Generating 3D-Consistent Videos from Unposed Internet Photos

Gene Chou[1] Kai Zhang[2] Sai Bi[2] Hao Tan[2] Zexiang Xu[2]
Fujun Luan[2] Bharath Hariharan[1] Noah Snavely[1]

[1]Cornell University [2]Adobe Research

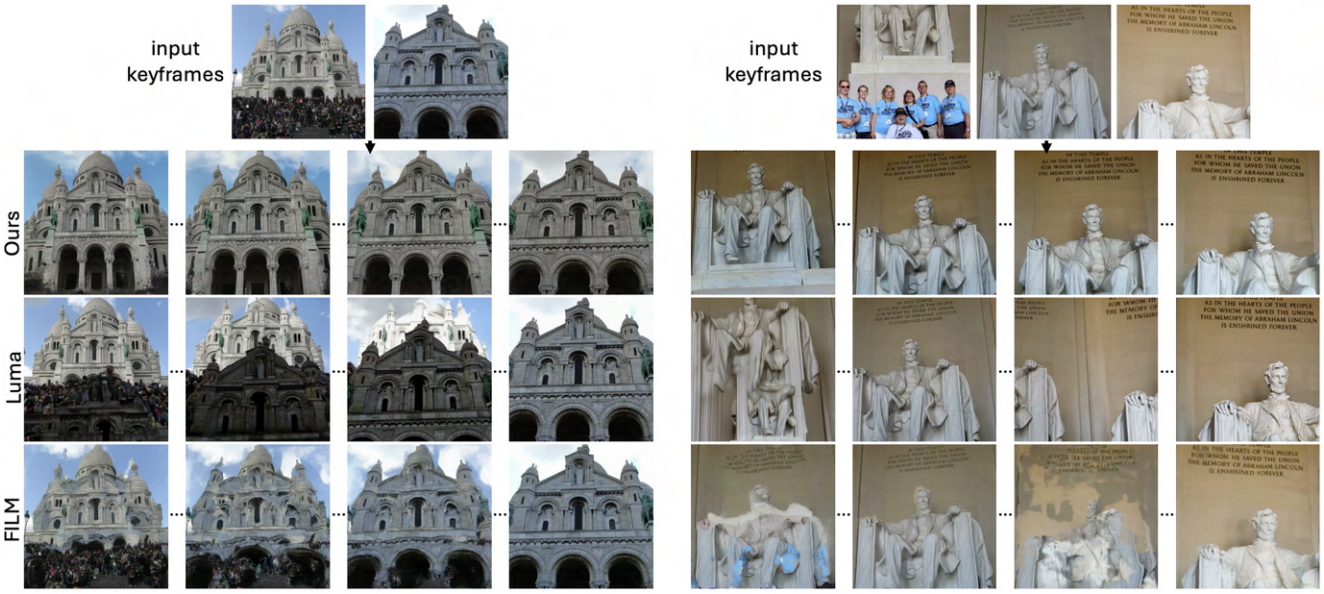

Figure 1. Given n unposed input keyframes, the goal is to generate a video of the scene with a realistic camera trajectory and consistent geometry. From top to bottom: Ours, Luma Dream Machine [41] (a commercial video generation model), FILM [50] (a frame interpolation method). Luma hallucinates new buildings (left scene) and statues (right scene) without understanding the scene layout. FILM is unable to handle wide-baseline inputs and produces blurry transitions. See our supplement for video playback.

Abstract

We address the problem of generating videos from unposed internet photos. A handful of input images serve as keyframes, and our model interpolates between them to simulate a path moving between the cameras. Given random images, a model's ability to capture underlying geometry, recognize scene identity, and relate frames in terms of camera position and orientation reflects a fundamental understanding of 3D structure and scene layout. However, existing video models such as Luma Dream Machine fail at this task. We design a self-supervised method that takes advantage of the consistency of videos and variability of multiview internet photos to train a scalable, 3D-aware video model without any 3D annotations such as camera parameters. We validate that our method outperforms all baselines in terms of geometric and appearance consistency. We also show our model benefits applications that enable camera control, such as 3D Gaussian Splatting. Our results suggest that we can scale up scene-level 3D learning using only 2D data such as videos and multiview internet photos.

1. Introduction

Recent advances in video foundation models [6, 18, 24, 30, 82] learn rich spatio-temporal representations that capture the underlying structure and dynamics of the visual world. It is not surprising that these models contain strong 3D priors that can be used for a variety of downstream applications through finetuning, such as 3D object generation [10, 19] and novel view synthesis [17, 87].

In this paper, we further investigate the capabilities of video models to understand 3D structure of real-world scenes. To this end, we propose the task of generating videos from a

handful (2-5) of unposed internet photos of the same scene. The generated frames should simulate a path moving between the locations of the cameras, from the first to the second, then from the second to the third, and so on. Given random images, a model's ability to capture underlying geometry, recognize scene identity, and relate frames in terms of camera position and orientation reflects a fundamental understanding of 3D structure and scene layout.

Interestingly, we find that this task is challenging for existing video models. Even when scaled to a commercial level (e.g. Luma Dream Machine [41]), we find that these models often perform creative morphing effects rather than produce realistic camera motion. We show examples in Fig. 1. Frame interpolation methods such as FILM [12, 31, 50] are unable to handle wide baselines because they are generally trained on videos with very little camera motion, leading to blurry transitions. Luma Dream Machine generates high-resolution, sharp videos, but ignores the identity of the scene and creates new structures to fit the input keyframes. In the scene on the left (Sacre Coeur), for instance, Luma transforms the crowd and stairs into a new building. In the scene on the right (Lincoln Memorial Statue), it creates a second statue.

Our main insight is that simply training for general video synthesis is not enough: we need to introduce scalable, 3D-aware objectives. Finetuning with camera poses is an option [39, 74], but collecting 3D annotations is costly [13, 65]. On the other hand, unstructured images and videos abound [1, 57].

Thus, we design two objectives. The first is multiview inpainting, which learns 3D priors without 3D annotations. The model takes a variable number of condition images, captured from random and wide-baseline viewpoints of a scene, and inpaints an 80% masked target. Through this process, it learns to extract structural information and scene identity from the condition images, and the illumination and scene layout from the remaining 20% of the target, in order to fill in the target image accurately. Multiview inpainting allows us to leverage the vast corpora of images, which, compared to existing video datasets, provides a greater variety of scenes, more diverse camera viewpoints, and is more accessible [65]. We train using internet photos to adapt to in-the-wild settings, but this objective can be trivially extended to other data sources, such as video, synthetic data, and self-driving datasets for further scaling.

Our second objective is view interpolation, which takes the start and end frames of a video, and generates intermediate frames. It requires no annotations from the videos.

We illustrate these two objectives in Fig. 2. These objectives are complementary in enabling our proposed task. Multiview inpainting addresses geometric understanding by training the model to extract 3D relationships from wide-baseline, unposed images. View interpolation addresses temporal coherence by training the model to generate smooth, consistent camera trajectories, which is our desired output.

Finally, we unify these two objectives under the same diffusion denoising objective. We finetune from a video diffusion model, which adds noise to and denoises selected patches of images: the masked pixels in the multiview inpainting objective, and the intermediate frames in the view interpolation objective. This allows us to jointly train both objectives without additional strategies such as pretraining or distillation. As a result, our model can take noisy internet photos and produce consistent trajectories, even though it has never seen this input-output pairing during training.

As shown in Fig. 1, our model produces a realistic camera path that captures the layout of the scene, even though the keyframes contain illumination variations and occlusions. For best viewing results, please visit our website, where we show side-by-side video playback for comparisons. Our approach and results suggest that multiview and video datasets can be complementary to each other even when disjoint.

Following recent works that refer to input images that condition videos as "keyframes" [41, 70, 78, 83], and with the added challenge that our keyframes are internet photos, we refer to our approach as **k**ey**f**rame-**c**onditioned video generation in-the-wild (KFC-W). In Sec. 4, we evaluate our method by 1) user studies that show it outperforms existing state-of-the-art video generation models; 2) downstream applications such as 3D reconstruction that validate its geometric and appearance consistency.

To sum up, our contributions are as follows:

1. We enable consistent video generation conditioned on wide-baseline, in-the-wild keyframes.
2. We propose a scalable self-supervised training scheme that takes advantage of both multiview and video datasets. The resulting model is 3D-aware without requiring 3D supervision such as camera poses.
3. We evaluate our method on multiple benchmarks and validate its geometric and appearance consistency. The generated videos can be converted into 3D models (e.g. 3DGS) for tasks requiring camera control.

2. Related Work

Reconstruction-based view synthesis. One way to synthesize videos is through 3D reconstruction and novel view synthesis, with NeRFs [44] and 3DGS [33] being popular methods from the past few years. However, even as follow-up work has improved rendering quality [3, 4, 67], speed [21, 45, 51], and has loosened the requirements on number of input views [14, 67, 72, 84] and poses [5, 11, 16, 27, 43], they are mostly confined to carefully curated captures. A smaller line of work extends these methods to in-the-wild setting [35, 37, 42, 62, 88] by optimizing from internet photos, but they require dense views, usually at least a few dozen, and preprocessing [56] to obtain camera parameters. Even then, the resulting methods lack the ability to fill in unseen

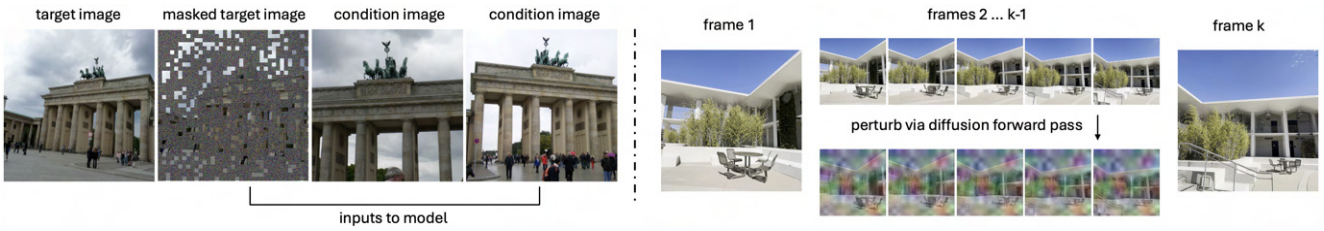

Figure 2. **Training objectives.** Left: Multiview inpainting. We provide n condition images and one target image to a diffusion model. We add noise to 80% of the target following the diffusion process. The condition images and remaining 20% of the target are kept clean. Note how some regions in the target are not seen in the conditions. The model learns priors such as symmetry to generate a plausible image. Right: View interpolation. We take k images from a video sequence and add noise to frame 2 to $k-1$ following the diffusion process. The model generates a sequence following a plausible camera path connecting the first and last frames.

viewpoints, leading to floaters and artifacts.

More recently, starting from LRM [25], transformer-based [66] feed-forward methods [28, 68, 75, 80, 89] aim to perform view synthesis and even reconstruction through learned priors. However, these methods require accurate camera poses during training and testing, and are currently limited to the object-level [13] or high-quality video captures [92].

Generation-based view synthesis. Finetuning image foundation models [22, 53] to learn 3D geometry has also proven effective [39, 40, 73, 79, 93]. For example, 3DiM [73] and Zero-1-to-3 [39] fine-tune diffusion models to generate novel views conditioned on input views and poses, leveraging and enhancing geometric priors that come from large-scale pretraining. Follow-up work like ZeroNVS [8, 55, 58, 65, 85] generalizes to full scenes, but jittering, blurriness, and inconsistency across viewpoints are common in their outputs. A number of methods [17, 52, 64, 86, 87] generates image sets or videos to improve consistency, but 3D supervision is still required during training. This limits scalability since camera parameters for in-the-wild data are typically obtained through SfM [56, 60], which can be computationally expensive and unreliable for sparse views.

Video generation and 3D learning. Video foundation models [2, 6, 7, 18, 23, 24, 30, 41, 59, 63, 82] have gained traction due to their potential for immersive storytelling. They also learn spatial and temporal priors useful for various downstream applications [9, 10, 17, 81, 87, 91]. For instance, VFusion3D [19] leverages the consistency in video models to generate 3D assets. 4DiM [74] jointly trains on video and 3D data for 4D reconstruction. In the same spirit, our work finetunes video foundation models with 3D-aware objectives and internet photos to scale up 3D learning in-the-wild.

Relevant to our method is frame interpolation [12, 15, 26, 31, 41, 50, 70]. However, these models are trained on frames with little to no camera motion, but focus on moving objects. Later, we show these methods cannot handle wide-baseline views nor produce realistic camera motion.

3. Method

We first provide an overview of keyframe-conditioned video generation in-the-wild (KFC-W). The goal is to generate a sequence of consistent frames given a handful (2-5) of unposed internet photos. One core challenge is the absence of supervised training pairs, i.e., internet photos and clean videos from the same scene that we can use as ground truth to resemble our input-output during inference. Our approach is we instead take *unpaired* corpora of internet photos and videos, and jointly train two subtasks on each kind of data with the same model: multiview inpainting and view interpolation, shown in Fig. 2.

For multiview inpainting, the only data annotation is that we take images from the same scene. Sources such as Wikimedia Commons provide millions of such multiview images with detailed indices and labels. Thus, we use MegaScenes [65], which contains 8M internet photos from 430k scenes labeled by Wikimedia Commons (we do not use any 3D supervision in the dataset).

For view interpolation, we use RealEstate10k [92] and DL3DV [38]. Both datasets capture sequences of frames within a short time frame, without noticeable changes in illumination. They also do not contain dynamic objects, which are outside the scope of this paper.

Both objectives are jointly trained on a latent Diffusion Transformer (DiT) [47, 53]. All images are passed through a pretrained VAE [34] encoder before further processing, including patchifying and adding noise based on diffusion processes. The image patches are then passed through the transformer. This process is shown in Fig. 3.

Next, we explain our self-supervised approach (Sec. 3.1, 3.2) and inference and training details (Sec. 3.3, 3.4).

3.1. Multiview Inpainting

We develop a task that 1) learns 3D priors without 3D annotations, and 2) learns from unstructured image collections such as internet photos.

The model takes a variable number of condition images, captured from random and wide-baseline viewpoints of a

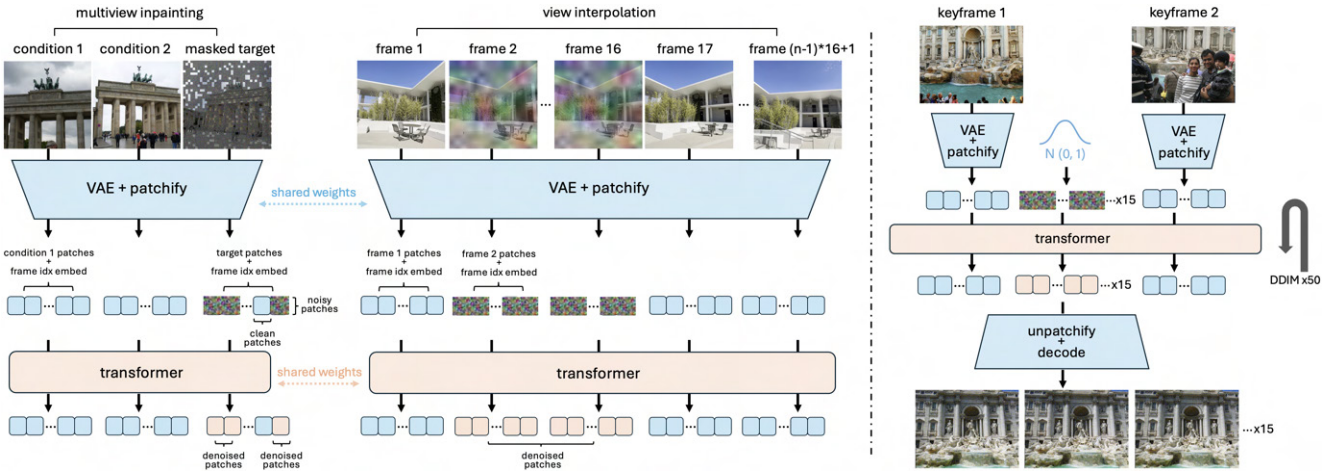

Figure 3. Multiview inpainting of internet photos and view interpolation of videos can be unified under the same denoising objective. Left: **Training.** We denoise the noisy patches (masked patches in multiview inpainting and intermediate frames in view interpolation), while extracting visual information from clean patches (blue patches) via self-attention. Then, we calculate a loss between the denoised (orange) and ground-truth patches. This process operates in latent space.
Right: **Inference.** Given unposed images of the same scene, we initialize and denoise a fixed number of frames via DDIM.

scene, and inpaints an 80% masked target. Through this process, it learns to extract structural information and scene identity from the condition images, and the illumination and scene layout from the remaining 20% of the target, allowing the model to fill in the target image accurately. In the example in Fig. 2, the target image is taken from an angle rotated counter-clockwise, where some of the structure is not directly observed in the conditions. The model must understand priors such as symmetry to fill in the target. This is important for our final task, as input keyframes are sparse and the model must plausibly fill in content unseen in inputs.

This approach is inspired by CroCo [76, 77], which observed that cross-view masked image modeling teaches strong 3D priors. Our implementation differs from CroCo in a few ways. CroCo only operates on two images (i.e. cross-view) because it creates a transformer decoder for each image, which becomes memory-heavy for multiple images. It obtains training pairs from in-the-wild images using camera poses and heuristics. It uses a deterministic MAE [20] objective which does not allow for probabilistic generation. In contrast, our method leverages self-attention for all images to take an arbitrary number of inputs, requires minimal annotations on data, and generates diverse outputs.

We implement the patchify and masking operations in latent space [34, 53] since we use a latent DiT. To all patches of the same image, we add a frame index embedding (since transformers are order-agnostic). For three images, the frame index embeddings would be $(0, 1, 2)$ passed through a linear layer. The patches of the condition images and 20% of the target are kept clean, while we apply the diffusion forward process to the remaining 80%. We only calculate the loss between these 80% of patches with ground truth. We also use a frozen semantic segmentation model on the RGB images to ignore transient objects such as people and vehicles; see our supplement for details.

3.2. View Interpolation

This task teaches the model to produce smooth, consistent frames given start and end condition images.

From a video, we randomly sample $16 \times (n-1)+1$ sequential frames, where n is the number of condition images and $2 \leq n \leq 5$. Every 16th frame is a condition image; i.e. the first frame and 17th frame are condition images, and the 15 intermediate frames in between two conditions are targets that we add noise to using the forward diffusion process. To all patches of the same image, we add a frame index embedding. We simply pass the order of the images $(0, 1, 2, ..., (n-1) * 16)$ through a linear layer. We experimented with normalizing frame indices as well as learnable positional encodings, but found no difference. This simple objective alone teaches the model to interpolate between input images, though as we show in Sec. 4 as an ablation, it is not sufficient for wide-baseline and in-the-wild images.

One important aspect of generating consistent in-the-wild videos is controlling illumination. During training, we condition every image on its own CLIP embedding [49]. We pass all images through a frozen CLIP encoder, and reshape the global feature map to the same shape as the image patches (recall these image patches come from encoding the RGB images through a VAE, then patchifying). Then, we simply perform addition before passing the patches into the transformer. We also experimented with modulation [48] but did not observe any differences. During inference, we condition all initialized noisy frames on a CLIP embedding of an

image with a desired illumination.

However, this alone does not force the model to use these embeddings, because the model can simply extract the illumination of the condition images, which, in videos, would be roughly the same as the targets'. Thus, we apply extreme color jittering to the condition images using PyTorch's `ColorJitter` transformation (details in supplement). The model must then extract illumination features from the CLIP embeddings to denoise the intermediate frames, since there is no other source that provides this information. Shown in Fig. 4, this mechanism allows us to specify a desired illumination during inference. In the top two rows, we randomly select one image (red-bordered), and condition all intermediate frames on its CLIP embedding. The output frames reflect the illumination of the red-bordered image, and remain consistent throughout the sequence. On the other hand, when we do not condition on CLIP embeddings (bottom row), generated frames do not maintain a consistent appearance.

It is worth noting that CLIP embeddings likely only contain coarse information, since its training captions contain terms like "cloudy," "sunny," rather than physical properties such as sun angles. Thus, even though we show that this method is capable of controlling coarse illumination, such as the general color scale, there are many possibilities for future work for fine-grained control.

Finally, we simulate segmentation by randomly masking regions of the condition images, since our input keyframes during inference will also be segmented for transient objects. See our supplement for details.

3.3. Inference

We show our inference pipeline on the right side of Fig. 3. It combines the two training objectives, with internet photos as input conditions (keyframes), and intermediate video frames as output. More formally, our input is a set of unposed images of the same scene $\mathbf{x} = \{x_1, ..., x_n\}$ where $2 \leq n \leq 5$. Our model generates 15 frames in between each input pair (x_i, x_{i+1}). We concatenate the frames into a video sequence $\mathbf{y} = \{y_1, y_2, ..., y_k\}$ where $\{y_1 ... y_{15}\}$ represents a video path between x_1 and x_2, and $k = 15 \times (n-1)$. Thus, the order of the inputs affects the video sequence. To the inputs, we perform segmentation to ignore transient objects such as people and vehicles. To the intermediate frames, we condition them on a CLIP embedding. If not otherwise specified, we use the embedding of the first input.

We run 50 DDIM [61] steps and decode only the intermediate frames. When showing generated videos, we do not include input keyframes as they may contain occlusions.

3.4. Training Details

We use a latent Diffusion Transformer (DiT) [47, 53]. When training on view interpolation we input sequences of 17, 33, 49, or 65 frames, corresponding to 2, 3, 4, and 5 keyframes, respectively. Our approach extends to longer videos or denser interpolated frames at the cost of compute (i.e., transformer sequence length). We chose 15 intermediate frames as a balance between compute and visual quality.

Figure 4. Top two rows: We control illumination by conditioning on the CLIP embedding of the red-bordered image during inference. Bottom: Without this condition, illumination varies across frames.

An interesting observation is the model's ability to infer which task to perform based on the format of the input. We do not provide flags for context switching, but simply pass in a number of images corresponding to the objective. Intuitively, both objectives are sufficiently similar to the model, as condition images convey scene identity, and CLIP embeddings supply illumination information. The only difference is the output format and length which can be determined based on the frame indices.

We finetune an internal pretrained text-to-video model, though our approach can be applied to any text-to-image or text-to-video model based on a transformer architecture. We inject our extra condition information – frame index embedding and CLIP embedding – by simply adding them to image patch tokens. Observations suggest that the specific model and checkpoint we use is comparable to state-of-the-art open-source video models such as CogVideoX-5B [82]. We train our model using MegaScenes, Re10k, and DL3DV on 32 NVIDIA A100 80G GPUs for 3 days. In the supplement, we provide details for reproducing our method.

4. Experiments and Evaluation

In the following experiments, we test on the Phototourism dataset [29] from the Image Matching Challenge, which contains images of 21 landmarks around the world collected from the internet, as well as the official test split of RealEstate10k (Re10k) [92].

From each scene in the Phototourism dataset, we randomly sample 3 sets each of 2, 3, 4, and 5 views, leading to 12 sets of sparse views per scene, and 252 sets in total. We call this our *Phototourism* test set. From the first 50 scenes from the Re10k test set, we randomly sample 3 sets each of 2, 3, 4, and 5 views, leading to 12 sets of sparse views per scene, and 600 sets in total. We call this our *Re10k* test set.

We perform the following evaluations:

1. We run video generation conditioned on the testing

Table 1. User study results. For a given scene, users vote between ours and a baseline. Our method outperforms commercial models such as Luma on all three criteria.

	Win Rate of Ours (Full) on Phototourism		
vs.	Consistency	Camera Motion	Aesthetics
FLAVR [31]	100.0%	100.0%	100.0%
LDMVFI [12]	100.0%	100.0%	100.0%
FILM [50]	100.0%	100.0%	100.0%
Ours (Video-Only)	100%	96.67%	96.67%
Luma [41]	60.21%	73.63%	60.21%

	Win Rate of Ours (Full) on Re10k		
vs.	Consistency	Camera Motion	Aesthetics
FLAVR [31]	100.0%	100.0%	100.0%
LDMVFI [12]	100.0%	100.0%	100.0%
FILM [50]	99.50%	81.00%	99.50%
Ours (Video-Only)	76.55%	71.31%	77.02%
Luma [41]	83.65%	84.86%	68.27%

images across multiple baselines, and conduct a user study that ask users to express a preference between pairs of results according to three separate criteria: "Consistency," "Camera path," and "Aesthetics." We show the user study interface and detailed descriptions of each criterion in Fig. 8.

2. We validate the consistency of our generated frames in 3D geometry and appearance using two downstream applications. First, we run COLMAP [56] on the original sparse views, then include our generated frames. This experiment tests whether the generated frames are geometrically consistent with the original views and provide support for feature correspondences. Second, we optimize a 3D scene with 3D Gaussian Splatting (3DGS) [33] on the original sparse views, then on our generated frames. 3DGS minimizes a rendering-based reconstruction loss and requires input images be consistent in appearance.

4.1. Video Interpolation and Generation

Baselines. We compare to the following methods: FILM [50], FLAVR [31], LDMVFI [12], and Luma Dream Machine [41]. The first three are open-source frame interpolation methods, while Luma is a commercial video generation model and we use its paid Dream Machine API. We note that, to the best of our understanding, Luma is the only publicly accessible large-scale model that is capable of interpolating between wide-baseline views. Stable Video Diffusion [6], Pika Labs [36], Emu Video [18], and CogVideoX [82] do not support this feature. Runway's Gen-3 Alpha [54] supports "frame-interpolation," where input images must have little to no camera motion. We were not able to produce any reasonable results with it on our task.

For methods that can only take two input images, we simply concatenate the videos from sequential keyframes.

We also add an additional baseline: Ours (video-only). This is our method without the multiview inpainting training objective, trained solely on the view interpolation objective

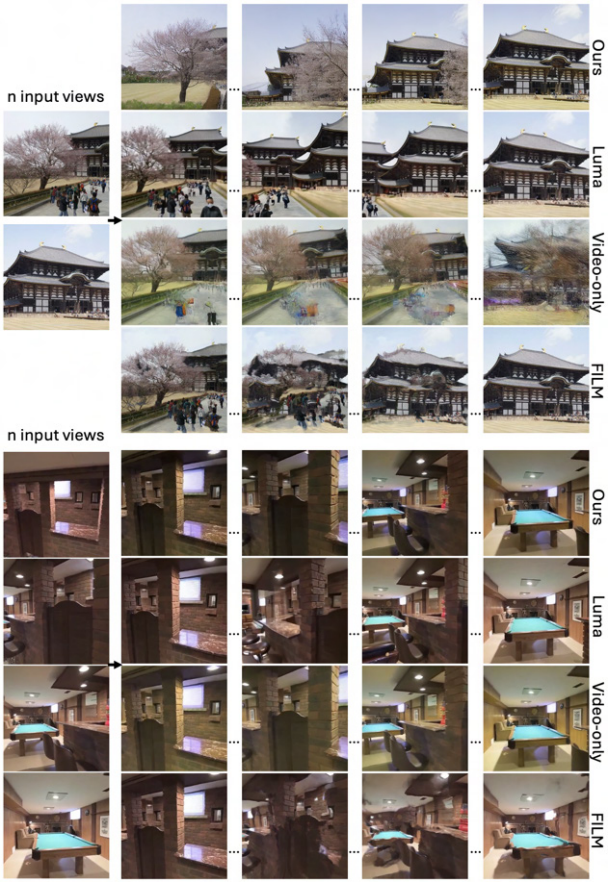

Figure 5. From top to bottom: Ours (Full), Luma, Ours (Video-only), FILM. Luma hallucinates new buildings (top scene) and produces jittering motions (bottom). Ours (Video-only) generates consistent videos on Re10k but not Phototourism. FILM is unable to handle wide-baseline inputs and produces blurry transitions.

using DL3DV and Re10k, with all other training details kept identical. This ablation allows us to understand the effect of the multiview inpainting objective. It also serves as a comparison to a second large-scale video generation model (apart from Luma), though finetuning is necessary because our internal model was pretrained with text conditions only. We denote our full method Ours (Full).

User study setup. For our study, we randomly sampled 25 scenes: 15 from the Phototourism dataset, and 10 from Re10k. Our method is compared to each baseline via pairwise comparisons. Users are shown two videos generated from the same input frames and are asked to select which method they prefer according to each evaluation criterion (or the user can select "Cannot Decide"). We show a visualization of the user study interface and detailed descriptions of each criterion in the supplement. When tallying the results, a direct vote counts as 1 point, and a "Cannot Decide" option counts as 0.5 votes each. For instance, two identical videos should both get 0.5 votes on each criterion, leading to a 50%

preference rate for that specific matchup. We calculate how often our method is preferred over each baseline, shown in Tab. 1, with results on the two datasets shown separately.

Comparing to frame interpolation methods. For comparisons to frame interpolation baselines (FILM, FLAVR, LDMVFI), we collected responses from 10 users, with a total of 150 votes on the Phototourism test set, and 100 votes on the Re10k test set.

As shown in Tab. 1, none of the baselines produce competitive video sequences on the Phototourism dataset. Their training data consists of small baselines or fixed cameras, and are trained to model dynamics rather than camera motion. On the Re10k dataset, only FILM can produce logical generated frames (and only when the camera motion is small), leading to a few "Cannot Decide" votes. See Fig. 5 (bottom scene is from Re10k) and our website for examples. .

Comparing to our ablated model and Luma. For comparisons to Ours (Video-only) and Luma, we collected reponses from 42 users, with a total of 587 votes on the Phototourism test set, and 340 votes on the Re10k test set.

Our video-only baseline (i.e., our model trained only with view interpolation) works well on the Re10k test set, as expected. Images from Re10k scenes are generally closer together, and the model has seen imagery from this domain during training. Many videos are nearly identical to Ours (Full), leading to a number of "Cannot Decide" ratings, although viewpoints with greater distances still led to flickering. On the other hand, this method cannot handle the wide-baselines and variability that appear in the Phototourism dataset, even though we used augmentation techniques such as color jittering and random masking during training. This demonstrates that introducing internet photos into training is crucial for dealing with in-the-wild scenes.

For Luma, their Dream Machine API allows users to specify a prompt as well as a start and end keyframe. We set the prompt to "consistent illumination and smooth camera path" for all videos. Luma, as a commercial model, is sampled more densely and at a higher resolution, and trained on substantially more data on a larger model than ours. However, our method outperforms Luma on all metrics. Luma struggles in several ways on this task: there are noticeable illumination changes as Luma transitions from one source image to the next ("Consistency"). Luma also tends to hallucinate new buildings and structures ("Camera path"), shown in Fig. 1 and Fig. 5. This indicates that Luma does not understand the layout of the scene, and is instead generating morphing effects to match the condition images. We believe that Luma achieves a lower preference rate on "Aesthetics" due in part to these artifacts. In our own observations, Luma's frames can sometimes appear sharper, and dynamic objects such as people walking are modeled more realistically. This led Luma to be preferred for "Aesthetics" in many (but not most) comparisons.

Figure 6. We run an ablation "Long-Video" with only the view-interpolation objective, but trained on inputs with up to 5x wider baselines. It fails to generalize to inputs with minimal overlap, while Ours (Full) still understands the scene layout.

Table 2. Comparison of COLMAP reconstructions when run on original sets of sparse internet photos vs. adding in generated frames. Ours (Full) generates geometrically consistent images that provide feature correspondences.

COLMAP SfM	Success Rate	Registered Images
Only Internet Photos	115/252	378/882
+ Generated views (Full)	**235**/252	**741**/882
+ Generated views (Video-only)	179/252	589/882

Surprisingly, Luma was rated as worse on Re10k scenes than on the Phototourism scenes, according to how often our method is preferred. Upon observation, we saw that its artifacts tend to be more noticable in indoor scenes, marked by issues such as disappearing objects. Furthermore, Luma produces random jittering motions when the input viewpoints are very nearby, possibly to fill time. Please refer to our website for side-by-side comparisons of generated videos.

These observations show our multiview inpainting objective teaches our model to be 3D-aware and understand depth and symmetry, leading to more realistic video sequences.

Multiview inpainting improves general view interpolation. Since Ours (Video-only) works well on Re10k, we verify whether training only on videos, but with wide-baseline inputs, can replace multiview inpainting and internet photos. Thus, we run an ablation similar to Ours (Video-only), but with up to 5x wider baselines as conditions during training. Shown in Fig. 6, this model ("Long-Video") fails on inputs with minimal overlap, while Ours (Full) still understands the scene layout. This shows multiview inpainting combined with internet photos improves general view interpolation, not just on in-the-wild data. We argue that internet photos contain diverse viewpoints, such as extreme rotations and zooming, that cannot be easily learned from videos.

Table 3. We run InstantSplat on sparse input images vs our generated frames and compare rendering metrics on 10 random images from each scene. Our method produces densely sampled frames while being consistent in illumination and geometry, leading to improved metrics.

	Phototourism						Re10k					
	On Internet Photos			On Generated Frames			On Original Frames			On Generated Frames		
Method	PSNR(↑)	SSIM(↑)	LPIPS(↓)	PSNR(↑)	SSIM(↑)	LPIPS(↓)	PSNR(↑)	SSIM(↑)	LPIPS(↓)	PSNR(↑)	SSIM(↑)	LPIPS(↓)
InstantSplat	11.701	0.3510	0.5703	**13.960**	**0.4123**	**0.4864**	19.857	0.7269	0.2663	**21.798**	**0.7916**	**0.2190**

Figure 7. We run InstantSplat [14] on original input views and on generated frames. The rendered results on generated frames exhibit fewer artifacts and sharper content.

4.2. Application 1: SfM Reconstruction

We validate whether our generated frames are consistent in geometry, and therefore suitable for downstream applications such as 3D reconstruction. We run COLMAP [56] on the original input views, then include our generated frames. COLMAP is unreliable on sparse views because when viewpoints are too far apart, COLMAP struggles to find sufficient matching features between pairs of images. As shown in Tab. 2, only 45% of sets of sparse views are successfully reconstructed ("Only Internet Photos"). Of the 882 images total, 43% were registered. On the other hand, COLMAP successfully reconstructed 93% of sets of views when we included our video frames, which were generated from the same sets of sparse views ("+ Generated views (Full)"). Now, 84% of the original frames could be registered, nearly doubling the amount. This shows our generated frames provide reliable feature correspondences that connect distant views.

However, not all view interpolation methods achieve this effect. We also run this experiment by adding generated frames from the "Ours (Video-Only)" baseline, denoted "+ Generated views (Video-only)". There was significantly less improvement: 71% success rate and 67% registered images.

4.3. Application 2: 3D Gaussian Splatting

We experiment with running 3D Gaussian Splatting (3DGS) on our generated frames. In contrast to COLMAP, which is robust to illumination changes, the input images for (vanilla) 3DGS must be consistent in appearance and illumination. To validate our method, we compare running 3DGS on original inputs vs. our generated frames (without the original inputs). We use InstantSplat [14], a 3DGS method that builds on COLMAP-Free-3DGS [16] and DUSt3R [69] to generate Gaussian Splats given sparse, unposed images.

We show results in Tab. 3 and Fig. 7. "On Internet Photos" and "On Original Frames" refer to training each scene using the original sparse views in the test sets, while "On Generated Frames" refer to using our model's output when conditioned on those same sparse views.

For each scene, we train with the default settings in the InstantSplat open-source code. Then, we sample 10 images not used for training from the scene, register their poses to the same coordinate frame, and render them using the trained model to report PSNR, SSIM [71], and LPIPS [90] metrics. We provide details in the supplement.

Internet photos from the Phototourism dataset have wide baselines, significant occlusions, and varying illumination, which make it very difficult to train 3DGS methods based on a pixel-rendering loss. Our generated frames are denser and with more consistent illumination, leading to substantial improvements in reconstruction metrics. Even though the Re10K dataset has similar illumination conditions and smaller baselines across frames, we observe an improvement across all metrics as a result of training with denser frames. We provide rendered sequences on the website.

Note that the metrics fo the Phototourism dataset are much lower than those for Re10k because the testing ground-truth is also internet photos, which lead to high rendering losses (see the ground-truth column in Fig. 7).

Through these applications, we show that video models trained with 3D-aware objectives can be useful as 3D priors for various downstream tasks.

5. Future Work and Discussion

Potential future work include scaling up data to model dynamic objects, enforcing a fine-grained constraint on illumination, and extrapolating to unseen views. We posit that brute-force scaling will not help video models understand the physical world. However, rather than incorporating conditions that can be difficult to estimate at scale, we jointly train a scalable 3D-aware objective. We suggest that this concept can be applied to other tasks as well, such as modeling motion that respects physical constraints [32].

References

[1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Apostol (Paul) Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. In *arXiv:1609.08675*, 2016. 2

[2] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Yuanzhen Li, Tomer Michaeli, Oliver Wang, Deqing Sun, Tali Dekel, and Inbar Mosseri. Lumiere: A space-time diffusion model for video generation. In *Proceedings of the 41st International Conference on Machine Learning*, 2024. 3

[3] Jonathan T. Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P. Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields, 2021. 2

[4] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Zip-nerf: Anti-aliased grid-based neural radiance fields. *ICCV*, 2023. 2

[5] Wenjing Bian, Zirui Wang, Kejie Li, Jiawang Bian, and Victor Adrian Prisacariu. Nope-nerf: Optimising neural radiance field with no pose prior. In *CVPR*, 2023. 2

[6] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 1, 3, 6

[7] 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. 3

[8] Shengqu Cai, Eric Ryan Chan, Songyou Peng, Mohamad Shahbazi, Anton Obukhov, Luc Van Gool, and Gordon Wetzstein. Diffdreamer: Towards consistent unsupervised single-view scene extrapolation with conditional diffusion models. In *ICCV*, 2023. 3

[9] Boyuan Chen, Diego Marti Monso, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion, 2024. 3

[10] Zilong Chen, Yikai Wang, Feng Wang, Zhengyi Wang, and Huaping Liu. V3d: Video diffusion models are effective 3d generators. *arXiv preprint arXiv:2403.06738*, 2024. 1, 3

[11] Zezhou Cheng, Carlos Esteves, Varun Jampani, Abhishek Kar, Subhransu Maji, and Ameesh Makadia. Lu-nerf: Scene and pose estimation by synchronizing local unposed nerfs. In *CVPR*, 2023. 2

[12] Duolikun Danier, Fan Zhang, and David Bull. LDMVFI: Video frame interpolation with latent diffusion models. In *AAAI*, 2024. 2, 3, 6

[13] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, Eli VanderBilt, Aniruddha Kembhavi, Carl Vondrick, Georgia Gkioxari, Kiana Ehsani, Ludwig Schmidt, and Ali Farhadi. Objaverse-xl: A universe of 10m+ 3d objects. *arXiv preprint arXiv:2307.05663*, 2023. 2, 3

[14] Zhiwen Fan, Wenyan Cong, Kairun Wen, Kevin Wang, Jian Zhang, Xinghao Ding, Danfei Xu, Boris Ivanovic, Marco Pavone, Georgios Pavlakos, Zhangyang Wang, and Yue Wang. Instantsplat: Unbounded sparse-view pose-free gaussian splatting in 40 seconds, 2024. 2, 8

[15] Haiwen Feng, Zheng Ding, Zhihao Xia, Simon Niklaus, Victoria Abrevaya, Michael J. Black, and Zhang Xuaner. Explorative in-betweening of time and space. In *European Conference on Computer Vision*, 2024. 3

[16] Yang Fu, Sifei Liu, Amey Kulkarni, Jan Kautz, Alexei A. Efros, and Xiaolong Wang. Colmap-free 3d gaussian splatting. In *CVPR*, 2024. 2, 8

[17] Jeong gi Kwak, Erqun Dong, Yuhe Jin, Hanseok Ko, Shweta Mahajan, and Kwang Moo Yi. Vivid-1-to-3: Novel view synthesis with video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6775–6785, 2024. 1, 3

[18] Rohit Girdhar, Mannat Singh, Andrew Brown, et al. Emu video: Factorizing text-to-video generation by explicit image conditioning. *arXiv preprint arXiv:2311.10709*, 2023. 1, 3, 6

[19] Junlin Han, Filippos Kokkinos, and Philip Torr. Vfusion3d: Learning scalable 3d generative models from video diffusion models. *European Conference on Computer Vision (ECCV)*, 2024. 1, 3

[20] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv:2111.06377*, 2021. 4

[21] Peter Hedman, Pratul P. Srinivasan, Ben Mildenhall, Jonathan T. Barron, and Paul Debevec. Baking neural radiance fields for real-time view synthesis. *ICCV*, 2021. 2

[22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arxiv:2006.11239*, 2020. 3

[23] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, and David J. Fleet. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 3

[24] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022. 1, 3

[25] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. In *ICLR*, 2024. 3

[26] Siddhant Jain, Daniel Watson, Eric Tabellion, Ben Poole, Janne Kontkanen, et al. Video interpolation with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7341–7351, 2024. 3

[27] Kaiwen Jiang, Yang Fu, Mukund Varma T, Yash Belhe, Xiaolong Wang, Hao Su, and Ravi Ramamoorthi. A construct-optimize approach to sparse view synthesis without camera pose. *SIGGRAPH*, 2024. 2

[28] Haian Jin, Hanwen Jiang, Hao Tan, Kai Zhang, Sai Bi, Tianyuan Zhang, Fujun Luan, Noah Snavely, and Zexiang Xu. Lvsm: A large view synthesis model with minimal 3d inductive bias, 2024. 3

[29] Yuhe Jin, Dmytro Mishkin, Anastasiia Mishchuk, Jiri Matas, Pascal Fua, Kwang Moo Yi, and Eduard Trulls. Image Matching across Wide Baselines: From Paper to Practice. *International Journal of Computer Vision*, 2020. 5

[30] Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling, 2024. 1, 3

[31] Tarun Kalluri, Deepak Pathak, Manmohan Chandraker, and Du Tran. Flavr: Flow-agnostic video representations for fast frame interpolation. In *WACV*, 2023. 2, 3, 6

[32] Bingyi Kang, Yang Yue, Rui Lu, Zhijie Lin, Yang Zhao, Kaixin Wang, Gao Huang, and Jiashi Feng. How far is video generation from world model? – a physical law perspective, 2024. 8

[33] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023. 2, 6

[34] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3, 4

[35] Jonas Kulhanek, Songyou Peng, Zuzana Kukelova, Marc Pollefeys, and Torsten Sattler. Wildgaussians: 3d gaussian splatting in the wild. *arXiv*, 2024. 2

[36] Pika Labs. Pika labs: Ai video generation platform. https://pika.art/, 2024. Accessed: 2024-11-10. 6

[37] Zhengqi Li, Wenqi Xian, Abe Davis, and Noah Snavely. Crowdsampling the plenoptic function. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 178–196. Springer, 2020. 2

[38] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. *arXiv preprint arXiv:2312.16256*, 2023. 3

[39] 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 *ICCV*, 2023. 2, 3

[40] 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. *arXiv preprint arXiv:2309.03453*, 2023. 3

[41] LUMA. Luma dream machine, 2024. 1, 2, 3, 6

[42] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7210–7219, 2021. 2

[43] Andreas Meuleman, Yu-Lun Liu, Chen Gao, Jia-Bin Huang, Changil Kim, Min H. Kim, and Johannes Kopf. Progressively optimized local radiance fields for robust view synthesis. In *CVPR*, 2023. 2

[44] 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. *ICCV*, 2021. 2

[45] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, 2022. 2

[46] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 1

[47] William Peebles and Saining Xie. Scalable diffusion models with transformers. *arXiv preprint arXiv:2212.09748*, 2022. 3, 5

[48] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville. Film: Visual reasoning with a general conditioning layer. In *AAAI*, 2018. 4

[49] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 4

[50] Fitsum Reda, Janne Kontkanen, Eric Tabellion, Deqing Sun, Caroline Pantofaru, and Brian Curless. Film: Frame interpolation for large motion. In *European Conference on Computer Vision (ECCV)*, 2022. 1, 2, 3, 6

[51] Christian Reiser, Songyou Peng, Yiyi Liao, and Andreas Geiger. Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps. In *CVPR*, 2021. 2

[52] Xuanchi Ren and Xiaolong Wang. Look outside the room: Synthesizing a consistent long-term 3d scene video from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[53] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 3, 4, 5

[54] Inc. Runway AI. Introducing gen-3 alpha: A new frontier for video generation. https://runwayml.com/research/introducing-gen-3-alpha, 2024. Accessed: 2024-11-10. 6

[55] 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 real image. In *CVPR*, 2024. 3

[56] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2, 3, 6, 8

[57] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 2

[58] 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. *arXiv preprint arXiv:2405.17251*, 2024. 3

[59] Uriel Singer, Adam Polyak, Thomas Hayes, Dafna Shaham, Chitwan Saharia, William Chan, and Mohammad Norouzi. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 3

[60] Noah Snavely, Steven M. Seitz, and Richard Szeliski. Photo tourism: exploring photo collections in 3d. *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 2006. 3

[61] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv:2010.02502*, 2020. 5

[62] Jiaming Sun, Xi Chen, Qianqian Wang, Zhengqi Li, Hadar Averbuch-Elor, Xiaowei Zhou, and Noah Snavely. Neural 3d reconstruction in the wild. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–9, 2022. 2

[63] The Movie Gen team @ Meta. Movie gen: A cast of media foundation models. arXiv preprint arXiv:2410.13720, 2024. 3

[64] Hung-Yu Tseng, Qinbo Li, Changil Kim, Suhib Alsisan, Jia-Bin Huang, and Johannes Kopf. Consistent view synthesis with pose-guided diffusion models. In *CVPR*, 2023. 3

[65] Joseph Tung, Gene Chou, Ruojin Cai, Guandao Yang, Kai Zhang, Gordon Wetzstein, Bharath Hariharan, and Noah Snavely. Megascenes: Scene-level view synthesis at scale. In *ECCV*, 2024. 2, 3

[66] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. 3

[67] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T. Barron, and Pratul P. Srinivasan. Ref-NeRF: Structured view-dependent appearance for neural radiance fields. In *CVPR*, 2022. 2

[68] Peng Wang, Hao Tan, Sai Bi, Yinghao Xu, Fujun Luan, Kalyan Sunkavalli, Wenping Wang, Zexiang Xu, and Kai Zhang. Pf-lrm: Pose-free large reconstruction model for joint pose and shape prediction. *arXiv preprint arXiv:2311.12024*, 2023. 3

[69] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, 2024. 8

[70] Xiaojuan Wang, Boyang Zhou, Brian Curless, Ira Kemelmacher, Aleksander Holynski, and Steve Seitz. Generative inbetweening: Adapting image-to-video models for keyframe interpolation. *arXiv preprint arXiv:2408.15239*, 2024. 2, 3

[71] 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. 8

[72] Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. Nerf–: Neural radiance fields without known camera parameters. *arXiv preprint arXiv:2102.07064*, 2021. 2

[73] 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. 3

[74] Daniel Watson, Saurabh Saxena, Lala Li, Andrea Tagliasacchi, and David J Fleet. Controlling space and time with diffusion models. *arXiv preprint arXiv:2407.07860*, 2024. 2, 3

[75] Xinyue Wei, Kai Zhang, Sai Bi, Hao Tan, Fujun Luan, Valentin Deschaintre, Kalyan Sunkavalli, Hao Su, and Zexiang Xu. Meshlrm: Large reconstruction model for high-quality mesh. *arXiv preprint arXiv:2404.12385*, 2024. 3

[76] Philippe Weinzaepfel, Thomas Lucas, Vincent Leroy, Yohann Cabon, Vaibhav Arora, Romain Brégier, Gabriela Csurka, Leonid Antsfeld, Boris Chidlovskii, and Jérôme Revaud. CroCo v2: Improved Cross-view Completion Pre-training for Stereo Matching and Optical Flow. In *ICCV*, 2023. 4

[77] Weinzaepfel, Philippe and Leroy, Vincent and Lucas, Thomas and Brégier, Romain and Cabon, Yohann and Arora, Vaibhav and Antsfeld, Leonid and Chidlovskii, Boris and Csurka, Gabriela and Revaud Jérôme. CroCo: Self-Supervised Pre-training for 3D Vision Tasks by Cross-View Completion. In *NeurIPS*, 2022. 4

[78] Wenming Weng, Ruoyu Feng, Yanhui Wang, Qi Dai, Chunyu Wang, Dacheng Yin, Zhiyuan Zhao, Kai Qiu, Jianmin Bao, Yuhui Yuan, Chong Luo, Yueyi Zhang, and Zhiwei Xiong. Art•v: Auto-regressive text-to-video generation with diffusion models. *arXiv preprint arXiv:2311.18834*, 2023. 2

[79] Rundi Wu, Ben Mildenhall, Philipp Henzler, Keunhong Park, Ruiqi Gao, Daniel Watson, Pratul P Srinivasan, Dor Verbin, Jonathan T Barron, Ben Poole, et al. Reconfusion: 3d reconstruction with diffusion priors. *arXiv preprint arXiv:2312.02981*, 2023. 3

[80] Desai Xie, Sai Bi, Zhixin Shu, Kai Zhang, Zexiang Xu, Yi Zhou, Sören Pirk, Arie Kaufman, Xin Sun, and Hao Tan. Lrm-zero: Training large reconstruction models with synthesized data. *arXiv preprint arXiv:2406.09371*, 2024. 3

[81] Sherry Yang, Jacob C Walker, Jack Parker-Holder, Yilun Du, Jake Bruce, Andre Barreto, Pieter Abbeel, and Dale Schuurmans. Position: Video as the new language for real-world decision making. In *Proceedings of the 41st International Conference on Machine Learning*, pages 56465–56484. PMLR, 2024. 3

[82] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 1, 3, 5, 6

[83] Shengming Yin, Chenfei Wu, Huan Yang, Jianfeng Wang, Xiaodong Wang, Minheng Ni, Zhengyuan Yang, Linjie Li,

Shuguang Liu, Fan Yang, Jianlong Fu, Ming Gong, Lijuan Wang, Zicheng Liu, Houqiang Li, and Nan Duan. Nuwa-xl: Diffusion over diffusion for extremely long video generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023)*, Toronto, Canada, 2023. Association for Computational Linguistics. 2

[84] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images, 2021. 2

[85] Hong-Xing Yu, Haoyi Duan, Junhwa Hur, Kyle Sargent, Michael Rubinstein, William T Freeman, Forrester Cole, Deqing Sun, Noah Snavely, Jiajun Wu, et al. Wonderjourney: Going from anywhere to everywhere. *arXiv preprint arXiv:2312.03884*, 2023. 3

[86] Jason J. Yu, Tristan Aumentado-Armstrong, Fereshteh Forghani, Konstantinos G. Derpanis, and Marcus A. Brubaker. Polyoculus: Simultaneous multi-view image-based novel view synthesis. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 3

[87] 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. *arXiv preprint arXiv:2409.02048*, 2024. 1, 3

[88] Dongbin Zhang, Chuming Wang, Weitao Wang, Peihao Li, Minghan Qin, and Haoqian Wang. Gaussian in the wild: 3d gaussian splatting for unconstrained image collections. *arXiv preprint arXiv:2403.15704*, 2024. 2

[89] Kai Zhang, Sai Bi, Hao Tan, Yuanbo Xiangli, Nanxuan Zhao, Kalyan Sunkavalli, and Zexiang Xu. Gs-lrm: Large reconstruction model for 3d gaussian splatting. In *European Conference on Computer Vision*, 2024. 3

[90] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 8

[91] Tianyuan Zhang, Hong-Xing Yu, Rundi Wu, Brandon Y. Feng, Changxi Zheng, Noah Snavely, Jiajun Wu, and William T. Freeman. PhysDreamer: Physics-based interaction with 3d objects via video generation. In *European Conference on Computer Vision*. Springer, 2024. 3

[92] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *ACM Trans. Graph. (Proc. SIGGRAPH)*, 37, 2018. 3, 5

[93] Zhizhuo Zhou and Shubham Tulsiani. Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12588–12597, 2023. 3