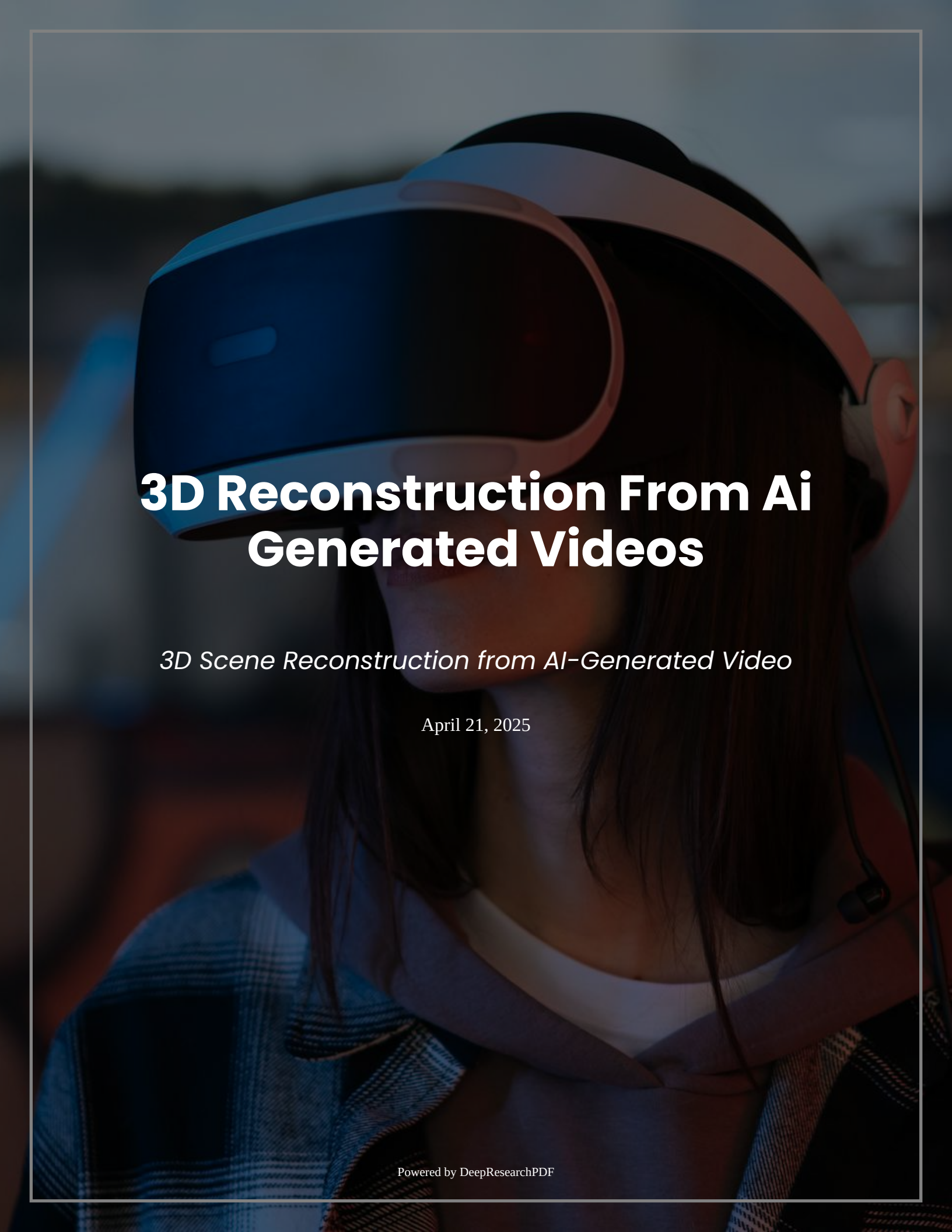

3D Reconstruction From Ai Generated Videos

3D Scene Reconstruction from AI-Generated Video

April 21, 2025

3D Scene Reconstruction from AI-Generated Video

Background: Text-to-Video Generation and 3D Consistency

Recent text-to-video (T2V) models like **Wan2.1** (Alibaba, 2025) and OpenAI's **Sora** (2024) can synthesize short video clips from text prompts. These models produce visually rich content, often a few seconds long (e.g. Wan2.1 generates ~5s of 480p video in a few minutes on consumer GPUs (GitHub - Wan-Video/Wan2.1: Wan: Open and Advanced Large-Scale Video Generative Models)). An important question is whether these AI-generated videos are *3D-consistent* – i.e. if the frames correspond to a coherent three-dimensional scene. OpenAI reported that Sora, when trained at scale, exhibits emergent 3D consistency: as the **camera** moves, objects and people in the scene maintain stable positions and shapes in 3D space (Video generation models as world simulators | OpenAI). This suggests it might be possible to recover a 3D scene from such a video. In fact, researchers have begun treating **video generation models as world simulators**, examining how well their outputs adhere to real-world physics and geometry (Sora Generates Videos with Stunning Geometrical Consistency). The idea is to leverage the frames of a generated video as multi-view images of a scene, then use 3D reconstruction techniques to build a spatial representation (like a point cloud or mesh) that can be viewed from novel angles.

Prior Work Combining Video Generation and 3D Reconstruction

Several early projects and research papers have explicitly explored pipelines that **generate a video from text and then reconstruct a 3D scene** from that video:

- **Geometry Benchmark via Reconstruction:** Li *et al.* (2024) introduced a benchmark to evaluate physical realism of AI videos by reconstructing them in 3D (Sora Generates Videos with Stunning Geometrical Consistency). They processed videos from models like Sora, Runway Gen-2, and Pika Labs through an off-the-shelf structure-from-motion (SfM) pipeline (COLMAP) and **3D Gaussian Splatting** reconstruction (Sora Generates Videos with Stunning Geometrical Consistency). The premise was that if a generated video truly respects real-world geometry (camera projection and rigid scene structure), it should yield a high-quality 3D model. In their setup, they did not customize COLMAP for AI content; they simply used SfM to compute camera poses and then ran a dense reconstruction (Gaussian splats) to get a 3D model (Sora Generates Videos with Stunning Geometrical Consistency). A higher number of feature matches and successful 3D points indicated better adherence to multi-view geometry (epipolar constraints) (Sora Generates Videos with Stunning Geometrical Consistency) (Sora Generates Videos with Stunning Geometrical Consistency). This work demonstrated that **some AI-generated videos can indeed be reconstructed into 3D models**, though quality varies with the video's consistency.

- **OpenAI Sora Experiments:** OpenAI's own Sora model outputs were tested in 3D reconstruction scenarios. Fu *et al.* (CVPR 2024) applied a *pose-free* reconstruction pipeline to Sora videos (COLMAP-Free 3D Gaussian Splatting). Their method, dubbed **COLMAP-Free 3D Gaussian Splatting**, does not require prior camera poses. Instead, it sequentially processes video frames, estimating camera trajectory and growing a set of 3D Gaussians (point primitives) on the fly (COLMAP-Free 3D Gaussian Splatting). They reported *encouraging results* – effectively turning Sora's text-generated videos into interactive 3D scenes (COLMAP-Free 3D Gaussian Splatting). The authors note this could enable "infinite 3D data from a few lines of text," hinting at the

- potential of text-to-video as a source of 3D training data or assets.

- **Community Prototype (Polycam):** In a practical demo, an author on Medium tried using **Polycam** (a mobile photogrammetry app) to convert Sora's sample videos into 3D models (Can OpenAI Sora Solve 3D Modeling? I Tried This… | by Cindy X. L. | Medium). Polycam likely extracted frames and applied a multi-view reconstruction (possibly leveraging COLMAP and Gaussian splatting under the hood). The results were mixed. If the AI video had a suitable camera motion (e.g. an *arc around an object*), a recognizable though fragmented 3D model could be obtained (Can OpenAI Sora Solve 3D Modeling? I Tried This… | by Cindy X. L. | Medium) (Can OpenAI Sora Solve 3D Modeling? I Tried This… | by Cindy X. L. | Medium). However, videos with only forward camera movement or inconsistent object appearance produced poor reconstructions (broken geometry or nonsensical output) (Can OpenAI Sora Solve 3D Modeling? I Tried This… | by Cindy X. L. | Medium) (Can OpenAI Sora Solve 3D Modeling? I Tried This… | by Cindy X. L. | Medium). This experiment highlighted key **quality factors**: the video must maintain object shape/texture across views, and the scene should remain mostly static (as if all frames were captured at the same moment) (Can OpenAI Sora Solve 3D Modeling? I Tried This… | by Cindy X. L. | Medium). Any violation (e.g. object changes, style flicker, or only seeing one side of an object) makes it hard to "stitch" a coherent model from the frames. The takeaway was that *yes*, some AI videos can be converted to 3D, but current text-to-video outputs were not optimized for reconstruction purposes (Can OpenAI Sora Solve 3D Modeling? I Tried This… | by Cindy X. L. | Medium).

- **ReCamMaster (2025):** Rather than explicitly reconstructing geometry, **ReCamMaster** offers a pipeline to *directly synthesize novel camera views from a single video*. Bai *et al.* (2025) developed this system to "re-capture" a given video with a new camera trajectory, effectively performing *viewpoint extrapolation*. In their research, they trained a video diffusion model with conditioning to control camera motion, enabling it to output a new video of the same scene from a different angle or path. While their proprietary model was not

- open-source, the authors released a version built on Wan2.1 to validate the idea ([GitHub - KwaiVGI/ReCamMaster: [ARXIV'25] ReCamMaster: Camera-Controlled Generative Rendering from A Single Video](https://github.com/KwaiVGI/ReCamMaster#:~:text=%E2%9A%99%EF%B8%8F%20Code%3A%20ReCamMaster%20%2B%20Wan2,Inference%20%26%20Training)). The open ReCamMaster code (using Wan2.1) allows users to input a video and choose a trajectory (pan, tilt, orbit, etc.), and the model generates a new video from that perspective. This is highly relevant because it **achieves novel view synthesis** of an AI-generated scene without explicitly producing a point cloud or mesh. Essentially, the text-to-video model itself is leveraged to imagine the scene from other angles. ReCamMaster's results showed plausible new views for many in-the-wild videos (and by extension could work on AI videos), though quality depends on the model's internal consistency. This approach indicates one way to avoid an intermediate reconstruction step: use generative networks to *render* new views directly. The limitation is that you don't get an explicit 3D representation for further use; the output is another video.

- **Text-to-Video with 3D Constraints:** A line of academic work is emerging that *fuses video generation with 3D reconstruction tasks during training*. One example is a 2025 paper by Harsh Jhamtani *et al.* (title: **JOint Generation and 3D Reconstruction**, placeholder name "\nameMethod"). They fine-tune a diffusion-based video generator (specifically an open-source Sora-like model) by introducing a secondary task: predicting a 3D point cloud and camera poses for the video (\nameMethod: Towards 3D-Consistent Video Generators). In their unified architecture, the model not only produces a video from a text prompt, but also learns to output a corresponding **3D point map** (via a depth/correspondence decoder) as it generates frames (\nameMethod: Towards 3D-Consistent Video Generators) (\nameMethod: Towards 3D-Consistent Video Generators). The added supervision (a multi-view photometric loss and pose consistency loss) forces the generator to become more 3D-consistent. As a result, the *trained model improved its 3D coherence significantly*, and could even estimate camera trajectories

- competitive with traditional SfM methods (\nameMethod: Towards 3D-Consistent Video Generators). This is a promising direction: the video model itself gains an understanding of 3D structure, making its outputs inherently easier to reconstruct. It effectively combines text-to-video **and** video-to-3D in one system. (Code is expected to be released upon paper acceptance.)

- **Frame Interpolation for Sparse Views:** Another notable work is by Gene Chou *et al.* (2024), who tackled *generating 3D-consistent videos from unposed images*. In their project (nicknamed **KFC-W**), they fine-tuned a text-to-video diffusion model on tasks like multiview inpainting and view interpolation using internet photo datasets (KFC-W) (KFC-W). Given a handful of unposed images (e.g. tourist photos of a landmark from different angles), their model can generate a smooth video that moves between those viewpoints. This is like doing "neural camera path interpolation." They showed that inserting the generated intermediate frames into a COLMAP reconstruction greatly improves the connectivity and completeness of the resulting point cloud (KFC-W). The model's frames provide reliable new correspondences between very different original views, yielding a better SfM solution. They also tried feeding the frames into a 3D Gaussian Splatting algorithm, finding that more consistent lighting and geometry (thanks to the AI frames) boosted reconstruction metrics (KFC-W). While this work starts from real images rather than text, it demonstrates the power of **AI-generated frames to aid 3D reconstruction**. A text-to-video model that is 3D-aware can act as a "multiview augmenter," filling in gaps in viewpoints.

- **Video-to-3D Distillation:** *Very* recently (2025), Wang *et al.* proposed **VideoScene (VideoScene: Distilling Video Diffusion Model to Generate 3D Scenes in One Step) (VideoScene: Distilling Video Diffusion Model to Generate 3D Scenes in One Step)**, which directly distills a pretrained video diffusion model into a 3D scene generator. They leverage a large text-to-video diffusion (CogVideoX) as a prior and guide it to produce a **NeRF-like representation** from a few input images. Their pipeline first uses a fast

- multi-view reconstruction module (a learned **MVSplat** network for sparse-view Gaussian splatting) to get a coarse 3D scene and render intermediate frames (VideoScene: Distilling Video Diffusion Model to Generate 3D Scenes in One Step) (VideoScene: Distilling Video Diffusion Model to Generate 3D Scenes in One Step). Those frames, along with the diffusion prior, then produce a refined video in one go, effectively generating a 3D-consistent video with novel views "baked in." This approach merges video generation and 3D understanding, aiming to **"bridge the gap from video to 3D" in one step** (VideoScene: Distilling Video Diffusion Model to Generate 3D Scenes in One Step). It's an advanced technique that indicates the trend of unifying these tasks for efficiency.

- **Dynamic Scene Reconstruction (4D):** Most efforts above assume a mostly static scene (only camera moves). But what if the video contains moving subjects (a person, an animal)? A very new development is **Vidu4D** by Han *et al.* (late 2024) (Vidu4D: Single Generated Video to High-Fidelity 4D Reconstruction with Dynamic Gaussian Surfels). Vidu4D is a reconstruction pipeline designed to take *a single AI-generated video* and recover a full **4D representation** – meaning it captures both the 3D geometry of the scene and the temporal evolution (motion) of dynamic elements (Vidu4D: Single Generated Video to High-Fidelity 4D Reconstruction with Dynamic Gaussian Surfels). They use a pretrained text-to-video model to generate a clip (examples include a cat turning its head, or a fantastical creature moving) (Vidu4D: Single Generated Video to High-Fidelity 4D Reconstruction with Dynamic Gaussian Surfels) (Vidu4D: Single Generated Video to High-Fidelity 4D Reconstruction with Dynamic Gaussian Surfels). Then, through two novel stages (initializing non-rigid warp fields and optimizing **Dynamic Gaussian Surfels**), they reconstruct each frame's geometry and link them over time (Vidu4D: Single Generated Video to High-Fidelity 4D Reconstruction with Dynamic Gaussian Surfels) (Vidu4D: Single Generated Video to High-Fidelity 4D Reconstruction with Dynamic Gaussian Surfels). The result is a **time-varying point-based model** that can render the scene from any viewpoint at any time, with high

- fidelity in both appearance and motion. Vidu4D addresses challenges like deformation and frame-by-frame distortion, which are prevalent in diffusion-generated videos (Vidu4D: Single Generated Video to High-Fidelity 4D Reconstruction with Dynamic Gaussian Surfels). Their samples show realistic novel-view renderings of moving subjects (Vidu4D: Single Generated Video to High-Fidelity 4D Reconstruction with Dynamic Gaussian Surfels). This work is on the cutting edge, effectively combining text-to-video, 3D reconstruction, and even animation capture. A *project page* and demos are available (Vidu4D: Single Generated Video to High-Fidelity 4D Reconstruction with Dynamic Gaussian Surfels), indicating rapid progress toward *text-to-4D* generation.

Several other related projects exist (e.g. Meta's MAV3D, "text to 3D video" demos, and text-to-4D pipelines like **4Dynamic** and **Motion4D** that integrate image, video, and 3D generation), but the examples above cover the key ideas in current research. In summary, there is a clear emerging workflow: **use a text-to-video model to generate a scene, then apply 3D reconstruction (classical or neural) to get a structured 3D/4D model, and finally render new views or even new motions**. Some researchers have even remarked that *as of 2024, the best way to get a text-to-3D scene is to use text-to-video followed by 3D reconstruction*, since direct text-to-3D is still maturing (Ben Poole on X: " the best current method for text-to-3d scenes is text ...).

3D Reconstruction Techniques for Monocular Video

Reconstructing a 3D scene from a monocular video is a long-studied problem in computer vision. In the context of AI-generated videos, we can leverage many of the same techniques developed for real footage. The goal is to obtain a representation such as a point cloud, mesh, or implicit field that encodes the scene's geometry (and possibly appearance).

- *1. Structure-from-Motion (SfM) and Multi-View Stereo: **A traditional pipeline would first estimate the** camera poses **for each frame of the video, then reconstruct 3D points. SfM algorithms like** COLMAP** are well-suited here – they take a set of images (frames) with unknown viewpoints and find both the camera parameters and a sparse set of 3D keypoints by matching features across frames. This does *not* require any prior knowledge of poses; it is “pose-free” in the sense that the algorithm solves for the poses using feature correspondence and bundle adjustment. SfM has been successfully applied to AI-generated frames (Sora Generates Videos with Stunning Geometrical Consistency). Once camera extrinsics and intrinsics are recovered, a **multi-view stereo** (MVS) step can densify the reconstruction: it computes depth maps for the key frames (by finding dense matches or using photometric consistency) and then fuses them into a dense point cloud or mesh. COLMAP has built-in MVS, and newer methods like **3D Gaussian Splatting** can take the sparse COLMAP output and directly optimize a dense point-based radiance field (Sora Generates Videos with Stunning Geometrical Consistency). The result is often a colored point cloud or mesh of the scene. Many tools exist in this domain:

- *Open-source Photogrammetry:* Tools like COLMAP, OpenMVG, or AliceVision (Meshroom) could, in principle, be fed with frames from a text-to-video model. They will attempt to reconstruct as if it were a real video. The success depends on the video’s consistency (feature matching has to work). If the AI video has sharp, trackable details that persist, these tools can yield a decent point cloud. The above-mentioned studies used COLMAP unmodified on AI videos and got measurable reconstructions (Sora Generates Videos with Stunning Geometrical Consistency).

- *Luma AI / Nerfstudio:* These are modern toolkits that streamline the photo-to-3D pipeline. **Luma** (by Luma Labs) is a product that uses NeRF under the hood; users can upload a casually captured video and get a 3D asset. For a stable video, Luma’s pipeline (which likely uses SfM to get poses, then trains a Neural Radiance Field) could be repurposed for AI content. **Nerfstudio** is an open-source framework that lets you input

- a video or images, it runs COLMAP internally for poses, and then trains a NeRF or similar representation. An AI-generated video could be input to Nerfstudio to produce a NeRF model for that scene. Indeed, one could imagine generating a panorama video with Wan2.1 and then using Nerfstudio to obtain a 3D radiance field that can be rendered at any view.

- *2. SLAM (Simultaneous Localization and Mapping): **If the video is longer or we want online reconstruction, SLAM methods are useful. Monocular SLAM (like the classic** ORB-SLAM**) will track visual features frame-to-frame to estimate the camera trajectory on the fly, and build a sparse map of 3D feature points. Some SLAM systems also integrate stereo or depth cues to produce semi-dense maps (e.g.** LSD-SLAM **yields a semi-dense depth map by propagating depth from keyframes). New learning-based SLAM such as** DROID-SLAM** (ICCV 2021) use deep networks to predict optical flow and depth, achieving very robust tracking and dense reconstruction in real time. These could potentially handle the less stable details of AI videos better than pure feature-based methods. If one treated the AI video as coming from a moving camera in an unknown scene, a SLAM algorithm could recover the camera motion (poses) and a sparse cloud. Then a densification (again via MVS or depth fusion) could give the full 3D. The advantage of SLAM is it doesn’t require all frames upfront (it can work as the video plays) and can loop-refine if needed. The challenge is that SLAM assumes *physical consistency* – any hallucinated changes or lack of persistent features in the video will break the tracking. Still, for short, coherent clips SLAM should work similarly to SfM.

- *3. Depth-Estimation + Fusion: **Another avenue is to apply monocular** depth prediction **to each frame and then register those depth maps in space. There are many CNNs that predict per-frame depth (MiDaS, DPT, etc.), even some specifically trained for video consistency. If we also obtain camera poses (from SfM or SLAM, or even by assuming some known motion), we can back-project the depths to build a point cloud.**

- NeuralRecon** (NeuralRecon: Real-Time Coherent 3D Reconstruction from Monocular Video) is a system that does this: it takes known camera poses (e.g. from ARKit or SLAM) and uses a neural network to predict and fuse depth in real time, using a learned TSDF (truncated signed distance field) volume representation (NeuralRecon: Real-Time Coherent 3D Reconstruction from Monocular Video). It achieves *dense, coherent reconstruction* of room-scale scenes from a single moving camera at interactive rates (NeuralRecon: Real-Time Coherent 3D Reconstruction from Monocular Video). Adapting this to AI videos would require the video to have enough viewpoint change and a method to get the approximate poses. If available, the combination of learned depth (which can fill in details even with textureless regions, etc.) and a running fusion (which accumulates a 3D model as frames come in) could produce a pretty good 3D scene quickly. The catch is that depth predictors might not be reliable on fantastical AI content unless they were trained on similar imagery.

- *4. Neural Radiance Fields (NeRF) and Variants: **NeRF has revolutionized novel view synthesis by representing a scene as a continuous field learned from multi-view images. To use NeRF for reconstruction, one typically needs camera poses (again, from SfM or known trajectory). The frames from the AI video can supervise a NeRF model such that it learns to emit the correct colors for any ray (thus capturing geometry implicitly via the density field). Once trained, the NeRF can render the scene from arbitrary viewpoints with photorealistic quality. If precise geometry is needed, one can extract a point cloud or mesh from the NeRF (e.g. by sampling points at a certain density threshold). A major benefit is that NeRF can handle subtle effects like reflections or fine textures if the data allows, and it can optimize even with slightly imperfect pose estimates (there are techniques like BARF – Bundle-Adjusting NeRF – that refine pose during NeRF training ([[PDF] BARF : Bundle-Adjusting Neural Radiance Fields - Chen-Hsuan Lin](https://chenhsuanlin.bitbucket.io/bundle-adjusting-NeRF/paper.pdf**

- **#:~:text=%5BPDF%5D%20BARF%20%3A%20Bundle,))). Recent NeRF variants that** do not require known poses **could be very relevant for AI videos: for instance,** NeRF-- (NeRF minus minus)** by Wang *et al.* (2021) and subsequent works allow jointly optimizing camera pose parameters alongside the radiance field (NeRF--: Neural Radiance Fields Without Known Camera Parameters) (UP-NeRF: Unconstrained Pose Prior-Free Neural Radiance Field). These methods add extra regularization or coarse-to-fine strategies to avoid getting stuck in bad minima. More explicitly, **NoPe-NeRF, SCNeRF, BARF, UP-NeRF**, etc., have all demonstrated pose-free or pose-robust NeRF training by using approximate pose initialization or extra cues like depth priors (TD-NeRF: Novel Truncated Depth Prior for Joint Camera Pose and ...) (UP-NeRF: Unconstrained Pose Prior-Free Neural Radiance Field). In practice, this means you could feed in the frames of a generative video without knowing the camera path, and the system will figure out both the camera trajectory and the 3D volume. Fu *et al.'s* **NoPo-Splat** (COLMAP-Free 3DGS) is one example in the explicit domain: by using an **explicit point cloud representation (Gaussian splats)**, it becomes easier to incrementally solve for pose and geometry (COLMAP-Free 3D Gaussian Splatting). The continuity of the video (small inter-frame movement) is a strong cue that these methods exploit. This is particularly handy for AI videos, where you typically don't have ground-truth camera data. As long as the video frames have enough parallax and consistency, a NeRF or point-based model can be fitted in a self-supervised manner.

In summary, there is a **toolbox of 3D reconstruction methods** that can handle monocular video input, with or without known camera poses:

- *If you can estimate poses* (via SfM or SLAM), you can directly apply MVS, depth fusion, or NeRF training to get a 3D scene.
- *If pose estimation is tricky* (e.g. due to textureless or inconsistent content), newer end-to-end methods can jointly estimate structure and motion, albeit with more computational cost and some risk of failure on complex scenes.

Notably, many off-the-shelf solutions (COLMAP, Meshroom, Luma, Nerfstudio) could be tried on AI videos without modification. The bigger question is how **feasible** and effective these reconstructions are, given the characteristics of AI-generated footage.

Feasibility and Challenges

Using current video generation models as input for 3D reconstruction is **feasible**, but there are several challenges and nuances:

- **3D Consistency of AI Videos:** The biggest factor is whether the AI-generated frames correspond to a single coherent 3D scene. Early text-to-video models (like CogVideo 2022 or basic latent diffusion videos) often had *flickering details*, wobbling objects, or inconsistent lighting – all of which break multi-view geometry assumptions. Newer models (e.g. diffusion transformers like ModelScope and Wan2.1, or Sora) have improved temporal consistency and even demonstrate a degree of 3D awareness. Sora, for instance, can maintain a person's appearance as the camera moves and respects occlusion continuity (Video generation models as world simulators | OpenAI). Still, these models are not explicitly trained on multi-view consistency. As observed in practice, an object might subtly change shape or size between frames because the model isn't constrained by true geometry. Such changes will confuse a reconstruction algorithm: feature matching might find false correspondences, and the solver may either fail to converge or produce a distorted point cloud. One study noted that Sora's demo videos were *not* originally intended for 3D capture, and so the reconstructions came out "fragmented" or imperfect (Can OpenAI Sora Solve 3D Modeling? I Tried This… | by Cindy X. L. | Medium) (Can OpenAI Sora Solve 3D Modeling? I Tried This… | by Cindy X. L. | Medium). We can expect this to rapidly improve as research like \nameMethod and others enforce geometry during generation. But as of 2024, one must choose the AI prompts and models carefully to obtain reconstruction-friendly videos. Good strategies include prompting for a slow, smooth camera pan around a scene, or a 360-degree view of an

- object, rather than chaotic camera motions or videos with many moving parts.

- **Limited View Coverage:** A video only shows a scene from certain angles along its trajectory. If the text prompt doesn't explicitly instruct the model to show *all* sides of an object or room, the generated video might have a narrow baseline. For example, imagine a 5-second video where the camera just moves slightly to the right of a subject – you'll get frames from maybe a 30° span. Trying to reconstruct a full 3D model of the subject's backside or hidden areas is impossible from those frames alone. Real multi-view datasets handle this by requiring the photographer to move all around. With AI videos, one could try prompting something like "a rotating view around the object" or use a tool like ReCamMaster to generate additional views. But in a single pass, you might have to accept partial reconstructions. Some pipeline ideas (like VideoScene and KFC-W above) mitigate this by generating in-between frames or even hallucinating additional views. However, if the model hasn't seen the backside, any hallucination might be pure guesswork (not reliable for reconstruction). **In practice**, if one wants a complete 3D capture, you might generate multiple videos from different angles, or a video that explicitly circles the scene. This is an area where having control over the camera path in the generative model is very useful.

- **Feature Quality and Textures:** Traditional SfM relies on detecting and matching visual features (corners, blobs) across frames. AI-generated imagery sometimes has a **painterly** or smooth style that lacks high-frequency detail, or conversely, it might have inconsistent random textures (e.g. "visual noise") that confuse matches. If the video frames are too uniform or too random, pose estimation can fail. Some AI videos might also suffer from minor aliasing or warping artifacts frame to frame (e.g. a building's windows don't line up exactly). This reduces the number of stable correspondences. That said, methods like DROID-SLAM or using learned features (SuperPoint, etc.) could improve robustness on AI data. The Sora evaluation benchmark actually measured things like number of matching points and inlier ratios as a

- metric (Sora Generates Videos with Stunning Geometrical Consistency) – indicating that with a high-fidelity generative video, you can get plenty of matches. So, **feasibility depends on the fidelity of the video generation**. Models that output higher resolution and sharper details will produce better reconstruction results.

- **Camera Intrinsics Mismatch:** One subtle issue is that AI models do not explicitly know the camera's focal length or lens parameters of their output. They generally assume a pinhole camera model by default (as learned from training data), but the effective focal length may vary scene to scene (diffusion models can implicitly change perspective). If the reconstruction pipeline assumes a fixed focal length and the AI frames violate that (say, the model unintentionally zooms or changes FOV), the SfM could struggle. Some works have added adjustable intrinsics estimation to NeRF optimization to account for this. It might be wise to allow the solver to estimate focal length, or use wide-baseline methods that are less sensitive to exact intrinsics. In most cases, this is a minor concern; AI videos appear to have roughly consistent perspective per scene.

- **Dynamic Elements:** If the AI video contains moving objects or fluid elements (e.g. people walking, trees swaying), classical reconstruction which assumes a static scene will treat those as outliers. You'd typically get blobby artifacts or multiple ghosted positions for moving stuff. One would need to either mask out dynamic elements in the frames or use a dynamic reconstruction approach. The Vidu4D work is one solution, but it's quite involved. For most use cases focusing on static scene layout or a single object, it's best to prompt for a mostly static environment (or use a model that can separate foreground motion from background). There are also techniques like reconstructing only the static background via multi-view and handling the moving object separately (e.g. reconstructing a character via motion capture or template model).

- **Integration Challenges:** Putting together a full pipeline (text → video → 3D → new view) requires expertise across domains. There may not be many out-of-the-box tools yet that take a generative video and pop out a 3D model automatically. However, with some scripting, one

- could integrate a HuggingFace text-to-video model (like Wan2.1) to produce frames, then feed those into COLMAP or Nerfstudio. **Open-source code and demos** are appearing: for instance, the *ReCamMaster* GitHub (KwaiVGI/ReCamMaster) provides inference code using Wan2.1 ([GitHub - KwaiVGI/ReCamMaster: [ARXIV'25] ReCamMaster: Camera-Controlled Generative Rendering from A Single Video](https://github.com/KwaiVGI/ReCamMaster#:~:text=%E2%9A%99%EF%B8%8F%20Code%3A%20ReCamMaster%20%2B%20Wan2,Inference%20%26%20Training)), and the *Wan2.1* repository itself is public (GitHub - Wan-Video/Wan2.1: Wan: Open and Advanced Large-Scale Video Generative Models). For reconstruction, COLMAP and other SfM tools are readily available, and Fu et al. have released their COLMAP-Free 3D Gaussian Splatting code (likely on their project page). So researchers and developers can experiment with these building blocks. The primary challenge is that each component may need tuning when used on AI data (e.g., adjusting COLMAP settings to handle lower contrast, or ensuring the video generator's output format is compatible, etc.).

- **Quality of Results:** Even when everything works, the quality of the 3D reconstruction from AI videos currently might be **lower than from real photographs**. You might end up with a sparse or noisy point cloud if the video wasn't perfect. Cindy's Medium article showed that the reconstructed Japanese street was fragmented with missing parts (Can OpenAI Sora Solve 3D Modeling? I Tried This... | by Cindy X. L. | Medium). This is partly due to the video's limited sweep and partly due to generation artifacts. Another example she gave: a video of a castle rotation *should* have yielded a decent model, but the output "doesn't make sense at all" (Can OpenAI Sora Solve 3D Modeling? I Tried This... | by Cindy X. L. | Medium) – likely because the AI frames were inconsistent. So, while feasibility is proven, **the reliability is not guaranteed**. Each scene might require a few attempts (maybe try different prompts or slight model variations) to get a reconstruction-friendly result. The community is actively working on improving this: by making video models more physics-aware, by

- designing reconstruction algorithms that can handle AI imperfection, and by increasing resolution/length of AI videos.

Despite these challenges, the overall trajectory is very promising. As one Medium article concluded, using AI videos for 3D scene creation is a “big step” and could eventually make virtual content creation much easier (Can OpenAI Sora Solve 3D Modeling? I Tried This… | by Cindy X. L. | Medium) (Can OpenAI Sora Solve 3D Modeling? I Tried This… | by Cindy X. L. | Medium). We are essentially witnessing the convergence of generative modeling and classical 3D vision. In practical terms, a possible workflow today might be:

1. Use an open text-to-video model (e.g. Wan2.1) to generate a scene with a well-planned camera motion (perhaps a 360° orbit or a steady forward/backward pan).

2. Run a 3D reconstruction tool (COLMAP + MVS, or a NeRF pipeline, or a depth fusion method) on the extracted frames of that video.

3. Obtain a 3D representation (point cloud, mesh, or NeRF) of the scene.

4. Use that representation to render novel images or animations – effectively achieving text-to-3D via the intermediate of video.

Each step has multiple method options as discussed, and prior projects have validated parts of this pipeline. The **feasibility** is underscored by research prototypes that have already done end-to-end demonstrations (e.g., Vidu4D’s text→video→4D or the Sora→NoPoSplat example). The **challenges** serve as caveats that current results may require careful case-by-case handling and that truly seamless text-to-3D integration is still in development.

Conclusion and Outlook

Combining AI video generation with 3D reconstruction is an active frontier. Early work in 2023-2024 has shown that it’s possible to generate point clouds and even full 3D radiance fields from the outputs of text-to-video models. We have:

- **Academic Papers & Demos:** Approaches like Sora's geometry benchmark (Sora Generates Videos with Stunning Geometrical Consistency), the joint generation + 3D training framework (\nameMethod: Towards 3D-Consistent Video Generators), and pipelines like VideoScene (VideoScene: Distilling Video Diffusion Model to Generate 3D Scenes in One Step) and Vidu4D (Vidu4D: Single Generated Video to High-Fidelity 4D Reconstruction with Dynamic Gaussian Surfels), each pushing the envelope on integrating these technologies. Many of these come with project pages or code repositories (e.g., *Wan2.1* and *ReCamMaster* on GitHub, *NoPoSplat* on the authors' website, etc.), which interested readers can explore.

- **Open-Source Projects:** Wan2.1 itself is open-source (GitHub - Wan-Video/Wan2.1: Wan: Open and Advanced Large-Scale Video Generative Models), providing a foundation to experiment with text-to-video. The community has built tools (ComfyUI workflows, Diffusers integration) around it. On the reconstruction side, we have mature tools like COLMAP and emerging neural tools like Nerfstudio, which can be readily applied to image sets from any source. While not a one-click solution yet, developers have started chaining these: for instance, generating an AI video and then using Nerfstudio to create a NeRF (some Reddit users have discussed trying exactly this). The code released with research like ReCamMaster or KFC-W could be repurposed to test your own prompts and videos.

- **Method Comparisons:** If we compare methods, classical photogrammetry (SfM/MVS) provides **accuracy** when the video frames are clean, while neural implicit methods (NeRF, Gaussian splats) provide **smooth, complete renderings** even with moderate noise, and can be faster to render novel views. Pose-free methods are invaluable when you have no clue about the camera (which is the case for a random AI video), but they may need good initialization or assume small motion between frames. Two-stage pipelines (video then recon) are easier to set up with existing tools, whereas one-stage pipelines (joint training) promise better end results but are mostly in research phase. In terms of output, point clouds and meshes are great for integration into 3D

- engines or games, while NeRFs or surfel clouds are great for **visual fidelity** (photorealistic novel views) but harder to edit or export to standard modeling software. Depending on the end-use, one might choose one or the other. For example, if the goal is to create a VR scene, one might prefer to get a mesh via MVS. If the goal is to create an artistic render from a new angle, training a NeRF might be fine.

- **Challenges & Future:** The current challenges revolve around consistency and quality, but these are being rapidly addressed. It's reasonable to expect that in the near future, text-to-video models will incorporate *3D scene representation under the hood*, effectively doing "NeRF in the loop" generation so that every frame is multi-view consistent. When that happens, the reconstruction step might become trivial (or even unnecessary, if the model can output a 3D asset directly). In the meantime, hybrid workflows (generate video → reconstruct → refine with generative model for details) are a practical way forward. Some companies and researchers are already eyeing this for content creation: for instance, one blog noted the potential of using AI to generate virtual environments for filmmakers, who can then "use a camera in a 3D-rendered scene for filming" (Can OpenAI Sora Solve 3D Modeling? I Tried This… | by Cindy X. L. | Medium).

- *In conclusion**, prior work strongly suggests that the pipeline of *Text-to-Video → 3D Reconstruction → Novel View Synthesis* is not only possible but is one of the most promising approaches to achieving text-to-3D scene generation with today's technology. We have references to multiple papers, codebases, and demos (as provided above) that explore this idea from different angles. With ongoing improvements in both video generation fidelity and 3D reconstruction algorithms, the gaps are closing fast. It's an exciting interdisciplinary area, and experimenting with existing models like Wan2.1 in conjunction with reconstruction tools could yield surprisingly rich 3D results. Each method has its pros and cons, but together they are paving the way toward on-demand 3D scene synthesis from mere text descriptions.

- *Sources:**

- OpenAI, *“Video generation models as world simulators”* – Sora technical report (2024) (Video generation models as world simulators | OpenAI) (Sora Generates Videos with Stunning Geometrical Consistency)

- Xuanyi Li et al., *“Sora Generates Videos with Stunning Geometrical Consistency”* – arXiv:2402.17403 (2024), used COLMAP+GS to evaluate AI video geometry (Sora Generates Videos with Stunning Geometrical Consistency) (Sora Generates Videos with Stunning Geometrical Consistency).

- Yang Fu et al., *“COLMAP-Free 3D Gaussian Splatting”* – CVPR 2024. Project page demonstrates 3D reconstructions from Sora videos (COLMAP-Free 3D Gaussian Splatting) (COLMAP-Free 3D Gaussian Splatting).

- Cindy X. Liu, *“Can OpenAI Sora solve 3D modeling? I tried this…”* – Medium article (Oct 2023) (Can OpenAI Sora Solve 3D Modeling? I Tried This… | by Cindy X. L. | Medium) (Can OpenAI Sora Solve 3D Modeling? I Tried This… | by Cindy X. L. | Medium).

- Jianhong Bai et al., *“ReCamMaster: Camera-Controlled Generative Rendering from a Single Video”* – arXiv:2503.11647 (2025). Code on GitHub; uses Wan2.1 for open-source implementation ([GitHub - KwaiVGI/ReCamMaster: [ARXIV'25] ReCamMaster: Camera-Controlled Generative Rendering from A Single Video](https://github.com/KwaiVGI/ReCamMaster#:~:text=%E2%9A%99%EF%B8%8F%20Code%3A%20ReCamMaster%20%2B%20Wan2,Inference%20%26%20Training)).

- Harsh Jhamtani et al., *“JOint Generation and 3D Reconstruction (\nameMethod)”* – arXiv:2501.01409 (2025). Unified video diffusion + 3D point map estimation (\nameMethod: Towards 3D-Consistent Video Generators) (\nameMethod: Towards 3D-Consistent Video Generators).

- Gene Chou et al., *“Generating 3D-Consistent Videos from Unposed Photos”* – Self-supervised T2V model, improves SfM and 3DGS with AI frames (KFC-W) (KFC-W).

- Hanyang Wang et al., *“VideoScene: Distilling Video Diffusion Model to Generate 3D Scenes in One Step”* – arXiv:2504.01956 (2025). Introduces MVSplat and 3D-aware distillation (VideoScene: Distilling Video Diffusion Model to Generate 3D Scenes in One Step) (VideoScene: Distilling Video Diffusion Model to Generate 3D Scenes in One Step).

- K. Han et al., *“Vidu4D: Single Generated Video to High-Fidelity 4D Reconstruction with Dynamic Gaussian Surfels”* – arXiv:2405.16822 (2024). Dynamic scene recon from text-to-video (Vidu4D: Single Generated Video to High-Fidelity 4D Reconstruction with Dynamic Gaussian Surfels) (Vidu4D: Single Generated Video to High-Fidelity 4D Reconstruction with Dynamic Gaussian Surfels).

- Wan2.1 official GitHub and Technical Report – details on the open-source video model and performance (GitHub - Wan-Video/Wan2.1: Wan: Open and Advanced Large-Scale Video Generative Models).

- NeuralRecon (ZJU, CVPR 2021) – monocular video to real-time dense reconstruction (requires poses) (NeuralRecon: Real-Time Coherent 3D Reconstruction from Monocular Video) (NeuralRecon: Real-Time Coherent 3D Reconstruction from Monocular Video).

- COLMAP (Schönberger et al. 2016) – SfM and MVS pipeline widely used for 3D reconstruction (Sora Generates Videos with Stunning Geometrical Consistency).

- BARF (Lin et al. 2021) – Bundle-Adjusting NeRF, tackling unknown pose in NeRF optimization ([[PDF] BARF : Bundle-Adjusting Neural Radiance Fields - Chen-Hsuan Lin](https://chenhsuanlin.bitbucket.io/bundle-adjusting-NeRF/paper.pdf#:~:text=%5BPDF%5D%20BARF%20%3A%20Bundle,)).

- Additional context from Reddit discussions and Twitter (Ben Poole, 2023) on text-to-3D via video (Ben Poole on X: "the best current method for text-to-3d scenes is text ...).

References

[1] GitHub - Wan-Video/Wan2.1: Wan: Open and Advanced Large-Scale Video Generative Models:
https://github.com/Wan-Video/Wan2.1#:~:text=,Video%2C%20Video

[2] Video generation models as world simulators | OpenAI:
https://openai.com/index/video-generation-models-as-world-simulators/#:~:text=3D%20consistency,dimensional%20space

[3] Sora Generates Videos with Stunning Geometrical Consistency:
https://sora-geometrical-consistency.github.io/#:~:text=quantitatively,world%20physics%20rules

[4] Sora Generates Videos with Stunning Geometrical Consistency:
https://sora-geometrical-consistency.github.io/#:~:text=

[5] Sora Generates Videos with Stunning Geometrical Consistency:
https://sora-geometrical-consistency.github.io/#:~:text=observation%20cameras%20must%20sufficiently%20meet,The%20closer%20the%20two

[6] Sora Generates Videos with Stunning Geometrical Consistency:
https://sora-geometrical-consistency.github.io/#:~:text=of%20high,to%20the%20total%20number%20of

[7] COLMAP-Free 3D Gaussian Splatting:
https://oasisyang.github.io/colmap-free-3dgs/#:~:text=Reconstructing%203D%20scenes%20from%20Sora,Videos

[8] COLMAP-Free 3D Gaussian Splatting:
https://oasisyang.github.io/colmap-free-3dgs/#:~:text=While%20neural%20rendering%20has%20led,the%20input%20frames%20in%20a

[9] Can OpenAI Sora Solve 3D Modeling? I Tried This… | by Cindy X. L. | Medium:
https://tiancaixinxin.medium.com/can-openai-sora-solve-3d-modeling-i-tried-this-47c1576a4e8d#:~:text=To%20answer%20the%20question%2C%20yes%2C,representation%20in%20a%20latent%20space

[10] Can OpenAI Sora Solve 3D Modeling? I Tried This… | by Cindy X. L. | Medium:
https://tiancaixinxin.medium.com/can-openai-sora-solve-3d-modeling-i-tried-this-47c1576a4e8d#:~:text=I%20think%20it%E2%80%99s%20because%20the,to%20learn%20a%203D%20representation

[11] Can OpenAI Sora Solve 3D Modeling? I Tried This… | by Cindy X. L. | Medium:
https://tiancaixinxin.medium.com/can-openai-sora-solve-3d-modeling-i-tried-this-47c1576

a4e8d#:~:text=I%20thought%20this%20one%20would,doesn%E2%80%99t%20make%20sense%20at%20all

[12] Can OpenAI Sora Solve 3D Modeling? I Tried This… | by Cindy X. L. | Medium: https://tiancaixinxin.medium.com/can-openai-sora-solve-3d-modeling-i-tried-this-47c1576a4e8d#:~:text=3D%20consistency

[13] Can OpenAI Sora Solve 3D Modeling? I Tried This… | by Cindy X. L. | Medium: https://tiancaixinxin.medium.com/can-openai-sora-solve-3d-modeling-i-tried-this-47c1576a4e8d#:~:text=Can%20text,perspective%2C%20there%20is%20a%20possibility

[14] \nameMethod: Towards 3D-Consistent Video Generators: https://arxiv.org/html/2501.01409v2#:~:text=To%20summarize%2C%20we%20present%20a,trained%20generator

[15] \nameMethod: Towards 3D-Consistent Video Generators: https://arxiv.org/html/2501.01409v2#:~:text=that%20unifies%20the%20two,aware%20tasks

[16] KFC-W: https://genechou.com/kfcw/#:~:text=We%20address%20the%20problem%20of,validate%20that%20our%20method%20outperforms

[17] KFC-W: https://genechou.com/kfcw/#:~:text=We%20design%20two%20objectives%3A%201,which%20is%20our%20desired%20output

[18] KFC-W: https://genechou.com/kfcw/#:~:text=3D%20Reconstruction%20via%20COLMAP

[19] KFC-W: https://genechou.com/kfcw/#:~:text=3D%20Gaussian%20Splatting%20via%20InstantSplat

[20] VideoScene: Distilling Video Diffusion Model to Generate 3D Scenes in One Step: https://arxiv.org/html/2504.01956v1#:~:text=Recovering%203D%20scenes%20from%20sparse,limited%20by%20slow%20inference%20time

[21] VideoScene: Distilling Video Diffusion Model to Generate 3D Scenes in One Step: https://arxiv.org/html/2504.01956v1#:~:text=and%20the%20lack%20of%203D,results%20than%20previous%20video%20diffusion

[22] VideoScene: Distilling Video Diffusion Model to Generate 3D Scenes in One Step: https://arxiv.org/html/2504.01956v1#:~:text=reconstruction,enabling%20us%20to%20leap%20over

[23] VideoScene: Distilling Video Diffusion Model to Generate 3D Scenes in One Step: https://arxiv.org/html/2504.01956v1#:~:text=match%20at%20L332%20fast%2C%20feedforward,the%20coarse%20scene%20as%20follows

[24] VideoScene: Distilling Video Diffusion Model to Generate 3D Scenes in One Step: https://arxiv.org/html/2504.01956v1#:~:text=are%20capable%20of%20generating%20video,aware%20leap%20flow%20distillation

[25] Vidu4D: Single Generated Video to High-Fidelity 4D Reconstruction with Dynamic Gaussian Surfels: https://arxiv.org/html/2405.16822v1#:~:text=In%20this%20work%2C%20we%20introduce,fidelity%20appearance%20and%20accurate%20geometry

[26] Vidu4D: Single Generated Video to High-Fidelity 4D Reconstruction with Dynamic Gaussian Surfels: https://arxiv.org/html/2405.16822v1#:~:text=Image%3A%20Refer%20to%20caption%20,reminiscent%20of%20a%20Dutch%20cap

[27] Vidu4D: Single Generated Video to High-Fidelity 4D Reconstruction with Dynamic Gaussian Surfels: https://arxiv.org/html/2405.16822v1#:~:text=Figure%201%3A%20Text,in%20our%20project%20page%20for

[28] Vidu4D: Single Generated Video to High-Fidelity 4D Reconstruction with Dynamic Gaussian Surfels: https://arxiv.org/html/2405.16822v1#:~:text=dynamic%20transformation%20accurately%20represents%20motion,Secondly

[29] Vidu4D: Single Generated Video to High-Fidelity 4D Reconstruction with Dynamic Gaussian Surfels: https://arxiv.org/html/2405.16822v1#:~:text=However%2C%20achieving%20high,coherent%20depiction%20of%20dynamic%20subjects

[30] Ben Poole on X: "the best current method for text-to-3d scenes is text ...: https://x.com/poolio/status/1758234056251334711#:~:text=,video%20followed%20by%203D%20reconstruction

[31] Sora Generates Videos with Stunning Geometrical Consistency: https://sora-geometrical-consistency.github.io/#:~:text=We%20refrain%20from%20modifying%20the,are%20described%20in%20the%20following

[32] NeuralRecon: Real-Time Coherent 3D Reconstruction from Monocular Video: https://zju3dv.github.io/neuralrecon/#:~:text=,time

[33] NeuralRecon: Real-Time Coherent 3D Reconstruction from Monocular Video: https://zju3dv.github.io/neuralrecon/#:~:text=reconstruction%20from%20a%20monocula

r%20video,Scenes

[34] NeuralRecon: Real-Time Coherent 3D Reconstruction from Monocular Video: https://zju3dv.github.io/neuralrecon/#:~:text=propose%20to%20directly%20reconstruct%20local,knowledge%2C%20this%20is%20the%20first

[35] NeRF--: Neural Radiance Fields Without Known Camera Parameters: https://nerfmm.active.vision/#:~:text=NeRF,computed%20camera

[36] UP-NeRF: Unconstrained Pose Prior-Free Neural Radiance Field: https://openreview.net/forum?id=UvBwXdL95b#:~:text=UP,and%20various%20light%20conditions

[37] TD-NeRF: Novel Truncated Depth Prior for Joint Camera Pose and ...: https://arxiv.org/abs/2405.07027#:~:text=TD,by%20jointly%20optimizing

[38] Video generation models as world simulators | OpenAI: https://openai.com/index/video-generation-models-as-world-simulators/#:~:text=Long,their%20appearance%20throughout%20the%20video

[39] GitHub - Wan-Video/Wan2.1: Wan: Open and Advanced Large-Scale Video Generative Models: https://github.com/Wan-Video/Wan2.1#:~:text=Wan%3A%20Open%20and%20Advanced%20Large,Video%20Generative%20Models

[40] Can OpenAI Sora Solve 3D Modeling? I Tried This… | by Cindy X. L. | Medium: https://tiancaixinxin.medium.com/can-openai-sora-solve-3d-modeling-i-tried-this-47c1576a4e8d#:~:text=This%20Japanese%20street%20model%20is,details%2C%20still%20not%20that%20ideal

[41] Can OpenAI Sora Solve 3D Modeling? I Tried This… | by Cindy X. L. | Medium: https://tiancaixinxin.medium.com/can-openai-sora-solve-3d-modeling-i-tried-this-47c1576a4e8d#:~:text=Another%20difference%20I%20wanna%20point,well%20into%20the%20industry%E2%80%99s%20needs

[42] Can OpenAI Sora Solve 3D Modeling? I Tried This… | by Cindy X. L. | Medium: https://tiancaixinxin.medium.com/can-openai-sora-solve-3d-modeling-i-tried-this-47c1576a4e8d#:~:text=Generating%203D%20scenes%20is%20still,without%20scanning%20any%20real%20objects

[43] \nameMethod: Towards 3D-Consistent Video Generators: https://arxiv.org/html/2501.01409v2#:~:text=themselves%20are%20not%20truly%203D,aware%20tasks

[44] VideoScene: Distilling Video Diffusion Model to Generate 3D Scenes in One Step: https://arxiv.org/html/2504.01956v1#:~:text=overlap%20across%20input%20views%20w

ith,model%20to%20generate%203D%20scenes