Controlling Space and Time with Diffusion Models

Daniel Watson* **Saurabh Saxena*** **Lala Li*** **Andrea Tagliasacchi** **David J. Fleet**
Google DeepMind

Abstract

We present 4DiM, a cascaded diffusion model for 4D novel view synthesis (NVS), conditioned on one or more images of a general scene, and a set of camera poses and timestamps. To overcome challenges due to limited availability of 4D training data, we advocate joint training on 3D (with camera pose), 4D (pose+time) and video (time but no pose) data and propose a new architecture that enables the same. We further advocate the calibration of SfM posed data using monocular metric depth estimators for metric scale camera control. For model evaluation, we introduce new metrics to enrich and overcome shortcomings of current evaluation schemes, demonstrating state-of-the-art results in both fidelity and pose control compared to existing diffusion models for 3D NVS, while at the same time adding the ability to handle temporal dynamics. 4DiM is also used for improved panorama stitching, pose-conditioned video to video translation, and several other tasks.

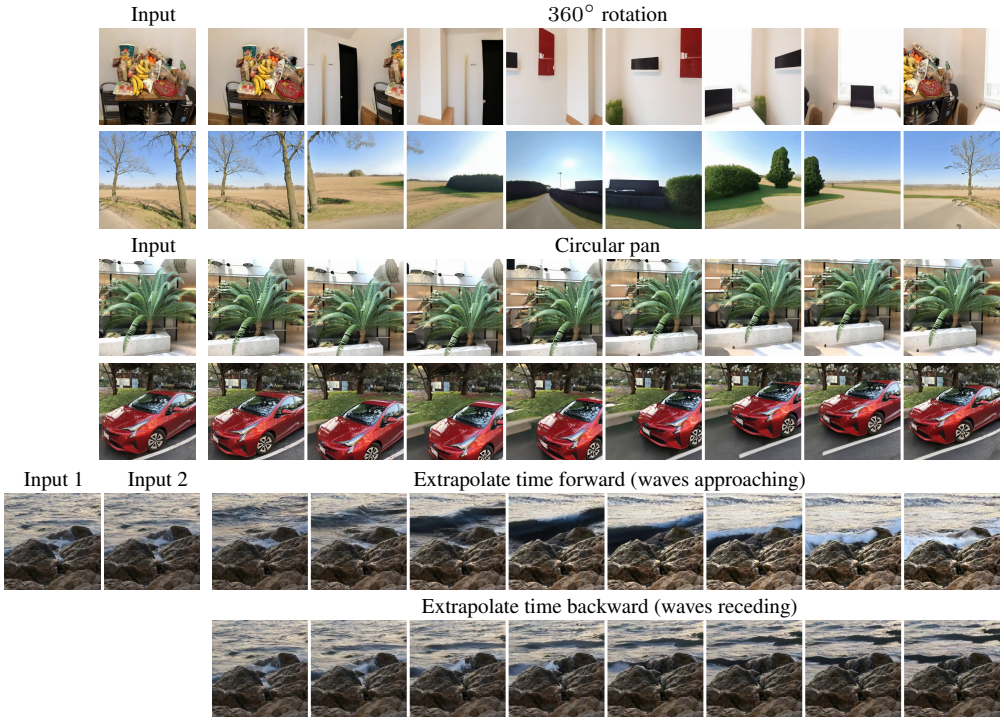

Figure 1: Zero-shot samples from 4DiM on LLFF [Mildenhall et al., 2019] and an internal video dataset given one or two input images. We show samples on different camera and time trajectories. 4DiM is a 32-frame model, so here we evenly sub-sample outputs. Samples are best viewed as video in our website: `https://4d-diffusion.github.io`

Preprint. Under review.

1 Introduction

Novel view synthesis (NVS) and 3D generative models have emerged as a new frontier in generative models, enabling the synthesis of novel views of 3D objects and scenes with control over camera pose. These models compliment text-to-image/video models, with potential uses in the generation of 3D assets, augmented reality, synthetic data for model training (e.g., in robotics), and view interpolation. Notable recent examples using diffusion models include 3DiM [Watson et al., 2022] and Zero-1-to-3 [Liu et al., 2023b]. They do not do generate explicit 3D scenes, rather, they leverage *implicit* 3D knowledge for conditional view generation, as a form of image+pose conditioned image generation.

To date, these models have focused mainly on *objects* with blank backgrounds, rather than generic natural scenes, and on limited camera poses, like those over the surface of the view-sphere fixated on objects centered at the origin. While recent work [Wang et al., 2023a, Yu et al., 2023a] has attempted to train models for generic scenes and 3D viewpoints, zero-shot performance on out of distribution scenes remains fraught. The biggest challenge is that 3D scene data is scarce. Much of the existing data also relies on COLMAP [Schönberger and Frahm, 2016, Schönberger et al., 2016] for the estimation of camera pose, yielding noisy poses with unknown scale. This is additionally problematic for effective control in the single-image to 3D task because models must sample from the distribution of plausible scales at inference time.

The goal of this work is to extend NVS diffusion models in three respects: 1) from objects to *scenes*; 2) to free-form camera poses specified in meaningful physical units; and 3) to simultaneously allow spatial and temporal control via camera pose and timestamp conditioning. To this end, we introduce *4DiM*, a diffusion model for NVS conditioned on (one or more) images of arbitrary *scenes*, camera *pose*, and *time*. 4DiM is trained on a mixture of data sources, including posed 3D images/video and unposed video, of both indoor and outdoor scenes. We make this possible through various technical innovations that enable training from missing data (e.g., images without pose or time annotations) and sampling with separate guidance weights on different conditioning images, poses or times. We also introduce calibrated versions of video datasets posed via COLMAP. Calibrated data makes it easier to learn metric regularities in the world, like the typical sizes of everyday objects and spatial relations, and it enables one to specify camera poses in meaningful physical units. As such, 4DiM generates 3D consistent, multi-view images or video of dynamic scenes. Beyond sample generation 4DiM can be used to unlock myriad applications, like video-to-video translation, improved panoramic stitching, or training explicit 3D models with score-distillation sampling [Poole et al., 2022].

To summarize, our main contributions include:

- 4DiM, a pixel-based diffusion model for novel view synthesis conditioned on one or more images of arbitrary *scenes*, camera *pose*, and *time*. 4DiM comprises a base model that generates 32 images at 64×64, and a multi-view super resolution model that up-samples to $32 \times 256 \times 256$;
- An effective data mixture for 4D models comprising posed and unposed video of indoor and outdoor scenes, enabling zero-shot application with fine-grained pose control and dynamics;
- Novel architectural elements to facilitate time and pose conditioning as well as training with incomplete data, and *multi-guidance* to simultaneously guide on camera pose and timestamps;
- A *Calibrated* version of RealEstate10K to improve model fidelity and enable metric pose control;
- Extensive evaluation and comparisons to prior work, including novel SfM-based metrics, namely, *keypoint distance* and *SfM distances* to better quantify pose alignment and dynamics.

2 Related work

We next provide a brief overview of prior work on camera control and 3D NVS with diffusion models. We also briefly discuss some recent work on 3D extraction using geometry-aware methods, which has emerged as an area of key interest to the research community. Here, the use of geometry-free pose-conditional diffusion models (like 4DiM) has been the key building block.

Unlike the typical setting of Neural Radiance Fields [Mildenhall et al., 2021] (NeRF) where tens-to-hundreds of images are used as input for 3D *reconstruction*, pose-conditional diffusion models for NVS aim to extrapolate plausible, diverse, 3D consistent samples with as few as a single image input. Conditioning diffusion models on an image and relative camera pose was introduced by Watson et al. [2022] and Liu et al. [2023b] as an effective alternative to prior few-view NVS methods, overcoming severe blur and floater artifacts [Sitzmann et al., 2019, Yu et al., 2021, Niemeyer et al., 2022, Sajjadi et al., 2022]. However, they rely on neural architectures unsuitable for jointly modeling more than two

views, and consequently require heuristics such as *stochastic conditioning* or Markovian sampling with a limited context window [Yu et al., 2023a], for which it is hard to maintain 3D consistency.

Subsequent work proposed attention mechanisms leveraging epipolar geometry to improve the 3D consistency of image-to-image diffusion models for NVS [Tseng et al., 2023], and more recently, fine-tuning text-to-video models with temporal attention layers or attention layers limited to overlapping regions [He et al., 2024, Wang et al., 2023a] to model the joint distribution of generated views and their camera extrinsics. These models however still suffer from several issues: they have difficulty with static scenes due to persistent dynamics from the underlying video models; they still suffer from 3D inconsistencies and low fidelity; and they exhibit poor generalization to out-of-distribution image inputs. Alternatives have been proposed to improve fidelity in multi-view diffusion models of 3D scenes, albeit sacrificing the ability to model free-form camera pose; e.g., see MVDiffusion and follow-ups [Tang et al., 2023, 2024] where the data distribution is limited in its pose trajectories.

A concurrent trajectory of closely related work on *3D extraction* has recently emerged following DreamFusion [Poole et al., 2022], where instead of directly training diffusion models for NVS, new techniques for sampling views parametrized as a NeRF with volume rendering are proposed such as *Score Distillation Sampling* (SDS) and *Variational Score Distillation* [Wang et al., 2024] (VSD). Here, a pre-existing diffusion model acts as a prior that drives the generation process. This enables, for example, text-to-3D using a text-to-image diffusion model. As demonstrated by MVDream [Shi et al., 2023], ReconFusion [Wu et al., 2023] and CAT3D [Gao* et al., 2024], using a pose-conditional diffusion model offers the unique advantage that the diffusion model can produce samples at the exact viewpoint specified for volume rendering during score distillation or NeRF postprocessing, which results in dramatically improved sample quality. All of these works establish further interest on improving pose-conditional diffusion models, which is the focus of our work, thereby enabling improvements in 3D extraction methods that build upon such models.

3 4D novel view synthesis models from limited data

4DiM uses a continuous-time diffusion model to learn the joint distribution over multiple views:

$$p(\boldsymbol{x}_{C+1:N} \mid \boldsymbol{x}_{1:C},\, \boldsymbol{p}_{1:N},\, t_{1:N}) \tag{1}$$

where $\boldsymbol{x}_{C+1:N}$ are generated images, $\boldsymbol{x}_{1:C}$ are conditioning images, $\boldsymbol{p}_{1:N}$ are relative camera poses (extrinsics and intrinsics) and $t_{1:N}$ are scalar, relative timestamps.[1] Following Ho et al. [2020], we choose our loss function to be the error between the predicted and actual noise (L_{simple} in their paper), though we use the L1 rather than L2 norm, because, like prior work [Saharia et al., 2022a, Saxena et al., 2023], we found that it improves sample quality. Our models use the "v-parametrization" [Salimans and Ho, 2022], which helps stabilize training, and we adopt the noise schedules proposed by Kingma et al. [2021]. Our current models process $N = 8$ or 32 images at resolution 256×256 (N is the number of conditioning plus generated frames). To this end, we decompose the task into two models similar to prior work [Ho et al., 2022a,b]; we first generate images at 64×64, and then up-sample using a $4\times$ super-resolution model trained with noise conditioning augmentation [Saharia et al., 2022c]. We also finetuned the 32-frame model to condition on 2 and 8 frames to enable more comparisons and example applications of 4DiM models.

Training data. While 3D assets, multi-view image data, and 4D data are limited, video data is available at scale and contains rich information about the 3D world, despite not having camera poses. One of our key propositions is thus to train 4DiM on a large-scale dataset of 30M videos without pose annotations, jointly with 3D and 4D data. As shown in Section 5.1, video plays a key role in helping to regularize the model. The 3D datasets used to train 4DiM include ScanNet++ [Yeshwanth et al., 2023] and Matterport3D [Chang et al., 2017], which have *metric* scale, and more free-form camera poses compared to other common 3D datasets in the literature (e.g., CO3D [Reizenstein et al., 2021] and MVImgNet [Yu et al., 2023b]). We also use 1000 scenes from Street View with permission from Google, comprising posed panoramas with timestamps (i.e., it is a "4D" dataset). During training, we randomly sample views from Street View from the set of panorama images within K consecutive timesteps (K=5 for our 8-view models, and K=20 for 32-view models). We sample unposed videos

[1] We neither require sequential temporal ordering as in video models as our architecture is permutation-equivariant over frames. All N images (conditioning and generated) are processed by the diffusion model.

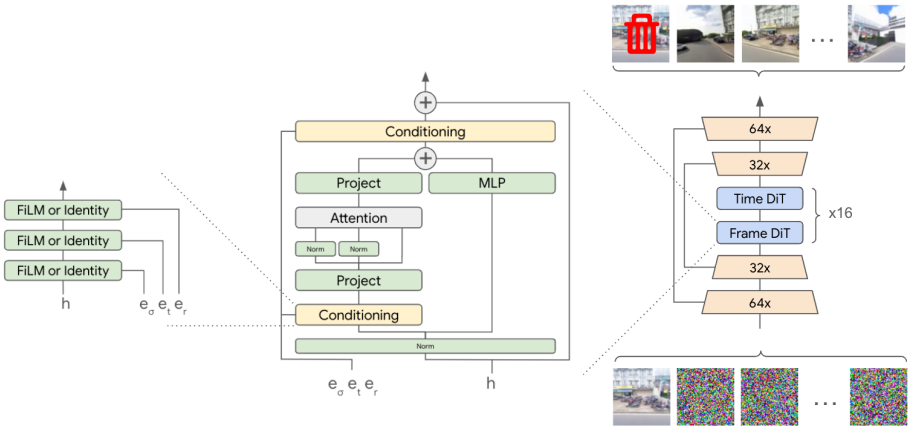

Figure 2: **Architecture** – We illustrate the 4DiM base model architecture, including choices for attention blocks and our conditioning mechanism. The 4DiM super-resolution model follows the same choices modulo minor differences to condition on low-resolution images. Outputs corresponding to the conditioning frames are always discarded from the training objective.

with probability 0.3 and views from posed datasets otherwise. 3D datasets are sampled in proportion to the number of scenes in each dataset.

Calibrated RealEstate10K. One particularly rich dataset for training 3D models is RealEstate10K [Zhou et al., 2018] (RE10K). It comprises 10,000 video segments of static scenes for which SfM [Schönberger and Frahm, 2016] has been used to infer per-frame camera pose, but only up to an unknown length scale. The lack of metric scale makes training more difficult because metric regularities of the world are lost, and it becomes more difficult for users to specify the target camera poses or camera motions in any intuitive units. This is especially problematic when conditioning on a single image, where scale itself otherwise becomes ambiguous. We therefore created a calibrated version of RealEstate10K by regressing the unknown metric scale from a monocular depth estimation model [Saxena et al., 2023]. (For details, see Supplementary Material A.) The resulting dataset, **cRE10K**, has a major impact on model performance (see Section 4 below). We follow the same procedure to calibrate the LLFF dataset [Mildenhall et al., 2019] for evaluation purposes.

Architecture. Relatively little training data has both time and camera pose annotations; most 3D data represent static scenes, while video data rarely include camera pose. Finding a way to effectively condition on both camera pose and time in a way that allows for incomplete training data is essential. We thus propose to chain "Masked FiLM" layers [Dumoulin et al., 2018] for (positional encodings of) diffusion noise levels, per-pixel ray origins and directions, and video timestamps. When any of these conditioning signals is missing (due to incomplete training data or random dropout for classifier-free guidance [Ho and Salimans, 2022]), the FiLM layers are designed to reduce to the identity function, rather than simply setting missing values to zero. This avoids the network from confusing a timestamp of zero with dropped or missing data. In practice, we replace the FiLM shift with zeros and scale with ones. We illustrate the overall choices in Fig. 2. (For details see Supplementary Material B.)

Sampling. Steering 4DiM with the correct sampling hyperparameters is essential, especially the guidance weights for classifier-free guidance [Ho and Salimans, 2022] (CFG). In its usual formulation, CFG treats all conditioning variables as "one big variable". In practice, placing a different weight on each variable is important; e.g., text requires high guidance weights, but high guidance weights on images can quickly lead to unwanted artifacts. We thus propose *multi-guidance*, where we generalize CFG to do exactly this, without making independence assumptions between conditioning variables. We start from a classifier-guided formulation [Dhariwal and Nichol, 2021], where we wish to sample with k conditioning signals, $v_1, v_2, ...v_k$, using the score

$$\nabla_{\boldsymbol{x}} \log p(\boldsymbol{x}) \; + \; w_1 \nabla_{\boldsymbol{x}} \log p(v_1|\boldsymbol{x}) \; + \; \sum_{j=2...k} w_j \nabla_{\boldsymbol{x}} \log p(v_j|v_{1:j-1}, \boldsymbol{x}) \; . \tag{2}$$

In the classifier-free formulation, this is equivalent to

$$(1 - w_1) \nabla_{\boldsymbol{x}} \log p(\boldsymbol{x}) \; + \; \sum_{j=2...k} (w_{j-1} - w_j) \nabla_{\boldsymbol{x}} \log p(\boldsymbol{x}|v_{1:j-1}) \; + \; w_k \nabla_{\boldsymbol{x}} \log p(\boldsymbol{x}|v_{1:k}) \; . \tag{3}$$

If we train our model such that it drops out only v_k, or v_{k-1} and v_k, ... , or all of $v_{1:k}$, we can sample with different guidance weights w_i on each conditioning variable v_i. 4DiM is trained to drop out conditioning signals with probability 0.1, and it drops either $t_{1:N}$, $t_{1:N}$ and $\boldsymbol{p}_{1:N}$, or all of $t_{1:N}$, $\boldsymbol{p}_{1:N}$ and $\boldsymbol{x}_{1:C}$. This is a natural choice, since poses and timestamps without corresponding conditioning images do not convey useful information. For best results, 4DiM uses a guidance weight 2.0 on conditioning images, a weight of 4.0 on camera poses, and 1.0 for timestamps.

4 Evaluation

Evaluating 4D generative models is challenging. Typically, methods for NVS are evaluated using generation quality. In addition, for pose-conditioned generation it is desirable to find metrics that capture 3D consistency (rendered views should be grounded in a consistent 3D scene) and pose alignment (the camera should move as expected). Furthermore, evaluating temporal conditioning requires some measure that captures the motion of dynamic content. We leverage several existing metrics and propose new ones to cover all these aspects. Below we cover each one in detail.

Image and Video fidelity. FID [Heusel et al., 2017] is a key metric for image quality, but if used in isolation it can be uninformative and a poor objective for hyper-parameter selection. For example, Saharia et al. [2022b] found that FID scores for text-to-image models are optimal with low CFG weights [Ho and Salimans, 2022], but text-image alignment suffers under his setting. In what follows we report FID as well as the improved Frechet distance with DINOv2 features (FDD) [Oquab et al., 2023], which appears to correlate with human judgements better than InceptionV3 features [Szegedy et al., 2016]. For video quality, we also report FVD scores [Unterthiner et al., 2018].

3D consistency. The work of Yu et al. [2023a] introduced the *thresholded symmetric epipolar distance* (TSED) as a more computationally efficient alternative to training separate NeRF models [Mildenhall et al., 2021] on every sample to evaluate 3D consistency [Watson et al., 2022]. First, SIFT is used to obtain keypoints between a pair of views. For each keypoint in the first image, we can compute the epipolar line in the second image and its minimum distance to the corresponding keypoint. This is repeated for each keypoint in the second image to obtain a SED score. A hyperparameter threshold is selected, and the percentage of image pairs whose median SED is below the threshold is reported. Below, we report TSED with a threshold of 2.0 and include the TSED score of ground truth data, as poses in the data are noisy and do not achieve a perfect score.

Pose alignment (new metric: SfM distances). We propose a set of metrics which we name *SfM distances*. Here, we run COLMAP pose estimation on the generated views, and compare the camera extrinsics predicted by COLMAP to the target poses. To get best results from COLMAP, we also feed it the input views. We report the relative error in camera positions ($\textbf{SfMD}_{\textbf{pos}}$) and the angular deviation of camera rotations in radians ($\textbf{SfMD}_{\textbf{rot}}$). Like we do for TSED scores, we also report the SfM distances for the ground truth data. Due to the inherent scale and rotational ambiguity of COLMAP, additional care must be taken to align the estimated poses to the original ones before comparing their differences as described above. Please see Supplementary Material D for more detail.

Metric scale pose alignment. The aforementioned metrics for 3D consistency and pose alignment are invariant to the scale of the camera poses as they rely on epipolar distances and SfM. In order to evaluate the metric scale alignment of 4DiM we report PSNR, SSIM [Wang et al., 2004] and LPIPS [Zhang et al., 2018]. These reconstruction-based metrics are not generally suitable to measure sample quality, as generative models can produce different but plausible samples. But for the same reason (i.e., that reconstruction metrics favor content alignment), we instead find them useful as an indirect indicator of *metric scale alignment*; i.e., given metric-scale poses, we would like models to not over/undershoot in position and rotation.

Dynamics (new metric: keypoint distance). One common failure mode that we observed early in our work is a tendency for certain models to copy input frames instead of generating temporal dynamics, and prior work has noted that good FVD scores can still be obtained with static video [Ge et al., 2024]. We propose a new metric called *keypoint distance* (**KD**), where we compute the average motion of SIFT keypoints[Lowe, 2004] across image pairs with $k \geq 10$ matches. We report results on generated and reference views to assess whether generated images have a similar motion distribution.

5 Experiments

As we mainly focus on camera control, we consider both in-distribution and OOD evaluations for 3D NVS capabilities in our ablations and comparisons to prior work. We use the RealEstate10K dataset [Zhou et al., 2018] as a common in-distribution evaluation. To maximize the amount of training data,

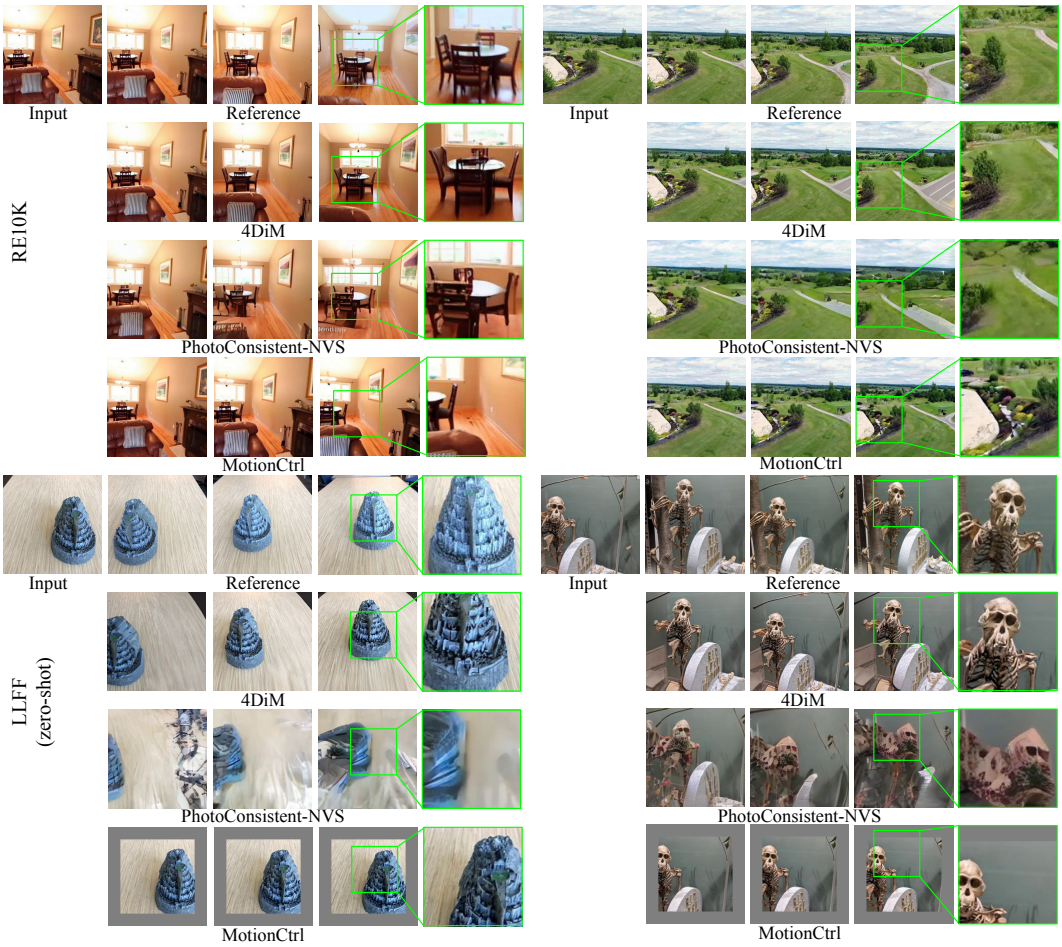

Figure 3: Qualitative comparison of 3D NVS for various diffusion models on RE10K (in-distribution) and LLFF (OOD). Note, this is not our final 4DiM model, but rather, an 8-frame 4DiM model trained only with RE10K for 3D data, so it is trained with the same data as prior work for fair comparison. See our website for more samples: `https://4d-diffusion.github.io`

we use 1% of the dataset as our validation split. We run inference on all baselines ourselves in this split, noting that they might instead be advantaged as our test data may exist in their training data (for PNVS, all our test data is in fact part of their training dataset, giving them a distinct advantage). For OOD evaluation, we use the LLFF dataset [Mildenhall et al., 2019].

We present our main results on 3D novel view synthesis conditioned on a *single* image in Tab. 1 and Fig. 3, comparing the capabilities of our 4DiM cascade against PhotoConsistentNVS[2] [Yu et al., 2023a] (PNVS) and MotionCtrl [Wang et al., 2023b]. These two are the strongest 3D scene NVS diffusion models with code and checkpoints available. For LLFF, we load the input trajectory of views in order, subsample frames evenly, and then generate 7 views given a single input image. RealEstate10K trajectories are much longer, so following PNVS, we subsample with a stride of 10. We use the 7-view NVS task conditioned on a single image, in order to avoid weakening the baselines: MotionCtrl can only predict up to 16 frames, and PNVS performance degrades with the length of the sequence as it is an image-to-image model. We trained versions of 4DiM that process 8-frames to this end (as opposed to using the main 32-frame 4DiM model and subsampling the output) for a more apples-to-apples comparison. Quantitative metrics are computed on 128 scenes from our RealEstate10K test split, and on all 44 scenes in LLFF.

We find that 4DiM achieves superior results in fidelity (FID, FDD, FVD) and metrc-scale pose alignment (LPIPS, PSNR, SSIM), by a wide margin compared to the baseline diffusion models.

[2]For PNVS, we follow Yu et al. [2023a] and use a Markovian sliding window for sampling, as they find it is the stronger than stochastic conditioning [Watson et al., 2022].

	FID (↓)	FDD (↓)	FVD (↓)	$\text{TSED}_{t=2}$ (↑)	SfMD_{pos}(↓)	SfMD_{rot}(↓)	LPIPS (↓)	PSNR (↑)	SSIM (↑)
cRE10k test									
PNVS	51.41	1007.	248.9	0.9961 (1.000)	**0.9773** (1.024)	0.3068 (0.3638)	0.3899	16.07	0.3878
MotionCtrl	43.07	370.6	411.7	0.4193 (1.000)	1.027 (1.032)	**0.2549** (0.3634)	0.5003	12.74	0.2668
4DiM-R	**31.23**	**306.3**	**195.1**	**0.9974** (1.000)	1.023 (1.075)	0.3029 (0.3413)	**0.2630**	**18.09**	**0.5309**
4DiM	31.96	314.9	221.9	0.9935 (1.000)	1.008 (1.034)	0.3326 (0.3488)	0.3016	17.08	0.4628
cLLFF zero-shot									
PNVS	185.4	2197.	1751.	0.7235 (0.9962)	0.9346 (0.8853)	0.2213 (0.1857)	0.5969	**12.04**	0.1311
MotionCtrl	106.0	531.5	1754.	0.1515 (0.9962)	0.9148 (0.9415)	**0.1366** (0.2216)	0.7016	9.722	0.06756
4DiM-R	**63.48**	**353.2**	**841.6**	**0.9659 (0.9962)**	0.9265 (0.8951)	0.2011 (0.1934)	**0.5403**	11.55	**0.1444**
4DiM	63.78	356.8	864.4	0.9167 (0.9962)	**0.9131 (0.8960)**	0.1838 (0.1943)	0.5415	11.58	0.1408

Table 1: **Comparison to prior work in 3D NVS.** We compare 8-frame 4DiM models against prior work, one only using cRE10K (4DiM-R) for a more fair comparison, and one trained with our full dataset mixture (4DiM). We evaluate in-distribution and zero-shot on LLFF. For 3D consistency and pose alignment, we include scores on real views (in parenthesis) to indicate plausible upper bounds (which may be imperfect as poses are noisy and SfM may fail). We highlight the best result in bold, and underline results where models score better than real images. We do not report scores when SfM fails, denoted with n/a. Our models substantially outperform baselines across all metrics.

	FID (↓)	FDD (↓)	FVD (↓)	$\text{TSED}_{t=2}$ (↑)	SfMD_{pos}(↓)	SfMD_{rot}(↓)	LPIPS (↓)	PSNR (↑)	SSIM (↑)
RE10k test									
4DiM-R (RE10k)	32.40	335.9	212.0	**1.000** (1.000)	**1.021** (1.022)	0.3508 (0.3737)	0.3002	16.64	0.4704
4DiM-R (cRE10k)	**31.23**	**306.3**	**195.1**	0.9974 (1.000)	1.023 (1.075)	**0.3029** (0.3413)	**0.2630**	**18.09**	**0.5309**
LLFF zero-shot									
4DiM-R (RE10k)	71.07	521.2	**740.1**	0.7727 (0.9962)	**0.9003** (0.8876)	0.2108 (0.1915)	**0.4568**	**12.87**	**0.2016**
4DiM-R (cRE10k)	**63.48**	**353.2**	841.6	**0.9659** (0.9962)	0.9265 (0.8951)	**0.2011** (0.1934)	0.5403	11.55	0.1444
ScanNet++ zero-shot									
4DiM-R (RE10k)	23.68	248.0	137.9	0.9512 (0.9815)	1.885 (1.858)	**1.534** (1.520)	0.1938	20.74	0.6716
4DiM-R (cRE10k)	**22.89**	**243.2**	**130.6**	**0.9685** (0.9815)	**1.851** (1.917)	1.541 (1.550)	**0.1809**	**21.25**	**0.6952**

Table 2: **Ablation showing the advantage of using calibrated data (metric scale camera poses).** We compare 8-frame 4DiM models trained on cRE10K vs uncalibrated RE10K. For RE10K and LLFF, models trained on cRE10K are tested on calibrated data, and models trained on uncalibrated RE10K are tested on uncalibrated data for a fair comparison. The exception is ScanNet++, which is already scale-calibrated. We find that training on cRE10K substantially improves in-domain and OOD performance both in terms of alignment and image fidelity. This is especially clear on reference-based image metrics such as PSNR, LPIPS, SSIM (with the exception of LLFF, where scores may be too low to be significant).

	FID (↓)	FDD (↓)	FVD (↓)	$\text{TSED}_{t=2}$ (↑)	SfMD_{pos}(↓)	SfMD_{rot}(↓)	LPIPS (↓)	PSNR (↑)	SSIM (↑)
cRE10k test									
4DiM (no video)	32.45	326.5	**207.7**	**0.9974** (1.000)	1.023 (1.029)	0.3356 (0.3692)	**0.2713**	**17.80**	**0.5209**
4DiM	**31.96**	**314.9**	221.9	0.9935 (1.000)	**1.008** (1.034)	**0.3326** (0.3488)	0.3016	17.08	0.4628
ScanNet++ test									
4DiM (no video)	24.61	250.7	**129.2**	0.9615 (0.9815)	1.895 (1.764)	**1.446** (1.431)	**0.1824**	**21.34**	**0.6969**
4DiM	**23.48**	**223.7**	146.0	n/a (0.9815)	**1.706** (1.889)	1.550 (1.503)	0.1965	20.67	0.6434
cLLFF zero-shot									
4DiM (no video)	64.34	401.4	884.1	**0.9886** (0.9962)	0.9157 (0.9067)	0.1872 (0.1965)	0.5437	11.45	**0.1430**
4DiM	**63.78**	**356.8**	**864.4**	0.9167 (0.9962)	**0.9131** (0.8960)	**0.1838** (0.1943)	**0.5415**	**11.58**	0.1408

Table 3: **Ablation showing the advantage of co-training with video data.** We train an identical 8-frame 4DiM model without video data to reveal the net-positive impact of co-training with video. Co-training with video data improves fidelity and generalization. Qualitatively, this is evident in 4DiM's ability to extrapolate realistic content.

Qualitatively, we find that MotionCtrl has significant trouble aligning with the conditioning images. We also observe that PNVS exhibits more artifacts, with a substantial drop in quality in the out-of-distribution setting. Surprisingly, PNVS achieves the best TSED in LLFF but upon closer inspection, we find that the number of keypoint matches is $\sim$3x less than for 4DiM. This suggests potential improvements for TSED, such as a higher threshold for the number of minimum keypoint matches, or requiring more spatial coverage of keypoints to discard examples for which the matches are too localized.

5.1 3D datasets significantly affect NVS quality

One of the most critical ingredients for 3D NVS, especially in the out-of-distribution setting, is the training dataset. We ablate this in three ways:

More diverse 3D data is helpful. Table 1 shows that adding more diverse 3D data beyond RealEstate10K to the training mixture improves pose alignment (SfM distances) on LLFF (zero-shot).

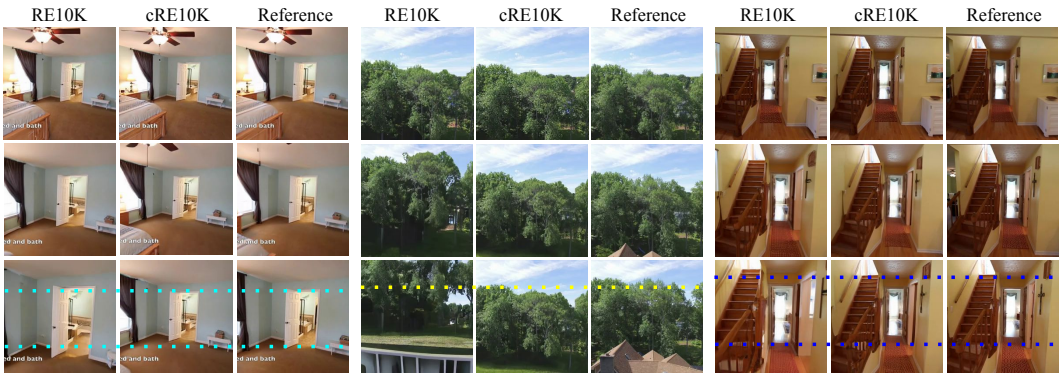

Figure 4: Qualitative comparison of 4DiM trained on calibrated vs uncalibrated RealEstate10K. The latter tends to 'overshoot' along the trajectory.

	30M video dataset test split			
	FID (↓)	FDD (↓)	FVD (↓)	KD (↑)
4DiM (1-frame conditioned)	62.73	355.4	934.0	2.214 (11.64)
4DiM (2-frame conditioned)	56.92	334.7	809.8	2.820 (11.51)
4DiM (8-frame conditioned)	43.34	320.7	400.2	3.181 (6.269)

Table 4: **Comparing different numbers of input frames for video extrapolation**. With 32-frame 4DiM models, we find that increasing the number of conditioning frames greatly improves dynamics. Note that, because this is an extrapolation task with constant video frame rate, the true keypoint distance decreases with more conditioning frames (i.e., less frames to generate).

The evaluation suggests some loss in fidelity when using the full mixture which is expected since the mixture is more diverse and includes both indoor and outdoor scenes.

Scale-calibrated data resolves ambiguity. To determine the impact of consistent, metric scale training data, we compare a model trained with on the original RE10k data against one with calibrated metric scale poses, with all else unchanged. Neither model includes the other 3D data sources in their training data used to train our best 4DiM model, because these are scale-calibrated and would confound this study otherwise. Quantitative results on 3D NVS from a single conditioning image are given in Tab. 2 (including zero-shot performance on LLFF and ScanNet++), and qualitative results are given in Fig. 4. We find that models trained on uncalibrated data often overshoot or undershoot when scale is ambiguous, and that models trained on scale-calibrated data solve this problem.

Large-scale video data improves generalization. We also show the importance of leveraging video as a training data source. Table 3 presents results comparing 8-frame 4DiM trained on our final data mixture to the otherwise identical model but trained without video data. We find that including video data improves fidelity and generalization ability which is expected since available 3D/4D data is not as diverse. The improvement is more apparent qualitatively (see Supplementary Material F for more samples).

5.2 Emergence of temporal dynamics

Early in our experiments, we observed that video from a single image can often yield samples with little motion. While this can be remedied by lowering the guidance weight on the conditioning image, this can entail a cost in fidelity. Nevertheless, with two or more input frames, 4DiM can successfully interpolate and extrapolate video. We quantitatively illustrate these results through our keypoint distance metric (along with other standard metrics) in Tab. 4. We suspect that future work including additional controls (e.g., text) on models like 4DiM, as well as larger scale, will likely play a key role in breaking modes for temporal dynamics, given that text+image-to-video has been successful for large models [Brooks et al., 2024].

5.3 Multiframe Conditioning

While we (and most baselines) focus on single frame conditioning, using more than one frame unlocks interesting capabilities. Here we show two applications. First, we consider 4DiM on the task of panorama stitching (rendering novel viewpoints), given discrete images that collectively cover a full 360° FOV (see Fig. 5). For reference we stitch the images using a naive homography warp and

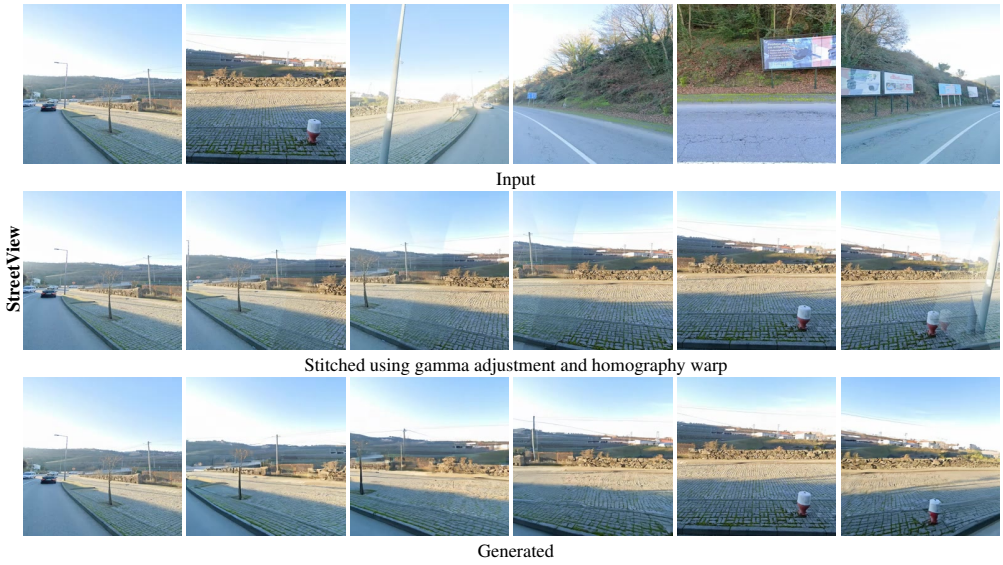

Figure 5: Stitching panoramas is non-trivial due to exposure differences, which is hard to fix with simple gamma/gain adjustment. Here we show novel views rendered by 4DiM conditioned on 6 frames covering 360 ° fov. We use our 8-frame conditioned model - conditioning frames are randomly replicated to achieve desired arity. (Note this is a subset of the 24 generated frames by 4DiM.)

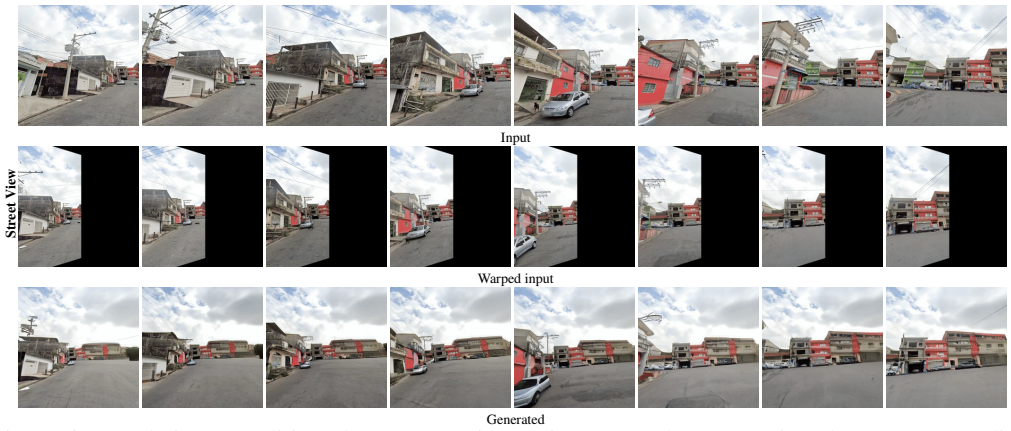

Figure 6: Novel views conditioned on a space-time trajectory on the Street View dataset. We condition on 8 consecutive time steps from the front-left camera and generate views at the same time steps with a camera facing the front. We find that the hallucinated regions (right half of the images) are consistent across space-time. Warped input images are provided for reference.

# conditioning frames	Street View panoramas					Matterport3D 360° panoramas				
	FID (↓)	FDD (↓)	PSNR (↑)	SSIM (↑)	LPIPS (↓)	FID (↓)	FDD (↓)	PSNR (↑)	SSIM (↑)	LPIPS (↓)
1 (60°FOV)	119.7	1601	12.16	0.2925	0.5438	86.41	1080	9.366	0.1638	0.6467
6 (360°FOV)	37.43	354.1	17.76	0.5249	0.2663	24.86	239.4	15.64	0.4974	0.2987

Table 5: Quantitative performance on generating panos in the extrapolation (number of conditioning frames is 1) and interpolation (number of conditioning frames is 6) regime. Using more conditioning frames improves image fidelity which is to be expected. See Fig. 5 for samples.

gamma adjustment [Brown and Lowe, 2007]. We find that outputs from 4DiM have higher fidelity than this baseline, and they avoid artifacts like those arising from incorrect exposure adjustment.

We next consider the task of rendering an existing space-time trajectory from novel viewpoints (i.e., video to video translation). We condition on images from a single camera at 8 consecutive frames from a Street View sequence and render a different view yielding 3D-consistent hallucinated regions in the generated images. Results are shown in Fig. 6.

6 Discussion and future work

To the best of our knowledge, 4DiM is the first model capable of generating multiple, approximately consistent views over simultaneous camera and time control from as few as a single input image. 4DiM achieves state-of-the-art pose alignment and much better generalization compared to prior work on 3D NVS. This new class of models opens the door to myriad downstream applications, some of which we demonstrate, e.g., changing camera pose in videos and seamlessly stitching panoramas. While still not perfect, 4DiM also achieves previously unseen quality in extremely challenging camera trajectories such as 360° rotation from a single image in the wild, indoor or outdoor. We expect 4DiM and follow-up works to have a major impact on methods focused on 3D model extraction, which rely heavily on pose-conditional diffusion models.

Limitations and future work. Results with 4DiM are promising, but there is room for improvement, with more calibrated 3D/4D data and larger models, with which the improved capacity is expected to improve image fidelity, 360° camera rotation, and dynamics with single-frame conditioning.

Societal impact. Like all generative image and video models, it is important to develop and release models with care. 4DiM is trained largely on data of scenes without people (or anonymized where present), and does not condition on text prompts, which mitigates many of the safety issues that might otherwise arise.

References

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. URL `https://openai.com/research/video-generation-models-as-world-simulators`.

Matthew A. Brown and David G. Lowe. Automatic panoramic image stitching using invariant features. *International Journal of Computer Vision*, 74:59–73, 2007. URL `https://api.semanticscholar.org/CorpusID:5328928`.

Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017.

Mostafa Dehghani, Josip Djolonga, Basil Mustafa, Piotr Padlewski, Jonathan Heek, Justin Gilmer, Andreas Peter Steiner, Mathilde Caron, Robert Geirhos, Ibrahim Alabdulmohsin, et al. Scaling vision transformers to 22 billion parameters. In *International Conference on Machine Learning*, pages 7480–7512. PMLR, 2023.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Vincent Dumoulin, Ethan Perez, Nathan Schucher, Florian Strub, Harm de Vries, Aaron Courville, and Yoshua Bengio. Feature-wise transformations. *Distill*, 3(7):e11, 2018.

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

Songwei Ge, Aniruddha Mahapatra, Gaurav Parmar, Jun-Yan Zhu, and Jia-Bin Huang. On the content bias in fr\'echet video distance. *arXiv preprint arXiv:2404.12391*, 2024.

Justin Gilmer, Andrea Schioppa, and Jeremy Cohen. Intriguing properties of transformer training instabilities, 2023. *To appear*.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

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, 2022a.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022b.

Emiel Hoogeboom, Jonathan Heek, and Tim Salimans. simple diffusion: End-to-end diffusion for high resolution images. In *International Conference on Machine Learning*, pages 13213–13232. PMLR, 2023.

Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707, 2021.

Hao Liu, Matei Zaharia, and Pieter Abbeel. Ring attention with blockwise transformers for near-infinite context. *arXiv preprint arXiv:2310.01889*, 2023a.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9298–9309, 2023b.

David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004.

Ben Mildenhall, Pratul P. Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)*, 2019.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

Michael Niemeyer, Jonathan T Barron, Ben Mildenhall, Mehdi SM Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5480–5490, 2022.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023.

Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv*, 2022.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–16. IEEE, 2020.

Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10901–10911, 2021.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–10, 2022a.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022b.

Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4713–4726, 2022c.

Mehdi SM Sajjadi, Henning Meyer, Etienne Pot, Urs Bergmann, Klaus Greff, Noha Radwan, Suhani Vora, Mario Lučić, Daniel Duckworth, Alexey Dosovitskiy, et al. Scene representation transformer: Geometry-free novel view synthesis through set-latent scene representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6229–6238, 2022.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022.

Saurabh Saxena, Junhwa Hur, Charles Herrmann, Deqing Sun, and David J. Fleet. Zero-shot metric depth with a field-of-view conditioned diffusion model. *arXiv:2312.13252*, 2023.

Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-Motion Revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016.

Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023.

Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. *Advances in Neural Information Processing Systems*, 32, 2019.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

Shitao Tang, Fuyang Zhang, Jiacheng Chen, Peng Wang, and Yasutaka Furukawa. Mvdiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion. *arXiv*, 2023.

Shitao Tang, Jiacheng Chen, Dilin Wang, Chengzhou Tang, Fuyang Zhang, Yuchen Fan, Vikas Chandra, Yasutaka Furukawa, and Rakesh Ranjan. Mvdiffusion++: A dense high-resolution multi-view diffusion model for single or sparse-view 3d object reconstruction, 2024.

Hung-Yu Tseng, Qinbo Li, Changil Kim, Suhib Alsisan, Jia-Bin Huang, and Johannes Kopf. Consistent view synthesis with pose-guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16773–16783, 2023.

Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in Neural Information Processing Systems*, 36, 2024.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Zhouxia Wang, Ziyang Yuan, Xintao Wang, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. *arXiv preprint arXiv:2312.03641*, 2023a.

Zhouxia Wang, Ziyang Yuan, Xintao Wang, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. *arXiv preprint arXiv:2312.03641*, 2023b.

Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. *arXiv preprint arXiv:2210.04628*, 2022.

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.

Yifeng Xiong, Haoyu Ma, Shanlin Sun, Kun Han, and Xiaohui Xie. Light field diffusion for single-view novel view synthesis. *arXiv preprint arXiv:2309.11525*, 2023.

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2023.

Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4578–4587, 2021.

Jason 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)*, pages 7071–7081. IEEE, 2023a.

Xianggang Yu, Mutian Xu, Yidan Zhang, Haolin Liu, Chongjie Ye, Yushuang Wu, Zizheng Yan, Chenming Zhu, Zhangyang Xiong, Tianyou Liang, et al. Mvimgnet: A large-scale dataset of multi-view images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9150–9161, 2023b.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in Neural Information Processing Systems*, 32, 2019.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018.

Controlling Space and Time with Diffusion Models
Supplementary Material

A Calibrated RealEstate10K

To calibrate RealEstate10K, we predict a single length scale per scene with the help of a recent zero-shot model for monocular depth [Saxena et al., 2023] that predicts metric depth from RGB and FOV. For each image we compute metric depth using the FOV provided by COLMAP. The length-scale is then obtained by regressing the 3D points obtained by COLMAP to metric depth at image locations to which the COLMAP points project. An L1 loss provides robustness to outliers in the depth map estimated by the metric depth models. The mean and variance of the per-frame scales yields a per-sequence estimator. We use the variance as a simple measure of confidence to identify scenes for which the scale estimate may not be reliable, discarding $\sim$30% of scenes with the highest variance. Table 6 shows the importance of filtering out scenes with less reliable calibration. The calibrated dataset will be made public to facilitate quantitative comparisons.

	FID (↓)	FDD (↓)	FVD (↓)	$\text{TSED}_{t=2}$ (↑)	SfMD_{pos}(↓)	SfMD_{rot}(↓)	LPIPS (↓)	PSNR (↑)	SSIM (↑)
cRE10k test									
4DiM-R (no filtering)	31.81	313.8	237.6	0.9815 (1.000)	1.035 (1.049)	0.3328 (0.3529)	0.3107	16.49	0.4465
4DiM-R (w/ filtering)	**31.23**	**306.3**	**195.1**	**0.9974** (1.000)	**1.023** (1.075)	**0.3029** (0.3413)	**0.2630**	**18.09**	**0.5309**
ScanNet++ zero-shot									
4DiM-R (no filtering)	24.35	**232.0**	143.1	n/a (0.9815)	1.881 (1.774)	1.639 (1.512)	0.1990	20.15	0.6436
4DiM-R (w/ filtering)	**22.89**	243.2	**130.6**	0.9685 (0.9815)	**1.851** (1.917)	**1.541** (1.550)	**0.1809**	**21.25**	**0.6952**
cLLFF zero-shot									
4DiM-R (no filtering)	**62.55**	369.5	848.5	0.9167 (0.9962)	**0.9214** (0.8844)	0.2194 (0.1932)	0.5412	11.48	0.1389
4DiM-R (w/ filtering)	63.48	**353.2**	**841.6**	**0.9659** (0.9962)	0.9265 (0.8951)	**0.2011** (0.1934)	**0.5403**	**11.55**	**0.1444**

Table 6: **Data filtering ablation.** We show the importance of discarding scenes where scale calibration is least reliable. Beyond the clear impact on fidelity, metric scale alignment itself improves by a wide margin. This is quantitatively shown through the reference-based image metrics (LPIPS, PSNR, SSIM) which favor content alignment, as discussed in Section 4.

B Neural Architecture

UViT backbone. As hinted by Figure 2, 4DiM uses a UViT architecture following Hoogeboom et al. [2023]. Compared to the commonly employed UNet architecture [Ronneberger et al., 2015], UViT uses a transformer backbone at the bottleneck resolution with no convolutions, leading to improved accelerator utilization. Information across frames mixes exclusively in temporal attention blocks, following [Ho et al., 2022b]. We find that this is key to keep computational costs within reasonable limits compared to, say, 3D attention [Shi et al., 2023]. Because the temporal attention blocks have a limited sequence length (32 frames), we insert them at all UViT resolutions. To limit memory usage, we use per-frame self-attention blocks only at the 16x16 bottleneck resolution where the sequence length is the product of the current resolution's dimensions.

Transformer block design. Our transformer blocks combine helpful choices from several prior work. We use parallel attention and MLP blocks following Dehghani et al. [2023], as we found in our early experiments that this leads to slightly better hardware utilization while achieving almost identical sample quality. We also inject conditioning information in a similar fashion to Peebles and Xie [2023], though we use fewer conditioning blocks due to the use of the parallel attention + MLP. We additionally emply query-key normalization following the findings of Gilmer et al. for improved training stability, but with RMSNorm [Zhang and Sennrich, 2019] for simplicity.

Conditioning. We use the same positional encodings for diffusion noise levels and relative timestamps as [Saharia et al., 2022b]. For relative poses, we follow 3DiM [Watson et al., 2022], and condition generation with per-pixel ray origins and directions, as originally proposed by SRT [Sajjadi et al., 2022] in a regressive setting. Xiong et al. [2023] has compared this choice to conditioning via encoded extrinsics and focal lengths a-la Zero-1-to-3 [Liu et al., 2023b], finding it advantageous. While a thorough study of different encodings is missing in the literature, we hypothesize that rays are a natural choice as they encode camera intrinsics in a way that is independent of the target resolution, yet gives the network precise, pixel-level information about what contents of the scene are visible and which are outside the field of view and therefore require extrapolation by the model. Unlike Plucker coordinates, rays additionally preserve camera positions which is key to deal with occlusions in the underlying 3D scene.

C Compute, training time, and sharding

We train 8-frame 4DiM models for $\sim$1M steps. Using 64 TPU v5e chips, we achieve a throughput of approximately 1 step per second with a batch size of 128. The model has $\sim$2.6B parameters, and available HBM is maximized with various strategies: first, we use bfloat16 activations (but still use float32 weights to avoid instabilities). We also use FSDP (i.e., zero-redundancy sharding [Rajbhandari et al., 2020] with delayed and rematerialized all-gathers). We then finetune this model to its final 32-frame version. This strategy allows us to leverage large batch-size pre-training, which is not possible with too many frames because the amount of activations stored for backpropagation scales linearly respect the number of frames per training example. The model is finetuned with the same number of chips, albeit only for 50,000 steps and at batch size of 32. This allows us to shard the frames of the video and use more than one chip per example at batch size 1. When using temporal attention, we simply all-gather the keys and values and keep the queries sharded over frames (one of the key insights from Ring Attention [Liu et al., 2023a]), though we don't decompose the computation of attention further as we find we have sufficient HBM to fully parallelize over all-gathered keys and values. FSDP is also enabled on the frame axis to maximize HBM savings so the number of shards is truly the number of chips (as opposed to the number of batch-parallel towers). This is possible because frame sharding requires an all-reduce by mean on the loss over the frame axis. Identically to zero-redundancy sharding, this can be broken down into a reduce-scatter followed by an all-gather.

D Further details on proposed metrics

We follow Yu et al. [2023a] and compute TSED only for contiguous pairs in each view trajectory, and discard sequences where we find less than 10 two-view keypoint matches. We use a threshold of 2.0 in all our experiments. For our proposed SfM distances, in order to get a scale-invariant metric, we first align the camera positions predicted by COLMAP by relativizing the predicted and original poses with respect to the first conditioning frame. This should resolve rotation ambiguity. Then, we resolve scale ambiguity by analytically solving for the least squares optimization problem that aligns the original positions to re-scaled COLMAP camera positions. Finally, we report the *relative error* in positions under the L_2 norm, i.e., we normalize by the norm of the original camera positions.

E Further details on baselines from prior work

For PNVS, the conditioning image is the largest square center crop of the original conditioning frame, resized to 256×256 pixels. This is the same preprocessing procedure used by all our 4DiM models. PNVS generates output frames sequentially (one at a time), using 2000 denoising steps for each frame. All 4DiM samples are produced with 256 denoising steps.

For MotionCtrl, we use the checkpoint based on Stable Diffusion, which generates 14 frames conditioning on one image and 14 poses. We set the sampling hyper-parameter speed to 1, use 128 ddim denoising steps, and keep the other hyper-parameters as default values in the released script from MotionCtrl. To avoid any loss of quality due to an out-of-distribution resolution or aspect ratio, we use the resolution 576×1024, i.e. the same resolution as MotionCtrl was trained on. RE10K images already have the desired aspect ratio. LLFF images have resolution 756×1008; we are thus forced to crop them to 567×1008 so they have the same aspect ratio MotionCtrl was trained with. For comparability with 4DiM and PNVS, we *postprocess* MotionCtrl samples at 567×1008 by taking the largest center crop and then resizing to 256×256. Note that, in order to preserve the aspect ratio of LLFF, the resulting square outputs are for a smaller region than 4DiM and PNVS, hence the gray padding for MotionCtrl LLFF samples in Figure 3.

F Samples

We include more 4DiM samples below, though we strongly encourage the reader to browse our website: `https://4d-diffusion.github.io`

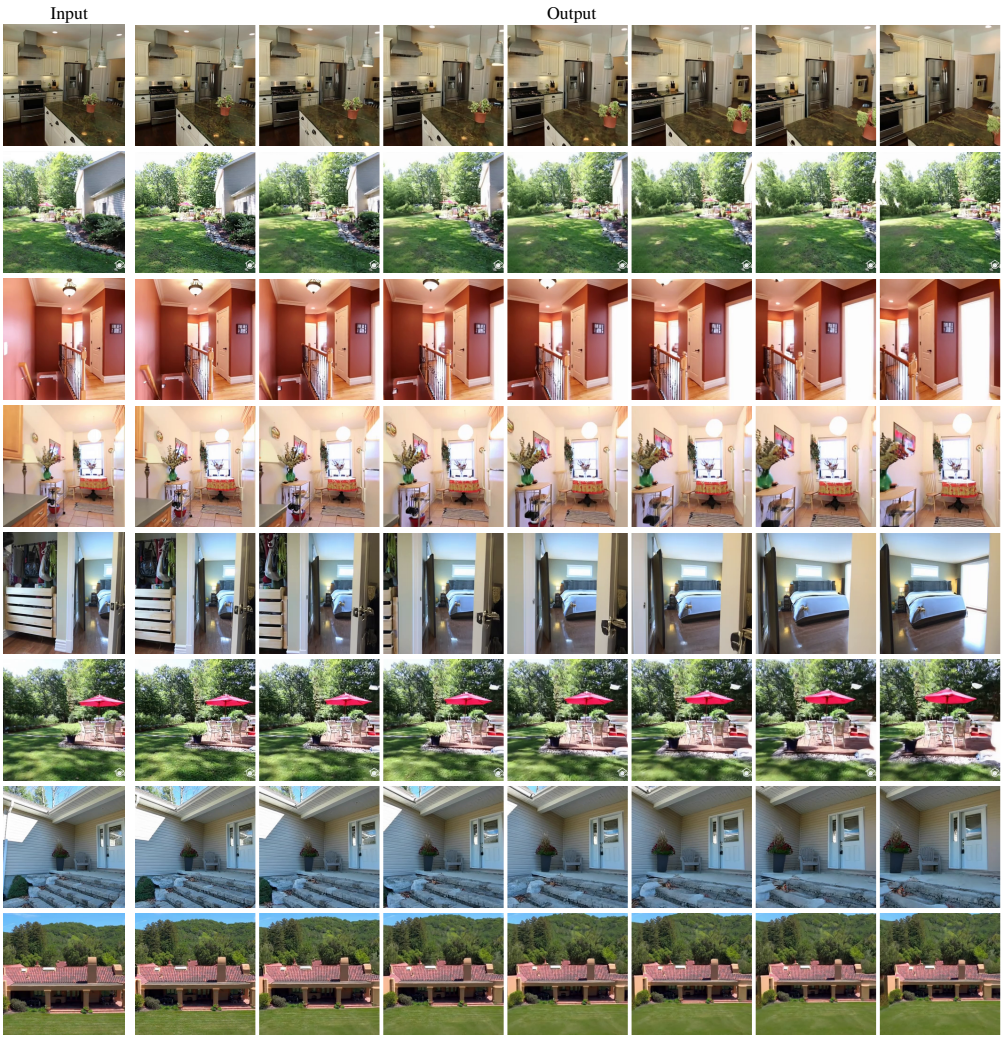

Figure 7: More 4DiM samples from the RealEstate10K dataset.

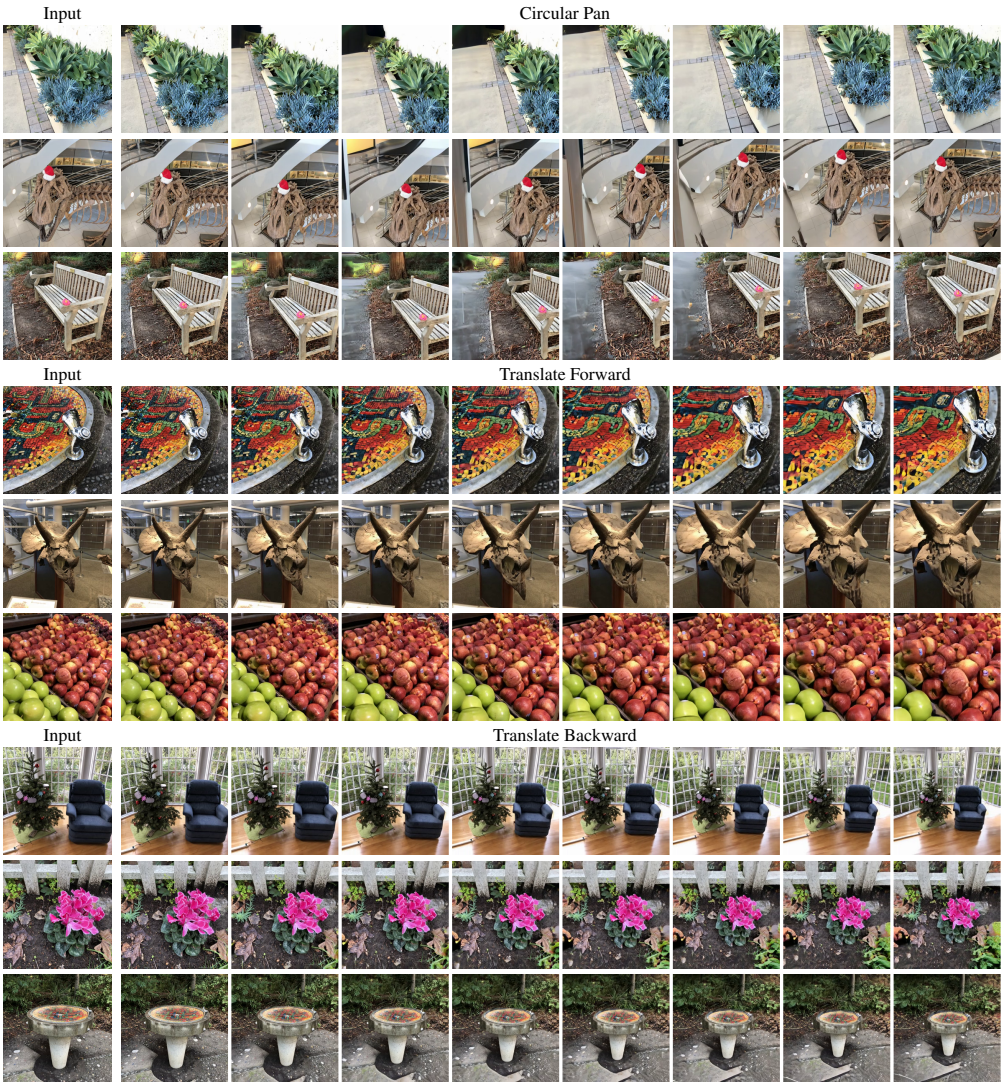

Figure 8: More 4DiM samples with custom trajectories from the LLFF dataset.