Hybrid Tours: A Clip-based System for Authoring Long-take Touring Shots

XINRUI LIU, Cornell University, USA
LONGXIULIN DENG, Cornell University, USA
ABE DAVIS, Cornell University, USA

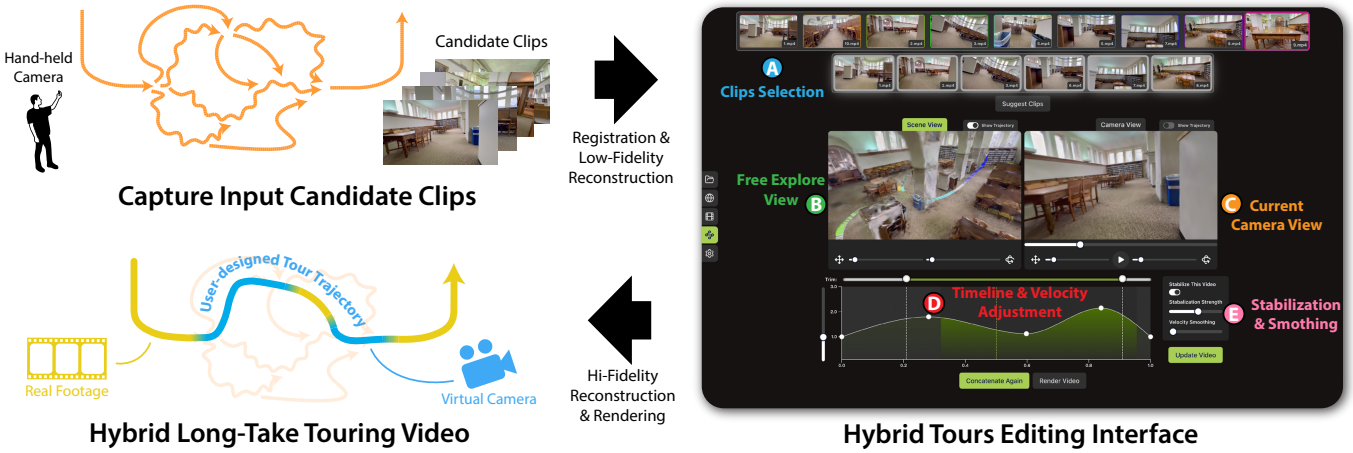

Fig. 1. **Hybrid Tours**. We present *Hybrid Tours*, a tool for creating long-take touring shots from short hand-captured video clips. Users start by capturing *candidate clips* (top-left) that approximate different segments of potential touring camera trajectories. Then, a coarse subset of these frames is used to reconstruct a low-cost high-speed pre-visualization of the scene for path planning. Our *editing interface* (right) then lets users design longer camera trajectories by filtering, combining, and re-timing candidate clips. Finally, once the user is satisfied with the pre-visualized video, we optimize additional reconstruction of the scene around their chosen camera trajectory to render a final high-quality *hybrid long-take touring video* (bottom-left).

Long-take touring (LTT) shots are characterized by smooth camera motion over a long distance that seamlessly connects different views of the captured scene. These shots offer a compelling way to visualize 3D spaces. However, filming LTT shots directly is very difficult, and rendering them based on a virtual reconstruction of a scene is resource-intensive and prone to many visual artifacts. We propose *Hybrid Tours*, a hybrid approach to creating LTT shots that combines the capture of short clips representing potential tour segments with a custom interactive application that lets users filter and combine these segments into longer camera trajectories. We show that Hybrid Tours makes capturing LTT shots much easier than the traditional single-take approach, and that clip-based authoring and reconstruction leads to higher-fidelity results at a lower cost than common image-based rendering workflows.

CCS Concepts: • **Computing methodologies** → **Image-based rendering**; **Graphics systems and interfaces**; **Computational photography**; • **Human-centered computing** → **Interactive systems and tools**; *User studies*; *Usability testing*; Graphical user interfaces; Activity centered design.

Additional Key Words and Phrases: Long-take Shots, 3D Gaussian Splatting, Video Editing

ACM Reference Format:
Xinrui Liu, Longxiulin Deng, and Abe Davis. 2025. Hybrid Tours: A Clip-based System for Authoring Long-take Touring Shots. *ACM Trans. Graph.* 44, 4 (August 2025), 13 pages. https://doi.org/10.1145/3731423

Authors' addresses: Xinrui Liu, Cornell University, 107 Hoy Road, Ithaca, New York, USA, 14853, xinrui@cs.cornell.edu; Longxiulin Deng, Cornell University, 107 Hoy Road, Ithaca, New York, USA, 14853, longxiulin@cs.cornell.edu; Abe Davis, Cornell University, 107 Hoy Road, Ithaca, New York, USA, 14853, abedavis@cornell.edu.

ACM 0730-0301/2025/8-ART
https://doi.org/10.1145/3731423

1 INTRODUCTION

Long-take touring (LTT) videos have become an increasingly popular way to showcase large physical spaces. These videos are characterized by long, smooth camera trajectories that seamlessly connect different parts of a complex scene in one uninterrupted shot. However, capturing LTT video is very difficult; typically, the camera trajectory must be choreographed ahead of time and captured in one take by a skilled drone or steadicam operator. This traditional workflow, which we call the *all-real* workflow, often requires several attempts to successfully capture the desired camera path in a single shot, which can quickly become expensive and, in some environments, dangerous.

Image-based rendering (IBR) techniques like NeRF [Mildenhall et al. 2020] and 3D Gaussian splatting (3DGS) [Kerbl et al. 2023] have made alternative *all-virtual* workflows increasingly popular. In such workflows, the user first scans and reconstructs an environment, then plans and renders a camera trajectory using their

virtual reconstruction of the scene. This approach relaxes the need to perform difficult camera maneuvers in the physical world and lets users explore different camera trajectories after data has been captured. In theory, the ability to synthesize unrecorded views of a scene should make capturing data easier, overall. However, this benefit is complicated by the time and computational resources needed to reconstruct a scene, which often make it difficult to know when enough data has been captured or what parts of a scene will reconstruct well until it is too late to capture more data. In addition to computational costs, this is a major reason why all-real approaches are still favored by professionals.

In this work, we observe that all-real and all-virtual workflows sit at extreme ends of an under-explored design space that balances capture-time challenges against the cost and uncertainty of reconstructing large scenes. Building on this observation, we introduce *Hybrid Tours*, a new workflow and interactive editing tool that uses short video clips as a basic primitive for capture, path-planning, and rendering. By designing our workflow around short video clips, we can make capture much simpler and more flexible than all-real approaches, and generate higher-quality visual results at lower computational costs than all-virtual approaches. The code and data for this paper are at https://github.com/liuxr0831/Hybrid-Tours.

1.1 A Hybrid Clip-Based Workflow

We refer to Hybrid Tours as a workflow, and not just an interactive system, because it changes the entire process of creating LTT shots to combine aspects of all-real and all-virtual workflows:

Workflow	Output of Capture	Path Planning
All-Real	Single Video	Before Capture
All-Virtual	Set of Images	After Capture
Hybrid Tours	Set of Videos	Before+After

The observation that combining these workflows can be useful is not entirely new; many users of IBR know that rendering camera trajectories close to captured video input is a good way to ensure high-quality visual results. However, that insight is scarcely found in the design of existing IBR tools and interfaces, which mostly treat videos the same as an unordered collection of images. By making video clips an explicit primitive for capture, path planning, and reconstruction, Hybrid Tours makes it much easier to coordinate decisions made during capture, registration, editing, and rendering.

2 RELATED WORK

Much of our approach's strength comes from considering the design of data capture, reconstruction, editing, and rendering together in one workflow. By contrast, most previous efforts have focused on either capture for all-real approaches or reconstruction for all-virtual approaches.

2.1 All-Real Path Planning & Capture

Some past work has focused on developing tools to help users capture more stable video by hand. For example, Sayed et al. [Sayed et al. 2022] present a system that uses a scriptable actuated gimbal to help film long cinematic shots. However, most work on capturing long-take shots has focused on path planning and automated capture of video with unmanned aerial vehicles (UAVs). Some commercial tools offer a gallery of template UAV shots that can be captured in outdoor settings [DJI 2024]. Other research focuses on interactive tools for piloting drones [Chen et al. 2021; Inoue et al. 2023; Temma et al. 2019], or path planning in virtual settings before executing all-real capture with a drone [Galvane et al. 2018; Gebhardt et al. 2016; He et al. 1996; Joubert et al. 2016, 2015; Nägeli et al. 2017; Roberts and Hanrahan 2016; Xie et al. 2018].

2.2 All-Virtual Reconstruction and Rendering

While IBR has been an active area of research for several decades [Gortler et al. 1996; Levoy and Hanrahan 1996], its use for creating LTT-like videos is comparatively recent and mostly limited to tools that reconstruct scenes using some variant of NeRF [Mildenhall et al. 2020] or 3DGS [Kerbl et al. 2023]. The recent progress of these rendering methods has made all-virtual approaches to LTT more viable, but data capture and reconstruction for large or complex scenes are still very expensive and prone to visual artifacts, which is part of the challenge we address. Some past works have also looked at interactive guidance for the efficient capture of IBR data in limited settings, such as bounded subjects [Davis et al. 2012] or a planar window into a scene [Mildenhall et al. 2019]. Our work lets users capture more arbitrary camera paths in short segments, which we use in a novel interactive system for composing and rendering longer shots virtually.

There is a great deal of recent and ongoing work in computer graphics and computer vision that focuses on improving IBR reconstructions (e.g., [Diolatzis et al. 2024; Fang et al. 2022; Fridovich-Keil et al. 2023; Li et al. 2023; Wu et al. 2023; Ye et al. 2024; Zhao et al. 2024]). However, our work is largely orthogonal to this line of research, as we treat the IBR algorithm used in our workflow as a largely interchangeable component. Our current implementation builds on the original 3DGS code from [Kerbl et al. 2023], but is designed to be agnostic to use with other IBR methods.

The closest related work on all-virtual approaches is NerfStudio [NerfStudio 2024], an open-source project described as "a simple API that allows for a simplified end-to-end process of creating, training, and testing NeRFs" [Team 2022] in its documentation. While Nerfstudio was initially designed to facilitate research, its modularity and support for interactive viewing, path planning, and rendering, have made it a go-to tool for practitioners as well. Its interactive tools offer free viewpoint control over a virtual camera, which users can keyframe and interpolate to render video [nerfstudio 2022]. However, this interface has no notion of input camera trajectories. Instead, captured video is treated as an unordered collection of independent images. This accommodates applications that assume data will be found, rather than captured directly by the user (e.g., rendering based on Internet photo collections [Li et al. 2020; Martin-Brualla et al. 2021]), but discarding input trajectories forgoes several opportunities to better integrate decisions made during capture, which we explore at length in our work.

2.3 Video Stabilization

Lastly, a third space of related work is video stabilization, which warps the result of all-real capture to smooth the motion in a video.

Stabilization is typically treated as a post-processing effect applied to existing video, rather than a workflow for creating new video. However, it offers a strong motivation for exploring hybrid workflows that sit between all-real and all-virtual approaches. Stabilization can indirectly make all-real capture easier by making the quality of results more tolerant of camera shake during filming. And while most approaches do not involve a full reconstruction of the captured scene, some of the most powerful variants register and reconstruct sparse features in the scene to anchor the warping of input video frames (e.g., [Joshi et al. 2015; Kopf et al. 2014; Li et al. 2023; Liu et al. 2011, 2021; Meuleman et al. 2023; Peng et al. 2024; Zhao et al. 2023]). If we interpret this warping as a type of local view synthesis, the lesson we take from stabilization techniques is that reconstruction and rendering become much easier for views that are very close to the trajectory of a captured video. Hybrid Tours extends that benefit to enable a workflow where users can explore a space of alternative trajectories after filming.

3 BACKGROUND

3.1 The Real-Virtual Spectrum of LTT Shots Workflows

We can interpret all-real and all-virtual approaches as two extremes on a spectrum of workflows that balances how much control a user has after data has been captured against the costs of sampling and reconstructing an entire scene. At one end, all-real approaches limit path planning to before and during capture but do not require any reconstruction. At the other end of the spectrum, all-virtual approaches let users render arbitrary camera paths after capture, but require dense sampling and costly reconstruction of the scene.

Unfortunately, the tradeoffs made by these two extremes are often baked into the design of tools that facilitate their use. All-real capture tools typically focus on capturing data in a single take, with no support for incorporating reconstructions or merging distinct video segments after capture. Meanwhile, virtual tools like NerfStudio [NerfStudio 2024] treat input video as a collection of independent image frames and discard information about the recording camera's trajectory during capture. Virtual camera paths are then specified by keyframing and interpolating a free-viewpoint virtual camera. Critically, this leaves the interface for specifying virtual camera paths agnostic to the quality of the reconstructed views that they contain. To guarantee high-quality outputs, a user must either reconstruct the entire scene with high quality or manually search for camera paths that reconstruct well given the available data (as shown in Figure 2).

3.2 The Costs of Virtual Approaches

All-virtual approaches are limited by the set of images that are actually recorded. As such, much of their added flexibility is contingent upon recording more views (or at least a broader distribution of views) than one would need to achieve comparable quality in an all-real workflow. In addition to any burden this places on capture, the need to process this data creates significant computational costs that are a common barrier to many potential users. Once data has been captured, there are four steps involved in creating videos with IBR: registration, reconstruction, path planning, and rendering. Registration involves using some variant of Structure from Motion (SfM) such as COLMAP [Schönberger and Frahm 2016] to solve for camera poses and sparse geometry. Reconstruction then uses this information to train an IBR representation like 3DGS or NeRF. In path planning, users specify a camera trajectory, and in rendering, that trajectory is sampled at each frame and rendered to create a video. The most expensive step in this process is usually registration, a non-convex optimization problem. For large scenes, most of the cost comes from global bundle adjustments, which scale poorly with the number of images ($O(n^3)$ in general, and super-linear even in very optimistic settings) and are difficult to parallelize. LTT shots showcase large and complex scenes from many perspectives, which requires a large number of input images, making these global bundle adjustments prohibitively expensive for many users.

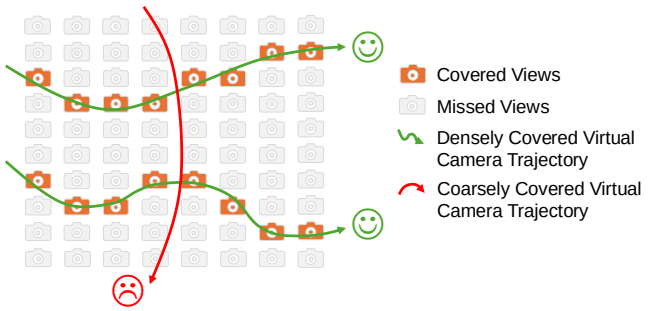

Fig. 2. **Different Output Video Quality for Different Virtual Trajectories**: Given a set of captured input views in a scene (orange cameras), some virtual camera paths can be reconstructed with much higher quality than others. When data is recorded by a moving camera, the density of captured views along the paths of camera motion tends to be much higher. As a result, virtual camera trajectories that align with captured paths (green) tend to reconstruct much better than trajectories that cut across captured paths (red). However, existing tools like NerfStudio do not factor input camera trajectories or estimated reconstruction quality into the virtual path planning process.

3.3 The Hybrid Tours Approach

The goal of Hybrid Tours is to balance the strengths of real and virtual workflows to make it easier for users to create high-quality LTT shots from hand-captured video. To do this, we need a way to integrate decisions made before, during, and after capture. Hybrid Tours accomplishes this by reframing capture around *candidate clips*, which are short video clips that represent pieces of longer potential camera trajectories. The "real" part of our workflow involves capturing these candidate clips. Then, in the virtual part of our workflow, we use these clips as the basis for defining and rendering longer virtual trajectories that the user can explore after capture.

Hybrid Tours offers a few key advantages over all-real workflows. First, it makes capture easier by letting users break long camera trajectories into smaller, more manageable segments. In some cases, these segments can even be combined to create camera paths that would be impossible to capture in a single take (e.g., flying a camera through narrow openings or transparent surfaces). Second, our tool can stabilize and re-time camera paths after capture, which lets users record data with regular hand-held cameras. Third, Hybrid Tours

lets users explore a combinatorial space of alternative virtual paths after data has been captured. More specifically, our interactive editor lets users build virtual camera paths by sequencing and filtering candidate clips in different ways (Figure 8). This means that at capture time, users can record multiple options for a given tour segment. Later, the user can compare and choose between those options in our interactive editor. Our editor also offers automatic suggestions based on an analysis of captured clips and partially specified user objectives. For example, if the user chooses a clip they would like to start with, and a clip they would like to finish on, Hybrid Tours can suggest an optimal path through other captured clips to connect the two.

Hybrid Tours also offers advantages over all-virtual approaches. First, it makes it easier to leverage framing and composition decisions that a user makes during capture. Second, it allows for a much more efficient registration and rendering pipeline. More specifically, we can aggressively subsample input clips for an initial reconstruction and path planning. While this can also be done with existing all-virtual workflows, by framing path planning as a recombination of captured clips, Hybrid Tours guarantees that authored virtual trajectories will stay close to yet unregistered images, and tells us precisely which images those are. This makes it easy to render full-quality results after path planning, often without registering a majority of captured video frames (i.e., by ignoring frames from clips that are not used in the final trajectory).

Finally, Hybrid Tours makes it easy to combine filtered virtual camera paths with real and unfiltered ones, offering a way to integrate live-action clips with motion into output videos. We demonstrate an example application of this in tabletop/day_to_night.mp4.

4 HYBRID TOURS

We start by outlining our full workflow, then describe our interface for editing and visualizing virtual camera paths, which is the main interactive component of Hybrid Tours.

4.1 Workflow

4.1.1 Capture. Users begin by capturing data (Figure 3 step 1). There is no specialized software involved in this step, but there are instructions for users to follow. Users should capture clips that represent portions of longer potential camera paths, and adjacent clips should "connect", meaning subsequent clips should start near the ends of previous clips and face in roughly the same direction. During editing, users will have the flexibility to filter and re-time camera trajectories, so roughly covering the right set of viewpoints during capture is more important than making sure the camera's path is smooth. For particularly complex scenes or input camera trajectories, users may optionally capture additional images or footage to aid with registration.

4.1.2 Pre-visualization Reconstruction. Our interactive editor for designing virtual camera paths uses an initial low-cost reconstruction of the scene for pre-visualization (Figure 3 step 2). For this reconstruction, we divide each clip into three segments—a beginning, middle, and end—with configurable durations and sampling densities. By default, our pre-visualization uses all frames from the first and last seconds of each clip, as well as a subsampling of the middle frames at 10fps. Users can also designate a clip to be included as real footage to allow for dynamic content in the scene. In this case, we only register frames from the beginning and end sections of the clip, which are used to optimize for smooth camera paths leading into and out of the clip.

Once we have selected frames for pre-visualization, we register them using COLMAP [Schönberger and Frahm 2016]. Here, we match each frame with 5 previous and subsequent extracted frames within clips, and we perform exhaustive feature matching across clips to save computation. Once registered, we use the registered images, excluding any taken from clips marked as real footage, to reconstruct the scene. Our current implementation uses 3DGS [Kerbl et al. 2023] at 1/4 of our final resolution for pre-visualization, but this part of the system could be replaced with other reconstruction and rendering methods in the future.

4.1.3 Path Design and Editing. Our interface for designing and editing LTT paths uses our initial reconstruction of the scene to pre-visualize virtual LTT trajectories (Figure 3 step 3). Users explore different camera paths by concatenating and filtering a subset of the captured candidate clips. For each clip, the user can decide whether to use a filtered virtual rendering or the original video frames. Section 4.3 and the supplemental material describe this interface in greater detail.

4.1.4 Rendering Final Video. To render the final video, we densely sample all frames from candidate clips included in the user's designed trajectory (Figure 3 step 4). Then, we use COLMAP to match each extracted frame with 15 previous frames and 15 subsequent frames from the same clip. Since these new frames are very close to frames from the same clip that have already been registered, we found that additional global bundle adjustments were not necessary here. Then, we re-optimize a 3DGS reconstruction at the chosen output resolution and render all frames at that resolution. We also provide the option of ignoring registered frames from unused candidate clips, which can make reconstruction easier in some cases. For transitions between real and virtual segments in an output path, we use a linear cross-fade between virtual and real frames with the duration of 1 second.

4.2 The Frame Graph & Clip Graph

Multiple features of Hybrid Tours use graph representations built over captured data to assist with path planning. Firstly, for each ordered pair of candidate clips, we build a *frame graph* over all frames registered for previsualization. Each frame I_i is a node in this graph, and edges $I_i \rightarrow I_j$ represent the cost of transitioning from frame I_i to frame I_j in a rendered trajectory. We assume that the cost of transitioning between subsequent frames in a captured clip is zero, and set the cost of transitioning between frames from different clips based on their difference in position, orientation, and velocity. The shortest path in this graph between the first frame of a candidate clip c_i and the last frame of a candidate clip c_j predicts which portion of each clip to use in an unfiltered path that connects them. We can also use the cost of this shortest path as the edge cost in a separate *clip graph*, where each node represents a different candidate clip. In our interface, we use the clip graph to help the user

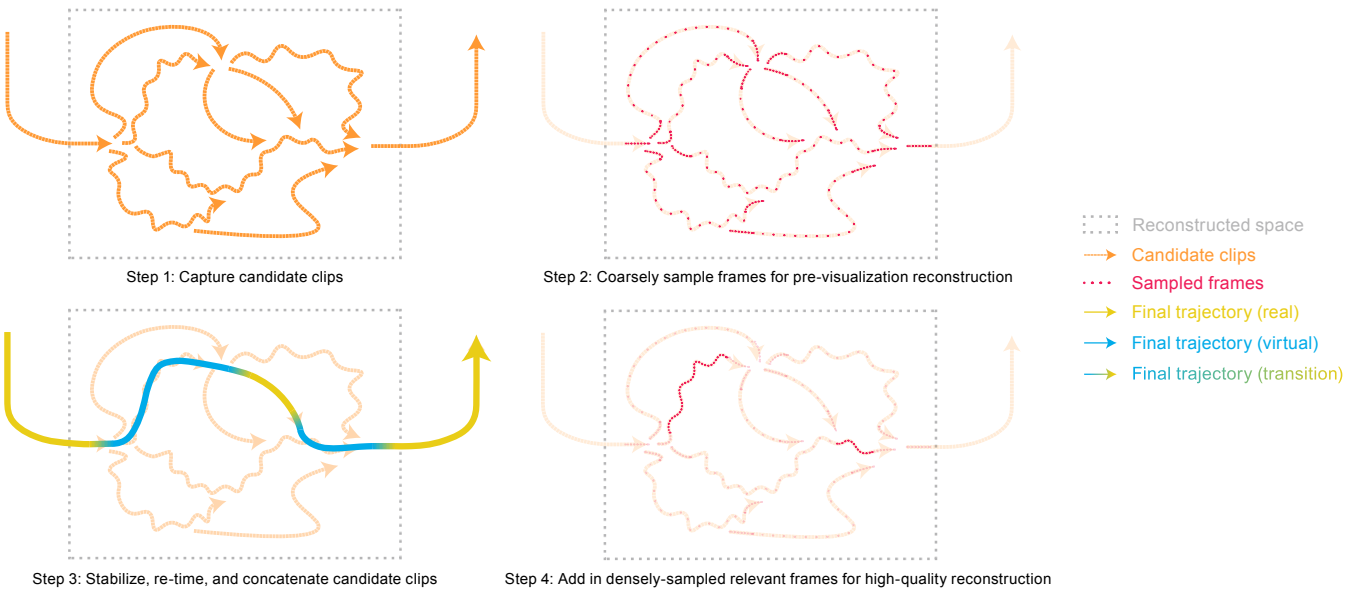

Fig. 3. **Hybrid Tours Workflow**: In step 1, the user captures multiple candidate clips that start, end, or fully stay within the clips concatenation space. In step 2, we coarsely sample the candidate clips to build the pre-visualization reconstruction. In step 3, the user uses the UI to stabilize, trim, re-time, and concatenate the candidate clips to create the final camera trajectory that could involve both the original high-quality 2D footage (shown in yellow) or edited clips whose frames are rendered from virtual reconstruction (shown in blue). The virtual and real frames are linearly blended at transitions (shown in the gradient from yellow to blue or blue to yellow) to achieve smooth visual results. In step 4, we densely sample all frames along the final camera trajectory where virtually rendered frames are used and build the high-quality reconstruction to render the final video.

with the initial selection and ordering of candidate clips, and we use the frame graph to optimize initial trajectories based on a given ordering of candidate clips. More details about the construction of each graph can be found in our supplemental material and code.

4.3 Editing Interface

Our editing interface is shown in Figure 4. The main components of the interface are a section for clip selection (Figure 4 A), two rendering contexts for visualizing the scene and currently selected trajectory (Figure 4 B and C), and controls for filtering and editing the positions and velocities of that trajectory (Figure 4 D and E). We summarize the features of our editing tool here, and a much more detailed description of how trajectories are calculated and filtered can be found in our supplemental material.

4.3.1 Clip Selection. Users add clips to the current trajectory by dragging them from the row of candidate clips at the very top to the row of currently selected clips just below it. As they do this, Hybrid Tours uses the clip graph to identify and highlight clips in the candidate row that connect well with the end of the current trajectory, which makes the combinatorial space of possible paths much easier to explore. Users can alternatively select a beginning clip and end clip, and ask Hybrid Tours to optimize for a sequence that connects these clips smoothly by finding the shortest path between them in our clip graph.

4.3.2 Pre-visualization. The current camera trajectory is represented by a viewpoint spline. By default, we fit this spline to the

Fig. 4. **Editing Interface**: Our editing interface has five basic elements: the top Clip Selection (A) part for selecting and ordering candidate clips, the left scene view (B) for a third-person view of the scene and camera trajectories, the right scene view (C) for the current camera view, the timeline window (D) for adjusting camera velocity and trimming the video, and the stabilization and smoothing controls (E) for adjusting camera pose and velocity smoothing. This screenshot shows the interface being used to edit three concatenated clips, but the user can alternatively select individual clips to edit prior to concatenation using the same interface.

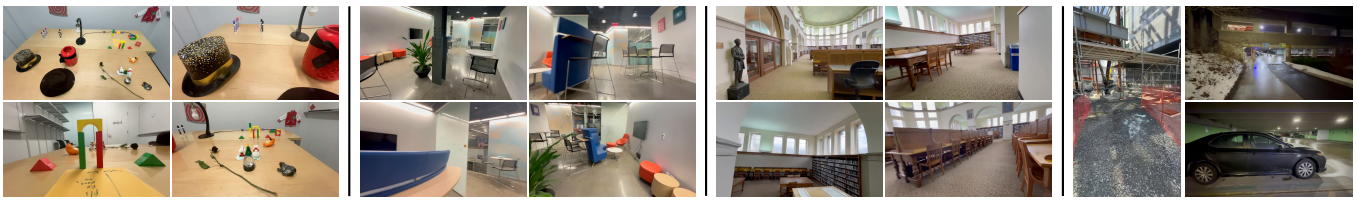

Fig. 5. **Example Tour Scenes**: The best way to appreciate our results is by watching the generated tour videos and demos in our supplemental material. Here we show five example scenes covering a large range of scales: (left) a small tabletop scene, (middle left a medium-sized floor scene, (middle right) a large indoor library scene, (right) a large outdoor stairway scene and a garage scene.

viewpoints taken from the shortest path through the sequence of selected clips in the frame graph. Users can change this path by changing the clip sequence, manually adjusting the start or endpoint of an included clip, or by filtering (described below). The left rendering context in our interface shows the current reconstruction from a free-viewpoint camera, visualizing the current trajectory with each segment color-coded by its source candidate clip. The right rendering context shows a first-person view of the current trajectory, controlled by the timeline below.

4.3.3 Filtering & Re-timing. A key aspect of our interface design is that virtual camera trajectories are limited to filtered combinations of captured trajectories, which helps ensure that synthesized views stay close to captured images. Once the user has selected a sequence, our stabilization and smoothing interface offers different ways to edit the resulting trajectory. The interface can operate on individual candidate clips, or on the current combined trajectory, and users can switch between these using the tab control on the left side of the screen. The stabilization strength slider controls how much smoothing is applied to the positions of the path being edited, and the remaining controls are used to edit velocity. Our interface actually contains two timelines: the timeline under our camera view visualization, which is parameterized by time, and the main timeline view at the bottom of the interface, which is parameterized by progress along the currently selected trajectory spline. On this main timeline view, we display a curve representing the rate of progress along the spline at each point in its trajectory (i.e., velocity relative to the input video). Users can edit this curve locally by adding and manipulating control points to make relative adjustments to velocity. Alternatively, the velocity smoothing slider interpolates this curve toward one that has uniform absolute (i.e., arc-length) velocity over time, offering a way for users to edit velocity relative to a uniform baseline. Finally, there is a "Trim" slider above the main timeline that can be used to crop the current trajectory or candidate clip in time.

5 RESULT & ANALYSIS

Our supplemental material includes videos showing how our tool is used in different scenes and examples of the videos that it generates. Fig. 5 shows four example scenes that vary significantly in scale and detail. We also include all videos mentioned in this section. Note that these videos are compressed for space, which results in some reduction of visual quality.

5.1 Stabilization, Re-timing, and Concatenation

In Figure 6, we use stairway/3.mp4 to illustrate the effect of camera trajectory stabilization and re-timing. We stabilized stairway/3.mp4 at three different stabilization strengths, 1, 4, and 7. We also gradually slowed it down from the original velocity at the start to half of the original velocity at the end, which we visualize in Figure 6.

In Figure 7 we use the concatenation of stairway/5.mp4 and stairway/6.mp4 to illustrate the result of camera trajectory concatenation. The camera's position and orientation are continuous with the last and next videos at the start and the end of the concatenating clip. The line charts for the camera's movement velocity, angular velocity, and angular acceleration also show that we preserve the continuity of these values.

Figure 8 visualizes an example of how different orderings of different sets of candidate clips can be used to explore different LTT camera trajectories. Note that this visualization only includes a subset of the clips from one video. Our supplemental material also includes results that blend real footage captured at different times and virtually-rendered frames (see tabletop/day_to_night.mp4).

5.2 Reconstruction Cost & Quality

The best way to evaluate the visual quality of our final results is by watching videos in our supplemental material. Using the real frames as references, we compare the quality of rendered views in Figure 9 between the all-virtual approach and Hybrid Tours both visually and with LPIPS scores based on the AlexNet architecture [Zhang et al. 2018]. To approximate the traditional all-virtual approach, we built two reconstructions, one by uniformly subsampling only the extra scan video (quailty/extra_scan.mp4 in the supplemental material) to exclude all frames of candidate clips, the other by uniformly subsampling both the extra scan video and all candidate clips (1.mp4 to 10.mp4 in tabletop folder of supplemental material). Then, using these two reconstructions as well as the Hybrid Tours workflow, we rendered along a camera trajectory very close to tabletop/7.mp4, aligned frames in the rendered videos with the closest real frames in quality/7_real.mp4, and computed the average LPIPS distance across all frames against real frames. Even though both all-virtual reconstructions involved the same amount of input frames (2259 images) as Hybrid Tours final rendering, they still produce worse results than Hybrid Tours final rendering, regardless of whether we include the candidate clips in the subsampled input videos or not. Here, Hybrid Tours achieved better quality by concentrating

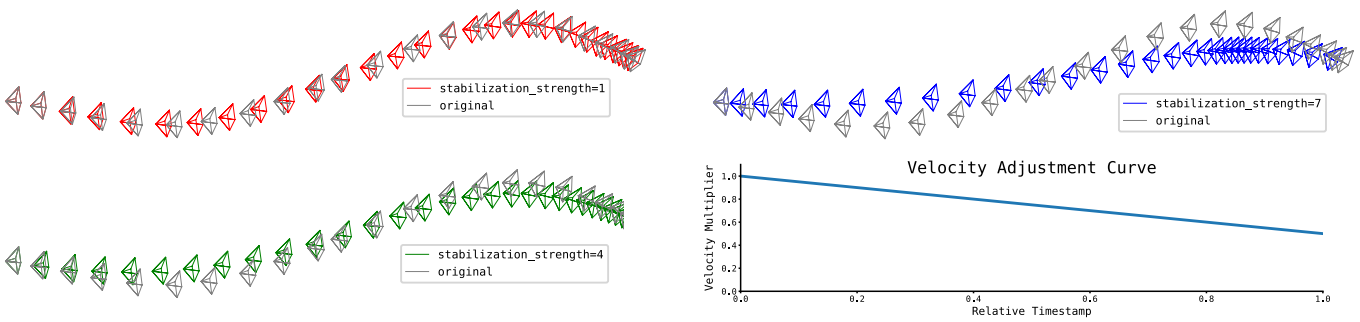

Fig. 6. **Effect of Camera Trajectory Stabilization and Re-timing**: In this figure, the camera moves from left to right. As the stabilization strength increases, the stabilized camera trajectory gets farther away from the original trajectory and contains fewer fluctuations. We plot out the camera frustums along both the original and the stabilized trajectory at 10 fps. Here, we use the velocity adjustment curve to gradually slow down the camera's velocity relative to the original video. The camera behaves as expected as we can see there are more camera frustums positioned closely with each other on the right side of each trajectory.

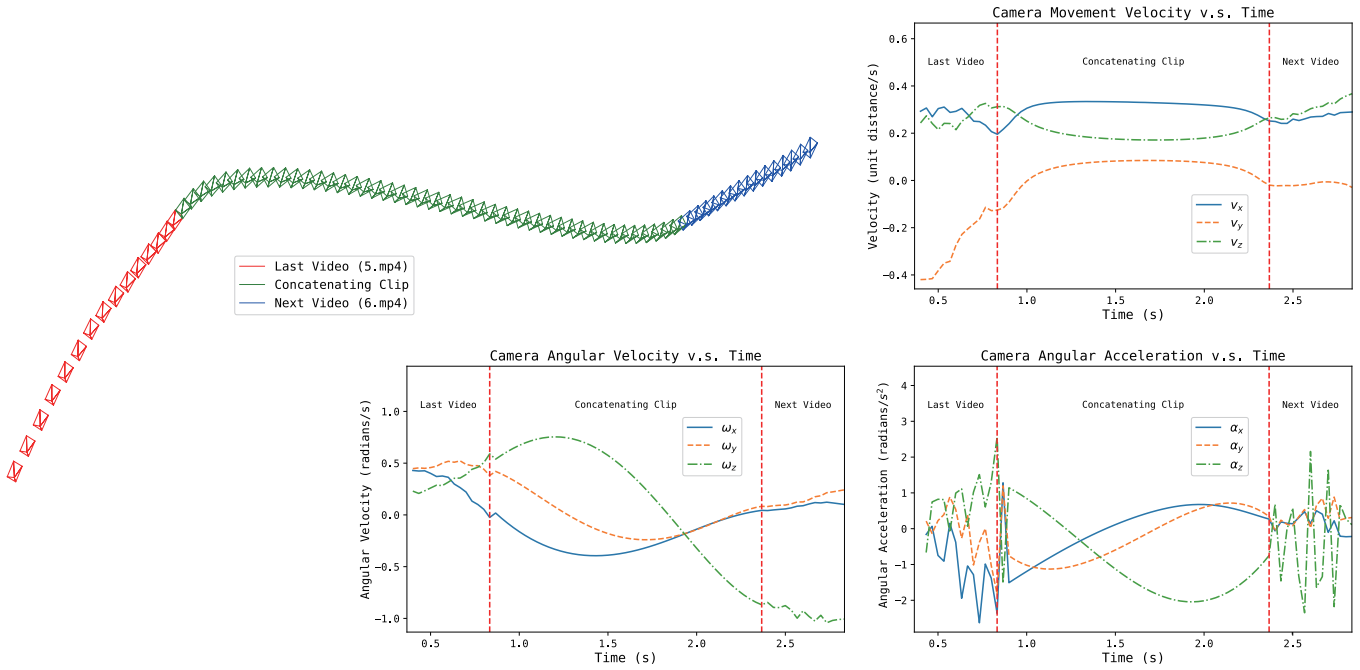

Fig. 7. **Camera Movement during Concatenating Clip**: In this figure, the camera moves from left to right. The last 15 frames of the last video, all frames in the concatenating clip, and the first 15 frames of the next video are plotted. As shown by the plotted frustum, the camera's position and orientation are continuous with the last and the next video during the concatenating clip. The camera movement velocity along the three axes shows that we do enforce continuous movement velocity with our approach. The camera's angular velocity and acceleration during the concatenating clip are also continuous with the last and the next video, showing that we do ensure smooth camera orientation transition. The video corresponding to the plotted camera frustums is stairway/56.mp4 in the supplemental material.

the budget of frames closely around the final camera trajectory (i.e. by densely sampling the candidate clip quality/7_real.mp4). However, this comparison significantly underestimates the benefits of our workflow to visual quality, because the most significant improvement to quality comes not from improving the reconstruction globally, but from limiting the LTT paths that a user can create to trajectories that are close to input video paths, and can then be reconstructed well.

Table 1 shows the time and memory costs of Hybrid Tours and the all-virtual approach that samples every frame. These numbers illustrate the large cost benefits of using Hybrid Tours, especially when the traditional all-virtual approach makes the resource requirements impractical for many users. For example, in scenes with more images (tabletop and floor), we see more than 20 times speedup in overall runtime compared to all-virtual approaches that register

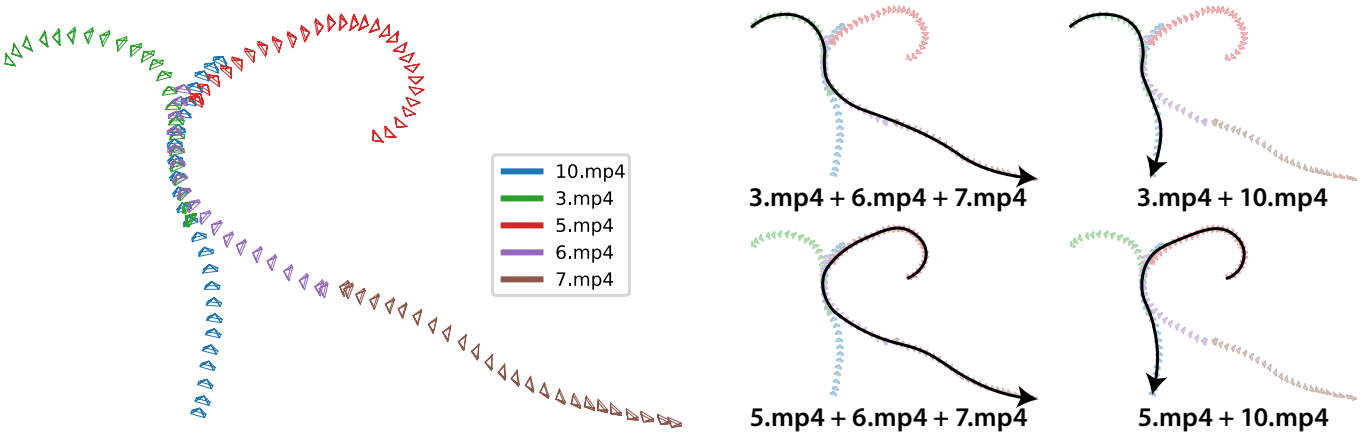

Fig. 8. **Possible Trajectories based on Candidate Clips**: This figure shows possible camera trajectories that could be constructed from a collection of candidate clips. Here, we plot a subset of candidate clips from the library scene and show the possible camera trajectories that could be constructed. The left part shows five candidate clips plotted in different colors. The black arrowed curve shows the camera trajectory, and the caption below shows the involved candidate clips. As four of these clips either begin or end at roughly the same place, by trimming and concatenating these clips, we can create four different camera trajectories.

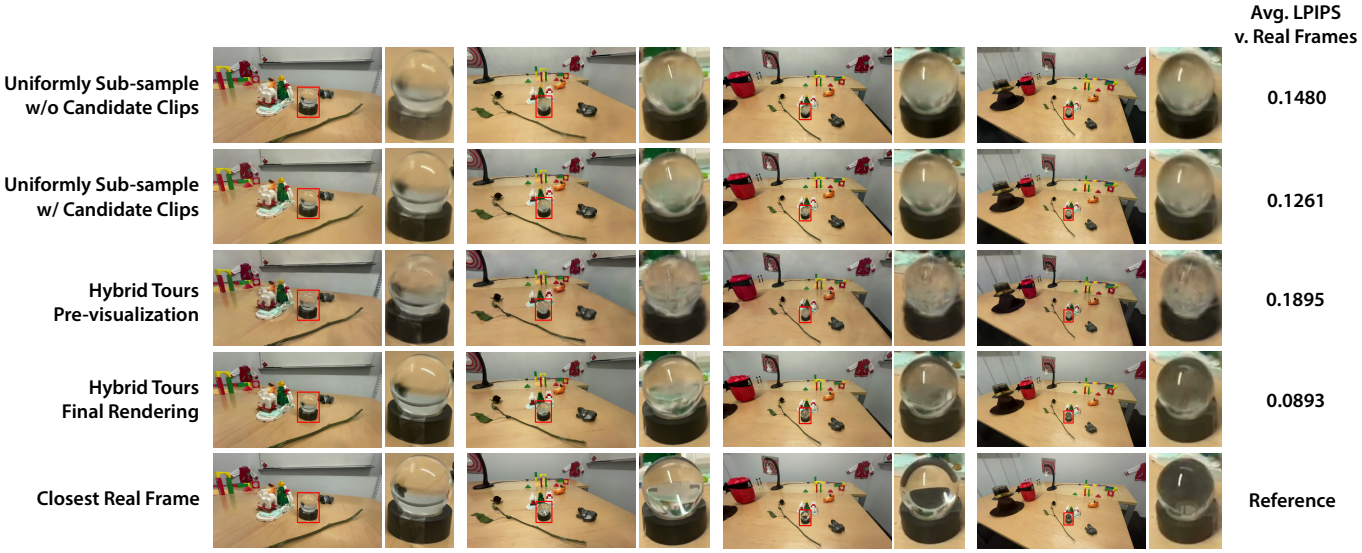

Fig. 9. **Novel View Rendering Quality Comparison**: We compare the novel view rendering quality between all-virtual reconstructions built from naive subsampling of input videos, our Hybrid Tours pre-visualization, and the Hybrid Tours final high-quality rendering using LPIPS based on the AlexNet architecture [Zhang et al. 2018]. The first row comes from uniform subsampling of the extra scan video, which is not included among the candidate clips. The second row uniformly subsamples both this extra scan video and the candidate clips. These two rows approximate the current all-virtual approach. The third and fourth rows represent the Hybrid Tours previsualization and final rendering, respectively. The reconstructions from the first two rows involved the same total number of images (2259 frames) as the Hybrid Tours final rendering. We rendered videos along camera trajectories very close to quality/7_real.mp4 and calculated the average LPIPS score across all frames. These rendered videos, the extra scan video, and relevant candidate clips can be found in the quality folder of supplemental material. We selected some frames from the rendered videos and highlighted the glass ball to better visualize the quality differences.

every frame. Additional details of runtime comparison can be found in the supplemental material.

Figure 10 compares our default pre-visualization pipeline with one that uses a much more aggressive subsampling strategy to perform our initial reconstruction in well under an hour. We were able to do this for three out of four of our scenes, but the stairway scene required more images to make registration successful. The ability to pre-visualize this quickly makes it feasible to do on-site, which opens up the potential to iterate on capture after an initial session of editing. Additional details such as the number of involved frames and runtime comparison for the low-cost pre-visualization can be found in the supplemental material.

Table 1. **Memory Requirement and Runtime Comparison.** In this table, we compare the memory requirements and overall runtime between the traditional all-virtual approach that samples every frame and Hybrid Tours based on a representative video for each scene. The integers in the RAM requirements columns represent the number of involved images, and the memory requirement is shown below the integer. The overall run time for Hybrid Tours includes the time spent on pre-visualization reconstruction and the high-quality reconstruction. The percentage shown next to the Hybrid Tours statistic is relative to the statistics for the traditional all-virtual approach that samples every frame. Hybrid Tours requires much less memory to load training images for final rendering thanks to its frame sampling strategy. Hybrid Tours only performs global bundle adjustment on 6.8% to 36.2% of all frames, which significantly reduces the runtime of COLMAP registration and results in much lower overall runtime. Under the Hybrid Tours workflow, the stairway scene was processed with 4 CPU cores, 20 GB of RAM, and an RTX A5000, and the remaining three scenes were processed with 4 CPU cores, 30 GB of RAM, and an RTX A6000. Under the traditional all-virtual workflow that samples every frame, everything was performed under the same spec except that the 3DGS step was performed with extra RAM to load all images. Besides, under the traditional all-virtual approach, we matched the video frames in the same way as Hybrid Tours except that each video frame of the floor scene is matched with the previous 20 and next 20 frames as the floor scene's videos contain a lot of motion blur. For the extra videos that are not candidate clips in the Hybrid Tours workflow, we sample them at 2 fps for the library, tabletop, and floor scene and at 3 fps for the stairway scene.

Scene and Video	RAM Requirements & Number of Involved Frames			COLMAP+3DGS Run Time	
	Sample Every Frame	Hybrid Tours Pre-visualization	Hybrid Tours Final Rendering	Sample Every Frame	Hybrid Tours
stairway (245678.mp4)	2630 (20.72 GB)	947 (36.0%) (0.47 GB)	1133 (43.1%) (8.93 GB)	1804 min 02 s	533 min 08 s (29.6%)
library (1236789.mp4)	2842 (49.11 GB)	1030 (36.2%) (1.11 GB)	1275 (44.9%) (22.03 GB)	2739 min 35 s	283 min 16 s (10.3%)
tabletop (1234567.mp4)	18831 (325.40 GB)	2070 (11.0%) (2.24 GB)	2343 (12.4%) (40.49 GB)	> 1 week	502 min 44 s (<5.0%)
floor (1234.mp4)	20323 (351.18 GB)	1384 (6.8%) (1.49 GB)	1479 (7.3%) (25.56 GB)	> 1 week	311 min 53 s (<3.3%)

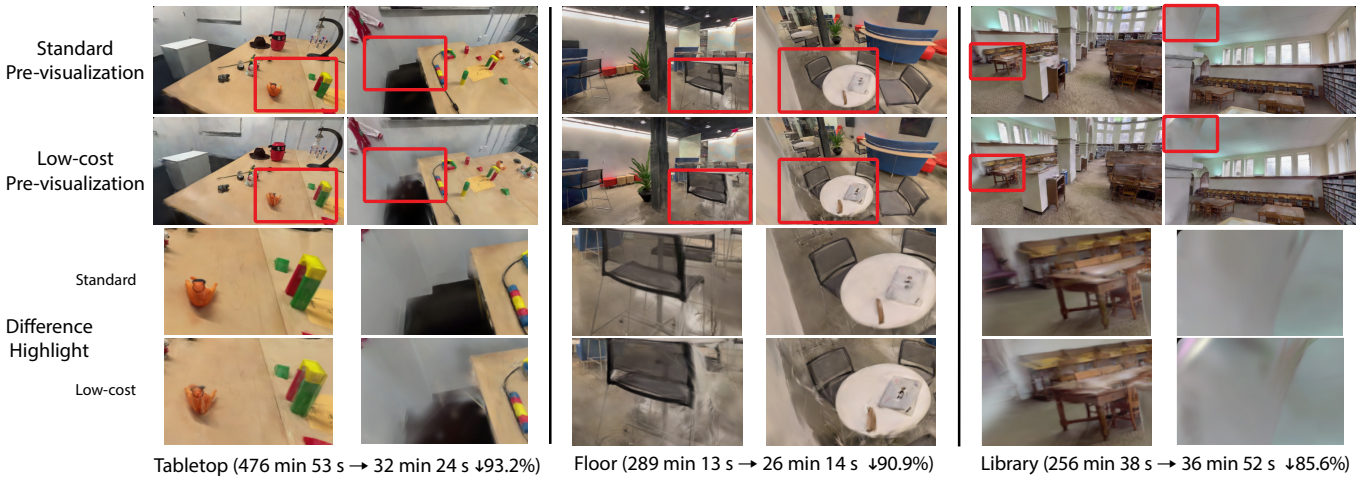

Fig. 10. **Standard and Low-cost Pre-visualization Comparison**: We visually compare the registration and reconstruction quality of the Hybrid Tours pre-visualization step between the standard approach and a low-cost approach that involves aggressively sub-sampling candidate clips under 10 fps, only sampling first and last 7 frames of candidate clips, and removing or trimming extra videos. Although the low-cost pre-visualization does have more artifacts, the registration accuracy and reconstruction quality of the low-cost approach are still comparable to the standard approach, indicating that the low-cost pre-visualization reconstruction still lets the user know the camera pose in the scene and supports the downstream shots authoring tasks. In terms of run time, the low-cost previsualization takes 85.6% to 93.2% less time to build than the standard pre-visualization.

6 INFORMAL CAPTURE STUDY

We conducted a small informal IRB-approved study to explore whether users found it easy to understand and adapt to a clip-based capture workflow, and whether this changed how users thought about capturing LTT shots. In particular, we wanted to see how users adapted to the idea of breaking longer shots into shorter segments and whether they could recognize a simple scenario where this lets them capture an otherwise impossible trajectory.

During the study, the participants were first asked to film along the camera trajectory shown in Figure 11 using both the traditional

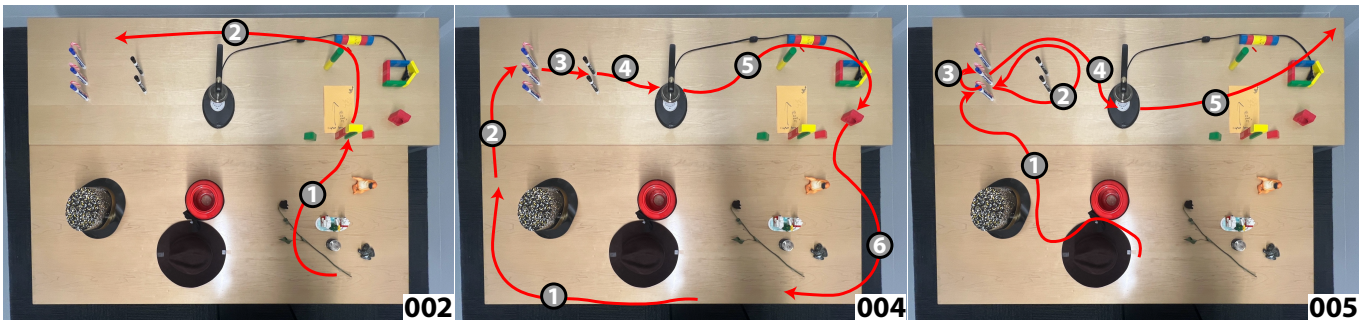

Fig. 12. **Camera Trajectories Created by Participants**: This figure shows the camera trajectories designed by Participant 002, 004, and 005. The curved red arrows represent the short clips filmed by the participants with the numbers on them being their orders in the full camera trajectory. All participants including them designed trajectories that cannot be captured in a single take due to the and must be created using Hybrid Tours, indicating that they understood the Hybrid Tours approach within the short period of the user study and applied it in creating their own shots.

all-real approach and the Hybrid Tours approach. Then, the participants were asked to create their own camera trajectories using either the traditional all-real approach or the Hybrid Tours approach.

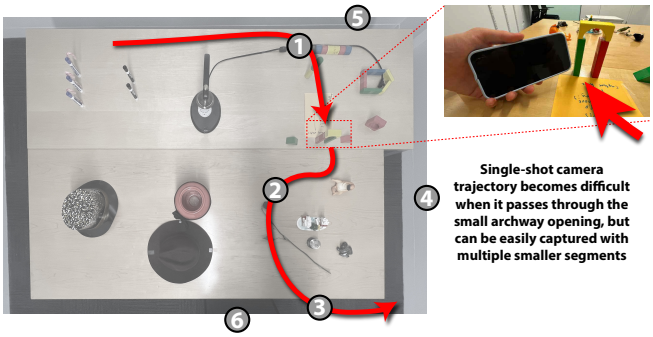

Fig. 11. **User Study Trajectory**: During the shot that the participant needed to film, the camera should move smoothly along the trajectory 1+2+3, but this camera trajectory was infeasible in real life as the smartphone was too big to go through the arch. During the study, most users broke this shot into three separate pieces (shown in red) and filmed them separately. Participants also made diverse choices in terms of where to stand while filming segment 2 and 3.

We found that all seven participants (6 males, 1 female, all of them between the ages of 19 and 30) easily figured out that they could break the long take shot into three parts as shown in Figure 11, although one of the participants chose to film segments 2 and 3 together in a single take at the end, which made the task slightly more physically difficult. This shows that users could understand the idea of breaking long trajectories into short segments.

The participants created highly diverse camera trajectories, and some of them are shown in Figure 12. While every participant's camera trajectory was unique, they had a few things in common. Every user chose to design a shot that was infeasible to capture in a single take. They also achieved their shots by recording multiple clips. In addition, one participant explicitly mentioned that breaking a long-take shot into pieces allows for more creative camera trajectory design. This indicates that compared to the all-real approach, Hybrid Tours allows for and inspires more creative camera trajectory designs.

While informal, this study provided some confirmation that users could understand and adapt to the clip-based capture workflow of Hybrid Tours. More details of the informal capture study can be found in the supplemental material.

7 END-TO-END CASE STUDY

We also conducted more comprehensive IRB-approved end-to-end case studies with two users to compare Hybrid Tours with the all-real and all-virtual approaches on a common scene. Each participant created long-take touring shots using all three workflows. Then, they were asked to evaluate their experience with each workflow as well as the quality of the LTT videos they produced. Based on our observations and participant feedback, we addressed six research questions. We also made improvements to the UI and the tutorial video based on user feedback. The detailed plan of the study can be found in supplemental material.

7.1 Participants

We completed the study with two participants. Participant 001 is a 27-year-old male experienced in filmmaking and cinematography, but minimal experience with IBR (he had used NeRF once informally and controlled virtual cameras in Unity). Participant 002 is a 23-year-old make who uses 3DGS regularly, has captured data for 3DGS multiple times, but had minimal experience with real or virtual cinematography or authoring camera trajectories. Neither of the two participants had captured LTT shots before, or filmed with tools like a FPV drone or steadicam.

We note that we also recruited one other participant (26-year-old male), with experience using NeRF and Blender. However, the data captured by participant 003 for all-virtual and HybridTours workflows failed to yield a usable reconstruction of the scene, making it impossible to complete the study. Specifically, the all-virtual scan captured by that participant was not correctly registered by COLMAP, leading to an output video with quite noticeable artifacts that resulted from registration error. While capturing candidate

clips, that participant failed to interpret candidate clips as segments of the final shots. After capturing, he mentioned the following:

> "I was thinking about the candidate clips as scans to cover the 3D scene from different angles, so I captured them like that (repeating the motion sequence of hand forward then rotate around several times). At the moment you mentioned that candidate clips should be sequences of clips that will later get concatenated together, I finally realized that I should capture like that (moves hands forward stably as if the participant is holding a smartphone/camera)."

Based on his words, we updated the capture tutorial to explicitly mention that the user should think about candidate clips as sequences of clips that will later get concatenated together.

7.2 Study Design

We began by asking users to answer questions about prior experience, which can be found in section the supplemental material. We then explained LTT shots with example videos. Then we asked them to author shots using the all-virtual and all-real workflows. Finally, we explained the Hybrid Tours workflow and tasked them with using it to author a LTT video.

We evaluated all three approaches in the same scene because the scene itself also impacts users' creativity, and we want to evaluate how workflows themselves, not the scene, affect the users' creativity. For example, a straight hallway presents very limited creativity opportunities as the only LTT shot possible is to walk through it. On the other hand, a library with chairs, tables, and bookshelves allows the user to perform cool shots through small holes and gives the user many different camera trajectory segments to build the final shots. Therefore, letting participants use three different filming approaches on the same scene is the only feasible way to evaluate the three filming approaches fairly.

7.2.1 All-Virtual Task. We tasked them with capturing data that would be used to create an LTT shot with NerfStudio. Here, we let them review NerfStudio tutorial material if desired. Participant 001 chose to capture the scene with video, while Participant 002 captured many individual photos. We processed Participant 001's videos using ns-process-data, the official command-line tool of NerfStudio for registering custom video input data. For Participant 002's image data, we registered with COLMAP. Registered frames were finally fed into NerfStudio's 3DGS pipeline to reconstruct the scene, and each participant was asked to author long-take touring shots from their captured data using NerfStudio. We did not give participants an explicit time limit for this last step, and each completed it in under half an hour.

7.2.2 All-Real Task. The participants were brought back to the same scene and given 10 minutes to design and capture long-take touring shots using the all-real approach. We provided the participants with a handheld actuated smartphone stabilizer that they could optionally use. After the 10 minutes capturing time, the participants were told to submit their final long-take touring shots.

7.2.3 Hybrid Tours Task. Finally, we brought each participant back to the same scene again and explained the Hybrid Tours workflow, including showing them a demo of the interface and how their data would be used. We also let them briefly try out the interface with a pre-loaded sample dataset from a different scene. We then asked them to capture data for authoring an LTT video using the Hybrid Tours. Participants were given 10 minutes to capture candidate clips. We then registered their clips and asked them to author an LTT video using Hybrid Tours. We did not give participants an explicit time limit for this last step, and each spent under half an hour.

7.2.4 Workflow Assessment. After they finished authoring all the shots, we surveyed each participant about their experience with each workflow. We then showed them the final rendered videos from each approach and asked them to evaluate the subjective quality of each result. The evaluation questions, answers, and videos created by each participant with each workflow are included in our supplemental material. Note that due to the limited size of the supplemental material, the end-to-end study output videos went through heavy compression which reduces their visual quality.

7.3 Research Questions

7.3.1 Compared to the all-real approach, does Hybrid Tours address challenges during capturing? Both Participant 001 and 002 recognized advantages of Hybrid Tours during capturing. Participant 002 felt that using the all-real approach, he needed the handheld stabilizer to produce videos without visible camera shake, but he did not need it using Hybrid Tours. Both participants liked that Hybrid Tours allows users to retry segments along the whole trajectory without breaking the continuity of the final output. Participant 002 also mentioned that Hybrid Tours "definitely made certain shots possible that you couldn't have shot without hybrid tours such as flying through narrow passages." However, Participant 002 believed that Hybrid Tours added some cognitive load to filming "since there were more things you have to keep in mind when filming and designing your sequence of shots."

7.3.2 Compared to the all-real approach, does Hybrid Tours help users explore different camera trajectories by recombining candidate clips? Participant 002 utilized this feature and commented that "this feature was very useful as it allowed me to try out different trajectories and see which ones looked better while still being able to reuse certain sequences of shots." Though, he again noted some increased cognitive load related to keeping track of where candidate clips were. Both participants also appreciated the clip suggestion system, with Participant 001 writing that it "saves me time from looking through all the videos."

7.3.3 Compared to the all-virtual approach, does Hybrid Tours make it easier or harder to create and edit camera trajectories? Participant 002 mentioned that Hybrid Tours makes it easier to make use of the camera pose and timing of the candidate clips, "making the entire end-to-end process much more intuitive." Both Participant 001 and 002 mentioned that the velocity adjustment chart was very useful. Using this feature, Participant 001 corrected undesirably slow camera movement and sped up the camera as it turned a corner around a shelf, and Participant 002 accelerated the camera while it was descending "to create a feeling of movement" resulting in a roller-coaster effect. Participant 002 casually commented during

editing that his manual edits seemed to make camera movement less smooth than the automatic velocity smoothing. Participant 002 commented that Hybrid Tours made it difficult for him to deviate from captured clips. Though, based on his experience with NeRF/3DGS, he correctly speculated that this aspect of Hybrid Tours could "vastly improve the final reconstruction quality for complicated trajectories, especially trajectories that pass near objects."

7.3.4 Compared to the other two approaches, does Hybrid Tours inspire or limit users' creativity? Both Participant 001 and 002 agreed that Hybrid Tours inspired their creativity. Participant 001, who had more of a film background, enjoyed capturing candidate clips in the real world and felt it let him explore more creative shots, while adding keyframes in NerfStudio did not give him the same feeling. Participant 002 liked that Hybrid Tours "allows the creation of shots that cannot be filmed with an all-real approach."

7.3.5 How would users choose among the all-real, all-virtual, and Hybrid Tours approaches in different scenarios? Participant 001 would choose Hybrid Tours for "real estate touring and maybe Game of Thrones intro inspired (tour) of the university campus." Participant 002 noted that his preference might depend on the type of video being created, preferring Hybrid Tours for long or complicated trajectories, and the all-real workflow for simple camera trajectories that are easy to film. Participant 002 also noted that the free-viewpoint control of NerfStudio might be preferable for exploring camera trajectories when output video quality is not a concern.

7.3.6 How does the output video quality of all three approaches compare to each other? The output videos of the all-virtual approach for both participants contain noticeable artifacts. While creating camera trajectories in NerfStudio, Participant 001 repeatedly complained about visual artifacts in the reconstruction and the difficulty of finding a camera path that contained fewer artifacts. In the survey, both Participant 001 and 002 believed that the output videos of Hybrid Tours had fewer visual artifacts. Participant 002 also commented that the quality of the Hybrid Tours' output videos are close to the quality of the all-real approach output.

7.4 Learning Curve

Creating LTT videos is inherently difficult, and while Hybrid Tours offers significant advantages, it still has a learning curve. Both participants found the quality of their Hybrid Tours results satisfying, but Participant 002 commented that "if I were to use Hybrid tours again, there would be drastic improvement both in quality and efficiency." To partially illustrate this, and to further demonstrate Hybrid Tours's editing capabilities, we created new edits of the same sequences used by each participant with smoothing and timing parameters further adjusted in the Hybrid Tours interface. These more polished results, which can be found in our supplemental, point to some of the learning curve associated with our interface, and suggest that results may improve with further training or repeated use.

8 DISCUSSION

8.1 Potential Users and Application Scenarios

Hybrid Tours reduces many common barriers to creating LTT shots. It makes filming much easier and more flexible, especially when professional stabilization equipment is not available, or when capturing uninterrupted shots is difficult. It also significantly reduces the computational resources necessary for virtual workflows. Notably, some of the results in this work would be to difficult for the authors to capture using an all-real workflow, and require too many compute resources to produce with an all-virtual workflow.

For professional cameramen and drone operators, Hybrid Tours is particularly useful for addressing camera trajectory feasibility issues, especially unexpected ones. For example, a professional drone operator may find the designed trajectory infeasible due to a smaller-than-expected window and windy weather. With Hybrid Tours, the drone operator can cover different segments of the path in different flights and explore different combinations of clips post-capture.

Hybrid Tours could make communicating and coordinating with customers easier for professional filmmakers. Customers may have uncertain and changing requirements for what the final shot should cover. With Hybrid Tours, the cameraman can film many clips to cover all potential trajectories that the customer is interested in and decide on the final shots after filming. The pre-visualization reconstruction and editing interface also provide a low-cost solution for cameraman to edit shots and receive feedback from customers in a timeframe where additional data may be captured before leaving the capture site.

8.2 Limitations and Future Directions

While it significantly reduces costs compared to existing all-virtual approaches, Hybrid Tours still requires some time to reconstruct a scene. Even our low-cost pre-visualization, which takes less than an hour to compute, takes long enough that in outdoor environments, for example, the angle of shadows may change before users have an opportunity to capture more data based on pre-visualization results.

The rendering quality of our current implementation could also be improved by integrating more recent IBR methods [Diolatzis et al. 2024; Guédon and Lepetit 2023; Ye et al. 2024; Yu et al. 2023]. Besides, the IBR algorithm itself could be modified to better support Hybrid Tours. For example, we could increase the weight of frames along the final camera trajectory during training to further improve the final rendering quality.

8.3 Conclusion

Our work proposes Hybrid Tours, a safe, low-cost, and flexible solution for producing long-take touring shots without any specialized capture device. Hybrid Tours improves on many aspects of all-real capture workflows and all-virtual IBR-based workflows for creating LTT video by integrating capture, path planning, and reconstruction. We believe this space of hybrid workflows offers a rich space of opportunities to explore in the future, and hope to facilitate work in this direction by making Hybrid Tours available to the public.

ACKNOWLEDGMENTS

This work was partially supported by a National Science Foundation Faculty Early Career Development Grant under award #2340448 and by a generous gift from Meta. We thank the user study participants for their participation and feedback. We thank the Durland Alternatives Library for letting us conduct the end-to-end case study.

REFERENCES

Linfeng Chen, Kazuki Takashima, Kazuyuki Fujita, and Yoshifumi Kitamura. 2021. PinpointFly: An Egocentric Position-control Drone Interface using Mobile AR. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 150, 13 pages. https://doi.org/10.1145/3411764.3445110

Abe Davis, Marc Levoy, and Fredo Durand. 2012. Unstructured Light Fields. *Comput. Graph. Forum* 31, 2pt1 (may 2012), 305–314. https://doi.org/10.1111/j.1467-8659.2012.03009.x

Stavros Diolatzis, Tobias Zirr, Alexander Kuznetsov, Georgios Kopanas, and Anton Kaplanyan. 2024. N-Dimensional Gaussians for Fitting of High Dimensional Functions. In *ACM SIGGRAPH 2024 Conference Papers* (Denver, CO, USA) *(SIGGRAPH '24)*. Association for Computing Machinery, New York, NY, USA, Article 126, 11 pages. https://doi.org/10.1145/3641519.3657502

DJI. 2024. Introduction to QuickShots via DJI Fly. Retrieved April 3, 2024 from https://support.dji.com/help/content?customId=en-us03400006482&spaceId=34&re=US&lang=en&documentType=artical&paperDocType=paper

Jiemin Fang, Taoran Yi, Xinggang Wang, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Matthias Nießner, and Qi Tian. 2022. Fast Dynamic Radiance Fields with Time-Aware Neural Voxels. In *SIGGRAPH Asia 2022 Conference Papers* (Daegu, Republic of Korea) *(SA '22)*. Association for Computing Machinery, New York, NY, USA, Article 11, 9 pages. https://doi.org/10.1145/3550469.3555383

Sara Fridovich-Keil, Giacomo Meanti, Frederik Warburg, Benjamin Recht, and Angjoo Kanazawa. 2023. K-Planes: Explicit Radiance Fields in Space, Time, and Appearance. arXiv:2301.10241 [cs.CV] https://arxiv.org/abs/2301.10241

Quentin Galvane, Christophe Lino, Marc Christie, Julien Fleureau, Fabien Servant, François-Louis Tariolle, and Philippe Guillotel. 2018. Directing Cinematographic Drones. *ACM Trans. Graph.* 37, 3, Article 34 (jul 2018), 18 pages. https://doi.org/10.1145/3181975

Christoph Gebhardt, Benjamin Hepp, Tobias Nägeli, Stefan Stevšić, and Otmar Hilliges. 2016. Airways: Optimization-Based Planning of Quadrotor Trajectories according to High-Level User Goals. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 2508–2519. https://doi.org/10.1145/2858036.2858353

Steven J. Gortler, Radek Grzeszczuk, Richard Szeliski, and Michael F. Cohen. 1996. The lumigraph. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques (SIGGRAPH '96)*. Association for Computing Machinery, New York, NY, USA, 43–54. https://doi.org/10.1145/237170.237200

Antoine Guédon and Vincent Lepetit. 2023. SuGaR: Surface-Aligned Gaussian Splatting for Efficient 3D Mesh Reconstruction and High-Quality Mesh Rendering. arXiv:2311.12775 [cs.GR]

Li-wei He, Michael F. Cohen, and David H. Salesin. 1996. The virtual cinematographer: a paradigm for automatic real-time camera control and directing. In *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '96)*. Association for Computing Machinery, New York, NY, USA, 217–224. https://doi.org/10.1145/237170.237259

Maakito Inoue, Kazuki Takashima, Kazuyuki Fujita, and Yoshifumi Kitamura. 2023. BirdViewAR: Surroundings-aware Remote Drone Piloting Using an Augmented Third-person Perspective. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 31, 19 pages. https://doi.org/10.1145/3544548.3580681

Neel Joshi, Wolf Kienzle, Mike Toelle, Matt Uyttendaele, and Michael F. Cohen. 2015. Real-time hyperlapse creation via optimal frame selection. *ACM Trans. Graph.* 34, 4, Article 63 (July 2015), 9 pages. https://doi.org/10.1145/2766954

Niels Joubert, Jane L. E, Dan B Goldman, Floraine Berthouzoz, Mike Roberts, James A. Landay, and Pat Hanrahan. 2016. Towards a Drone Cinematographer: Guiding Quadrotor Cameras using Visual Composition Principles. arXiv:1610.01691 [cs.GR]

Niels Joubert, Mike Roberts, Anh Truong, Floraine Berthouzoz, and Pat Hanrahan. 2015. An interactive tool for designing quadrotor camera shots. *ACM Trans. Graph.* 34, 6, Article 238 (nov 2015), 11 pages. https://doi.org/10.1145/2816795.2818106

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics* 42, 4 (July 2023). https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/

Johannes Kopf, Michael F. Cohen, and Richard Szeliski. 2014. First-person hyperlapse videos. *ACM Trans. Graph.* 33, 4, Article 78 (jul 2014), 10 pages. https://doi.org/10.1145/2601097.2601195

Marc Levoy and Pat Hanrahan. 1996. Light field rendering. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques (SIGGRAPH '96)*. Association for Computing Machinery, New York, NY, USA, 31–42. https://doi.org/10.1145/237170.237199

Zhengqi Li, Qianqian Wang, Forrester Cole, Richard Tucker, and Noah Snavely. 2023. DynIBaR: Neural Dynamic Image-Based Rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhengqi Li, Wenqi Xian, Abe Davis, and Noah Snavely. 2020. Crowdsampling the Plenoptic Function. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I* (Glasgow, United Kingdom). Springer-Verlag, Berlin, Heidelberg, 178–196. https://doi.org/10.1007/978-3-030-58452-8_11

Feng Liu, Michael Gleicher, Jue Wang, Hailin Jin, and Aseem Agarwala. 2011. Subspace video stabilization. *ACM Trans. Graph.* 30, 1, Article 4 (feb 2011), 10 pages. https://doi.org/10.1145/1899404.1899408

Yu-Lun Liu, Wei-Sheng Lai, Ming-Hsuan Yang, Yung-Yu Chuang, and Jia-Bin Huang. 2021. Hybrid Neural Fusion for Full-frame Video Stabilization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Ricardo Martin-Brualla, Noha Radwan, Mehdi S. M. Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. 2021. NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections. arXiv:2008.02268 [cs.CV] https://arxiv.org/abs/2008.02268

Andreas Meuleman, Yu-Lun Liu, Chen Gao, Jia-Bin Huang, Changil Kim, Min H. Kim, and Johannes Kopf. 2023. Progressively Optimized Local Radiance Fields for Robust View Synthesis. In *CVPR*.

Ben Mildenhall, Pratul P. Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. 2019. Local light field fusion: practical view synthesis with prescriptive sampling guidelines. *ACM Trans. Graph.* 38, 4, Article 29 (jul 2019), 14 pages. https://doi.org/10.1145/3306346.3322980

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. arXiv:2003.08934 [cs.CV] https://arxiv.org/abs/2003.08934

Tobias Nägeli, Lukas Meier, Alexander Domahidi, Javier Alonso-Mora, and Otmar Hilliges. 2017. Real-time planning for automated multi-view drone cinematography. *ACM Trans. Graph.* 36, 4, Article 132 (jul 2017), 10 pages. https://doi.org/10.1145/3072959.3073712

nerfstudio. 2022. Nerfstudio Viewer Tutorial. https://www.youtube.com/watch?v=nSFsugarWzk

NerfStudio. 2024. NerfStudio. Retrieved April 3, 2024 from https://docs.nerf.studio/

Zhan Peng, Xinyi Ye, Weiyue Zhao, Tianqi Liu, Huiqiang Sun, Baopu Li, and Zhiguo Cao. 2024. 3D Multi-frame Fusion for Video Stabilization. arXiv:2404.12887 [cs.CV] https://arxiv.org/abs/2404.12887

Mike Roberts and Pat Hanrahan. 2016. Generating dynamically feasible trajectories for quadrotor cameras. *ACM Trans. Graph.* 35, 4, Article 61 (jul 2016), 11 pages. https://doi.org/10.1145/2897824.2925980

Mohamed Sayed, Robert Cinca, Enrico Costanza, and Gabriel Brostow. 2022. LookOut! Interactive Camera Gimbal Controller for Filming Long Takes. *ACM Transactions on Graphics* 41, 3 (March 2022), 30:1–30:16. https://doi.org/10.1145/3506693

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

Nerfstudio Team. 2022. https://docs.nerf.studio/

Ryotaro Temma, Kazuki Takashima, Kazuyuki Fujita, Koh Sueda, and Yoshifumi Kitamura. 2019. Third-Person Piloting: Increasing Situational Awareness using a Spatially Coupled Second Drone. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology* (New Orleans, LA, USA) *(UIST '19)*. Association for Computing Machinery, New York, NY, USA, 507–519. https://doi.org/10.1145/3332165.3347953

Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 2023. 4D Gaussian Splatting for Real-Time Dynamic Scene Rendering. arXiv:2310.08528 [cs.CV] https://arxiv.org/abs/2310.08528

Ke Xie, Hao Yang, Shengqiu Huang, Dani Lischinski, Marc Christie, Kai Xu, Minglun Gong, Daniel Cohen-Or, and Hui Huang. 2018. Creating and chaining camera moves for quadrotor videography. *ACM Trans. Graph.* 37, 4, Article 88 (jul 2018), 13 pages. https://doi.org/10.1145/3197517.3201284

Keyang Ye, Qiming Hou, and Kun Zhou. 2024. 3D Gaussian Splatting with Deferred Reflection. In *ACM SIGGRAPH 2024 Conference Papers* (Denver, CO, USA) *(SIGGRAPH '24)*. Association for Computing Machinery, New York, NY, USA, Article 40, 10 pages. https://doi.org/10.1145/3641519.3657456

Zehao Yu, Anpei Chen, Binbin Huang, Torsten Sattler, and Andreas Geiger. 2023. Mip-Splatting: Alias-free 3D Gaussian Splatting. *arXiv:2311.16493* (2023).

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*.

Boming Zhao, Yuan Li, Ziyu Sun, Lin Zeng, Yujun Shen, Rui Ma, Yinda Zhang, Hujun Bao, and Zhaopeng Cui. 2024. GaussianPrediction: Dynamic 3D Gaussian Prediction for Motion Extrapolation and Free View Synthesis. In *ACM SIGGRAPH 2024 Conference Papers* (Denver, CO, USA) *(SIGGRAPH '24)*. Association for Computing Machinery, New York, NY, USA, Article 84, 12 pages. https://doi.org/10.1145/3641519.3657417

Weiyue Zhao, Xin Li, Zhan Peng, Xianrui Luo, Xinyi Ye, Hao Lu, and Zhiguo Cao. 2023. Fast Full-frame Video Stabilization with Iterative Optimization. arXiv:2307.12774 [cs.CV] https://arxiv.org/abs/2307.12774