MultiDiff: Consistent Novel View Synthesis from a Single Image - Supplementary Document -

Norman Müller[1] Katja Schwarz[1] Barbara Rössle[2] Lorenzo Porzi[1] Samuel Rota Bulò[1]
Matthias Nießner[2] Peter Kontschieder[1]

Meta Reality Labs Zurich[1] Technical University of Munich[2]

Appendix

In this appendix, we provide additional qualitative and quantitative results and discuss training and evaluation details.

A. Additional qualitative results

We provide additional qualitative comparisons with the baselines on RealEstate10K [13] in Fig. 2 as well as on ScanNet [5] in Fig. 4. While PhotoNVS [12] accumulates errors over the autoregressive sampling process, our model synthesizes realistic images for all target poses jointly. In comparison to DFM [9], our approach leverages strong image- and video-priors to achieve noticeably higher image fidelity.
Furthermore, we demonstrate the stochasticity of our approach in Fig. 3 where using the same reference image and target poses, our probabilistic method synthesizes multiple plausible novel views.
We also present an in-the-wild and 360° trajectory in Fig. 1.

Figure 1. Out-of-distribution examples: In-the-wild image with unknown camera parameters and 360° camera rotation.

B. Implementation details

B.1. Baselines

DFM We use the official implementation of the authors (https://github.com/ayushtewari/DFM.git). On Realestate10K, we evaluate the provided pre-trained checkpoint. On ScanNet, we train a model from scratch, following the official instructions for RealEstate10k. More specifically, we first train the model at resolution 64×64 for 75K iterations with a total batch size of 16 on 8 NVIDIA A100-SXM4-80GB GPUs. Next, we fine-tune the model at resolution 128×128 for 60K iterations using a total batch size of 8. At both resolutions, larger batch sizes did not fit in the 80GB memory of the GPUs.

PhotoNVS We use the official implementation of the authors https://github.com/YorkUCVIL/Photoconsistent-NVS.git and the provided checkpoint on Realestate10K. On ScanNet, we use the pre-trained VQGAN provided by the authors and train the model at 256×256 resolution for 500K iterations using 8 NVIDIA A100-SXM4-80B GPUs with an effective batch size of 64.

MVDiffusion Since MVDiffusion is a purely text-conditional model, we adapt the official implementation (https://github.com/Tangshitao/MVDiffusion.git) to accept a reference image at inference time. We encode the reference image into latent space and then encode it into the diffusion model's Gaussian prior space using DDIM inversion. During sampling, the encoded reference image is added to the batch. Since MVDiffusion uses attention layers that operate on all images in the batch jointly, the reference frame affects the sampling for all images. However, during sampling the score estimate is calculated using the full batch, while for DDIM inversion we can only obtain the score estimate for the reference image. In practice, sampling does hence not reproduce the reference image faithfully. We address this issue by additionally optimizing the reference latent after each denoising step to match the reference image. For the optimization at each sampling step, we use Adam [6] with a learning rate of 0.1 and train with an L2-Loss and a perceptual loss for 10 iterations.

Text2Room We use the official implementation of the authors (https://github.com/lukasHoel/

Figure 2. Additional qualitative comparison results on RealEstate10K [13].

Figure 3. Different samples generated by our probabilistic approach using the same reference image and target trajectory.

text2room) which also supports image-conditional generation. We follow the original setup and use IronDepth [1] for depth prediction and StableDiffusion2 inpainting (https://huggingface.co/stabilityai/stable-diffusion-2-inpainting) for image inpainting. Since Text2Room formulates the problem as pure depth-to-image/inpainting task, the same pretrained checkpoints can be used for both datasets, RealEstate10K and ScanNet, and no additional training is required.

B.2. MultiDiff

Training details For the encoding of the warped reference images, we use the encoder layers of the pre-trained text-to-image model Stable Diffusion 1.5 [7] and use the provided VQ-VAE for latent encoding and decoding. We initialize the denoising layers of our U-Net model with the pre-trained weights of VideoCrafter [4], a latent video diffusion model trained on large scale video data [2]. The temporal attention layers serve as strong prior for consistency - see performance of "MultiDiff no vid." in Table 3 of the main paper and Fig. 9 for a qualitative comparison. Nevertheless, we fine-tune all layers of the U-Net for the novel view synthesis task to enable the attention layers to learn correspondences between multiple views. For training, we use Adam [6] with a learning rate of 1e-05 and batch size of 6 with 16 target views per batch at a resolution of 256×256. Using 8 NVIDIA A100-SXM4-80B GPUs with an effective batch size of 48, we train for 300K iterations. We use DDPM [8] noise scheduling using $t = 1000$ time steps for denoising and perform evaluation using DDIM[8] sampling with 35 steps.

For noise warping, we found that using nearest-neighbor with a receptive field size of 4px at 256px resolution gave the best results. This limited receptive range ensures that the noise distribution remains roughly normal, preventing strong zooms from resulting in a few pixels covering large image portions.

Inference details Using the estimated depth maps with nearest-neighbor interpolation, we calculated the average

Figure 4. Additional qualitative comparison results on ScanNet [5]. Note that MVDiffusion requires the scene mesh for inference.

warping overlap of the initial image with the last frame in the sequence: 20.4% (24.7%) on RealEstate10K (ScanNet). The described refinement is applied on poses where the warping overlap is below 20%, which occurs in 51.8% (47.1%) of cases.

C. Evaluation

Data processing On RealEstate10K, we randomly select 1K sequences with at least 200 frames. For evaluation, for each sequence, we choose a random starting frame at least 200 frames ahead of the last frame. We select 16 frames for evaluation that we uniformly distribute within the interval of 200 views from the starting frame. Following previous evaluation protocols, for short-term evaluation, we set the 5th view to be 50 frames after the starting view in the original video. For long-term evaluation, the last view corresponds to the 200th frame after the starting frame.
On ScanNet, for each of the 100 test scenes, we sample 10 starting views ensuring at least 100 frames offset from the last frame in the recordings, resulting in a set of 1K test sequences. Since the camera movement in ScanNet recordings is considerably higher and the frame rate noticeably slower compared to RealEstate10K, we consider sequence lengths reduced by 50% in the original video. Therefore, we consider the 25th frame for short-term and the 100th view for long-term evaluation relative to the starting frame.

TSED To compute TSED scores [12], we use the official implementation from (https://github.com/YorkUCVIL/Photoconsistent-NVS) and provide additional quantitative results in Fig. 6 and Fig. 5. While DFM achieves the highest consistency scores due to its PixelNeRF [11] formulation, it suffers from noticeably worse image generation quality compared to MultiDiff (see Table 1 and 2 of the main paper as well as Fig. 2 and Fig. 4). As DFM does not support higher image resolutions, we measure TSED at 128×128 resolution.

Inference speed We report inference performances of PhotoNVS, DFM, and MultiDiff in Tab. 1. As our approach infers multiple frames in parallel and uses an efficient attention architecture, we observe noticeably shorter inference times while achieving higher image fidelity and consistency than the baselines. We note that our approach also scales to larger resolutions as the underlying latent video prior can easily be tuned for image sizes. This is in stark contrast to baselines like PhotoNVS, DFM for which the computational costs quickly become too high and require infeasible amounts of memory when trained on larger resolutions.

128px	s/frame	256px	s/frame
PhotoNVS	45.6	PhotoNVS	183
DFM	17.4	DFM	-
MultiDiff	1.02	MultiDiff	1.94

Table 1. Comparison of the inference speed evaluated in seconds per frame using FP32 on an NVIDIA A100-SXM4-80GB. By jointly inferring multiple frames in parallel and using efficient attention architecture, we achieve noticeably shorter inference times.

Fitting a NeRF Results for fitting a NeRF with Nerfacto are shown in Fig. 7, yielding small pixel-level inconsistencies with floating artifacts. As in *ReconFusion* [10], we use distillation to obtain a cleaner representation (second row in Fig. 7).

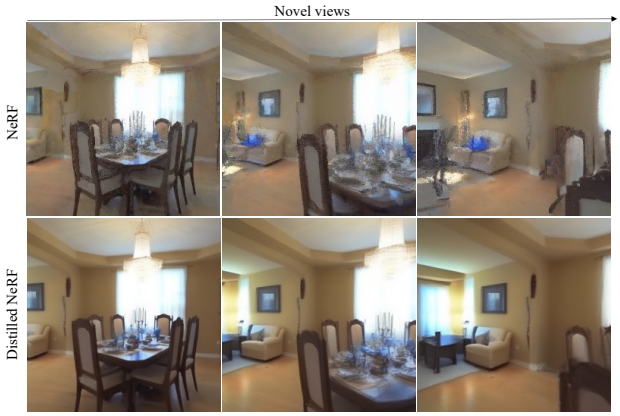

Figure 7. NeRF fitting from 16 synthesized views. Row 1: Nerfacto, row 2: Nerfacto + Distillation

D. Ablations

We show additional qualitative results ablating our design decisions in Fig. 9. We note that training our model from scratch ("MultiDiff no prior") leads to over-smoothed results that do not closely follow the target trajectory. Furthermore, we showcase the effect of using the image prior but not initializing the weights of our correspondence attention layer with the weights of the pre-trained video prior (see "MultiDiff no vid."): The results are overall less consistent as e.g. the floor changes from carpet to wood. Our method uses depth-based image warps to reproject the reference image to the target poses, providing strong cues about the target views. We ablate the importance of this in Table 3 of the main table (see "MultiDiff no warp") and show an example of in Fig. 9. Without using the warps of the reference image, our model is not able to faithfully follow the target trajectory. As under strong camera motion, there is little to no overlap with the reference image, we also learn an embedding of the target pose and show the effect of removing this information ("MultiDiff no pose") in Fig. 9. Using

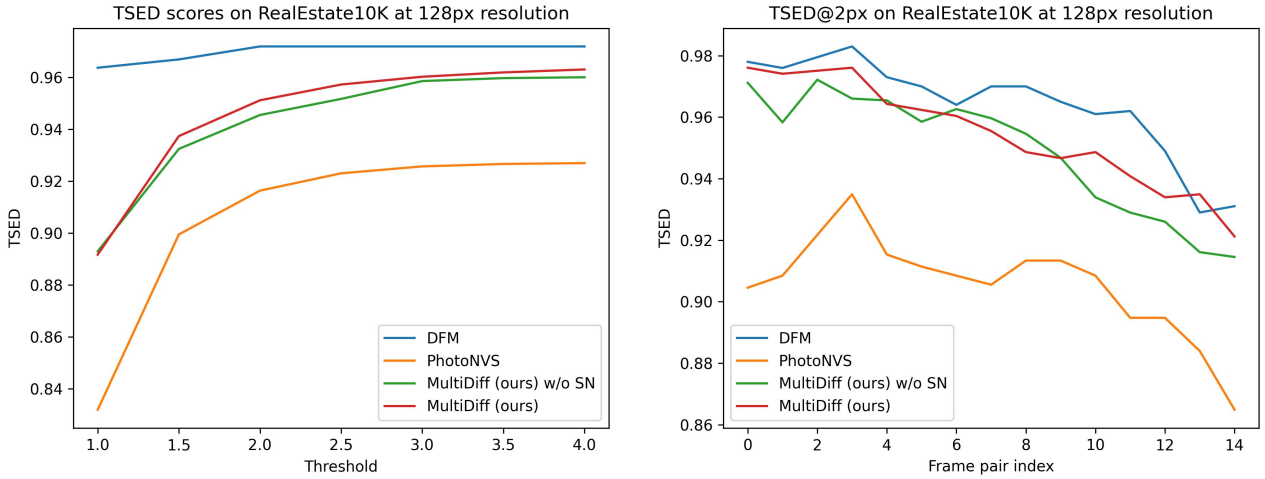

Figure 5. TSED scores evaluated on RealEstate10K at a resolution of 128×128. The left chart shows the TSED evaluated at different thresholds, on the right we plot the TSED scores over the pairs of frame indices along the trajectory.

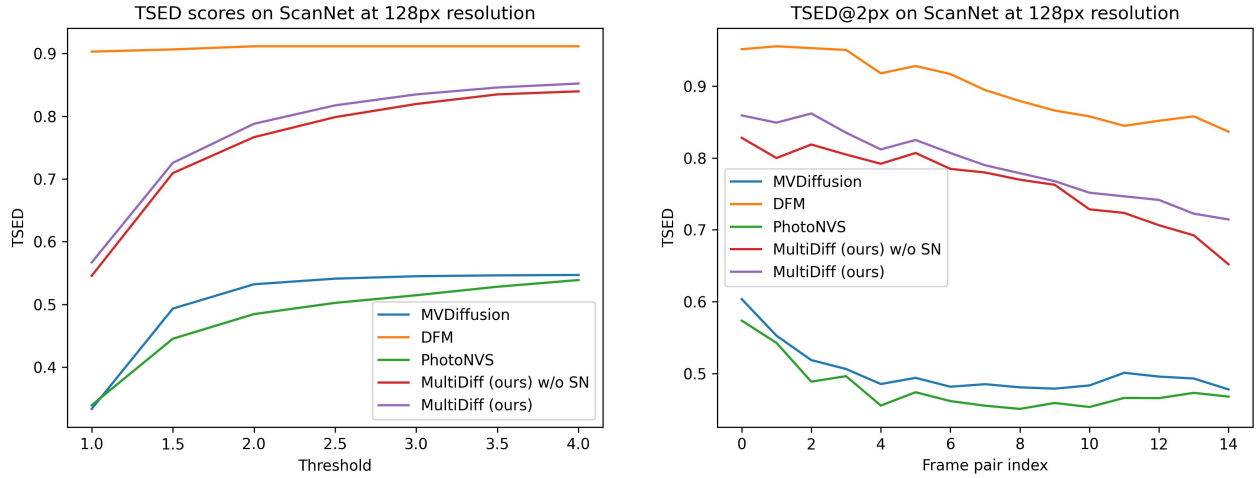

Figure 6. TSED scores evaluated on ScanNet at a resolution of 128×128. On the left, we show the TSED evaluated at different thresholds. The right chart plots the TSED scores over the pairs of frame indices along the trajectory.

Datset	Method	Short-term				Long-term			
		PSNR ↑	LPIPS ↓	FID ↓	KID ↓	FID ↓	KID ↓	FVD ↓	mTSED ↑
RE10K	MultiDiff + IronDepth	15.49	0.402	28.36	0.005	34.91	0.006	115.72	0.797
	MultiDiff + ZoeDepth	**15.65**	**0.393**	**25.90**	**0.004**	**30.15**	**0.006**	**105.9**	**0.855**
ScanNet	MultiDiff + IronDepth	**15.05**	0.435	44.15	0.010	**46.87**	0.013	118.3	0.503
	MultiDiff + ZoeDepth	15.00	**0.431**	**43.84**	**0.010**	47.11	**0.013**	**114.9**	**0.576**

Table 2. Qualitative comparison of using IronDepth trained on ScanNetv2 as alternative depth estimator evaluated 256×256 resolution. We notice that the results are comparable for the short-term metrics. For long-term evaluation, we observe that the non-metric scaling of IronDepth leads to worse mTSED scores.

the additional pose embeddings provides additional guidance about the target poses leading to better TSED scores.

In addition, we ablate the effect of using an alternative depth estimator to ZoeDepth [3] in Tab. 2. For this, we use IronDepth [1] pretrained on ScanNetv2 and report qualitative results on RealEstate10K and ScanNet. While we observe comparable results in image quality performance, we note that using IronDepth leads to worse consistency scores. As IronDepth does not provide estimates in metric scale, using these depth estimates to warp the reference image leads to less accurate conditional information. Ultimately, this results in lower consistency scores - see e.g., $mTSED$ that decreases by $\approx 6\%$ on RealEstate10K and $\approx 13\%$ on ScanNet using IronDepth compared to using ZoeDepth.

Furthermore, we qualitatively show the effect of using structured noise in Fig. 8 on a ScanNet test sequence. We note that by structuring the noise using the depth estimates, we obtain more realistic and consistent synthesis results.

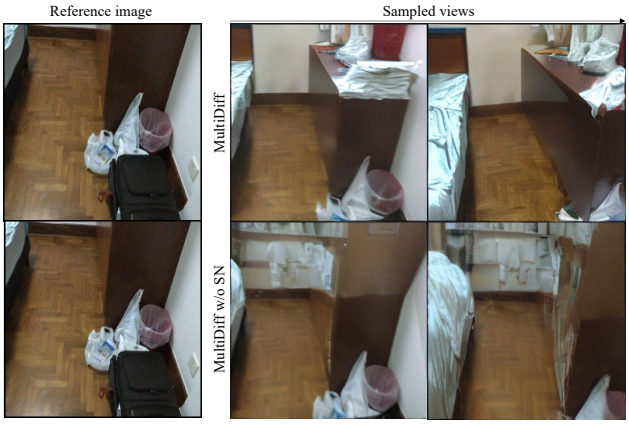

Figure 8. Additional qualitative comparison of applying structured noise on a ScanNet test sequence. Applying structured noise leads to more consistent and overall more realistic sampling results.

References

[1] Gwangbin Bae, Ignas Budvytis, and Roberto Cipolla. Irondepth: Iterative refinement of single-view depth using surface normal and its uncertainty. In *British Machine Vision Conference (BMVC)*, 2022. 3, 6

[2] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *IEEE International Conference on Computer Vision*, 2021. 3

[3] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth, 2023. 6

[4] VideoCrafter contributors. Videocrafter. *Github*. Accessed October 15, 2023 [Online] https://github.com/AILab-CVC/VideoCrafter. 3

[5] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 1, 4

[6] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 1, 3

[7] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 3

[8] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3

[9] Ayush Tewari, Tianwei Yin, George Cazenavette, Semon Rezchikov, Joshua B. Tenenbaum, Frédo Durand, William T. Freeman, and Vincent Sitzmann. Diffusion with forward models: Solving stochastic inverse problems without direct supervision. *NeurIPS*, 2023. 1

[10] Rundi Wu, Ben Mildenhall, Philipp Henzler, Keunhong Park, Ruiqi Gao, Daniel Watson, Pratul P. Srinivasan, Dor Verbin, Jonathan T. Barron, Ben Poole, and Aleksander Holynski. Reconfusion: 3d reconstruction with diffusion priors. *arXiv*, 2023. 5

[11] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural radiance fields from one or few images. In *CVPR*, pages 4578–4587, 2021. 5

[12] Jason J. Yu, Fereshteh Forghani, Konstantinos G. Derpanis, and Marcus A. Brubaker. Long-term photometric consistent novel view synthesis with diffusion models. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2023. 1, 5

[13] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. In *ACM TOG*, 2018. 1, 2

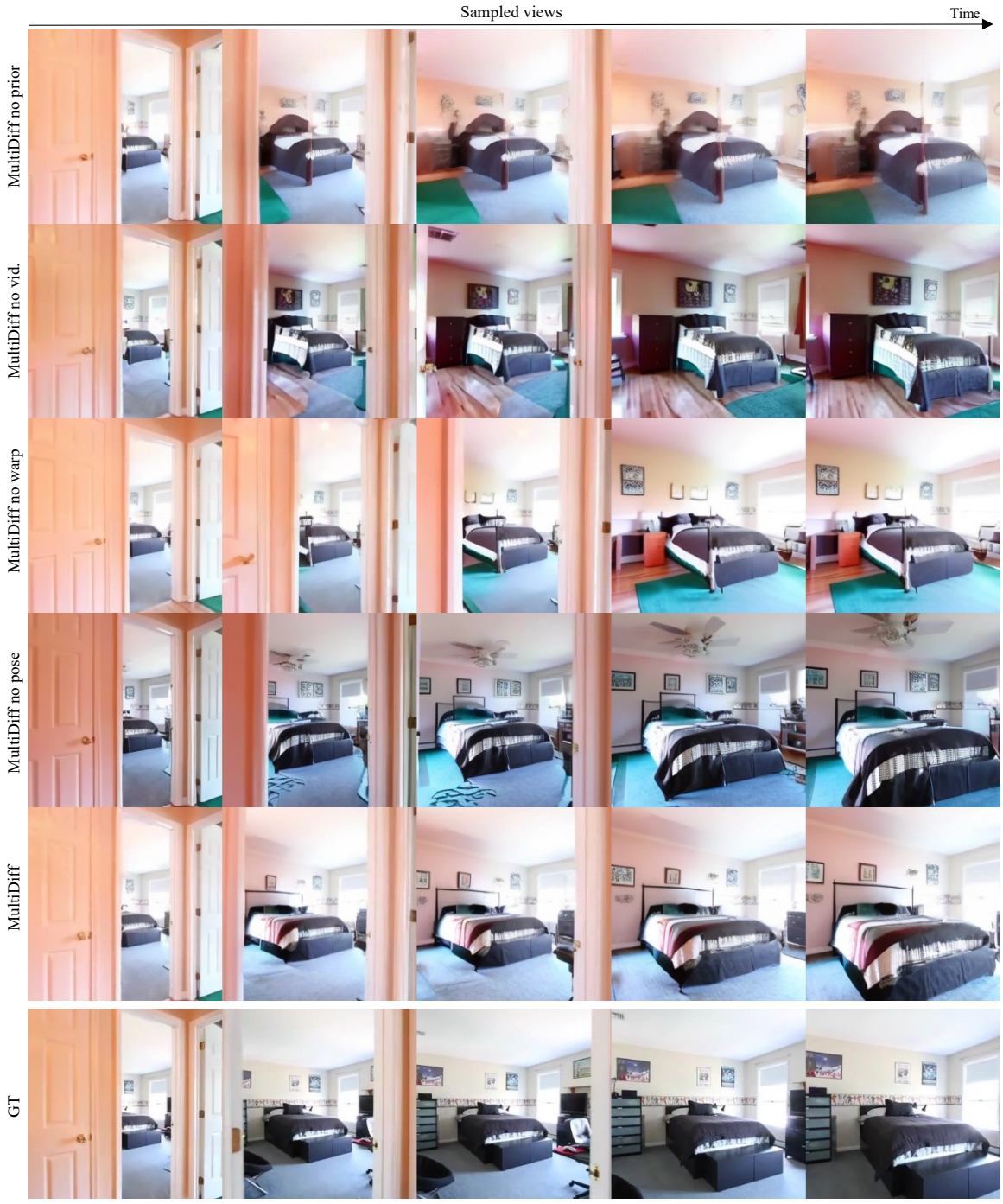

Figure 9. Qualitative comparison of the different ablations of our method on a RealEstate10K test sequence.