RETHINKING GEOMETRIC PRIORS IN NOVEL VIEW SYNTHESIS THROUGH MULTI-AXIS RELIABILITY FIELDS

Anonymous authors
Paper under double-blind review

ABSTRACT

Geometric priors are widely used in novel view synthesis (NVS) to incorporate structural guidance. However, recent results show that this conditioning often provides marginal gains under strong backbones. Does this mean that geometric priors are no longer important once the model can already infer cross-view correspondence on its own? Our controlled corruption experiments show otherwise: geometry remains useful, but its value is under-utilized when reliable and unreliable guidance are not distinguished. Motivated by this finding, we introduce Reliability-Calibrated Geometric Attention (RCGA), which models geometric reliability as a multi-axis field to control how explicit geometric evidence modifies the backbone's intrinsic correspondence. The field identifies where and under which model states geometry contributes complementary information. Specifically, the partition-preserving product-of-experts uses the predicted field to calibrate geometric evidence in joint attention for selective correspondence refinement. A comparative oracle supervises the field by measuring the incremental utility of geometry beyond the backbone during training. Across RealEstate10K and ScanNet, RCGA consistently improves epipolar, depth-projection, and coordinate priors over a strong diffusion backbone, achieving state-of-the-art performance while suppressing unreliable geometric guidance.

1 INTRODUCTION

Novel view synthesis (NVS) aims to generate an unobserved view of a scene from one or more source observations and a target camera transformation. In single-image NVS, the model must reconcile two competing requirements: preserving source-consistent geometry in regions supported by the observation, while plausibly synthesizing content that becomes newly visible from the target viewpoint Wiles et al. (2020). Geometric priors provide a natural inductive bias for the former, whereas increasingly powerful generative models offer strong priors for the latter. Recent progress, however, has increasingly relied on expressive generative backbones to learn cross-view correspondence implicitly. Geometry-free methods Rombach et al. (2021); Yu et al. (2023); Zhou et al. (2025) demonstrate competitive NVS without explicit geometric priors. VIVID Elata et al. (2025) further reports only marginal gains from additional encodings beyond camera-pose conditioning under a strong diffusion backbone.

These observations expose a gap between the structural information in geometric priors and the actual gains realized by NVS models. Existing geometry-guided NVS approaches incorporate structure through epipolar constraints Kant et al. (2024), depth-based projection Li et al. (2027), or warped features Jiang et al. (2024), but typically use these cues without explicitly modeling how their reliability varies across spatial queries and model states. This is problematic because geometric correspondence is inherently non-uniform Seo et al. (2024). Some regions and feature scales provide accurate source-target cues, whereas depth errors, occlusions, projection ambiguity, and limited visibility can make the same guidance misleading Liu et al. (2026). Our controlled corruption analysis (Fig. 6) explains why geometric priors yield only limited gains: when reliable and unreliable correspondences are integrated uniformly, the benefits of valid geometry are offset by the adverse effects of misleading guidance. This suggests that the key issue is whether the model can exploit them according to their reliability.

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.

Consequently, this work explicitly models the reliability of geometric guidance when integrating it into generative NVS models. Rather than remaining fixed, the trust placed in geometry should adapt to the model's evolving correspondence estimates. Crucially, reliability is assessed relative to the backbone's own correspondence, so geometry is trusted only when it provides complementary evidence beyond what the model already infers. The goal is therefore to rethink how geometric priors are used in generation, conditioning their influence on when, where, and under which model states they provide genuinely informative guidance.

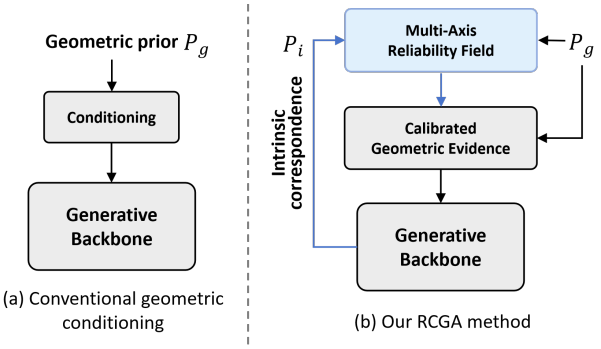

Figure 1: Instead of directly injecting geometric priors as conditioning signals, RCGA uses a multi-axis reliability field to adapt their influence to the backbone's intrinsic correspondence.

To realize this principle, we introduce Reliability-Calibrated Geometric Attention (RCGA). At its core, RCGA learns an explicit multi-axis reliability field that dynamically calibrates the integration of geometric priors with the backbone's implicit correspondence (Fig. 1). Instead of predicting a single geometry confidence, the reliability field varies across layer, attention head, spatial query, and diffusion noise level. Its value is inferred from three complementary sources of evidence, including the certainty of the backbone correspondence, the support of the geometric prior, and their mutual agreement. The resulting reliability controls a partition-preserving product-of-experts update, allowing geometry to redistribute source correspondence without perturbing the backbone's learned source-versus-target attention allocation. A comparative oracle further supervises whether geometry provides better correspondence than the backbone itself during training.

Experimental results support our view at both the mechanism and performance levels. Controlled corruption shows that reliable geometry improves correspondence, while unreliable geometry introduces opposing errors and weakens the overall benefit of explicit guidance. RCGA separates these effects through reliability calibration, consistently improving epipolar, depth-projection, and coordinate priors and achieving state-of-the-art performance on both RealEstate10K Zhou et al. (2018) and ScanNet Dai et al. (2017). These results highlight that effectively exploiting explicit geometry depends critically on distinguishing when and where it provides reliable evidence.

Our contributions are summarized as follows:

- The lack of relative geometric reliability modeling is identified as a key reason for the limited gains of explicit geometry in generative NVS, and is validated through controlled corruption experiments.
- A unified RCGA is introduced to model geometric reliability as a multi-axis field that controls how explicit geometric evidence modifies the backbone's intrinsic source correspondence, with comparative supervision learning its incremental utility relative to the backbone.
- Achieves state-of-the-art performance on RealEstate10K and ScanNet, with consistent gains across epipolar, depth-projection, and coordinate priors.

2 Related Work

Diffusion models operate in either pixel or latent space Ho et al. (2020); Rombach et al. (2022); Karras et al. (2022), with latent diffusion improving efficiency at the cost of an autoencoding bottleneck Ho et al. (2022); Karras et al. (2024). For NVS, VIVID Elata et al. (2025) shows that this compression can discard source details important for cross-view correspondence, motivating our use of a pixel-space backbone. Single-image NVS has historically coupled geometry and generation in several different ways. Geometry-first pipelines estimate scene structure, transform observed content to the target camera, and synthesize newly exposed regions Wiles et al. (2020); Shih et al. (2020);

Liu et al. (2021); Tucker & Snavely (2020); Rockwell et al. (2021). Other approaches construct explicit or learned 3D scene representations, including radiance fields and geometry-aware neural representations, to support novel-view rendering Yu et al. (2021); Johari et al. (2022).

In contrast, end-to-end generative approaches increasingly rely on learned image priors to model correspondence and hallucinate unobserved content. Geometry-Free View Synthesis Rombach et al. (2021) and SRT Sajjadi et al. (2022) demonstrate that expressive attention-based models can infer substantial cross-view structure without hand-designed 3D correspondence modules. Diffusion models further strengthen this paradigm: pose-guided diffusion Tseng et al. (2023), Photometric-NVS Yu et al. (2023), Zero-1-to-3 Liu et al. (2023), ZeroNVS Sargent et al. (2024), and EscherNet Kong et al. (2024) synthesize novel views through camera-conditioned generative models, while MultiDiff Müller et al. (2024), SyncDreamer Liu et al. (2024), CAT3D Gao et al. (2024), and SEVA Zhou et al. (2025) extend generative view synthesis toward jointly consistent or flexible multi-view generation. More recently, video-diffusion priors have also been adapted to scene-level view synthesis in ViewCrafter Yu et al. (2025) and NVS-Solver You et al. (2024).

Explicit geometry nevertheless remains valuable in modern generative NVS. Existing methods incorporate epipolar constraints Tseng et al. (2023); Kant et al. (2024), monocular depth and warped image or coordinate cues Seo et al. (2024); Müller et al. (2024); Park et al. (2025), or explicit 3D-aware representations Chan et al. (2023); Gu et al. (2023); Ye et al. (2024); Wu et al. (2024) to impose geometric structure on generation. More recent work points in a complementary direction: LagerNVS Szymanowicz et al. (2026) shows that strong 3D inductive bias can substantially benefit fully neural NVS, while MoAI Kwak et al. (2026) improves image-geometry alignment through cross-modal attention coupling. Recent methods such as GenWarp Seo et al. (2024), NVS-Solver You et al. (2024), and WAVE Park et al. (2025) soften rigid geometric conditioning through adaptive integration, whereas RCGA learns its utility relative to the backbone's evolving correspondence.

3 METHODOLOGY

3.1 OVERVIEW

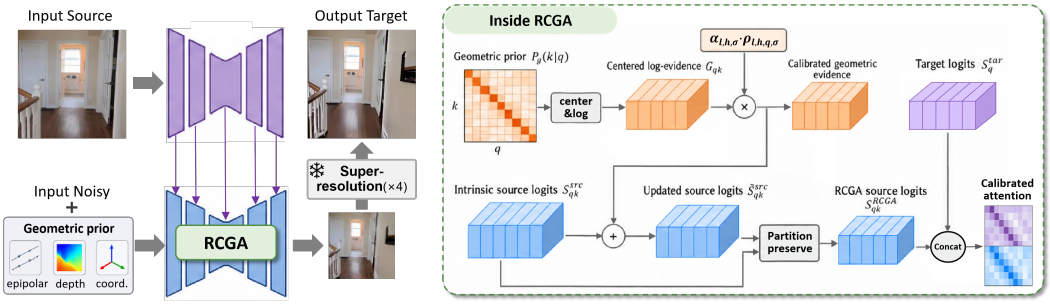

Figure 2: Overall framework. The proposed framework adopts a cascaded pixel-space diffusion model. RCGA dynamically calibrates explicit geometric priors against the backbone's intrinsic correspondence during target-view denoising.

RCGA is centered on a multi-axis reliability field that determines when and where explicit geometry should influence the backbone's intrinsic correspondence. Each geometric cue is first represented as a conditional source-correspondence prior $P_g(k \mid q)$ (Appendix A.1), while the generative backbone provides its intrinsic correspondence $P_i(k \mid q)$. The reliability field $\rho_{l,h,q,\sigma}$ is inferred from backbone certainty, geometric support, and backbone-geometry agreement, and measures the utility of geometry relative to the backbone's evolving correspondence (Sec. 3.2). RCGA then uses this field to modulate a partition-preserving product-of-experts update of the source correspondence, selectively refining where source information is retrieved without altering the backbone's source-versus-target attention allocation (Sec. 3.3). During training, a comparative oracle supervises the field by evaluating whether geometric correspondence provides more useful evidence than the intrinsic alternative (Sec. 3.4).

3.2 Multi-Axis Reliability Field

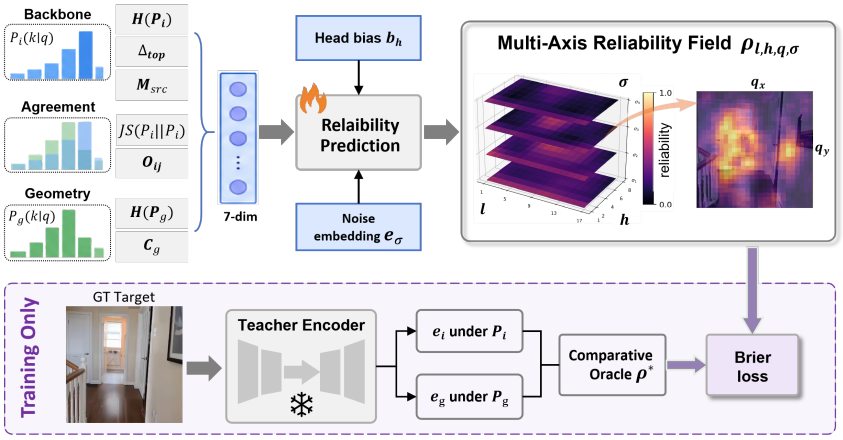

Figure 3: Multi-axis reliability field and its training supervision.

Let S^{src}_{qk} denote the original attention logit between a target query q and a source key k. We define the backbone's intrinsic source correspondence as

$$P_i^{l,h}(k \mid q, \sigma) = \text{Softmax}_{k \in \Omega_s^l}\left(S^{\text{src}}_{qk}\right), \tag{1}$$

where l, h, and σ denote the U-Net layer, attention head, and diffusion noise level. Since P_i is normalized only over source keys, it describes where the backbone retrieves source information independently of the total source-conditioning strength.

Geometric utility is inherently state-dependent: the same prior may be informative at one spatial location but misleading at another, and its usefulness can further vary across feature scales, attention heads, and denoising stages. RCGA therefore models reliability as a multi-axis relative field

$$\rho_{l,h,q,\sigma} \in [0, 1], \tag{2}$$

rather than a global confidence scalar. For fixed (l, h, σ), $\rho_{l,h,:,\sigma}$ forms a spatial reliability map over target queries; together, these maps define the reliability field illustrated in Fig. 3.

The field is inferred from three complementary sources of evidence: backbone certainty, geometric support, and backbone-geometry agreement. They are summarized by

$$\mathbf{z}_{l,h,q,\sigma} = [H(P_i),\ H(P_g),\ \text{JS}(P_i \parallel P_g),\ \mathcal{O}_{ig},\ \Delta_{\text{top}},\ \mathcal{M}_{\text{src}},\ \mathcal{C}_g], \tag{3}$$

where $H(P_i)$, Δ_{top}, and $\mathcal{M}_{\text{src}}$ characterize the certainty and source reliance of the backbone; $H(P_g)$ and $\mathcal{C}_g$ describe the concentration and support of the geometric prior; and $\text{JS}(P_i \parallel P_g)$ together with $\mathcal{O}_{ig}$ quantify their agreement. This design estimates the incremental value of geometry relative to the backbone, rather than geometry confidence in isolation. The exact definitions of these statistics are given in Appendix A.2.

The detached descriptor is mapped to reliability by a lightweight two-layer MLP f_l with hidden dimension 32, augmented with head-specific parameters and diffusion-noise conditioning:

$$\rho_{l,h,q,\sigma} = V_g^l(q) \cdot \text{sigmoid}\left(f_l(\text{sg}[\mathbf{z}_{l,h,q,\sigma}]) + b^\rho_{l,h} + r^\rho_{l,h}(\mathbf{e}(\sigma))\right), \tag{4}$$

where $V_g^l(q)$ indicates whether valid geometric support exists, $\text{sg}[\cdot]$ denotes stop-gradient, and $\mathbf{e}(\sigma)$ is the diffusion-noise embedding. Query-dependent statistics, head-specific parameters, block-specific f_l, and noise conditioning provide spatial, head, layer, and denoising adaptivity.

3.3 Partition-Preserving Product-of-Experts Attention

Given the predicted reliability field, RCGA integrates geometric evidence through a product-of-experts (PoE) update. We first center the geometric log-prior against the uniform correspondence

$$G_{qk} = \log P_g(k \mid q) - \log P_u(k \mid q), \qquad P_u(k \mid q) = \frac{1}{|\Omega_s^l|}. \tag{5}$$

For numerical stability, P_g is clamped to a minimum probability ϵ before evaluating the centered log-prior. A noise-aware gain

$$\alpha_{l,h,\sigma} = g\big(a_{l,h} + r^{\alpha}_{l,h}(\mathbf{e}(\sigma))\big) \tag{6}$$

controls how strongly each head responds to trusted geometry: ρ determines where geometry should be trusted, while α controls its overall strength, $g(\cdot)$ denotes the softplus function.

RCGA then updates only the source-attention logits:

$$\widetilde{S}^{\mathrm{src}}_{qk} = S^{\mathrm{src}}_{qk} + \alpha_{l,h,\sigma}\rho_{l,h,q,\sigma}G_{qk}, \tag{7}$$

which corresponds to the PoE form

$$\widetilde{P}(k \mid q) \propto P_i(k \mid q)\, P_g(k \mid q)^{\alpha_{l,h,\sigma}\rho_{l,h,q,\sigma}}. \tag{8}$$

Thus, geometry selectively sharpens or redirects intrinsic correspondence, while $\rho \to 0$ recovers the backbone behavior.

Direct logit modification would also change the competition between source and target keys. To isolate geometric guidance to *where* source information is retrieved, we preserve the original source log-partition:

$$S^{\mathrm{RCGA}}_{qk} = \widetilde{S}^{\mathrm{src}}_{qk} - \mathrm{LSE}_{k'}(\widetilde{S}^{\mathrm{src}}_{qk'}) + \mathrm{LSE}_{k'}(S^{\mathrm{src}}_{qk'}). \tag{9}$$

Consequently,

$$\sum_k e^{S^{\mathrm{RCGA}}_{qk}} = \sum_k e^{S^{\mathrm{src}}_{qk}}, \qquad \rho_{l,h,q,\sigma} = 0 \Rightarrow S^{\mathrm{RCGA}}_{qk} = S^{\mathrm{src}}_{qk}. \tag{10}$$

The final attention is therefore computed as

$$\mathrm{Softmax}\big([S^{\mathrm{tar}}_q; S^{\mathrm{RCGA}}_q]\big), \tag{11}$$

allowing RCGA to redistribute source correspondence without perturbing the source-versus-target attention allocation learned by the backbone.

3.4 Comparative Oracle Supervision

During training only, as shown in Fig. 3, the same frozen source encoder is additionally applied to the clean target x_t under no-gradient evaluation to obtain teacher features $F^l_t(q)$ matched to the source features $F^l_s(k)$. For each attention head, we compute the expected feature-matching errors induced by the implicit and explicit correspondence distributions:

$$e^{l,h}_i(q,\sigma) = \sum_k P^{l,h}_i(k \mid q,\sigma) \cdot d\left(\mathbf{F}^l_s(k), \mathbf{F}^l_t(q)\right), \tag{12}$$

$$e^{l,h}_g(q) = \sum_k P^{l,h}_g(k \mid q) \cdot d\left(\mathbf{F}^l_s(k), \mathbf{F}^l_t(q)\right), \tag{13}$$

where $d(\cdot,\cdot)$ is the squared Euclidean distance between ℓ_2-normalized features, equivalently twice the cosine distance. Expected matching error is used instead of a top-1 correspondence to reduce sensitivity to small attention shifts and multi-modal matches.

The resulting soft oracle is

$$\rho^*_{l,h,q,\sigma} = V^l_g(q) \cdot \mathrm{sigmoid}\left(\frac{e^{l,h}_i(q,\sigma) - e^{l,h}_g(q)}{\tau_o}\right). \tag{14}$$

where τ_o is a temperature hyperparameter set to 0.25 in our experiments. When explicit geometry explains the clean target feature better than the backbone correspondence, $e_g < e_i$ and the oracle assigns high reliability; when the intrinsic correspondence is superior, the target approaches zero. All quantities used to construct the oracle are detached from gradient computation.

We supervise the reliability prediction using the Brier loss

$$\mathcal{L}_{\mathrm{calib}} = \frac{1}{|\mathcal{I}|}\sum_{(l,h,q)\in\mathcal{I}} \big(\rho_{l,h,q,\sigma} - \rho^*_{l,h,q,\sigma}\big)^2, \tag{15}$$

where $\mathcal{I}$ denotes the set of supervised RCGA predictions. The complete training objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{diff}} + \lambda_{\mathrm{calib}}\mathcal{L}_{\mathrm{calib}}, \tag{16}$$

where $\mathcal{L}_{\mathrm{diff}}$ is the original diffusion objective. The teacher and clean target image are required only for constructing reliability supervision during training and are removed entirely at inference.

4 EXPERIMENTS

First, Sec. 4.1 benchmarks our method and examines whether it consistently improves different explicit geometric priors. Next, Sec. 4.2 investigates why explicit geometry was underutilized in previous methods and how our approach benefits view synthesis. Finally, Sec. 4.3 conducts ablation studies on the core components of RCGA and analyzes their underlying working mechanisms, with further reliability analysis in Appendix A.4. Metric computation follows VIVID Elata et al. (2025) evaluation protocol. Full experimental setup details are provided in Appendix A.3.

4.1 MAIN NVS RESULTS

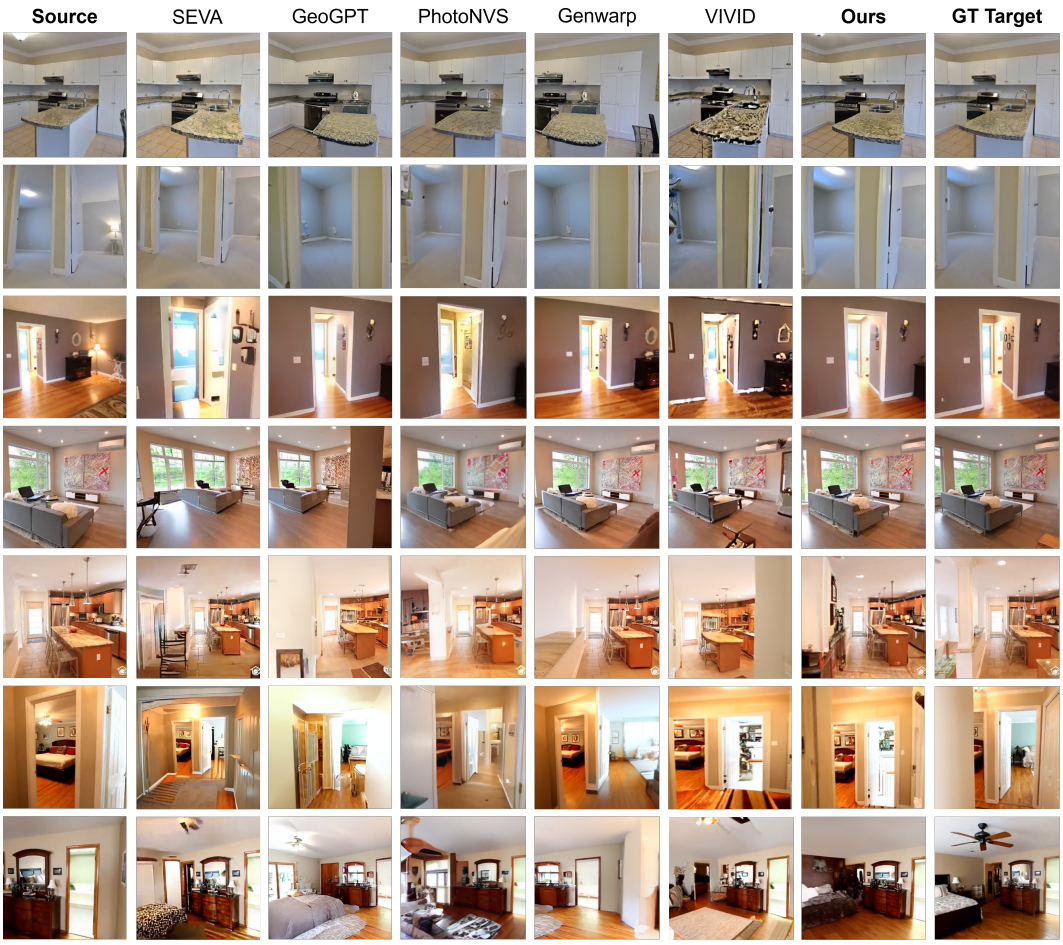

Figure 4: Qualitative comparison with state-of-the-art methods on RealEstate10K. From the top row to the bottom row, the examples contain progressively less source-visible information.

We compare RCGA with representative single-image NVS methods on RealEstate10K Zhou et al. (2018) and ScanNet Dai et al. (2017). As shown in Table 1, RCGA achieves the best performance on both datasets. On RealEstate10K, its advantage further increases from mid-range to long-range camera changes, where implicit cross-view matching becomes more challenging and reliable geometry can provide stronger complementary evidence. This trend is also reflected qualitatively in Fig. 4. Additional qualitative comparisons on ScanNet are provided in Appendix Fig. 9, showing consistent improvements on a second dataset.

Table 2 presents the effects of identical geometric information under different injection methods, thereby isolating the contribution of reliability calibration. Conventional injection mechanisms yield only marginal gains beyond pose conditioning. In contrast, RCGA substantially improves all three

Table 1: Quantitative comparison with state-of-the-art methods under mid-range and long-range settings at 256×256.

Method	RealEstate10K				ScanNet	
	Mid-range		Long-range		Mid-range	
	PSNR ↑	FID ↓	PSNR ↑	FID ↓	PSNR ↑	FID ↓
SEVA Zhou et al. (2025)	15.04	6.73	13.36	7.31	14.08	9.48
GeoGPT Rombach et al. (2021)	14.06	6.43	13.13	7.22	13.03	9.19
Photometric-NVS Yu et al. (2023)	13.32	7.12	12.05	9.22	12.36	9.91
GenWarp Seo et al. (2024)	13.43	5.91	12.10	7.38	12.51	8.74
VIVID Elata et al. (2025)	17.36	2.89	15.21	3.89	16.72	3.14
Ours (RCGA-Depth)	**18.34**	**2.66**	**16.96**	**3.35**	**17.84**	**2.78**

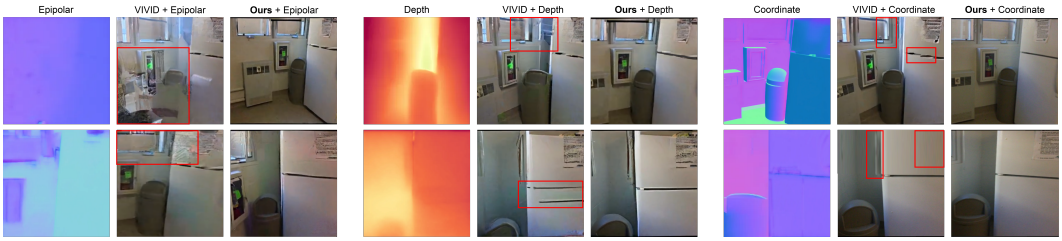

Figure 5: Structural detail comparison under the same geometric guidance.

priors, achieving PSNR gains of $+1.18$ dB, $+1.23$ dB, and $+0.92$ dB for epipolar, depth-projection, and coordinate priors, respectively. This consistent improvement demonstrates that RCGA effectively unleashes the utility of explicit geometry. The qualitative detail comparisons in Fig. 5 further show that, under the same depth prior, RCGA better preserves geometry-sensitive structures—such as sharp boundaries, thin objects, and planar intersections—while reducing double edges, bending, and local misalignment.

4.2 When Are Geometric Priors Reliable

This section investigates why explicit geometry yields only limited gains under strong generative backbones. We conduct a controlled corruption analysis on ScanNet using ground-truth depth and camera poses to construct a valid correspondence reference, and progressively perturb a subset of geometric correspondences while keeping the model checkpoint fixed. Examples of clean and corrupted queries are provided in Appendix Fig. 10.

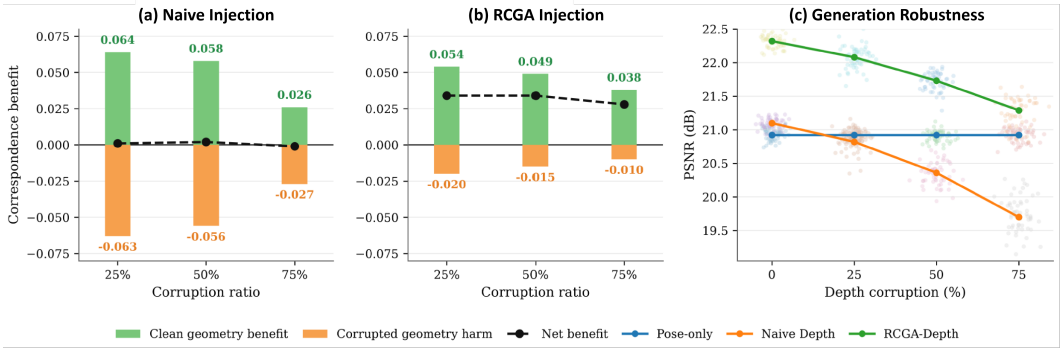

Figure 6: Effect of reliable and corrupted geometry under controlled correspondence perturbations on ScanNet.

Table 2: Comparison between conventional injection and RCGA under identical geometric priors at 64×64 resolution.

	Naive injection		RCGA	
Geometry prior	**PSNR ↑**	**FID ↓**	**PSNR ↑**	**FID ↓**
Pose_only	21.11 ± 0.04	3.00 ± 0.02	21.10 ± 0.04	3.02 ± 0.02
+ Epipolar	21.15 ± 0.04	2.87 ± 0.04	22.33 ± 0.03	2.58 ± 0.03
+ Depth	21.29 ± 0.04	2.99 ± 0.01	22.52 ± 0.02	2.65 ± 0.02
+ Coordinate	21.06 ± 0.04	3.08 ± 0.04	21.98 ± 0.02	2.82 ± 0.02

Table 3: Correspondence benefit and geometry influence under 50% depth-prior corruption.

	Naive projection		RCGA projection	
	Clean	**Corrupted**	**Clean**	**Corrupted**
Correspondence Benefit	$+0.0581$	-0.0565	$+0.0491$	-0.0147
Geometry Influence	0.1897	0.2911	0.1611	0.0892

Fig. 6(a)(b) supports a clear cancellation effect. Under naive geometric injection, reliable correspondences provide positive benefit, whereas corrupted correspondences introduce negative effects that increasingly offset these gains as the corruption ratio grows. Moreover, the benefit of clean geometry itself decreases, indicating that uniform injection cannot isolate useful geometric evidence from misleading cues. RCGA separates these effects through reliability calibration: harmful correspondences are strongly attenuated while useful geometry retains a positive contribution. This selective use of geometry also improves generation robustness under increasing corruption (Fig. 6(c)).

We further analyze the 50% depth-prior corruption setting at the correspondence level. Fig. 7 shows that RCGA assigns higher reliability to useful correspondences while reducing geometric influence where the prior is corrupted. Table 3 quantifies this behavior using Correspondence Benefit, which measures the improvement in ground-truth source correspondence relative to the intrinsic backbone attention, and Geometry Influence, which measures the magnitude of the geometry-induced attention change. Their formal definitions are provided in Appendix A.3 (Eqs. 38–39). Under naive projection, corrupted geometry produces a larger attention perturbation than clean geometry (0.2911 vs. 0.1897), while reversing the correspondence benefit from $+0.0581$ to -0.0565. RCGA suppresses the corrupted influence to 0.0892 while retaining 0.1611 on clean geometry, reducing the harmful effect to -0.0147 while preserving a positive $+0.0491$ benefit on reliable correspondence. These results provide correspondence-level evidence that RCGA selectively exploits geometric priors according to their reliability rather than uniformly weakening their influence.

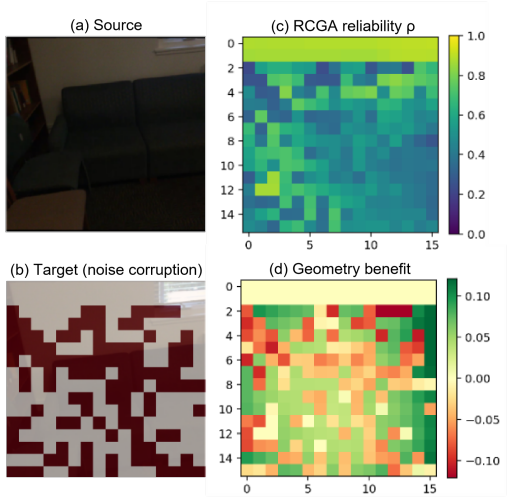

Figure 7: Visualization of reliability-calibrated correspondence under 50% depth-prior corruption.

4.3 Ablation and Analysis

All component ablations use the same pose-conditioned backbone and depth-projection prior, with the Brier score evaluating comparative reliability prediction. Results are summarized in Table 4, with additional ablations and reliability analysis provided in Appendix A.4. A fixed geometric weight yields only a limited gain, while progressively finer reliability modeling improves both gen-

Table 4: Component ablation of RCGA using the depth-projection prior in 64×64 on RealEstate10K

Variant	Query-adaptive	Head/Layer	Noise-aware	Partition preserve	PSNR ↑	FID ↓	Brier ↓
Pose baseline	×	×	×	×	21.10	3.02	–
Fixed geometry	×	×	×	✓	21.29	2.99	0.3889
Static calibration	×	✓	×	✓	21.58	2.90	0.2088
Non-spatial dynamic	×	✓	✓	✓	21.69	2.82	0.2053
No explicit noise	✓	✓	×	✓	22.35	2.78	0.1074
No partition	✓	✓	✓	×	20.74	3.04	0.2182
Full RCGA	✓	✓	✓	✓	**22.52**	**2.65**	**0.1055**

eration and calibration. Query-adaptive reliability provides the main improvement, with explicit noise conditioning bringing a smaller but consistent gain. Removing partition preservation substantially degrades performance, confirming the importance of refining source correspondence without altering the backbone's source-versus-target attention allocation. As shown in Fig. 8, the full RCGA further exhibits this dynamic behavior across denoising stages, concentrating reliability on geometrically transferable regions while suppressing ambiguous or occluded areas.

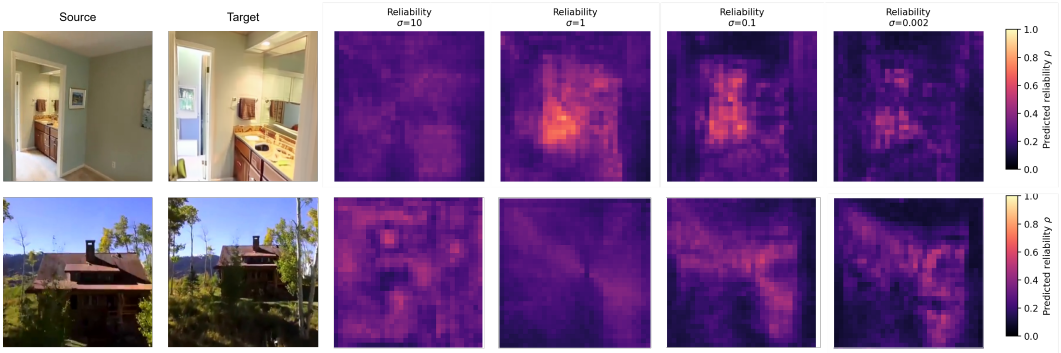

Figure 8: Spatial evolution of predicted reliability at 32×32 feature resolution across denoising stages.

5 Conclusion

Explicit geometry remains useful even when a strong generative backbone already forms meaningful cross-view correspondence, but its contribution depends on whether it adds information that the backbone does not already capture. Our results show that treating this usefulness as a relative, state-dependent quantity allows reliable geometry to refine correspondence while preventing misleading cues from dominating the generative process. This shifts the role of geometric priors from fixed conditioning signals to selectively used evidence that complements the model's intrinsic correspondence.

Several scope boundaries and technical limitations warrant further discussion. First, the frozen cascaded super-resolution stage can partially weaken the geometric fidelity established by the base generator. In addition, because our experiments are conducted within a cascaded pixel-space diffusion backbone, the applicability of this calibration behavior to latent or video-diffusion architectures has yet to be fully explored. Extending reliability calibration to learned high-resolution feature spaces and broader generative backbones therefore presents a promising direction for future work.

AI USE STATEMENT

Generative AI tools were utilized in this work solely for improving language readability and manuscript formatting. Generative AI tools were not used for primary idea generation, theoretical derivations, or critical experimental analyses, and other tasks requiring mandatory disclosure are not applicable to this work. All AI-assisted outputs were thoroughly reviewed and verified by the authors to ensure technical accuracy and consistency in tone. The authors take full responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide the information required to reproduce our method and experiments throughout the paper and appendix. The RCGA formulation, including the multi-axis reliability field, partition-preserving geometric update, and comparative supervision, is described in Secs. 3.2–3.4, with the geometric priors and reliability statistics detailed in Appendix A.1 and Appendix A.2. Experimental settings, datasets, evaluation resolutions, and controlled corruption protocol are documented in Appendix A.3. Additional reliability ablations are provided in Appendix A.4. All experiments are built on publicly available datasets and the official evaluation pipeline. Code and evaluation scripts will be released upon publication.

REFERENCES

Eric R. Chan, Koki Nagano, Matthew A. Chan, Alexander W. Bergman, Jeong Joon Park, Axel Levy, Miika Aittala, Shalini De Mello, Tero Karras, and Gordon Wetzstein. Generative novel view synthesis with 3d-aware diffusion models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4194–4206, 2023. doi: 10.1109/ICCV51070.2023.00389.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3d reconstructions of indoor scenes. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2432–2443, 2017. doi: 10.1109/CVPR.2017.261.

Noam Elata, Bahjat Kawar, Yaron Ostrovsky-Berman, Miriam Farber, and Ron Sokolovsky. Novel view synthesis with pixel-space diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26756–26766, 2025. doi: 10.1109/CVPR52734.2025.02492.

Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul Srinivasan, Jonathan T Barron, and Ben Poole. Cat3d: Create anything in 3d with multi-view diffusion models. *arXiv preprint arXiv:2405.10314*, 2024.

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 *International Conference on Machine Learning*, pp. 11808–11826. PMLR, 2023.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research*, 23(47):1–33, 2022.

Lei Jiang, Gerald Schaefer, and Qinggang Meng. Multi-scale feature fusion for single image novel view synthesis. *Neurocomputing*, 599:128081, 2024. ISSN 0925-2312. doi: 10.1016/j.neucom.2024.128081.

Mohammad Mahdi Johari, Yann Lepoittevin, and François Fleuret. GeoNeRF: Generalizing nerf with geometry priors. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18344–18347, 2022. doi: 10.1109/CVPR52688.2022.01782.

Yash Kant, Aliaksandr Siarohin, Ziyi Wu, Michael Vasilkovsky, Guocheng Qian, Jian Ren, Riza Alp Guler, Bernard Ghanem, Sergey Tulyakov, and Igor Gilitschenski. SPAD: Spatially aware multi-view diffusers. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10026–10038, 2024. doi: 10.1109/CVPR52733.2024.00956.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

Tero Karras, Miika Aittala, Jaakko Lehtinen, Janne Hellsten, Timo Aila, and Samuli Laine. Analyzing and improving the training dynamics of diffusion models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24174–24184, 2024. doi: 10.1109/CVPR52733.2024.02282.

Xin Kong, Shikun Liu, Xiaoyang Lyu, Marwan Taher, Xiaojuan Qi, and Andrew J. Davison. Eschernet: A generative model for scalable view synthesis. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9503–9513, 2024. doi: 10.1109/CVPR52733.2024.00908.

Min-Seop Kwak, Junho Kim, Sangdoo Yun, Dongyoon Han, Taekyung Kim, Seungryong Kim, and Jin-Hwa Kim. Aligned novel view image and geometry synthesis via cross-modal attention instillation. In *International Conference on Learning Representations*, volume 2026, pp. 16029–16050, 2026.

Weijia Li, Jun He, Junyan Ye, Huaping Zhong, Zhimeng Zheng, Zilong Huang, Dahua Lin, and Conghui He. CrossViewDiff: A cross-view diffusion model for satellite-to-street view synthesis. *Pattern Recognition*, 183:114735, 2027. ISSN 0031-3203. doi: 10.1016/j.patcog.2026.114735.

Andrew Liu, Ameesh Makadia, Richard Tucker, Noah Snavely, Varun Jampani, and Angjoo Kanazawa. Infinite nature: Perpetual view generation of natural scenes from a single image. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14438–14447, 2021. doi: 10.1109/ICCV48922.2021.01419.

Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9264–9275, 2023. doi: 10.1109/ICCV51070.2023.00853.

Yu Liu, Yingchao Feng, Yidan Zhang, et al. DIV: A visible-infrared object detection dataset for weakly aligned UAV scenarios. *Journal of Image and Graphics*, 2026. URL `https://link.cnki.net/urlid/11.3758.TB.20260806.1344.012`. Online first on CNKI.

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.

Norman Müller, Katja Schwarz, Barbara Rössle, Lorenzo Porzi, Samuel Rota Bulò, Matthias Nießner, and Peter Kontschieder. MultiDiff: Consistent novel view synthesis from a single image. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10258–10268, 2024. doi: 10.1109/CVPR52733.2024.00977.

Jiwoo Park, Tae Eun Choi, Youngjun Jun, and Seong Jae Hwang. WAVE: Warp-based view guidance for consistent novel view synthesis using a single image. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11906–11915, 2025. doi: 10.1109/ICCV51701.2025.01107.

Chris Rockwell, David F. Fouhey, and Justin Johnson. PixelSynth: Generating a 3d-consistent experience from a single image. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14084–14093, 2021. doi: 10.1109/ICCV48922.2021.01384.

Robin Rombach, Patrick Esser, and Björn Ommer. Geometry-free view synthesis: Transformers and no 3d priors. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14336–14346, 2021. doi: 10.1109/ICCV48922.2021.01409.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Mehdi S.M. Sajjadi, Henning Meyer, Etienne Pot, Urs Bergmann, Klaus Greff, Noha Radwan, Suhani Vora, Mario Lučić, Daniel Duckworth, Alexey Dosovitskiy, Jakob Uszkoreit, Thomas Funkhouser, and Andrea Tagliasacchi. Scene representation transformer: Geometry-free novel view synthesis through set-latent scene representations. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6219–6228, 2022. doi: 10.1109/CVPR52688.2022.00613.

Kyle Sargent, Zizhang Li, Tanmay Shah, Charles Herrmann, Hong-Xing Yu, Yunzhi Zhang, Eric Ryan Chan, Dmitry Lagun, Li Fei-Fei, Deqing Sun, and Jiajun Wu. ZeroNVS: Zero-shot 360-degree view synthesis from a single image. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9420–9429, 2024. doi: 10.1109/CVPR52733.2024.00900.

Junyoung Seo, Kazumi Fukuda, Takashi Shibuya, Takuya Narihira, Naoki Murata, Shoukang Hu, Chieh-Hsin Lai, Seungryong Kim, and Yuki Mitsufuji. GenWarp: Single image to novel views with semantic-preserving generative warping. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 80220–80243. Curran Associates, Inc., 2024. doi: 10.52202/079017-2550.

Meng-Li Shih, Shih-Yang Su, Johannes Kopf, and Jia-Bin Huang. 3D photography using context-aware layered depth inpainting. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8025–8035, 2020. doi: 10.1109/CVPR42600.2020.00805.

Stanislaw Szymanowicz, Minghao Chen, Jianyuan Wang, Christian Rupprecht, and Andrea Vedaldi. LagerNVS: Latent geometry for fully neural real-time novel view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15443–15453, 2026.

Hung-Yu Tseng, Qinbo Li, Changil Kim, Suhib Alsisan, Jia-Bin Huang, and Johannes Kopf. Consistent view synthesis with pose-guided diffusion models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16773–16783, 2023. doi: 10.1109/CVPR52729.2023.01609.

Richard Tucker and Noah Snavely. Single-view view synthesis with multiplane images. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 548–557, 2020. doi: 10.1109/CVPR42600.2020.00063.

Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. SynSin: End-to-end view synthesis from a single image. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7465–7475, 2020. doi: 10.1109/CVPR42600.2020.00749.

Rundi Wu, Ben Mildenhall, Philipp Henzler, Keunhong Park, Ruiqi Gao, Daniel Watson, Pratul P. Srinivasan, Dor Verbin, Jonathan T. Barron, Ben Poole, and Aleksander Hołyński. ReconFusion: 3d reconstruction with diffusion priors. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21551–21561, 2024. doi: 10.1109/CVPR52733.2024.02036.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10371–10381, 2024a. doi: 10.1109/CVPR52733.2024.00987.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in neural information processing systems*, 37:21875–21911, 2024b.

Jianglong Ye, Peng Wang, Kejie Li, Yichun Shi, and Heng Wang. Consistent-1-to-3: Consistent image to 3d view synthesis via geometry-aware diffusion models. In *2024 International Conference on 3D Vision (3DV)*, pp. 664–674, 2024. doi: 10.1109/3DV62453.2024.00027.

Meng You, Zhiyu Zhu, Hui Liu, and Junhui Hou. Nvs-solver: Video diffusion model as zero-shot novel view synthesizer. *arXiv preprint arXiv:2405.15364*, 2024.

Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural radiance fields from one or few images. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4576–4585, 2021. doi: 10.1109/CVPR46437.2021.00455.

Jason J. Yu, Fereshteh Forghani, Konstantinos G. Derpanis, and Marcus A. Brubaker. Long-term photometric consistent novel view synthesis with diffusion models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7071–7081, 2023. doi: 10.1109/ICCV51070.2023.00653.

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.

Jensen Zhou, Hang Gao, Vikram Voleti, Aaryaman Vasishta, Chun-Han Yao, Mark Boss, Philip Torr, Christian Rupprecht, and Varun Jampani. Stable virtual camera: Generative view synthesis with diffusion models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12405–12414, 2025. doi: 10.1109/ICCV51701.2025.01153.

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: learning view synthesis using multiplane images. *ACM Trans. Graph.*, 37(4), July 2018. ISSN 0730-0301. doi: 10.1145/3197517.3201323.

A Appendix



Figure 9: Additional qualitative comparison with state-of-the-art methods on ScanNet.

A.1 Geometry as a Conditional Correspondence Prior

RCGA represents different geometric cues through a common conditional correspondence prior over source keys,

$$P_g^{l,h}(k \mid q), \qquad \sum_{k \in \Omega_s^l} P_g^{l,h}(k \mid q) = 1, \tag{17}$$

together with a validity map $V_g^l(q,k) \in \{0,1\}$ indicating whether source key k provides valid geometric support for target query q. The corresponding query-level validity is

$$V_g^l(q) = \mathbf{1}\left[\sum_{k \in \Omega_s^l} V_g^l(q,k) > 0\right]. \tag{18}$$

The prior specifies where a target query should retrieve source information under a given geometric model. Each provider returns $\log P_g(k \mid q)$ in the same layout as the source-attention logits, allowing reliability prediction and correspondence refinement to remain independent of the particular geometric representation. We consider epipolar, depth-projection, and coordinate priors.

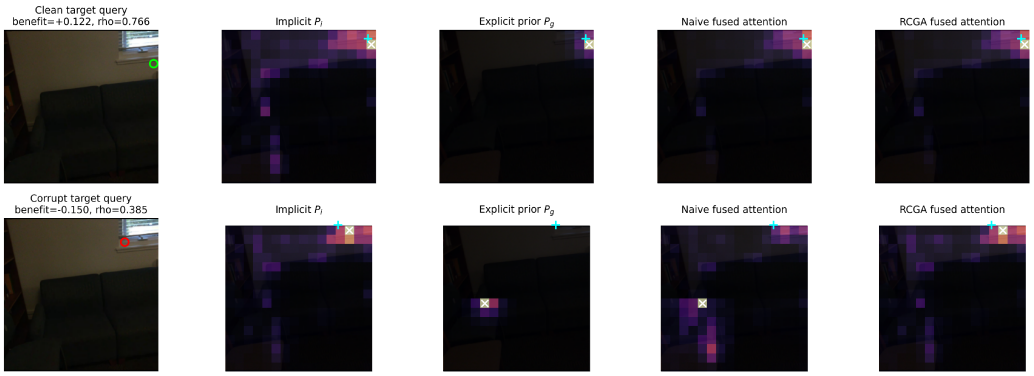

Figure 10: Illustration of clean and corrupted geometric correspondence for the same target query used in the controlled corruption analysis.

Epipolar Prior. Let $D_{\text{epi}}(q,k)$ denote the distance from source location k to the epipolar line induced by target query q. We define

$$P_g^{\text{epi},l,h}(k \mid q) \propto V_g^l(q,k) \exp\left(-\frac{D_{\text{epi}}(q,k)^2}{2(\sigma_{l,h}^{\text{epi}})^2}\right), \tag{19}$$

where $\sigma_{l,h}^{\text{epi}}$ is a learnable per-head bandwidth. The epipolar prior does not require monocular depth, but remains ambiguous along the epipolar line.

Depth-Projection Prior. Given predicted source depth $D_s(k)$, each source location is back-projected to 3D, transformed into the target camera, and reprojected:

$$\mathbf{X}_s(k) = D_s(k)\mathbf{K}_s^{-1}\bar{\mathbf{k}}, \qquad \mathbf{X}_t(k) = \mathbf{R}_{s\to t}\mathbf{X}_s(k) + \mathbf{t}_{s\to t}, \tag{20}$$

$$\hat{q}(k) = \pi\left(\mathbf{K}_t\mathbf{X}_t(k)\right), \tag{21}$$

where $\bar{\mathbf{k}}$ is the homogeneous source coordinate and π denotes perspective division. Invalid projections are removed, and an approximate z-buffer resolves competing projections at the current attention resolution. The correspondence prior is then

$$P_g^{\text{depth},l,h}(k \mid q) \propto V_g^l(q,k) \exp\left(-\frac{\|q-\hat{q}(k)\|_2^2}{2(\sigma_{l,h}^{\text{depth}})^2}\right). \tag{22}$$

Queries without valid geometric support have $V_g^l(q) = 0$, preventing unsupported geometry from affecting the reliability field.

Coordinate Prior. The coordinate prior uses the projected correspondence but evaluates similarity in a multi-frequency coordinate embedding space. Let $\gamma(\cdot)$ denote the Fourier coordinate encoding and d_{coord} the corresponding feature-space distance. We define

$$P_g^{\text{coord},l,h}(k \mid q) \propto V_g^l(q,k) \exp\left(-\frac{d_{\text{coord}}\big(\gamma(q),\gamma(\hat{q}(k))\big)}{2(\sigma_{l,h}^{\text{coord}})^2}\right). \tag{23}$$

Thus, the coordinate prior changes how projected correspondence is measured rather than introducing a separate coordinate feature branch.

The three variants differ in the construction of P_g and its valid support, while the reliability field and subsequent RCGA update remain unchanged.

For verification, we additionally use the uniform prior

$$P_u(k \mid q) = \frac{1}{|\Omega_s^l|}. \tag{24}$$

Its centered log-prior is identically zero, providing an exact identity case for the RCGA update.

A.2 Reliability Statistics and Field Prediction

In this section, we detail the seven statistics used to construct the multi-axis reliability field. Let $N_l = |\Omega_s^l|$ denote the number of source tokens at layer l. A reliability descriptor is formed for each RCGA block, attention head, target query, and diffusion state. Statistics involving the intrinsic correspondence are recomputed as denoising evolves, while geometry-only quantities remain fixed for a given geometric prior. All statistics are detached before field prediction.

Backbone correspondence entropy. The intrinsic source correspondence is obtained by normalizing the unmodified source-attention logits:

$$P_i^{l,h}(k \mid q, \sigma) = \mathrm{Softmax}_{k \in \Omega_s^l}\left(S_{qk}^{\mathrm{src}}\right). \tag{25}$$

We define its normalized entropy as

$$H(P_i) = -\frac{1}{\log N_l} \sum_{k \in \Omega_s^l} P_i(k \mid q, \sigma) \log\left(P_i(k \mid q, \sigma) + \epsilon\right). \tag{26}$$

Low entropy indicates concentrated intrinsic correspondence, whereas high entropy indicates greater ambiguity.

Geometric-prior entropy. The geometric correspondence is characterized similarly:

$$H(P_g) = -\frac{1}{\log N_l} \sum_{k \in \Omega_s^l} P_g(k \mid q) \log\left(P_g(k \mid q) + \epsilon\right). \tag{27}$$

This quantity measures the concentration of the geometric prior.

Backbone-geometry disagreement. We use the normalized Jensen-Shannon divergence

$$\mathrm{JS}(P_i \parallel P_g) = \frac{1}{2\log 2}\left[D_{\mathrm{KL}}(P_i \parallel M) + D_{\mathrm{KL}}(P_g \parallel M)\right], \qquad M = \frac{1}{2}(P_i + P_g), \tag{28}$$

which lies in $[0, 1]$.

Geometry-support mass. We define the support of the geometric prior as

$$\mathcal{S}_g(q) = \left\{k \in \Omega_s^l : P_g(k \mid q) > \frac{1}{N_l}\right\}. \tag{29}$$

The intrinsic probability mass assigned to this support is

$$\mathcal{O}_{ig} = \sum_{k \in \Omega_s^l} P_i(k \mid q, \sigma)\mathbf{1}[k \in \mathcal{S}_g(q)]. \tag{30}$$

Unlike the symmetric JS divergence, $\mathcal{O}_{ig}$ measures how much of the backbone correspondence already lies in geometry-supported locations.

Top correspondence margin. Let $p_{(1)}$ and $p_{(2)}$ denote the largest and second-largest probabilities in P_i. We define

$$\Delta_{\mathrm{top}} = p_{(1)} - p_{(2)}. \tag{31}$$

A larger margin indicates a more decisive intrinsic correspondence.

Source-attention mass. While P_i describes relative correspondence within the source branch, we also retain the total source mass from the original joint attention:

$$\mathcal{M}_{\mathrm{src}}^{l,h}(q, \sigma) = \sum_{k \in \Omega_s^l} \mathrm{Softmax}\left([S_q^{\mathrm{tar}}; S_q^{\mathrm{src}}]\right)_k. \tag{32}$$

This quantity measures how strongly the query relies on source information, independently of where that information is retrieved.

Geometric coverage. Let $V_g^l(q,k) \in \{0,1\}$ indicate whether source key k provides valid geometric support for query q. We define

$$\mathcal{C}_g = \frac{1}{N_l} \sum_{k \in \Omega_s^l} V_g^l(q,k). \tag{33}$$

The corresponding query-level validity gate is

$$V_g^l(q) = \mathbf{1}\left[\sum_{k \in \Omega_s^l} V_g^l(q,k) > 0\right]. \tag{34}$$

Thus, $\mathcal{C}_g$ measures the extent of geometric support, whereas $V_g^l(q)$ determines whether geometric calibration is enabled for the query.

Field prediction. The seven statistics form the descriptor

$$\mathbf{z}_{l,h,q,\sigma} = [H(P_i),\, H(P_g),\, \mathrm{JS}(P_i \parallel P_g),\, \mathcal{O}_{ig},\, \Delta_{\mathrm{top}},\, \mathcal{M}_{\mathrm{src}},\, \mathcal{C}_g]\,. \tag{35}$$

It is mapped to reliability by a lightweight block-specific function f_l, implemented as a $7 \to 32 \to 1$ MLP with a SiLU hidden activation and shared across heads within each block. The final reliability is

$$\rho_{l,h,q,\sigma} = V_g^l(q) \cdot \mathrm{sigmoid}\left(f_l(\mathrm{sg}[\mathbf{z}_{l,h,q,\sigma}]) + b_{l,h}^{\rho} + r_{l,h}^{\rho}(\mathbf{e}(\sigma))\right), \tag{36}$$

where $\mathrm{sg}[\cdot]$ denotes stop-gradient, $b_{l,h}^{\rho}$ is a learned head-specific bias, and $r_{l,h}^{\rho}(\mathbf{e}(\sigma))$ provides the head-specific response to the diffusion-noise embedding.

Query-dependent statistics provide spatial adaptation, f_l captures layer-wise variation, the head-specific terms capture attention-head variation, and $\mathbf{e}(\sigma)$ accounts for the denoising state. Together, these factors define the multi-axis reliability field $\rho_{l,h,q,\sigma}$.

A.3 Experimental Details

All experiments are built on the official 64×64 pose-conditioned diffusion model, with the backbone frozen and only RCGA parameters optimized using Adam with a learning rate of 10^{-3} and a global batch size of 32. RCGA-Depth uses source depth predicted by Depth Anything V2 Yang et al. (2024a;b). At inference, we follow the EDM sampling and classifier-free guidance settings of VIVID Elata et al. (2025), followed by its frozen super-resolution model when 256×256 outputs are required.

Main evaluation protocol. For the main comparison, all methods are evaluated on the same fixed source–target pairs using the same metric implementation. On RealEstate10K, we follow the VIVID protocol with 10K pairs, using frame gaps of 30–60 for mid-range and 60–120 for long-range evaluation. On ScanNet, we construct a fixed mid-range test set and reuse the same pairs for all methods. Baselines use their official implementations and publicly available checkpoints. All outputs are resized to 256×256 before metric computation, whereas RCGA-specific prior analysis and component ablations are evaluated directly on the 64×64 base predictions.

Controlled corruption protocol and metrics. For the ScanNet controlled analysis, ground-truth source/target depth and calibrated camera poses are used to construct visibility-consistent correspondences. Each valid query is assigned a deterministic 2D perturbation using a fixed seed, with horizontal and vertical offsets independently taking a signed magnitude of 25%, 37.5%, or 50% of the source-token grid width. Selected queries undergo a circular shift of the entire source-key distribution, preserving its shape and entropy. Queries are ranked by a fixed random score, and the lowest $\lfloor r N_{\mathrm{valid}} \rfloor$ are perturbed for $r \in \{0, 0.25, 0.5, 0.75\}$, yielding nested corruption subsets with fixed per-query offsets across ratios. Backbone, Naive, and RCGA use the same corrupted priors under otherwise identical settings, while Oracle replaces corrupted-query priors with a uniform distribution as a corruption-mask suppression upper bound.

The correspondence-level quantities reported in Table 3 are defined using an independent ground-truth source correspondence $k^*(q)$ for each visible target query. For a source-attention distribution

Table 5: Relative reliability versus geometry-only confidence. All variants use the same depth-projection prior and partition-preserving PoE update.

Variant	Backbone	Geometry	Agreement	PSNR ↑	FID ↓	Brier ↓
Geometry-only confidence	–	✓	–	21.91	2.85	0.176
Backbone-only confidence	✓	–	–	22.03	2.81	0.151
Backbone + geometry	✓	✓	–	22.21	2.75	0.132
Full relative reliability	✓	✓	✓	**22.52**	**2.65**	**0.1055**

$P_m^{l,h}(k \mid q, \sigma)$, its normalized expected correspondence error is

$$E_m^{l,h}(q,\sigma) = \sum_{k \in \Omega_s^l} P_m^{l,h}(k \mid q,\sigma) \frac{\|\mathbf{x}(k) - k^*(q)\|_2}{\sqrt{(H_s^l)^2 + (W_s^l)^2}}, \tag{37}$$

where $\mathbf{x}(k)$ denotes the two-dimensional source-token coordinate. Correspondence Benefit is then

$$B_m^{l,h}(q,\sigma) = E_i^{l,h}(q,\sigma) - E_m^{l,h}(q,\sigma), \qquad m \in \{\text{Naive}, \text{RCGA}\}, \tag{38}$$

so that positive values indicate improved correspondence relative to the intrinsic backbone attention.

Geometry Influence measures the magnitude of the geometry-induced change in source correspondence using total variation distance:

$$U_m^{l,h}(q,\sigma) = \frac{1}{2} \sum_{k \in \Omega_s^l} \left| P_m^{l,h}(k \mid q,\sigma) - P_i^{l,h}(k \mid q,\sigma) \right|. \tag{39}$$

Both quantities are computed per attention head and averaged over heads and the five noise probes $\sigma \in \{80, 10, 1, 0.1, 0.002\}$ at `enc.32x32_block1`, followed by pooled averaging over the corresponding visible clean or corrupted queries. These diagnostics use ScanNet ground-truth correspondence and are independent of the teacher-feature discrepancy used by the training-time comparative oracle.

A.4 Additional Ablations and Reliability Analysis

Unless otherwise specified, all experiments use the 64×64 diffusion model with the depth-projection prior and follow the same protocol as Sec. 4.3. We report PSNR, FID, and the Brier score against the comparative reliability target ρ^*.

Relative Reliability versus Geometry-Only Confidence. A central hypothesis of RCGA is that geometric reliability should not be estimated in isolation. The same prior may be useful when the backbone correspondence is uncertain, but provide little additional value when the backbone already forms a confident match. This comparison constructs the reliability field from progressively richer evidence. The geometry-only variant uses $H(P_g)$ and $\mathcal{C}_g$, while retaining the same validity gate V_g. The backbone-only variant uses $H(P_i)$, Δ_{top}, and $\mathcal{M}_{\text{src}}$. Geometric information is then added, followed by the explicit backbone–geometry agreement terms $\text{JS}(P_i \parallel P_g)$ and $\mathcal{O}_{ig}$.

Table 5 shows that geometry-only confidence already improves over fixed geometric injection, but remains clearly below the full relative field. Incorporating the backbone state further improves both generation and calibration, while explicit backbone-geometry agreement provides an additional gain. These results show that the three sources of evidence play complementary roles and, more importantly, support our central hypothesis that geometric utility depends on what the backbone has already inferred.

Individual Reliability Statistics. Finally, each statistic is removed individually from the full seven-dimensional reliability descriptor while keeping the field prediction function unchanged.

As shown in Table 6, no single statistic dominates the improvement. Removing either agreement statistic causes a relatively large degradation, supporting the use of complementary backbone certainty, geometric support, and backbone–geometry agreement cues.

Table 6: Leave-one-out ablation of the seven reliability statistics.

Removed Statistic	PSNR ↑	FID ↓	Brier ↓
$H(P_i)$	22.40	2.69	0.113
$H(P_g)$	22.39	2.70	0.115
$\mathrm{JS}(P_i \parallel P_g)$	22.32	2.73	0.124
$\mathcal{O}_{ig}$	22.35	2.72	0.120
Δ_{top}	22.44	2.68	0.111
$\mathcal{M}_{\mathrm{src}}$	22.37	2.71	0.121
$\mathcal{C}_g$	22.38	2.71	0.118
None (Full RCGA)	**22.52**	**2.65**	**0.1055**

Comparative Oracle Supervision and Sensitivity. Removing $\mathcal{L}_{\mathrm{calib}}$ and training with the diffusion objective alone reduces PSNR from 22.52 to 21.87 dB and increases FID from 2.65 to 2.87, while the Brier score degrades from 0.1055 to 0.241. We further examine whether the gain depends on the specific oracle configuration.

Table 7: Sensitivity to the comparative-oracle configuration.

Oracle Configuration	PSNR ↑	FID ↓
Cosine distance, $\tau_o = 0.125$	22.46	2.68
Squared distance, $\tau_o = 0.15$	22.47	2.67
Squared distance, $\tau_o = 0.40$	22.45	2.68
Default, $\tau_o = 0.25$	**22.52**	**2.65**

The default oracle uses squared Euclidean distance between normalized teacher and source features, $c(q,k) = [2 - 2\langle \hat{f}_t(q), \hat{f}_s(k) \rangle]_+$, with $\tau_o = 0.25$. To test whether the gain depends on this particular configuration, several nearby alternatives are evaluated in Table 7. There are only small variations across reasonable discrepancy metrics and temperatures, indicating that the improvement is not sensitive to the exact oracle configuration.