementary cues; together they improve rendering quality under complex corruptions. **3)** We validate the effectiveness of UG-Diff and its applicability to multiple diffusion-based restoration methods, providing a reliable conditional training paradigm.

2 Related Work

2.1 Distractor-Free neural rendering

Existing distractor-free neural rendering methods mainly focus on suppressing and filtering transient distractors Xu et al. (2024b); Zhang et al. (2024); Dahmani et al. (2024); Li et al. (2025b;a). One line of work, RobustNeRF Sabour et al. (2023) and SpotLessSplats Sabour et al. (2025) suppress distractors as statistical outliers using robust losses, but may fail under persistent occlusions where this assumption breaks down. Another line explicitly models distractors: Ha-NeRF Chen et al. (2022) and CR-NeRF Yang et al. (2023) instead introduce learnable embeddings to capture transient variations, while DeGauss Wang et al. (2025) and HybridGS Lin et al. (2025b) separate foreground

and background Gaussians to decouple dynamic and static components, at the cost of increased complexity and computation. Uncertainty-based methods detect distractors from a probabilistic perspective: NeRF on-the-go predicts pixel-wise uncertainty to down-weight unreliable observations, while WildGaussians and MU-GeNeRF extend this idea to 3DGS and feed-forward rendering, respectively. Nevertheless, occlusion-induced information loss still degrades their performance to varying degrees Fu et al. (2025); Tang et al. (2025).

2.2 Generative priors for scene reconstruction

Generative priors have been widely adopted in scene reconstruction Yang et al. (2024); Lin et al. (2025a); Wang et al. (2026). Early works use diffusion models as geometric and appearance regularizers: Nerfdiff Gu et al. (2023) generates novel views via diffusion and distills them into NeRF; SyncDreamer Liu et al. (2024) models cross-view spatial correlations to produce consistent observations. Later methods integrate generative priors into sparse-view reconstruction: SpatialCrafter Zhang et al. (2025) uses diffusion models to complete continuous scene observations along camera trajectories; ReconX Liu et al. (2026) generates missing views conditioned on scene geometry, alleviating under-constrained issues under sparse inputs. Recent works focus on efficient refinement: Difix3D uses a one-step diffusion model to restore neural rendering artifacts; AnyRecon Chen et al. (2026) and ArtiFixer De Lutio et al. (2026) leverage video diffusion priors to enforce multi-view consistency. However, these methods rely on clean and distractor-free supervision. CoherentGS Feng et al. (2026) and RogSplat address motion blur and distractor-corrupted observations, respectively, but their frozen generic priors struggle to adapt to scene-specific corruptions. In contrast, UG-Diff eliminates both clean-supervision dependency and frozen-prior limitations, achieving more robust rendering restoration via a dedicated conditional framework.

3 Preliminaries

3.1 Distractor-aware Uncertainty Modeling

Transient distractors violate multi-view consistency, resulting in misaligned rendering residuals. These residuals can be modeled as heteroscedastic noise and optimized by minimizing the negative log-likelihood Martin-Brualla et al. (2021); Kim et al. (2023):

$$\arg\min_{\beta} \left(\frac{\|I - \check{I}\|^2}{2\beta^2} + \log \beta \right) \tag{1}$$

where I and $\check{I}$ denote the ground-truth and rendered images, respectively, and β is the predicted uncertainty. As β is close to $\mathbf{0}$, it corresponds to stable scene structure; otherwise, it indicates more pronounced transient distractors, requiring reduced supervision weight in that region during gradient backpropagation. Essentially, β fits the residual distribution unexplained by geometric consistency, and thus serves as an indicator for distractor localization and an interpretable condition for subsequent degradation modeling.

3.2 Conditional Diffusion Models

Conditional diffusion models enable efficient image generation and restoration by performing the noising and denoising processes in latent space. Given an input image I, an encoder $\mathcal{E}$ first maps it to latent $z_0 = \mathcal{E}(I)$, and noise ϵ is injected at diffusion timestep t:

$$z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \qquad \epsilon \sim \mathcal{N}(0, \mathbf{I}) \tag{2}$$

where $\bar{\alpha}_t$ controls the relative weights of signal and noise. Conditioned on c and timestep t, the denoising network ϵ_θ predicts the injected noise, with the training objective defined as:

$$\mathbb{E}_{z_0,\epsilon,t} \left[\|\epsilon - \epsilon_\theta(z_t, t, c)\|_2^2 \right] \tag{3}$$

where c can consist of geometric information or other structured signals. During inference, the model progressively denoises the latent representation under the guidance of c and reconstructs the output image through the decoder $\mathcal{D}$.

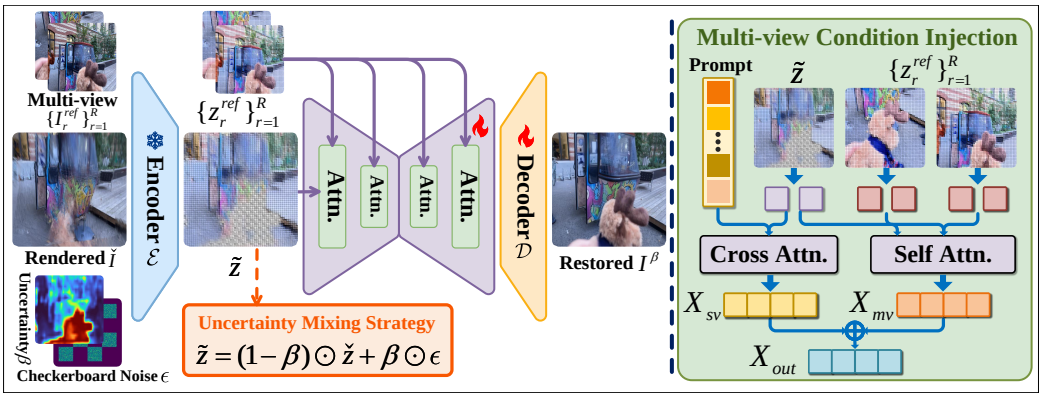

Figure 2: **Overview of UG-Diff.** Given a rendered image $\check{I}$ and R reference views $\{I_r^{ref}\}_{r=1}^R$, they are first mapped by the encoder $\mathcal{E}$ into latent features $\check{z}$ and $\{z_r^{ref}\}_{r=1}^R$. Then, the Uncertainty Mixing Strategy utilizes uncertainty β to modulate $\check{z}$ and perturbation ϵ to construct the degraded condition $\tilde{z}$, enabling the model to leverage generative priors to adapt to supervision containing distractors. Next, the Multi-view Condition Injection integrates text prompts for conditional interaction and aggregates complementary cues from $\tilde{z}$ and $\{z_r^{ref}\}_{r=1}^R$ via cross-view attention, providing sufficient geometric constraints for structural restoration. Finally, the decoder $\mathcal{D}$ maps the denoised features back to the image space, yielding the restored image I^β.

4 METHOD

Figure 2 presents the architecture of UG-Diff. We introduce Uncertainty Mixing Strategy (Sec. 4.1) and Multi-view Condition Injection (Sec. 4.2): The former modulates the diffusion input with uncertainty, enabling the model to adapt to distractor-corrupted supervision and learn robust restoration behavior; the latter aggregates complementary cues from multiple reference views to guide the model toward accurate restoration. Driven by the unified training objective (Sec. 4.3), UG-Diff achieves robust rendering restoration without requiring clean supervision.

4.1 UNCERTAINTY MIXING STRATEGY

Under the training paradigm in Sec. 3.2, the diffusion model is required to learn the mapping from the rendered image $\check{I}$ to the distractor-corrupted image I. However, since the rendered image $\check{I}$ lacks structural information about these distractors, optimizing Eq. 2 would force the model to absorb distractor-corrupted supervision, thereby corrupting its generative prior and severely degrading its structural restoration capability.

We propose an Uncertainty Mixing Strategy that treats the uncertainty β extracted in Sec. 3.1 as a continuous variable characterizing distractors, and accordingly models a conditional distribution $p(\cdot \mid \beta)$: higher uncertainty introduces more perturbations, thereby leveraging generative priors to fit distractors, whereas lower uncertainty preserves more of the original rendering for structural restoration, enabling the model to stably learn an uncertainty-conditioned graded restoration behavior. Since β is spatially continuous, a single training sample can simultaneously encompass varying states ranging from low to high uncertainty. Consequently, the model can jointly learn this continuous process within a unified latent space, spanning from structural restoration to distractor-dependent fitting, without explicitly partitioning different degradation levels. Specifically, given a rendered image $\check{I}$, the restored image I^β is sampled conditioned on β:

$$I^\beta \sim p_\theta(\cdot \mid \check{I}, \beta) \tag{4}$$

where $p_\theta(\cdot)$ denotes the parameterized generative distribution. For this purpose, we construct a degraded observation $\tilde{z}$ using uncertainty to serve as the conditional input:

$$\tilde{z} = (1 - \beta) \odot \check{z} + \beta \odot \epsilon, \quad \epsilon \sim p_\epsilon(\cdot) \tag{5}$$

where $\check{z} = \mathcal{E}(\check{I})$, $\odot$ is element-wise multiplication, and $p_\epsilon(\cdot)$ represents a predefined perturbation pattern. As shown in Figure 3, as β is close to $\mathbf{0}$, $\tilde{z}$ becomes exceptionally close to $\check{z}$, and I^β is gradually dominated by the distribution $p_\theta(\cdot \mid \check{I}, \mathbf{0})$; at this point, the model maximally preserves

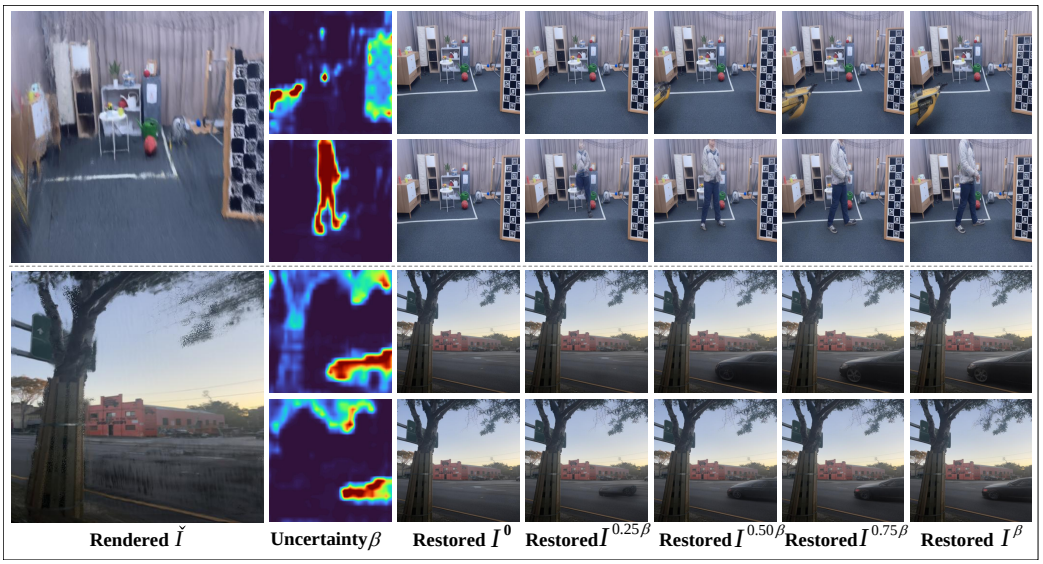

Figure 3: **Uncertainty-conditioned graded restoration behavior of UG-Diff.** As uncertainty β increases, the model progressively transitions from prioritizing static structure restoration to stronger distractor-dependent fitting.

the structural information in $\check{z}$ and gradually converts into generative restoration for static structures. Conversely, when the conditional input deviates from $\check{z}$, it relies more on the generative prior to fit distractors and align with real observations. Accordingly, Uncertainty Mixing establishes a β-controlled continuous conditional mapping between the two behaviors, thereby integrating the training and inference processes under the unified conditional modeling framework: At inference, when $\beta = \mathbf{0}$, the model naturally converts into a pure static structure restorer.

4.2 Multi-view Condition Injection

Local information loss caused by distractor corruption makes it difficult to accurately recover scene structure using single-view generative restoration alone. Therefore, we introduce Multi-view Condition Injection that allows the model to directly retrieve high-frequency textures and structural evidence from multiple reference views, providing sufficient cross-view geometric constraints for structural restoration. Specifically, in addition to constructing $\tilde{z}$, we also select R views as reference conditions $\{I_r^{ref}\}_{r=1}^{R}$ and encode them into latent features $z_r^{ref} = \mathcal{E}(I_r^{ref})$. To enable cross-view interaction, we compute two types of attention outputs in each self-attention block. First, each view independently performs cross-attention with the text prompt, yielding $\mathbf{X}_{sv} \in \mathbb{R}^{B\times V\times N\times C}$, where $V = 1 + R$ with the rendered view corresponding to $v = 0$, N is the number of tokens, and C is the feature dimension. Then, $\tilde{z}$ and $\{z_r^{ref}\}_{r=1}^{R}$ are concatenated along the sequence dimension to form $\mathbf{A} \in \mathbb{R}^{B\times L\times C}$, after which cross-attention is performed:

$$\mathbf{X}_{mv} = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V, \quad Q = W_Q\mathbf{A},\ K = W_K\mathbf{A},\ V = W_V\mathbf{A} \tag{6}$$

where W_Q, W_K, W_V are projection weight matrices, and d denotes the feature scaling dimension. To inject cross-view information while preserving the original generative prior of the diffusion model, we introduce a fixed fusion weight α and perform residual fusion between $\mathbf{X}_{sv}$ and $\mathbf{X}_{mv}$:

$$\mathbf{X}_{out} = \mathbf{X}_{sv} + \alpha(\mathbf{X}_{mv} - \mathbf{X}_{sv}) \tag{7}$$

All views pass through the entire denoising backbone; only the rendered view feature $\mathbf{x}_{re}$ at index $v = 0$ is extracted at the final output layer, while the reference view features are no longer involved in subsequent computation. Finally, $\mathbf{x}_{re}$ continues the denoising process and is mapped back to the image space by the decoder, yielding the final restored image $I^\beta = \mathcal{D}(\mathbf{x}_{re})$. This asymmetric design not only balances training efficiency and flexibility, but also further improves restoration performance by sufficiently aggregating complementary cues from the reference images.

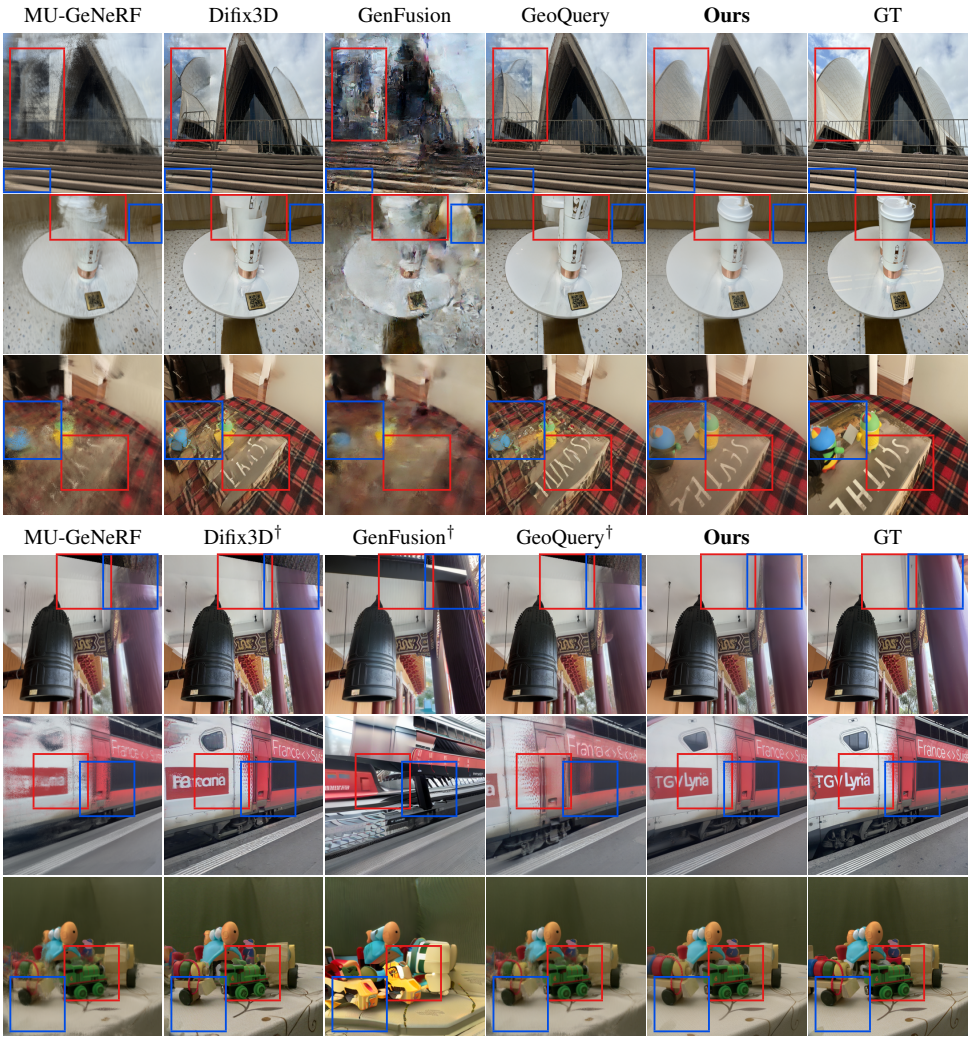

Figure 4: **Qualitative comparison with diffusion-based generative restoration methods.** † denotes inference using frozen pre-trained weights.

4.3 Training Objective

We adopt the pixel-level reconstruction loss $\mathcal{L}_{\text{recon}}$ and a structure-aware loss $\mathcal{L}_{\text{lpips}}$ to jointly train diffusion model: $\mathcal{L}_{\text{recon}}$ computes the ℓ_2 distance between restored image I^{β} and ground-truth image I to constrain structural recovery; $\mathcal{L}_{\text{lpips}}$ measures their perceptual discrepancy to enhance semantic consistency and improve high-frequency details affected by distractors. The overall loss is:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{recon}} + \lambda_2 \mathcal{L}_{\text{lpips}} \tag{8}$$

where λ_1 and λ_2 denote the loss weights. Notably, prior methods often introduce a Gram loss to enhance textures by modeling global feature correlations Wu et al. (2025a); Cao et al. (2026). However, in distractor-corrupted scenes, distractor components are also incorporated into the global statistics, which can impair the recovery of stable scene structures.

5 Experiments

5.1 Datasets and Baselines

We train on the distractor-corrupted datasets DF3DV-1K Lu et al. (2026) and On-the-go Ren et al. (2024), and evaluate on DF3DV-41 and On-the-go for in-domain, and on RobustNeRF Sabour et al. (2023) for cross-domain. We compare UG-Diff against recent diffusion-based restoration methods:

Table 1: **Quantitative comparison on in-domain and cross-domain test sets.** † denotes inference using frozen pre-trained weights; "w/ uncer" denotes that the Uncertainty Mixing Strategy is introduced during training.

Method	DF3DV-41				On-the-go				RobustNeRF			
	PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FID↓
MU-GeNeRF (CVPR 2026)	20.49	<u>0.675</u>	0.399	72.03	19.78	<u>0.609</u>	0.422	127.43	22.78	<u>0.781</u>	0.395	121.91
Difix3D (CVPR 2025)	20.08	0.646	0.231	28.50	19.03	0.569	0.303	48.98	22.39	0.762	0.272	47.87
Difix3D† (frozen)	20.38	0.654	0.293	25.86	19.44	0.586	0.297	45.77	22.77	0.771	0.272	44.53
Difix3D (w/ uncer)	20.49	0.664	0.286	26.42	19.59	0.597	0.296	47.03	22.91	0.783	0.269	46.11
GenFusion (CVPR 2025)	16.81	0.492	0.559	215.84	16.48	0.445	0.556	275.09	19.59	0.691	0.491	247.06
GenFusion† (frozen)	17.11	0.552	0.450	55.14	15.98	0.468	0.464	115.54	18.73	0.676	0.455	92.04
GenFusion (w/ uncer)	19.39	0.590	0.367	34.39	18.37	0.506	0.378	67.80	21.34	0.746	0.364	71.02
GeoQuery (SIGGRAPH 2026)	20.64	0.666	0.278	25.75	19.56	0.589	0.289	47.17	23.10	0.783	<u>0.260</u>	<u>44.05</u>
GeoQuery† (frozen)	17.82	0.591	0.428	52.17	16.73	0.503	0.448	87.98	21.91	0.770	0.363	86.84
GeoQuery (w/ uncer)	<u>20.76</u>	0.672	<u>0.275</u>	<u>24.96</u>	<u>19.81</u>	0.600	<u>0.285</u>	**44.97**	<u>23.15</u>	0.787	0.261	44.30
UG-Diff (Ours)	**21.60**	**0.694**	**0.246**	**24.77**	**20.70**	**0.634**	**0.250**	<u>45.40</u>	**23.62**	**0.800**	**0.236**	**43.69**

Difix3D Wu et al. (2025a), GenFusion Wu et al. (2025b) and GeoQuery Cao et al. (2026). All baselines are trained on the above datasets to evaluate robustness; we also directly perform inference with their pretrained weights to assess adaptability to scene-specific corruptions.

5.2 Experiment Settings

We adopt SD-Turbo Sauer et al. (2024) as the base diffusion model and optimize it using AdamW, with the learning rate set to 2×10^{-5} and the batch size to 1. The settings are $R = 2$, $\alpha = 0.7$, $\lambda_1 = 1.0$, $\lambda_2 = 1.0$, with an image resolution of 512×512. We use MU-GeNeRF Mu et al. (2026) as the neural rendering backbone to generate the rendered image $\check{I}$ and the uncertainty β. All experiments are conducted on a single NVIDIA A6000 GPU.

5.3 Comparison of Restoration Quality

As shown in Figure 4 and Table 1, compared with the original MU-GeNeRF renderings, UG-Diff effectively mitigates artifacts and missing textures caused by distractors, substantially enhancing overall rendering quality, with particularly notable gains in LPIPS/FID. In contrast, DIFIX3D and GeoQuery show unstable performance and even negative gains in PSNR/SSIM, failing to preserve global structure and local texture, while GenFusion suffers a marked performance drop. This is because they rely on clean and consistent supervision to constrain generation, and are thus prone to overfit distractor-corrupted targets during training while disrupting the original generative prior; furthermore, on unseen cross-domain data, frozen generative priors struggle to adapt to complex corruptions, causing restoration results to deviate from the actual structure. Conversely, UG-Diff utilizes uncertainty to modulate perturbations, enabling generative priors to adapt to distractor-corrupted supervision and learn restoration behaviors across degradation degrees, improving robustness in complex scenes.

Table 2: **Quantitative Comparison of Ablation Components.** "Ours" denotes the default setting: Uncertainty Mixing with Checkerboard Noise applied in latent space.

Setting	PSNR↑	SSIM↑	LPIPS↓	FID↓
Baseline	20.45	0.668	0.317	40.66
Unc. Modulation				
ControlNet	21.00	0.684	0.280	31.38
Concat	21.47	0.691	0.256	25.79
Cross Attention	21.32	0.694	0.255	26.23
Perturbation				
Uniform Noise	21.31	0.689	0.254	25.42
Gaussian Noise	21.24	0.687	0.256	25.88
Oper. Space				
Pixel Space	21.27	0.668	0.262	25.81
Ours	**21.60**	**0.694**	**0.246**	**24.77**

5.4 Ablation Study

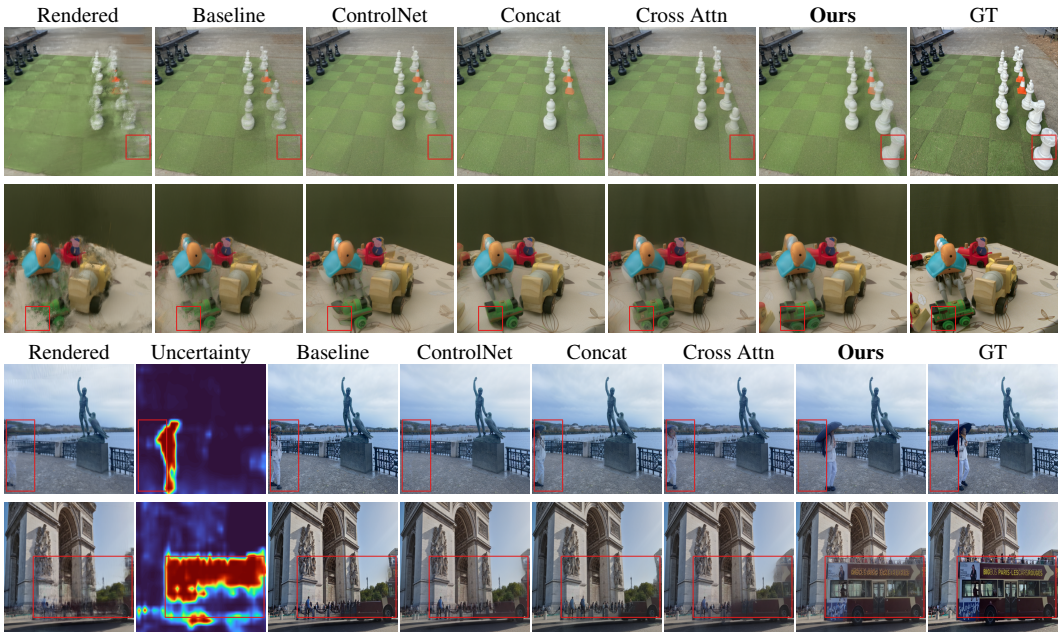

Figure 5: **Qualitative ablation of uncertainty modulation strategies.** "Baseline" denotes the setting without any uncertainty modulation.

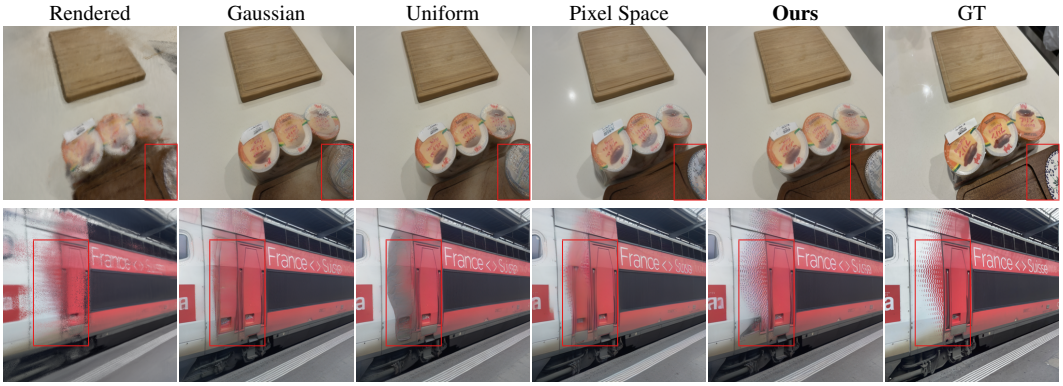

Figure 6: **Qualitative ablation of perturbation patterns and their operating spaces.**

As shown in Figure 5 and Table 2, we first compare different uncertainty modulation strategies (details in the Appendix). Compared with Uncertainty Mixing, the other strategies exhibit varying degrees of performance drop. Specifically, ControlNet struggles to sufficiently alter the original generation trajectory, thus limiting its generative capability in degraded regions; channel concatenation directly integrates uncertainty along the channel dimension, failing to provide explicit spatial modulation for corrupted regions; Cross Attention relies on token-level feature matching, weakening local spatial correspondences and hindering precise control over degraded regions. In contrast, Uncertainty Mixing smoothly incorporates uncertainty into the structural signal through continuous element-wise modulation, enabling more stable learning of restoration behaviors under distractor-corrupted supervision. As shown in Figure 6, further ablation on the perturbation patterns and their operating spaces reveals that checkerboard noise outperforms both uniform and Gaussian noise. This may be because its regular structured perturbations help the model better distinguish distractors from scene structure; applying perturbations in the latent space directly acts on the denoising process, making uncertainty modulation in latent space outperform that in pixel space.

Table 3: **Ablation on Multi-View Condition Injection.** $R = 2$ is used as the default setting.

Ref. Views	PSNR↑	SSIM↑	LPIPS↓	FID↓
$R = 0$	20.77	0.686	0.297	38.54
$\mathbf{R = 2}$	**21.60**	**0.694**	**0.246**	24.77
$R = 4$	21.40	0.692	0.252	24.84
$R = 6$	21.53	0.693	**0.246**	**24.46**

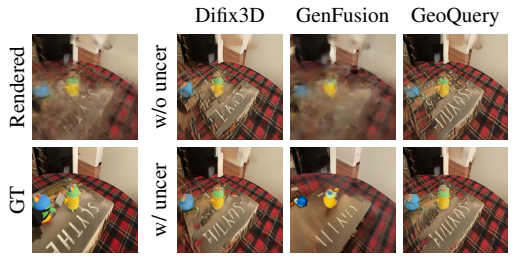

Figure 7: **Evaluation of the generalization of the Uncertainty Mixing Strategy.**

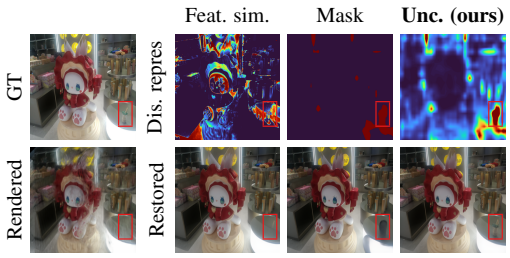

Figure 8: **Qualitative comparison of different distractor-aware representations.** The "cup" is a transient distractor.

As shown in Table 3, analysis of Multi-view Condition Injection reveals that only two views significantly improve restoration performance, indicating the sufficiency of a few cross-view observations for complementary cues.

5.5 Analysis of Generality

We further evaluate the generalizability of the Uncertainty Mixing Strategy. As shown in Figure 7 and Table 1, integrating this strategy improves the overall restoration quality of Difix3D and GenFusion, with particularly notable gains for GenFusion. This indicates that it effectively prevents distractor-corrupted supervision from disrupting the generative prior and encourages the model to learn graded restoration behaviors across varying degradation degrees. In contrast, GeoQuery shows relatively limited gains, possibly because its geometry-aware constraints conflict with uncertainty-based modulation in the latent space, thereby limiting the model's ability to adapt to distractors.

5.6 Comparison of Distractor-Aware

We compare continuous uncertainty with two alternative distractor-aware representations, with results shown in Figure 8 and Table 4. RogSplat Kong et al. (2025) perceives distractors via feature similarity between rendered and ground-truth images, but it cannot effectively distinguish rendering errors from distractors, thus failing to establish reliable conditional mappings. The binary mask identifies distractor regions using a hard threshold, whose discrete nature breaks the continuity of the conditional mapping and hinders stable learning of restoration behaviors across different degradation degrees. Conversely, MUGeNeRF's uncertainty not only characterizes distractor intensity more accurately, but also provides continuous and smooth modulation of perturbations, enabling more stable adaptation to complex real-world corruptions.

Table 4: **Quantitative comparison of different distractor-aware representations.**

Setting	PSNR↑	SSIM↑	LPIPS↓	FID↓
Feature Similarity	20.32	0.677	0.269	29.16
Binary Mask	21.10	0.682	0.264	26.50
Uncertainty (Ours)	**21.60**	**0.694**	**0.246**	**24.77**

6 Conclusion and Limitations

We present UG-Diff, an uncertainty-guided diffusion framework for robust neural rendering restoration. Specifically, the Uncertainty Mixing Strategy modulates the diffusion input with uncertainty, enabling stable learning of uncertainty-conditioned graded restoration behavior under distractor-corrupted supervision; meanwhile, Multi-view Condition Injection aggregates complementary cues from multiple reference views to alleviate distractor-induced information loss and improve structural restoration accuracy. However, UG-Diff still has several limitations: **1)** Under large-scale rendering collapse, limited structural information may cause the model to over-rely on generative priors, leading to restoration results that deviate from the real scene. **2)** The restoration performance of UG-Diff is affected by reference views; with large viewpoint discrepancies or insufficient co-visible regions between rendered and reference views, the model may struggle to acquire sufficient cross-view complementary cues. In future work, we will explore stricter geometric consistency constraints or uncertainty-based calibration mechanisms to further improve restoration quality.

AI USE STATEMENT

Generative AI tools were used to assist with language polishing, translation, and drafting or revising parts of the manuscript. All AI-assisted text was reviewed and edited by the authors. The research ideas, methodology, experimental design, implementation, results, and scientific conclusions were developed and verified by the authors, who take full responsibility for the content of the paper.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility of our work, we provide detailed descriptions of the method design, training settings, and experimental configurations in the main paper, along with the core implementation code in the supplementary material. We will further organize and release the relevant code and configurations in the final version to facilitate result verification and future research.

ETHICS STATEMENT

This research focuses on improving diffusion-based neural rendering and scene reconstruction algorithms. It does not involve sensitive personal information or private data, nor does it present apparent risks related to safety, bias, or potentially harmful applications. All datasets used in this work are publicly available for research purposes, and we follow their corresponding licenses and usage guidelines.

REFERENCES

Yanqi Bao, Tianyu Ding, Jing Huo, Yaoli Liu, Yuxin Li, Wenbin Li, Yang Gao, and Jiebo Luo. 3D Gaussian Splatting: Survey, Technologies, Challenges, and Opportunities. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(7):6832–6852, 2025. doi: 10.1109/tcsvt.2025.3538684.

Xiao Cao, Yuze Li, Youmin Zhang, Jiayu Song, Cheng Yan, Wen Li, and Lixin Duan. GeoQuery: Geometry-Query Diffusion for Sparse-View Reconstruction. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–11, 2026. doi: 10.1145/3799902.3811222.

Xingyu Chen, Qi Zhang, Xiaoyu Li, Yue Chen, Ying Feng, Xuan Wang, and Jue Wang. Hallucinated Neural Radiance Fields in the Wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12943–12952, 2022.

Yutian Chen, Shi Guo, Renbiao Jin, Tianshuo Yang, Xin Cai, Yawen Luo, Mingxin Yang, Mulin Yu, Linning Xu, and Tianfan Xue. AnyRecon: Arbitrary-View 3D Reconstruction with Video Diffusion Model. *arXiv preprint arXiv:2604.19747*, 2026. URL https://arxiv.org/abs/2604.19747.

Hiba Dahmani, Moussab Bennehar, Nathan Piasco, Luis Roldão, and Dzmitry Tsishkou. SWAG: Splatting in the Wild Images with Appearance-Conditioned Gaussians. In *European Conference on Computer Vision*, pp. 325–340. Springer, 2024. doi: 10.1007/978-3-031-73116-7_19.

Riccardo De Lutio, Tobias Fischer, Yen-Yu Chang, Yuxuan Zhang, Zhangjie Wu, Xuanchi Ren, Tianchang Shen, Katarína Tóthová, Zan Gojcic, and Haithem Turki. ArtiFixer: Enhancing and Extending 3D Reconstruction with Auto-Regressive Diffusion Models. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pp. 1–12, 2026. doi: 10.1145/3799902.3811060.

Yitong Dong, Qi Zhang, Minchao Jiang, Zhiqiang Wu, Qingnan Fan, Ying Feng, Huaqi Zhang, Hujun Bao, and Guofeng Zhang. One-Shot Refiner: Boosting Feed-forward Novel View Synthesis via One-Step Diffusion. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(5): 3659–3667, 2026. doi: 10.1609/aaai.v40i5.37365.

Yilun Du, Cameron Smith, Ayush Tewari, and Vincent Sitzmann. Learning To Render Novel Views From Wide-Baseline Stereo Pairs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4970–4980, 2023.

Chaoran Feng, Zhankuo Xu, Yingtao Li, Jianbin Zhao, Jiashu Yang, Wangbo Yu, Li Yuan, and Yonghong Tian. Breaking the Vicious Cycle: Coherent 3D Gaussian Splatting From Sparse and Motion-Blurred Views. *IEEE Transactions on Circuits and Systems for Video Technology*, 36(9): 13723–13737, 2026. doi: 10.1109/tcsvt.2026.3700148.

Chuanyu Fu, Yuqi Zhang, Kunbin Yao, Guanying Chen, Yuan Xiong, Chuan Huang, Shuguang Cui, and Xiaochun Cao. RobustSplat: Decoupling Densification and Dynamics for Transient-Free 3DGS. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 27126–27136. IEEE, 2025.

Jiatao Gu, Alex Trevithick, Kai-En Lin, Joshua M. Susskind, Christian Theobalt, Lingjie Liu, and Ravi Ramamoorthi. NerfDiff: Single-image View Synthesis with NeRF-guided Distillation from 3D-aware Diffusion. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 11808–11826. PMLR, 2023.

Shuting He, Peilin Ji, Yitong Yang, Changshuo Wang, Jiayi Ji, Yinglin Wang, and Henghui Ding. A Survey on 3D Gaussian Splatting Applications: Segmentation, Editing, and Generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–20, 2026. doi: 10.1109/tpami.2026.3704980. Early Access.

Yuzhou Ji, Ke Ma, Hong Cai, Anchun Zhang, Lizhuang Ma, and Xin Tan. LidarPainter: One-Step Away from Any Lidar View to Novel Guidance. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(7):5332–5340, 2026. doi: 10.1609/aaai.v40i7.37449.

Heng Jia, Linchao Zhu, and Na Zhao. H3R: Hybrid Multi-view Correspondence for Generalizable 3D Reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7655–7665, 2025.

Injae Kim, Minhyuk Choi, and Hyunwoo J Kim. UP-NeRF: Unconstrained Pose Prior-Free Neural Radiance Field. In *Advances in Neural Information Processing Systems*, volume 36, pp. 68184–68196, 2023. doi: 10.52202/075280-2981.

Hanyang Kong, Xingyi Yang, and Xinchao Wang. RogSplat: Robust Gaussian Splatting via Generative Priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 25735–25745. IEEE, 2025.

Jonas Kulhanek, Songyou Peng, Zuzana Kukelova, Marc Pollefeys, and Torsten Sattler. WildGaussians: 3D Gaussian Splatting In the Wild. In *Advances in Neural Information Processing Systems*, volume 37, pp. 21271–21288, 2024. doi: 10.52202/079017-0670.

Chengqi Li, Zhihao Shi, Yangdi Lu, Wenbo He, and Xiangyu Xu. Robust Neural Rendering in the Wild with Asymmetric Dual 3D Gaussian Splatting. In *Advances in Neural Information Processing Systems*, volume 38, pp. 50166–50191, 2025a. doi: 10.52202/085713-1672.

Mingrui Li, Shuhao Zhai, Zibing Zhao, Luyue Sun, Xinxiao Wang, Dong Li, Shuhong Liu, and Hongyu Wang. Wild3A: Novel View Synthesis from Any Dynamic Images in Seconds. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 7472–7480, 2025b. doi: 10.1145/3746027.3754899.

Chenguo Lin, Panwang Pan, Bangbang Yang, Zeming Li, and Yadong Mu. DiffSplat: Repurposing Image Diffusion Models for Scalable Gaussian Splat Generation. In *International Conference on Learning Representations*, volume 2025, pp. 40745–40767, 2025a.

Jingyu Lin, Jiaqi Gu, Lubin Fan, Bojian Wu, Yujing Lou, Renjie Chen, Ligang Liu, and Jieping Ye. HybridGS: Decoupling Transients and Statics with 2D and 3D Gaussian Splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 788–797. IEEE, 2025b.

Fangfu Liu, Wenqiang Sun, Hanyang Wang, Yikai Wang, Haowen Sun, Junliang Ye, Jun Zhang, and Yueqi Duan. ReconX: Reconstruct Any Scene From Sparse Views With Video Diffusion Model. *IEEE Transactions on Image Processing*, 35:2305–2319, 2026. doi: 10.1109/tip.2026.3666733.

Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. SyncDreamer: Generating Multiview-consistent Images from a Single-view Image. In *International Conference on Learning Representations*, volume 2024, pp. 27676–27697, 2024.

Cheng-You Lu, Yi-Shan Hung, Wei-Ling Chi, Hao-Ping Wang, Charlie Li-Ting Tsai, Yu-Cheng Chang, Yu-Lun Liu, Thomas Do, and Chin-Teng Lin. DF3DV-1K: A Large-Scale Dataset and Benchmark for Distractor-Free Novel View Synthesis. In *European Conference on Computer Vision*, pp. 185–206. Springer, 2026. doi: 10.1007/978-3-032-37429-5_12.

Ricardo Martin-Brualla, Noha Radwan, Mehdi S. M. Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7210–7219. IEEE, 2021.

Wenjie Mu, Zhan Li, Chuanzhou Su, Xuanyi Shen, Ziniu Liu, Fan Lu, Yujian Mo, Junqiao Zhao, Tiantian Feng, Chen Ye, and Guang Chen. MU-GeNeRF: Multi-view Uncertainty-guided Generalizable Neural Radiance Fields for Distractor-aware Scene. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 38049–38059, June 2026.

Weining Ren, Zihan Zhu, Boyang Sun, Jiaqi Chen, Marc Pollefeys, and Songyou Peng. NeRF On-the-go: Exploiting Uncertainty for Distractor-free NeRFs in the Wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8931–8940, 2024.

Sara Sabour, Suhani Vora, Daniel Duckworth, Ivan Krasin, David J. Fleet, and Andrea Tagliasacchi. RobustNeRF: Ignoring Distractors With Robust Losses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20626–20636, 2023.

Sara Sabour, Lily Goli, George Kopanas, Mark Matthews, Dmitry Lagun, Leonidas Guibas, Alec Jacobson, David Fleet, and Andrea Tagliasacchi. SpotLessSplats: Ignoring Distractors in 3D Gaussian Splatting. *ACM Transactions on Graphics*, 44(2):1–11, 2025. doi: 10.1145/3727143.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial Diffusion Distillation. In *European Conference on Computer Vision*, pp. 87–103. Springer, 2024. doi: 10.1007/978-3-031-73016-0_6.

Jiadong Tang, Yu Gao, Dianyi Yang, Liqi Yan, Yufeng Yue, and Yi Yang. DroneSplat: 3D Gaussian Splatting for Robust 3D Reconstruction from In-the-Wild Drone Imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 833–843. IEEE, 2025.

Rui Wang, Quentin Lohmeyer, Mirko Meboldt, and Siyu Tang. DeGauss: Dynamic-Static Decomposition with Gaussian Splatting for Distractor-free 3D Reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6294–6303. IEEE, 2025.

YuAn Wang, Xiaofan Li, Chi Huang, Wenhao Zhang, Hao Li, Bosheng Wang, Xun Sun, and Jun Wang. FaithFusion: Harmonizing Reconstruction and Generation via Pixel-wise Information Gain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1198–1209, 2026.

Jay Zhangjie Wu, Yuxuan Zhang, Haithem Turki, Xuanchi Ren, Jun Gao, Mike Zheng Shou, Sanja Fidler, Zan Gojcic, and Huan Ling. DIFIX3D+: Improving 3D Reconstructions with Single-Step Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26024–26035. IEEE, 2025a.

Sibo Wu, Congrong Xu, Binbin Huang, Andreas Geiger, and Anpei Chen. GenFusion: Closing the Loop between Reconstruction and Generation via Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6078–6088. IEEE, 2025b.

Haofei Xu, Anpei Chen, Yuedong Chen, Christos Sakaridis, Yulun Zhang, Marc Pollefeys, Andreas Geiger, and Fisher Yu. MuRF: Multi-Baseline Radiance Fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20041–20050, 2024a.

Haofei Xu, Songyou Peng, Fangjinhua Wang, Hermann Blum, Daniel Barath, Andreas Geiger, and Marc Pollefeys. DepthSplat: Connecting Gaussian Splatting and Depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16453–16463, 2025.

Jiacong Xu, Yiqun Mei, and Vishal M Patel. Wild-GS: Real-Time Novel View Synthesis from Unconstrained Photo Collections. In *Advances in Neural Information Processing Systems*, volume 37, pp. 103334–103355, 2024b. doi: 10.52202/079017-3283.

Yifan Yang, Shuhai Zhang, Zixiong Huang, Yubing Zhang, and Mingkui Tan. Cross-Ray Neural Radiance Fields for Novel-View Synthesis from Unconstrained Image Collections. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15901–15911, 2023.

Yunhan Yang, Yukun Huang, Xiaoyang Wu, Yuan-Chen Guo, Song-Hai Zhang, Hengshuang Zhao, Tong He, and Xihui Liu. DreamComposer: Controllable 3D Object Generation via Multi-View Conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8111–8120. IEEE, 2024.

Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. ViewCrafter: Taming Video Diffusion Models for High-fidelity Novel View Synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–18, 2025. doi: 10.1109/tpami.2025.3613256. Early Access.

Dongbin Zhang, Chuming Wang, Weitao Wang, Peihao Li, Minghan Qin, and Haoqian Wang. Gaussian in the Wild: 3D Gaussian Splatting for Unconstrained Image Collections. In *European Conference on Computer Vision*, pp. 341–359. Springer, 2024. doi: 10.1007/978-3-031-73116-7_20.

Songchun Zhang, Huiyao Xu, Sitong Guo, Zhongwei Xie, Hujun Bao, Weiwei Xu, and Changqing Zou. SpatialCrafter: Unleashing the Imagination of Video Diffusion Models for Scene Reconstruction from Limited Observations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 27794–27805. IEEE, 2025.

A Appendix

A.1 Additional Experimental Details

For in-domain evaluation, we use DF3DV-41 Lu et al. (2026) as the test set; from the On-the-go dataset Ren et al. (2024), we test on indoor and outdoor scenes including Corner, Train, Patio, and Patio-High, with the remaining scenes used for training. For cross-domain evaluation, we test on the Android, Crab1, Crab2, and Statue scenes from RobustNeRF Sabour et al. (2023). MU-GeNeRF Mu et al. (2026), which generates the rendered results $\check{I}$ and uncertainty estimates $\hat{\beta}$, is trained following its official configuration, including pre-extracted DINOv2 features, clutter/clean data splits, view selection, and hyperparameter settings. Since the uncertainty $\hat{\beta}$ predicted by MU-GeNeRF does not lie within the $[0, 1]$ interval, we apply quantile-based normalization prior to Uncertainty Mixing:

$$\beta = \text{clip}\left(\frac{\hat{\beta} - q_{0.01}}{q_{0.99} - q_{0.01}}, 0, 1\right) \tag{9}$$

where $q_{0.01}$ and $q_{0.99}$ denote the 1% and 99% quantiles of the original uncertainty, respectively. This operation mitigates the impact of extreme outliers and stably maps β into $[0, 1]$ for subsequent uncertainty-conditioned modulation.

A.2 Uncertainty Modulation Strategies

This section describes the implementation details of the uncertainty modulation strategies introduced in Sec. 5.4. For ControlNet, we employ the uncertainty map as an additional condition and inject conditional residuals via the ControlNet branch to guide the diffusion model for restoration; for Concat, we expand the RGB data into RGB-U data along the channel dimension to independently maintain uncertainty information; for Cross Attention, we first tokenize the uncertainty map and concatenate it with the text prompt, subsequently performing cross-attention with the latent feature $\check{z}$ to inject the uncertainty condition.

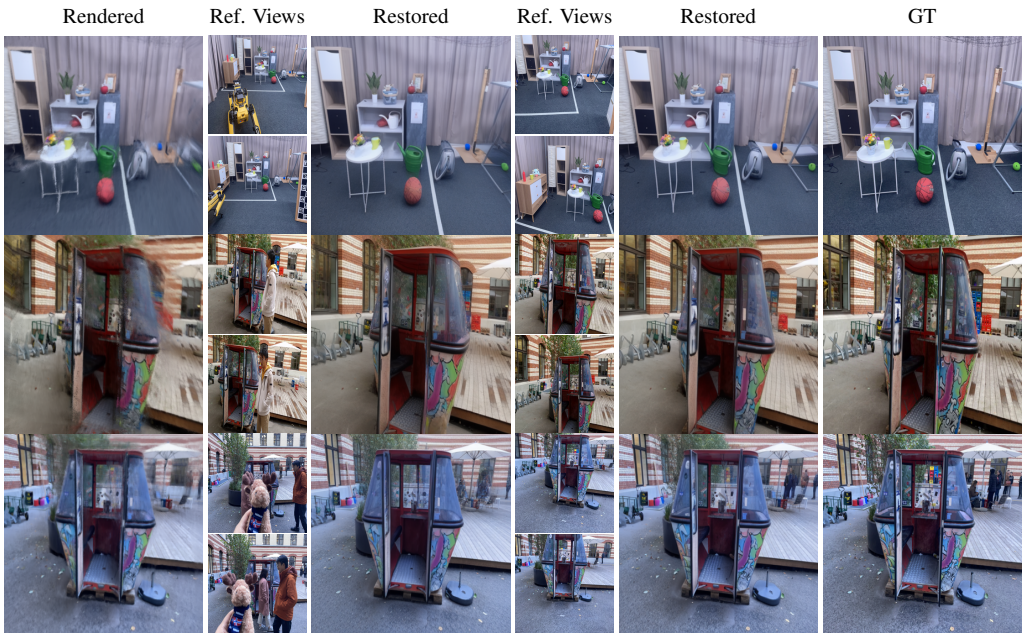

Figure 9: **Robustness of Multi-view Condition Injection to distractor-corrupted reference views.**

A.3 Robustness Evaluation of Reference Views

As shown in Figure 9, at inference, after replacing the default clean reference views with distractor-corrupted views, the model still maintains stable restoration performance. This indicates that Multi-

view Conditional Injection can effectively aggregate complementary information via cross-view feature interactions and adaptively extracts reliable cues to guide structural restoration, further validating the reliability of UG-Diff in practical applications.

A.4 Evaluation of Applicability

We replace the neural rendering backbone MU-GeNeRF Mu et al. (2026) with WildGaussians Kulhanek et al. (2024), and use its rendered results and uncertainty to train UG-Diff. As shown in Figure 10, UG-Diff still effectively enhances rendering quality and significantly improves structural defects caused by sparse views and distractor occlusions. This demonstrates that the training paradigm of UG-Diff is not tied to a specific neural rendering method and exhibits excellent applicability.



Figure 10: **Qualitative results of UG-Diff applied to WildGaussians.**

A.5 Limitations

UG-Diff still has the following limitations: **1)** Under large-scale rendering collapse, limited structural information may cause the model to over-rely on the generative prior, resulting in restorations that deviate from the real scene, as illustrated in Figure 11.



Figure 11: **Failure cases under large-scale rendering collapse.**

2) When the viewpoint discrepancy between the rendered and reference views is too large or their covisibility is insufficient, the model struggles to acquire sufficient cross-view complementary cues, thereby limiting its restoration performance. Figure 12 illustrates this sensitivity by comparing randomly chosen references with selected references.

A.6 Analysis of Gram Loss

The Gram loss matches global feature statistics by constraining the Gram matrix differences between generated and ground-truth images, and is commonly used to enhance texture details and perceptual

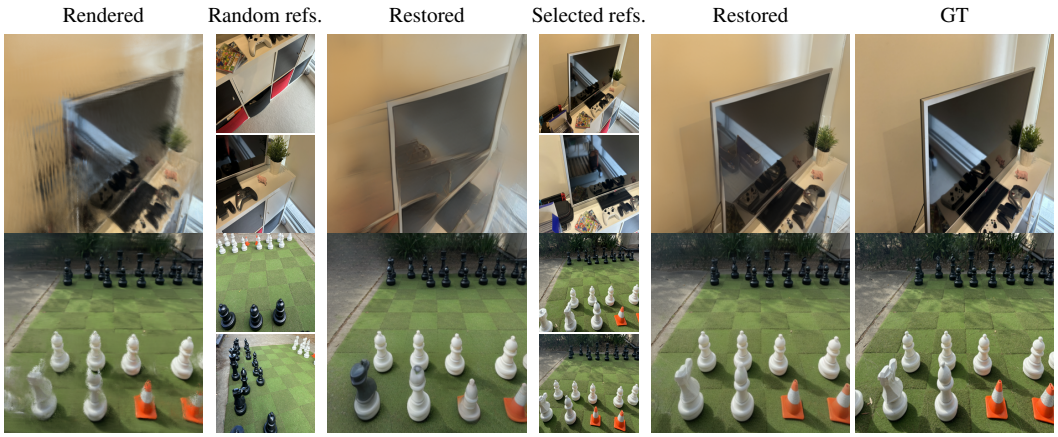

Figure 12: **Sensitivity to reference view selection.**

quality Wu et al. (2025a); Cao et al. (2026). However, this loss inherently lacks explicit spatial structural constraints. Under distractor-corrupted supervision, the model may also fit the statistical characteristics of distractors, making it prone to generating contents structurally inconsistent with the real scene, thereby impairing stable structural restoration, as shown in Figure 13.

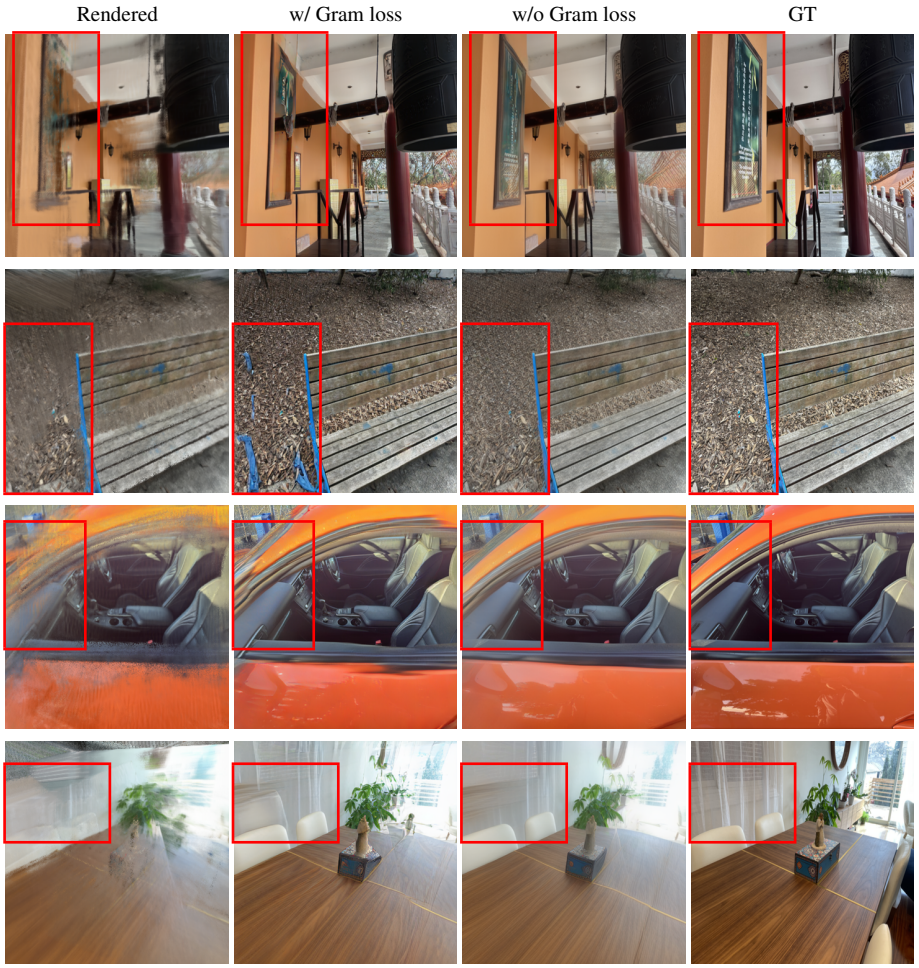

Figure 13: **Effect of Gram loss under distractor-corrupted supervision.**