SpatialSync: High-Frequency Cross-Verification Between 2D Images and 3D Reconstructions

Mose Sakashita
Fujitsu Research of America
Pittsburgh, Pennsylvania, USA
msakashita@fujitsu.com

Prajit Krisshna Kumar
Fujitsu Research of America
Pittsburgh, Pennsylvania, USA
pkkumar@fujitsu.com

David Lindlbauer
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
davidlindlbauer@cmu.edu

Koichiro Niinuma
Fujitsu Research of America
Pittsburgh, Pennsylvania, USA
kniinuma@fujitsu.com

Aakar Gupta
Fujitsu Research of America
Pittsburgh, Pennsylvania, USA
agupta@fujitsu.com

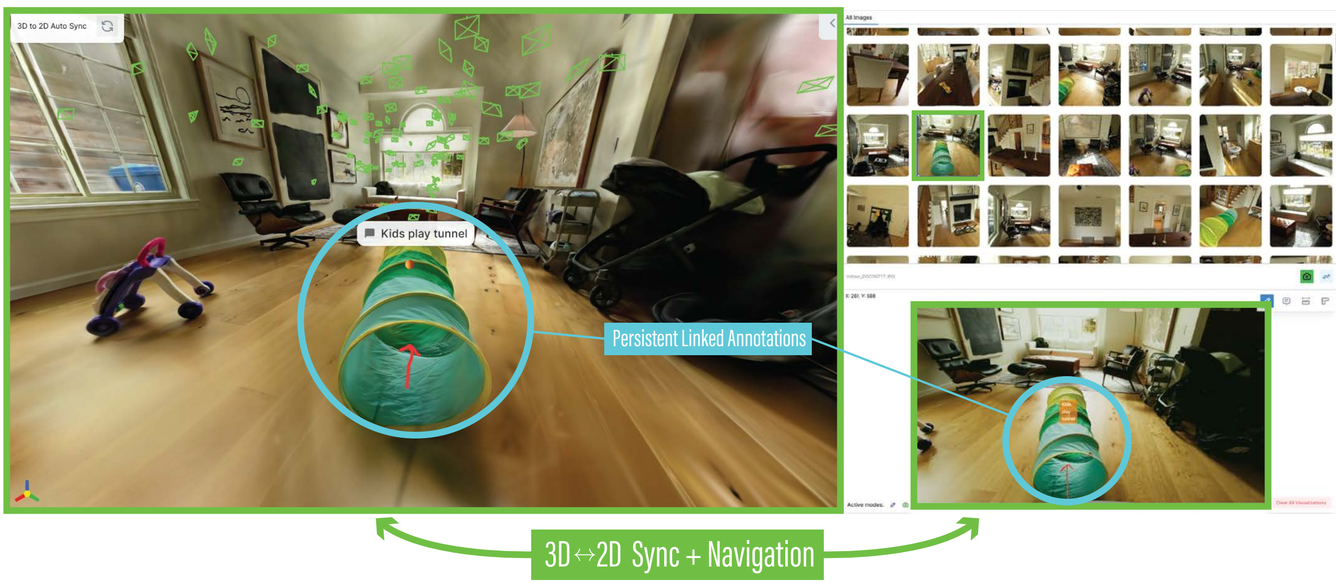

Figure 1: SpatialSync enables quick cross-modal verification through synchronized navigation between 3D reconstruction (e.g., Gaussian Splatting) and high-resolution 2D images. Users can transition between modalities while annotatiions such as labels or drawings remain bidirectionally linked and spatially anchored across representations.

Abstract

Recent advances in 3D reconstruction and view synthesis enable visually rich models from casually captured images and videos, making reconstructed environments increasingly useful for tasks such as visual inspection and spatial search. However, these reconstructions remain imperfect. They can be noisy or incomplete, smooth away fine detail, and introduce artifacts that limit reliable verification. As a result, users face a trade-off between pixel-accurate 2D images that lack spatial context and navigable 3D models that may be visually compelling but unreliable for confirming details. We present SpatialSync, an interactive system that tightly couples side-by-side 2D and 3D workspaces to reduce the cost of cross-modal validation. SpatialSync establishes bidirectional correspondence between modalities, allowing users to synchronize viewpoints, map 2D pixels to precise 3D locations, retrieve 2D evidence from 3D selections, and create persistent cross-modal annotations anchored in both representations. The system also supports multiple 3D representations, including dense point clouds and 3D Gaussian Splatting, while preserving correspondence. In a user study (N=11), participants reported that bidirectional transitions and linked annotations improved contextual grounding and reduced re-orientation effort when verifying fine details in reconstructions.

Conference acronym 'XX, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

CCS Concepts

• **Human-centered computing** → *Graphical user interfaces.*

Keywords

Spatial Interaction, 2D-3D Interaction, Cross-Modal Interaction

ACM Reference Format:
Mose Sakashita, Prajit Krisshna Kumar, David Lindlbauer, Koichiro Niinuma, and Aakar Gupta. 2026. SpatialSync: High-Frequency Cross-Verification Between 2D Images and 3D Reconstructions. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX).* ACM, New York, NY, USA, 12 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

3D reconstruction is shifting from sparse structure estimation to dense and visually convincing scene representations. For decades, Structure-from-Motion (SfM) [49] pipelines produced sparse point clouds primarily used to recover camera poses, but these reconstructions were visually impoverished and too incomplete for many visual tasks beyond coarse navigation and alignment. As a result, the images retained most of the scene's usable visual information, and interfaces treated the 3D model largely as a spatial index for navigation. Today, learning-based reconstruction and view synthesis methods can produce dense geometry and photorealistic novel views from casually captured photos and video frames. In particular, methods such as DUST3R [48] and VGGT [45], along with view synthesis techniques like Neural Radiance Fields (NeRF) [25] and Gaussian Splatting [16], can generate increasingly realistic 3D representations. This progress has expanded the role of reconstructed 3D content from a "navigational skeleton" to an increasingly central workspace for exploration and spatial reasoning.

Compared to earlier sparse reconstructions, today's dense or photorealistic models are more usable for visual inspection. They provide continuous surfaces, stronger spatial coherence, and enough appearance detail to support many first-pass judgments directly in 3D. However, dense reconstructions remain probabilistic approximations. Specifically, they may smooth away fine details (e.g., small text, labels, surface scratches), introduce view-dependent artifacts, or hallucinate plausible-but-incorrect structure. Even when visually compelling, reconstructed photorealism is frequently undermined by artifacts such as floaters and view inconsistencies [21, 30]. As a result, in many real workflows such as hardware inspection, real-estate review, infrastructure documentation, users rely on the 3D model for spatial context and coarse-to-midlevel visual assessment, while still needing the original 2D imagery as the most reliable source of ground-truth fine detail. In other words, dense reconstructions are often sufficient to surface candidate findings and localize them in context, but they are not consistently trustworthy to adjudicate those findings at pixel-level fidelity. Users therefore must repeatedly cross-check 3D observations against the original 2D images, which remain stable, high-resolution, and pixel-accurate evidence of scene appearance.

Existing interfaces still make this verification loop unnecessarily expensive. Prior systems (e.g., Photo Tourism [37]) go beyond pure camera-hopping by spatializing photo collections using recovered camera poses and supporting object-based photo browsing and annotations to be transferred across related views. However, their interaction model remains primarily organized around navigating among photos/camera viewpoints and sparse correspondences, which works well for browsing and storytelling but is less direct for high-frequency verification on dense reconstructions. Professional photogrammetry tools expose both images and reconstructions, but their interfaces are optimized for pipeline maintenance (e.g., alignment, masking, control points) and often require expertise or multi-step operations to answer simple verification questions such as: "Which image best shows this corner?" or "Where does this pixel land in 3D?"

We present **SpatialSync**, an interactive system designed to support frequent cross-modal validation between 3D reconstructions and 2D imagery during spatial inspection and analysis. Rather than spatializing photo collections into the 3D view, SpatialSync maintains separate but tightly coupled 2D and 3D workspaces presented side by side, allowing users to exploit the distinct strengths of each modality—high-resolution visual detail in images and global spatial context in reconstructions. The system supports bidirectional, content-centric transitions at the point and pixel levels. Users can map individual pixels in a 2D image to their precise 3D locations, or select a 3D region to retrieve all observing source images, ranked by visibility and viewpoint quality. To support iterative analysis and documentation, SpatialSync further enables persistent cross-modal annotations, including comments, freehand drawings, and metric measurements, that are anchored simultaneously in 2D imagery and 3D geometry and remain reusable across sessions. SpatialSync also supports multiple complementary 3D representations, including point clouds and Gaussian Splatting, allowing users to switch or blend representations while maintaining consistent bidirectional correspondence with the underlying 2D images. We conducted a preliminary user study (N = 11) in which participants completed information retrieval and spatial exploratory review tasks using SpatialSync. The study offers qualitative insights into how users engaged with bidirectional interactions and linked annotations when conducting cross-modal verification tasks between 2D images and 3D reconstructions.

In summary, our contributions include:

(1) We introduce cross-modal verification as a core interaction challenge in workflows involving high-fidelity 3D reconstructions, and present SpatialSync, an interaction framework that enables this workflow through a side-by-side workspace, enabling bidirectional synchronization, content-level 2D–3D transitions, and persistent cross-modal annotations.
(2) Findings from a preliminary user study showing how users combine 3D for spatial context and 2D for fine-detail verification, along with implications for cross-modal inspection systems.

2 Related Work

SpatialSync builds upon prior research in studies of visual detail versus spatial context, 3D reconstruction techniques from 2D images, and interactive systems that bridge 2D and 3D modalities. This section offers a comprehensive review of significant literature in these domains.

2.1 Visual Detail vs Spatial Context

Prior research has shown that 2D and 3D views offer complementary strengths depending on the task where 2D views typically support precise inspection and measurement, while 3D views aid in understanding global shape, structure, or layout [14, 39]. For example, Wickens et al. [38] observed that users rely on 3D views to gain a qualitative understanding of a scene and to present that understanding to others, while using 2D projections for tasks that demand quantitative comparisons or precise parameter tracking. Tory et al. [43] found that combining 2D and 3D views outperformed either in isolation for tasks requiring both high-level spatial reasoning and fine-grained accuracy. A range of prior focus+context visualization systems have explored techniques to help users balance detailed inspection with broader contextual awareness [5, 9, 15, 28, 29, 35]. These systems span diverse domains, including multi-resolution display techniques, augmented reality overlays, scientific data analysis, and navigation in synthetic 3D environments. Building on these studies, SpatialSync is designed to provide side-by-side workspaces supporting smooth cross modal verification between 3D reconstruction view to help users understand spatial context and 2D images for precise and detailed validation.

2.2 3D Reconstruction from 2D Images

Recent advances in computer vision have made it increasingly feasible to reconstruct 3D scenes from casually captured images and videos. Traditional pipelines such as Structure-from-Motion (SfM) [49] and Multi-View Stereo (MVS) [34] recover camera parameters and geometry through multi-stage optimization, typically producing sparse point clouds (and, when successful, coarse meshes) that are well-suited for pose estimation and navigation. For many visually grounded tasks, however, these outputs are too incomplete and visually impoverished to support inspection: fine-grained appearance remains largely "locked" in the source imagery, and the 3D model functions primarily as a spatial index over 2D views.

In contrast, recent learning-based reconstruction methods, including CUT3R [47], VGGT [45], and related approaches [46, 48, 52], have substantially increased reconstruction density and visual coherence. These feed-forward or hybrid methods can produce geometry that is useful not only for navigation, but also for first-pass visual inspection where users can perceive surfaces, continuity, and spatial relationships directly in 3D. View synthesis methods such as NeRF [25] and Gaussian Splatting [16] further improve apparent realism by representing scenes volumetrically and enabling photorealistic renderings from novel viewpoints. However, even these dense representations remain imperfect as sources of visual truth. Reconstructions may smooth away high-frequency detail or contain missing or inconsistent geometry, while view synthesis can introduce view-dependent artifacts such as floaters [21] and popping [30] and often fails to match the sharpness of the original 2D evidence. A growing body of work augments reconstruction with diffusion-based generative priors [20, 26, 50, 51], for example by synthesizing additional views to fill gaps and improve appearance. While these techniques can increase visual completeness, they also complicate verification since generated views may not correspond to any captured image, making it difficult for users to confirm whether a visually plausible detail reflects the physical scene or the model's prior.

Across both reconstruction and view synthesis, a persistent verification gap remains. Dense 3D representations are increasingly useful for inspection and spatial reasoning, but the original 2D images still provide the most reliable, pixel-accurate evidence for fine detail. SpatialSync addresses this gap by introducing interaction techniques that couple 3D reconstructions with their underlying 2D imagery for rapid, high-frequency validation.

2.3 Spatial and Cross-Modal Interaction

HCI systems have explored cross-modal interaction and annotation across 2D and 3D representations in domain-specific settings. For example, RealityEffects [19] supports authoring effects and annotations in volumetric video; VRSketchIn [10] and SymbiosisSketch [3] investigate pen- and tablet-based workflows that blend 2D sketching with 3D volumes; and Marcillat et al. [23] project 2D image annotations onto 3D models for underwater scenarios. Other systems connect modalities for collaboration and learning: Thing2Reality [13] links 2D whiteboards with 3D content in mixed reality, LoKi [40] combines high-fidelity 2D video with volumetric 3D data for learning, and TransceiVR [41] enables asymmetric communication where 2D drawings are projected into a shared 3D environment. Hybrid interfaces have also been explored in medical imaging workflows, integrating 2D annotations with 3D models to support diagnosis [22]. SharedNeRF [32] blends point cloud and view-synthesis rendering techniques to better support real-time interactions.

SpatialSync also builds on a long lineage of systems that link images to spatial representations for navigation, exploration, and annotation. Early work on spatializing photo collections, most notably Photo Tourism [37], demonstrated how recovered camera poses and sparse geometry can organize large collections and enable object-based navigation and transferring annotations across views. PhotoCloud [7] supports interactions with coupled 2D and 3D with spatially meaningful photo layout, while Nuernberger et al. [27] study "snapping-to-photos" techniques for virtual travel in reconstructed environments. Related systems use reconstruction as a scaffold to connect overlapping media for browsing and storytelling, such as Videoscapes [42], or to assist content creation and labeling during reconstruction, such as interactive architectural modeling from photo collections [36]. More recent work also aligns photo collections with terrain or map context for immersive navigation (e.g., Immersive Trip Reports [6]). Much of the foundational literature frames 2D-3D correspondence around browsing and viewpoint-based navigation rather than rapid, repeated verification of fine-grained visual claims against pixel-accurate source evidence. Commercial photogrammetry and spatial capture tools, including Autodesk ReCap [4], Agisoft Metashape [2], RealityScan [12], and Matterport [24], expose links between images and reconstructions. However, these systems are largely designed around photogrammetry and 3D maintenance workflows (such as alignment, masking, and control-point editing) and typically require expert knowledge or multi-step operations to resolve simple verification queries. In contrast, SpatialSync is designed for high-frequency

Figure 2: SpatialSync interface overview. SpatialSync combines a 3D Reconstruction Viewer (left), 2D Image List (top right), and 2D Image Viewer (bottom right). The system enables users to explore spatial context in 3D while referencing high-resolution 2D images, with support for bidirectional selection and image search.

2D–3D verification and visual analysis, presenting images and reconstructions in a side-by-side workspace with bidirectional and content-level linkage, enabling users to rapidly pivot between pixel-accurate 2D evidence and spatial 3D context in a few interactions when resolving routine verification queries.

3 SpatialSync

SpatialSync is designed to support tasks that require the complementary strengths of high-resolution 2D images and spatially coherent 3D reconstructions. In many real-world scenarios, users should be able to quickly move between these two modalities, leveraging the spatial overview and initial observations offered by 3D and relying on the visual fidelity of 2D images to inspect fine-grained details such as text, surface defects, or material properties. Alternatively, users may notice a detail in a single image and need to situate it within the broader scene, understanding its 3D location, its relationship to nearby structures, and whether similar features appear elsewhere. This process often requires quick access to the most informative supporting views and shifts between the two representations. SpatialSync is designed to make these transitions lightweight and accessible to non-experts, allowing users to move from 3D to 2D (*context* → *focus*), and from 2D to 3D (*focus* → *context*).

3.1 Overview

SpatialSync provides side-by-side 2D and 3D workspaces for high-frequency cross-verification between 2D images and 3D reconstructions through three core features. First, it provides Bidirectional View Synchronization, allowing users to align the 2D and 3D views, for example, switching the 3D camera to match a selected image's perspective or identifying the image corresponding to a selected 3D viewpoint. Second, it supports Modal Transition Navigation, enabling users to locate a point of interest in one modality and directly transition to its counterpart in the other. Users select a detail in an image to reveal its location in the 3D reconstruction, or start in 3D and reveal relevant original images. Lastly, the system supports Linked Annotations, user-authored annotations that link image regions and 3D locations, remaining accessible across sessions to support iterative analysis or documentation and collaboration. This workflow enables high-frequent **focus↔context** transitions, allowing users to seamlessly shift between detailed visual inspection and holistic spatial understanding, and to carry out complex tasks such as visual search, defect analysis, and spatial documentation.

A user can simply start SpatialSync by providing a set of images and the system automatically establishes the correspondences and shows the visualization interface. As illustrated in Fig. 2, the SpatialSync interface consists of three main components:

3D Reconstruction Viewer: Shows the reconstructed scene along with camera frustums, allowing users to freely navigate using mouse or keyboard input to examine spatial layout and identify regions of interest.

2D Image List: A scrollable panel of thumbnails representing all input images corresponding to the scene. The image selected by the user is highlighted in the panel, enabling quick switching between images.

2D Image Viewer: Displays the selected image in detail, with support for zooming, panning, and region-based annotations. This serves as the primary interface for detail-first workflows.

3.2 Bidirectional View Synchronization

When users identify an object or region of interest in a 2D image, aligning the 3D view to the same perspective typically requires manual navigation through a process that is time-consuming and cognitively demanding. Without clear spatial cues, users may struggle to establish a mental connection between the 2D image and its 3D location, leading to disorientation or additional effort. Conversely, users exploring a 3D reconstruction may want to inspect finer details via the original images, but locating the most relevant views requires guessing or tedious search.

To address these challenges, SpatialSync supports bidirectional view synchronization, which keeps 2D and 3D views spatially grounded. This functionality helps users rapidly align perspectives between modalities, making it easier to stay "on the same page" cognitively and visually. Below, we describe how synchronization works in both directions.

3.2.1 2D-to-3D synchronization. As demonstrated in Fig. 3, when a user hovers over an image in the 2D image list, the corresponding 3D camera pose rendered as a camera frustum is highlighted in the 3D view, providing immediate spatial grounding. These camera poses represent where these original images were captured. Selecting an image causes the 3D viewer to automatically transition to that image's original camera pose and camera parameters such as field-of-view, aligning the viewpoint accordingly. At the same time, the image is displayed in full in the 2D viewer for detailed inspection. This interaction allows users to move seamlessly from image-based exploration to spatial reasoning by situating the image within the broader 3D context.

3.2.2 3D-to-2D synchronization. In many workflows, users begin by exploring the 3D reconstruction to understand spatial layout or identify regions of interest, then need to inspect those regions in the original high-resolution images. In large-scale scenes with hundreds of source images, manually determining which images provide the most informative views can be time-consuming and error-prone. SpatialSync reduces this overhead by providing two complementary mechanisms for synchronizing from 3D to 2D. First, users can interact directly with camera frustums rendered in the 3D view (Fig. 4), each of which represents the position and orientation from which a 2D image was captured. Hovering over a frustum highlights its corresponding image, while selecting a frustum loads that image into the 2D viewer and aligns the 3D view to the same perspective. This interaction allows users to quickly connect spatial locations in the reconstruction with their source imagery. Second, users can enable an automatic image synchronization mode in which the 2D viewer continuously updates to display the image whose camera viewpoint most closely matches the user's current 3D view (Fig. 5). This mode supports fluid exploration by automatically surfacing the most relevant 2D evidence as users navigate the scene, without requiring explicit selection. Together, these mechanisms provide lightweight access to high-fidelity 2D imagery from within the 3D environment, supporting efficient verification of fine-grained visual details. In some scenarios, users may prefer to explore 2D and 3D content independently. For example, one might want to maintain a global overview in 3D while inspecting a series of images in detail, or keep a single image as a visual reference while freely navigating the scene. To support these workflows, SpatialSync allows users to toggle synchronization on or off at any time, providing flexibility in how views are coordinated.

3.3 Modal Transition Navigation

While bidirectional view synchronization provides spatial grounding at the image level, many tasks require more localized, fine-grained transitions between modalities. SpatialSync supports modal transition navigation, enabling users to move seamlessly between specific features in 2D and 3D—not just aligning camera viewpoints, but directly identifying and navigating to corresponding elements with pixel- or point-level precision. For instance, a user may wish to select a small visual detail (e.g., a label, a defect, or a tool) in a 2D image and examine its position and surroundings in 3D. Conversely, when exploring a 3D reconstruction, a user may encounter

Figure 3: Hovering over a 2D image highlights its corresponding camera frustum in the 3D reconstruction. Clicking a 2D image automatically orients the 3D viewer to the associated camera pose and loads the selected image into the focused 2D viewer.

Figure 4: Hovering over a camera pose in the 3D reconstruction highlights the corresponding 2D image in the image list. Clicking the camera pose loads the associated image in the 2D viewer and synchronizes the 3D view to match the selected image's viewpoint.

Figure 5: As the user navigates in the 3D Reconstruction Viewer (top), the system continuously updates and displays the closest 2D image (bottom) based on the current 3D viewpoint.

an object of interest and want to retrieve all images that captured it, to verify surface textures, lighting conditions, or image quality.

3.3.1 Focus → Context. As demonstrated in Fig. 6, by clicking on any pixel in the image, the system retrieves the corresponding 3D point from the reconstruction. The 3D viewer then navigates to center the camera on the selected point and updates the pivot for rotation accordingly, allowing users to rotate, pan, or zoom around the region of interest. This enables smooth transitions from fine-grained image inspection to spatial exploration at the pixel or point level.

3.3.2 Context → Focus. While users could manually search nearby frustums or synchronize the 3D viewpoint with the image viewer, it is often unclear which images contain the most useful views, especially when image quality or coverage varies. To support this, the user can click directly on the 3D model. As shown in Fig. 7, the system casts a ray from the screen-space cursor and identifies points within an adjustable radius of the clicked location. It then determines which input images observe those 3D points. These images are ranked based on the number of visible points, then presented as a sorted list in a dedicated tab within the 2D image list. The corresponding 3D points are also highlighted in the reconstruction viewer, offering immediate visual feedback. This mechanism allows users to quickly retrieve and compare relevant images, identify the most informative views, and transition into focused visual inspection with minimal effort.

Figure 6: Users can click on a specific region within a 2D image, and the 3D viewer will automatically navigate to the corresponding location in the reconstruction.

3.4 Linked Annotations Between Modalities

While transient correspondences presented above provide lightweight visual link between the modalities and support on-the-fly navigation and exploration, many tasks require users to preserve observations, mark important locations, or communicate insights across sessions. To support these needs, SpatialSync allows users to create **persistent correspondences** through spatially anchored

Figure 7: Clicking on a point in the 3D reconstruction triggers a spatial query that identifies nearby 3D points and retrieves all 2D images that capture that region. These relevant images are displayed in a new tab within the 2D image list, allowing users to select the optimal image for detailed visual inspection.

Figure 8: Persistent Linked Annotations. Users can create bidirectionally linked annotations, such as drawings, comments, distance, and angle measurements, directly within the 2D Image Viewer. These annotations are automatically synchronized with the corresponding location in the 3D Reconstruction Viewer and remain accessible across sessions.

annotations that link 2D image content to 3D reconstruction, as shown in Fig. 8.

3.4.1 Comments. Users can place a comment on any region of an image by clicking within the 2D viewer and typing a note. The comment is immediately visualized in both the 2D image and the corresponding 3D location. These comments serve as persistent reference points for revisiting observations, leaving contextual notes, or sharing information with collaborators.

3.4.2 Drawings. Users can draw directly onto the 2D image, such as sketching outlines, highlighting regions, or adding arrows. These drawings are projected into the 3D space by finding the corresponding 3D points along the pixel trajectory drawn by the user on the image, allowing spatially meaningful visual markings that retain their original context.

3.4.3 Measurements. For quantitative tasks, users can add measurements by selecting two or more points on the 2D image. SpatialSync retrieves the corresponding 3D coordinates and computes distances or angles, then overlays the results as visual elements in both the 2D and 3D views. This allows users to analyze object dimensions or spatial relationships.

These annotations—comments, drawings, and measurements—can be created directly from the 2D image viewer and are automatically projected into the 3D scene, maintaining a bidirectional link between modalities. All annotations in 3D are interactively linked back to their source image.

Because these annotations can persist across sessions, they support collaborative review, long-term documentation, and iterative analysis. By tightly linking observations across modalities, SpatialSync enables users to preserve, revisit, and share insights without losing spatial grounding or identified visual detail.

3.5 Implementation

SpatialSync employs a client-server architecture built on the Viser visualization framework [1]. The server, implemented in Python, handles 3D data processing and coordinates communication with clients, while the browser-based client renders visualizations and captures user interactions. Upon startup, the server processes a set of user-provided images to generate a 3D reconstruction using the VGGT model [45]. The server also manages a WebSocket server to transmit scenes to connected clients. We extend Viser's base functionality with custom message handlers for interactions such as point cloud selection, pixel-to-3D-point mapping, and visualization management. The client consists of React components that render the 3D scene using Three Fiber, with components for the image viewer and interactive visualization tools. Communication between components operates through a bidirectional WebSocket protocol, enabling real-time updates as users interact with either the 2D or 3D representations of the scene. Message handlers on both ends process events including camera movements, image selection, and the creation of measurements or annotations. The persistent annotations are linked across modalities, stored as JSON files, and can be reloaded in later sessions to support iterative analysis and documentation.

4 Usage Scenarios

We illustrate how SpatialSync can support various tasks that require both high-fidelity visual detail and spatial context. The following scenarios highlight how users can leverage bidirectional navigation, transient correspondences, and persistent annotations to perform spatial inspection, visual search, and collaborative analysis. Most of the datasets used in our scenarios and demonstrations are derived from the SMERF dataset [11] and the LERF dataset [17], with the exception of the hardware scene, which was created using our own custom data.

4.1 Reviewing Hardware Prototyping

Consider a scenario where an engineering student sends a video of a hardware prototype to their instructor for feedback. Fig. 9 illustrates that, using SpatialSync, the instructor can view a 3D reconstruction of the device and quickly identify components like microcontrollers or sensors based on spatial layout. To verify wiring or identify

Figure 9: Hardware Review Scenario. Users inspect components in 3D, retrieve relevant 2D images to view wiring details, and leave persistent annotations for documentation.

Figure 10: Label Information Retrieval. Users can select an item based on its spatial context and geometry in the 3D view, retrieve all images capturing that region, and examine the clearest image of the label or price tag in the 2D viewer.

specific modules, they can select a region in 3D and retrieve images that show the detail at high resolution. If the instructor notices a wiring issue or component mismatch, they can add a persistent comment directly onto the relevant image, which is linked to the 3D location for spatial context. These annotations make it easy to give targeted feedback and support iterative review across versions or submissions.

4.2 Label Information Retrieval

In retail or storage settings, users often need to verify product details—such as labels, prices, or placement—across a large, visually dense environment. With SpatialSync, a user can load image datasets of a store aisle and explore both the spatial layout and fine-grained visual details of products, as shown in Fig. 10. For example, to find the price of a specific fruit, the user can navigate through the 3D reconstruction to locate the item's position on the shelf, then retrieve the most relevant high-resolution image that clearly shows the price tag. Once identified, they can annotate the image with a comment for future reference, preserving both the visual and spatial context for documentation or review.

5 User Study

We conducted a preliminary user study to explore how SpatialSync supports tasks that require both spatial awareness and fine-grained visual inspection, and to better understand how users value and appropriate its bidirectional 2D–3D interaction techniques designed to support high-frequency cross-modal verification. Rather than benchmarking performance against an existing baseline, the goal of the study was to qualitatively assess whether SpatialSync's interaction techniques and its design support users' real-world inspection and review workflows, and to surface strengths, limitations, and opportunities for refinement. We recruited 11 participants (2 female, 9 male), aged 23 to 45 (Mdn = 27), all of whom received a $60 gift card for their participation. We did not require prior experience with 3D software or related tools.

5.1 Tasks

We designed two tasks to reflect common real-world scenarios in which users must repeatedly transition between spatial context and detailed visual evidence. These tasks were chosen to exercise different aspects of SpatialSync's interaction design.

5.1.1 Information Retrieval Task. Participants were asked to locate two specific products in a dense grocery store aisle using SpatialSync. They viewed a 3D reconstruction synchronized with a set of high-resolution 2D images (178 images), showing the same scene as demonstrated in Section 4.2. Their objective was to quickly find target items, identify them, and write down the corresponding prices directly on an aisle diagram (see Fig. 11). This task was designed to simulate scenarios that require both spatial orientation (e.g., finding the correct shelf) and visual detail (e.g., reading price tags).

5.1.2 Exploratory Home Review Task. Participants completed two subtasks simulating real estate scenarios: one from a homebuyer's perspective and one from a real estate agent's perspective.

Task A: Homebuyer Perspective. Participants imagined they were evaluating a room as a potential homebuyer. They were provided with a visualization that included several linked annotations in 3D reconstructions and 2D images (see Fig. 12). Their task was

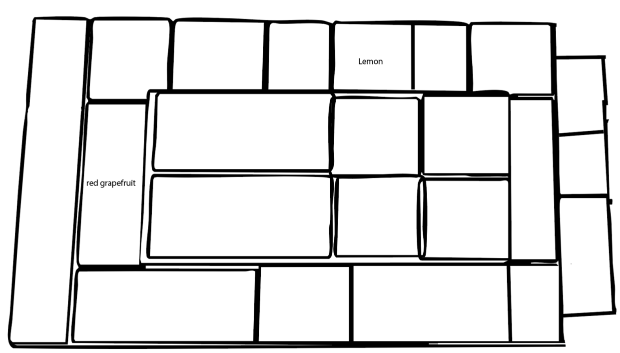

Figure 11: Diagram the participants used to place the items for the information retrieval task.

Figure 12: The house scene filled with persistent annotations created by the "real estate agent," which participants interacted with to explore and learn about the home during Task A of the Exploratory Home Review task.

to review key room specifications via annotations, examine highlighted design features such as furniture and decor, and use both the 3D view and 2D images (235 images) to develop a comprehensive understanding of the space. Participants were instructed to take notes during a 5-minute review and then provide a 1-minute verbal summary of their observations to the researcher.

Task B: Real Estate Agent Perspective. In this task, participants assumed the role of a real estate agent preparing a visual presentation for a client. Using SpatialSync, they explored and identified items or design features in the same scene (274 images). They were also asked measure and annotate key spatial metrics, for example, determining the dimensions of a couch and estimating how many people it could comfortably accommodate. In addition, participants identified two or more design features they would recommend to a client and added linked annotations directly to the visualization. After 5 minutes of exploration and annotation, participants gave a 1-minute summary of their findings and recommendations.

5.2 Procedure

The study began with participants completing an informed consent form, followed by a brief tutorial on the core features of SpatialSync. Participants were given time to practice these features before beginning the tasks. They first completed the *Information Retrieval Task*. Then they were introduced to the linked annotation feature. After practicing this feature, they proceeded with the *Exploratory Review Task*. After completing both tasks, participants filled out a post-study questionnaire, which included the System Usability Scale (SUS) [8] and open-ended questions for qualitative feedback.

5.3 Results and Feedback

SpatialSync received a mean SUS score of 73.0, placing it in the *"Good"* usability category and corresponding to the 70th–80th percentile based on industry benchmarks [18]. We now summarize the insights from the feedback of the participants.

Bidirectional View Synchronization. Participants consistently highlighted how bidirectional synchronization helped maintain spatial grounding and connect views across modalities. P1 found it *"beneficial to see the elements from both 2D and 3D views,"* while P6 noted they *"definitely switched back and forth between the two views, with the 3D view providing more of an overview of the space and the 2D view providing more details."* This complementarity was central to task performance, as P6 emphasized: *"since the two views provided different kinds of information, using both helped me complete the tasks better."*

For P5, the combination was not just helpful but essential: *"Without the 2D views, the 3D would likely be insufficient for identifying fine details. Similarly, without 3D, it would be more cumbersome to identify all of the features in a timely manner."* P1 described it as *"intuitive to get some local information from 2D images, and then mapping it to the corresponding 3D model helps to add its global position information,"* while P11 echoed this need by pointing out that *"the point cloud was too coarse to be able to capture the low-level details, hence it was useful having the synced 2D views."*

Participants also relied on bidirectional navigation for effective search: *"If the 3D view does not help, quickly looking through the images helps in navigating the 3D environment,"* said P4. P11 found it helpful to *"see the coarse location in the 3D view and then find the corresponding camera poses and inspect the 2D picture to read the details."* P1 summarized the general sentiment: *"To search for some spatial information, it is necessary to frequently exchange the views between the 2D image and 3D model."*

Modal Transition Navigation. Transient point-level correspondences were especially valued for exploration and verification. P7 described how they *"used the CTRL-click (3D to 2D image retrieval) feature a ton to find high-res images of objects I found in the 3D model,"* and P5 similarly noted, *"clicking on a part of the 3D model to find the corresponding image greatly improved my workflow."* This helped narrow the search space, as P4 explained: *"sometimes the 3D model is not accurate, but clicking on it gives the image representation as well."*

Participants found this transition from spatial context to detailed views highly effective. *"I found it very easy to match these 3D elements to the corresponding 2D images . . . because . . . anything in 3D can be tracked back to its 2D source,"* said P5. P10 praised the 2D-to-3D pipeline, saying it *"was seamless, made it super easy to understand the 3D point I was interacting with directly from 2D image."* P2 found that with modal navigation, *"I could easily find images I needed,"* and P6 emphasized how helpful it was *"to look through multiple 2D images to find one with a good angle to read the price."*

Linked Annotations. Persistent annotations—including comments, drawings, and measurements—enabled users to externalize their thinking and revisit observations across sessions. P5 described how the ability to *"add measurements and comments to the scene further solidified my overall understanding of the scene, allowing me to take notes along the way."* P11 appreciated the tight linkage across views, noting that *"the ability to add measurements or comments in one view and seeing those changes reflected in the other view is useful because then I don't have to find which specific image I annotated."*

This annotation linking supported personal reference as well as reviewing existing annotations in the scene. P9 found it helpful that annotations *"helped me keep track of notes to myself,"* and P2 noted that *"clicking comments and seeing the view from specific viewpoints are helpful"* in quickly regaining spatial context.

While overall feedback was highly positive, participants identified a few areas for improvement, particularly related to performance and visual quality. Several users reported noticeable lag when interacting with the 3D model. P11 noted, *"there was a bit more lag than I would've liked,"* while P3 mentioned the delay in the image retrieval process commenting that *"it would be nice if the (point-based image) search function was faster."* In addition, participants expressed a desire for higher-fidelity 3D reconstructions to better support detailed spatial reasoning. P5 suggested that *"a higher fidelity scene representation (like 3DGS) would have made it a bit easier,"* highlighting the limitations of the point cloud rendering in its current form. These reflections point to opportunities for improving rendering performance and visual realism in future iterations of the system (see Section 6.1).

The results from the user feedback validated the core premise of SpatialSync, that enabling bidirectional, synchronized interaction between 2D images and 3D reconstructions enhances users' ability to perform spatially and visually demanding tasks. Participants consistently leveraged the strengths of both representations—using 3D views for orientation and global context, and 2D images for fine-grained inspection and verification. The combination of transient navigation and persistent annotations supported not only efficient information retrieval but also meaningful spatial reasoning and documentation.

6 Discussion

We reflect on the limitations of SpatialSync and outline opportunities to expand its capabilities in future work.

6.1 Visual Fidelity and Complementary 3D Representations

Our current system primarily relies on VGGT point cloud reconstructions, which provide stable geometry and reliable correspondence to source images. However, participants in the user study noted limitations in visual fidelity, particularly when inspecting fine-grained spatial detail in the 3D view. In response to this feedback, we explored integrating 3DGS [16] as an optional high-fidelity rendering layer. This extension allows users to access photorealistic view synthesis while preserving bidirectional correspondence with the underlying 2D imagery and VGGT spatial structure (see Fig. 13). Users can toggle between or blend point cloud and 3DGS representations depending on task needs. The point cloud remains effective in regions with sparse coverage or uncertain geometry, while 3DGS provides smoother and more visually coherent renderings in well observed areas. This exploration highlights a broader design opportunity of supporting multiple complementary 3D representations within a cross-validation interaction framework proposed by this paper. Rather than replacing one representation with another, enabling users to fluidly switch or combine them may better accommodate varying data quality and inspection needs.

Figure 13: Users can switch or blend between aligned VGGT point cloud and 3D Gaussian Splatting (3DGS) renderings while preserving consistent spatial correspondence.

6.2 Accommodating Dynamic 3D Scenes

SpatialSync is currently designed for static scenes in which the 3D reconstruction is derived from a set of images or video frames that capture the static scene. However, many real-world scenarios involve dynamic elements, such as people or vehicles in motion. Extending the system to support dynamic 3D scenes could open up new workflows—for example, selecting a moving object in a 2D image to visualize its trajectory in 3D, or enabling camera-following behaviors to track dynamic targets across frames. Incorporating temporal correspondences and persistent 3D trajectories would enhance the utility of SpatialSync for analysis tasks involving movement, surveillance, or event review.

6.3 Supporting more modalities to engage with 3D and 2D representations

While SpatialSync currently supports bidirectional navigation between 2D images and 3D model, participants expressed interest in richer querying capabilities, particularly the ability to search using natural language or semantic attributes. For example, a user might want to search for "chair" or "car" and instantly locate all relevant 2D and 3D instances. This could be achieved by integrating semantic segmentation or visual grounding models into the 2D analysis pipeline, and propagating those labels to the 3D point cloud. Doing so would enable more flexible, task-driven access to content across modalities, facilitating efficient inspection, annotation, and retrieval.

6.4 Asynchronous and Synchronous Collaboration

Persistent annotations in SpatialSync are well-suited for asynchronous collaboration, allowing users to leave comments, drawings, or measurements for others to review later with both spatial context and source-image evidence preserved [44]. Future versions of SpatialSync could similarly support asynchronous handoff through annotated 2D-3D evidence trails that users can revisit and extend across sessions. Beyond asynchronous review, SpatialSync could also support real-time, synchronous collaboration. Since the system already operates with a client–server architecture, interactions such as viewpoint changes, 3D selections, retrieved evidence images, and annotation creation could be broadcast across multiple users in real time. Prior work on remote collaboration highlights the importance of preserving spatial awareness, and collaborators' locus of attention during distributed spatial tasks [31, 33]. Building on these ideas, SpatialSync could expose collaborators' active viewpoints, selected regions, inspected images, and annotation activity as shared awareness cues.

7 Conclusion

We present SpatialSync, an interactive system that unifies high-resolution 2D images and 3D reconstruction through bidirectiional synchronization, modal transition navigation, and persistent linked annotations. By allowing users to freely move between 2D details and 3D representations, SpatialSync can support new workflows for tasks that require high-frequency cross modal verification between visual detail and spatial awareness, such as spatial search, inspection, and collaborative documentation. Our user study showed that participants relied on the complementarity of 2D and 3D views to complete complex spatial reasoning tasks.

References

[1] 2025. Viser. https://viser.studio/main/. Accessed January 2026.

[2] Agisoft. 2026. Agisoft Metashape. https://www.agisoft.com/. Accessed January 2026.

[3] Rahul Arora, Rubaiat Habib Kazi, Tovi Grossman, George Fitzmaurice, and Karan Singh. 2018. SymbiosisSketch: Combining 2D & 3D Sketching for Designing Detailed 3D Objects in Situ. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–15. doi:10.1145/3173574.3173759

[4] Autodesk. 2026. Autodesk ReCap. https://www.autodesk.com/products/recap/overview. Accessed January 2026.

[5] Patrick Baudisch, Nathaniel Good, and Paul Stewart. 2001. Focus plus context screens: combining display technology with visualization techniques. In *Proceedings of the 14th Annual ACM Symposium on User Interface Software and Technology* (Orlando, Florida) *(UIST '01)*. Association for Computing Machinery, New York, NY, USA, 31–40. doi:10.1145/502348.502354

[6] Jan Brejcha, Michal Lukác, Zhili Chen, Stephen DiVerdi, and Martin Cadík. 2018. Immersive trip reports. In *Proceedings of the 31st Annual ACM symposium on user interface software and technology*. 389–401.

[7] Paolo Brivio, Luca Benedetti, Marco Tarini, Federico Ponchio, Paolo Cignoni, and Roberto Scopigno. 2012. PhotoCloud: Interactive remote exploration of joint 2D and 3D datasets. *IEEE computer graphics and applications* 33, 2 (2012), 86–96.

[8] John Brooke. 1996. SUS: A quick and dirty usability scale. Usability Evaluation in Industry, 189–194 pages.

[9] Jan Byška, Thomas Trautner, Sérgio M Marques, Jirí Damborskỳ, Barbora Kozlíková, and Manuela Waldner. 2019. Analysis of long molecular dynamics simulations using interactive focus+ context visualization. In *Computer Graphics Forum*, Vol. 38. Wiley Online Library, 441–453.

[10] Tobias Drey, Jan Gugenheimer, Julian Karlbauer, Maximilian Milo, and Enrico Rukzio. 2020. VRSketchIn: Exploring the Design Space of Pen and Tablet Interaction for 3D Sketching in Virtual Reality. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3313831.3376628

[11] Daniel Duckworth, Peter Hedman, Christian Reiser, Peter Zhizhin, Jean-François Thibert, Mario Lučić, Richard Szeliski, and Jonathan T. Barron. 2023. SMERF: Streamable Memory Efficient Radiance Fields for Real-Time Large-Scene Exploration. arXiv:2312.07541 [cs.CV]

[12] Epic Games. 2026. RealityScan. https://www.realityscan.com/. Accessed January 2026.

[13] Erzhen Hu, Mingyi Li, Jungtaek Hong, Xun Qian, Alex Olwal, David Kim, Seongkook Heo, and Ruofei Du. 2024. Thing2Reality: Transforming 2D Content into Conditioned Multiviews and 3D Gaussian Objects for XR Communication. arXiv:2410.07119 [cs.HC] https://arxiv.org/abs/2410.07119

[14] Mark St. John, Michael B. Cowen, Harvey S. Smallman, and Heather M. Oonk. 2001. The Use of 2D and 3D Displays for Shape-Understanding versus Relative-Position Tasks. *Human Factors* 43, 1 (2001), 79–98. doi:10.1518/001872001775992534 arXiv:https://doi.org/10.1518/001872001775992534 PMID: 11482314.

[15] Denis Kalkofen, Erick Mendez, and Dieter Schmalstieg. 2007. Interactive Focus and Context Visualization for Augmented Reality. In *Proceedings of the 2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR '07)*. IEEE Computer Society, USA, 1–10. doi:10.1109/ISMAR.2007.4538846

[16] 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/

[17] Justin* Kerr, Chung Min* Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. 2023. LERF: Language Embedded Radiance Fields. In *International Conference on Computer Vision (ICCV)*.

[18] James R Lewis and Jeff Sauro. 2009. The factor structure of the system usability scale. In *Human Centered Design: First International Conference, HCD 2009, Held as Part of HCI International 2009, San Diego, CA, USA, July 19-24, 2009 Proceedings 1*. Springer, 94–103.

[19] Jian Liao, Kevin Van, Zhijie Xia, and Ryo Suzuki. 2024. RealityEffects: Augmenting 3D Volumetric Videos with Object-Centric Annotation and Dynamic Visual Effects. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference* (Copenhagen, Denmark) *(DIS '24)*. Association for Computing Machinery, New York, NY, USA, 1248–1261. doi:10.1145/3643834.3661631

[20] Fangfu Liu, Wenqiang Sun, Hanyang Wang, Yikai Wang, Haowen Sun, Junliang Ye, Jun Zhang, and Yueqi Duan. 2024. ReconX: Reconstruct Any Scene from Sparse Views with Video Diffusion Model. arXiv:2408.16767 [cs.CV] https://arxiv.org/abs/2408.16767

[21] Xinhang Liu, Yu-Wing Tai, and Chi-Keung Tang. 2023. Clean-NeRF: Reformulating NeRF to account for View-Dependent Observations. arXiv:2303.14707 [cs.CV] https://arxiv.org/abs/2303.14707

[22] Veera Bhadra Harish Mandalika, Alexander I Chernoglazov, Mark Billinghurst, Christoph Bartneck, Michael A Hurrell, Niels de Ruiter, Anthony PH Butler, and Philip H Butler. 2018. A hybrid 2D/3D user interface for radiological diagnosis. *Journal of digital imaging* 31 (2018), 56–73.

[23] Marin Marcillat, Loic Van Audenhaege, Catherine Borremans, Aurélien Arnaubec, and Lenaick Menot. 2024. The best of two worlds: reprojecting 2D image annotations onto 3D models. *PeerJ* 12 (2024), e17557.

[24] Matterport. 2026. Matterport. https://matterport.com/. Accessed January 2026.

[25] 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. In *ECCV*.

[26] Niluthpol Chowdhury Mithun, Tuan Pham, Qiao Wang, Ben Southall, Kshitij Minhas, Bogdan Matei, Stephan Mandt, Supun Samarasekera, and Rakesh Kumar. 2025. Diffusion-Guided Gaussian Splatting for Large-Scale Unconstrained 3D Reconstruction and Novel View Synthesis. arXiv:2504.01960 [cs.CV] https://arxiv.org/abs/2504.01960

[27] Benjamin Nuernberger, Matthew Turk, and Tobias Höllerer. 2017. Evaluating snapping-to-photos virtual travel interfaces for 3D reconstructed visual reality. In *Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology* (Gothenburg, Sweden) *(VRST '17)*. Association for Computing Machinery, New York, NY, USA, Article 22, 11 pages. doi:10.1145/3139131.3139138

[28] H. Piringer, R. Kosara, and H. Hauser. 2004. Interactive focus+context visualization with linked 2D/3D scatterplots. In *Proceedings. Second International Conference on Coordinated and Multiple Views in Exploratory Visualization, 2004.* 49–60. doi:10.1109/CMV.2004.1319526

[29] Huamin Qu, Haomian Wang, Weiwei Cui, Yingcai Wu, and Ming-Yuen Chan. 2009. Focus+Context Route Zooming and Information Overlay in 3D Urban Environments. *IEEE Transactions on Visualization and Computer Graphics* 15, 6 (2009), 1547–1554. doi:10.1109/TVCG.2009.144

[30] Lukas Radl, Michael Steiner, Mathias Parger, Alexander Weinrauch, Bernhard Kerbl, and Markus Steinberger. 2024. StopThePop: Sorted Gaussian Splatting for View-Consistent Real-time Rendering. *ACM Transactions on Graphics* 4, 43, Article 64 (2024).

[31] Mose Sakashita, E. Andy Ricci, Jatin Arora, and François Guimbretière. 2022. RemoteCoDe: Robotic Embodiment for Enhancing Peripheral Awareness in Remote Collaboration Tasks. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW1, Article 63 (April 2022), 22 pages. doi:10.1145/3512910

[32] Mose Sakashita, Balasaravanan Thoravi Kumaravel, Nicolai Marquardt, and Andrew David Wilson. 2024. SharedNeRF: Leveraging Photorealistic and View-dependent Rendering for Real-time and Remote Collaboration. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 675, 14 pages. doi:10.1145/3613904.3642945

[33] Mose Sakashita, Ruidong Zhang, Xiaoyi Li, Hyunju Kim, Michael Russo, Cheng Zhang, Malte F. Jung, and François Guimbretière. 2023. ReMotion: Supporting Remote Collaboration in Open Space with Automatic Robotic Embodiment. 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 363, 14 pages. doi:10.1145/3544548.3580699

[34] Johannes L. Schonberger and Jan-Michael Frahm. 2016. Structure-From-Motion Revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[35] Horn-yeu Shiaw, Robert J. K. Jacob, and Gregory R. Crane. 2004. The 3D vase museum: a new approach to context in a digital library. In *Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries* (Tuscon, AZ, USA) *(JCDL '04)*. Association for Computing Machinery, New York, NY, USA, 125–134. doi:10.1145/996350.996381

[36] Sudipta N. Sinha, Drew Steedly, Richard Szeliski, Maneesh Agrawala, and Marc Pollefeys. 2008. Interactive 3D architectural modeling from unordered photo collections. In *ACM SIGGRAPH Asia 2008 Papers* (Singapore) *(SIGGRAPH Asia '08)*. Association for Computing Machinery, New York, NY, USA, Article 159, 10 pages. doi:10.1145/1457515.1409112

[37] Noah Snavely, Steven M. Seitz, and Richard Szeliski. 2006. Photo tourism: exploring photo collections in 3D. *ACM Trans. Graph.* 25, 3 (July 2006), 835–846. doi:10.1145/1141911.1141964

[38] R.R. Springmeyer, M.M. Blattner, and N.L. Max. 1992. A characterization of the scientific data analysis process. In *Proceedings Visualization '92*. 235–242. doi:10.1109/VISUAL.1992.235203

[39] Mark St. John, Michael B. Cowen, Harvey S. Smallman, and Heather M. Oonk. 2001. Information Availability in 2D and 3D Displays . *IEEE Computer Graphics and Applications* 21, 05 (Sept. 2001), 51–57. doi:10.1109/38.946631

[40] Balasaravanan Thoravi Kumaravel, Fraser Anderson, George Fitzmaurice, Bjoern Hartmann, and Tovi Grossman. 2019. Loki: Facilitating Remote Instruction of Physical Tasks Using Bi-Directional Mixed-Reality Telepresence. 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, 161–174. doi:10.1145/3332165.3347872

[41] Balasaravanan Thoravi Kumaravel, Cuong Nguyen, Stephen DiVerdi, and Bjoern Hartmann. 2020. TransceiVR: Bridging Asymmetrical Communication Between VR Users and External Collaborators. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '20)*. Association for Computing Machinery, New York, NY, USA, 182–195. doi:10.1145/3379337.3415827

[42] James Tompkin, Kwang In Kim, Jan Kautz, and Christian Theobalt. 2012. Videoscapes: exploring sparse, unstructured video collections. *ACM Trans. Graph.* 31, 4, Article 68 (July 2012), 12 pages. doi:10.1145/2185520.2185564

[43] M. Tory, A.E. Kirkpatrick, M.S. Atkins, and T. Moller. 2006. Visualization task performance with 2D, 3D, and combination displays. *IEEE Transactions on Visualization and Computer Graphics* 12, 1 (2006), 2–13. doi:10.1109/TVCG.2006.17

[44] Cheng Yao Wang, Mose Sakashita, Upol Ehsan, Jingjin Li, and Andrea Stevenson Won. 2020. Again, Together: Socially Reliving Virtual Reality Experiences When Separated. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3313831.3376642

[45] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. 2025. VGGT: Visual Geometry Grounded Transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[46] Jianyuan Wang, Nikita Karaev, Christian Rupprecht, and David Novotny. 2023. VGGSfM: Visual Geometry Grounded Deep Structure From Motion.

[47] Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. 2025. Continuous 3D Perception Model with Persistent State. *arXiv preprint arXiv:2501.12387* (2025).

[48] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. 2024. DUSt3R: Geometric 3D Vision Made Easy. In *CVPR*.

[49] Matthew J Westoby, James Brasington, Niel F Glasser, Michael J Hambrey, and Jennifer M Reynolds. 2012. 'Structure-from-Motion'photogrammetry: A low-cost, effective tool for geoscience applications. *Geomorphology* 179 (2012), 300–314.

[50] Jay Zhangjie Wu, Yuxuan Zhang, Haithem Turki, Xuanchi Ren, Jun Gao, Mike Zheng Shou, Sanja Fidler, Zan Gojcic, and Huan Ling. 2025. Difix3D+: Improving 3D Reconstructions with Single-Step Diffusion Models. *arXiv preprint arXiv: 2503.01774* (2025).

[51] Rundi Wu, Ben Mildenhall, Philipp Henzler, Keunhong Park, Ruiqi Gao, Daniel Watson, Pratul P. Srinivasan, Dor Verbin, Jonathan T. Barron, Ben Poole, and Aleksander Holynski. 2023. ReconFusion: 3D Reconstruction with Diffusion Priors. *arXiv* (2023).

[52] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. 2024. MonST3R: A Simple Approach for Estimating Geometry in the Presence of Motion. *arXiv preprint arxiv:2410.03825* (2024).

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009