Map2Video: Guiding Real-World-Grounded AI Video Generation

Anonymous Author(s)

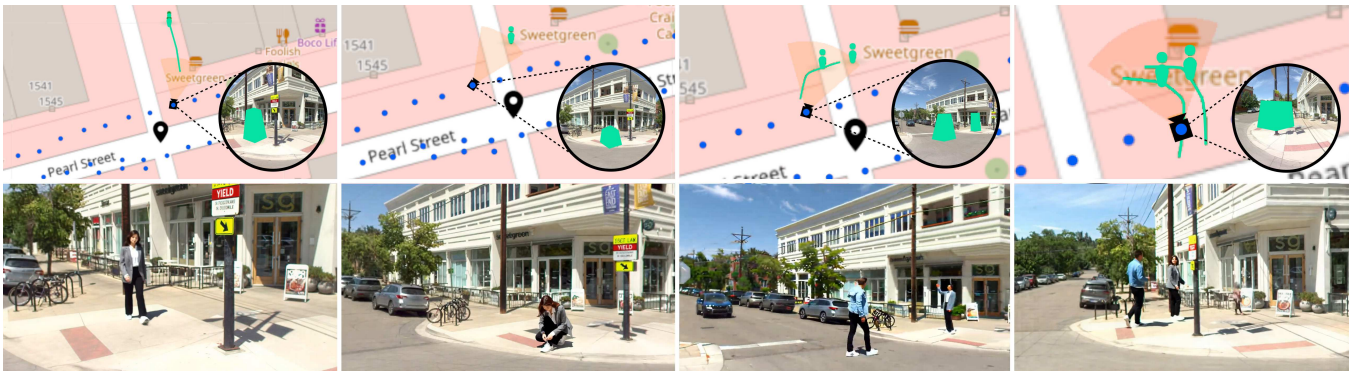

Figure 1: Map2Video helps users control video generation grounded in real-world locations by combining video inpainting with map-based annotation for character trajectory and masks and virtual camera framing on 360° street view imagery.

ABSTRACT

AI video generation lowers barriers to video creation, but existing tools provide limited support for grounding content in real-world locations, restricting their use in applications such as film previsualization and scenario generation. In practice, users rely verbose textual descriptions or reference images using on-site photos or street view imagery (SVI), which are cumbersome and unreliable across shots. Our formative study with five filmmakers revealed challenges in shot composition, character motion, and camera control when using SVI with an image-to-video tool. We present Map2Video, a system for guiding SVI-grounded AI video generation through map-based character trajectory annotation and direct manipulation of cameras and character masks defined in geodetic coordinates and projected onto panoramic screen coordinates. In an evaluation with twelve filmmakers, Map2Video provided stronger controllability for both scene replication and open-ended creative exploration, reduced cognitive effort, and improved spatial consistency across multiple shots compared to an image-to-video baseline.

CCS CONCEPTS

• **Human-centered computing**;

KEYWORDS

Generative Videos; AI-assisted Filmmaking; Human-AI Co-Creation

UIST '26, November 2–5, 2026, Michigan, USA

ACM ISBN 000-0-0000-0000-0/00/00…$15.00
https://doi.org/00.0000/0000000.0000000

ACM Reference Format:
Anonymous Author(s). 2026. Map2Video: Guiding Real-World-Grounded AI Video Generation. In *The 39th Annual ACM Symposium on User Interface Software and Technology (UIST '26), November 2–5, 2026, Michigan, USA.* ACM, New York, NY, USA, 15 pages. https://doi.org/00.0000/0000000.0000000

1 INTRODUCTION

Recent advances in text- and image-to-video generation have lowered barriers to video creation [7, 70, 90], potentially reshaping filmmaking practices [91] and democratizing creativity. However, our survey with 34 AI filmmakers revealed recurring challenges in controlling AI video generation, such as difficulties with camera control and maintaining consistency across clips. Guided by these reported challenges and the portfolio examples shared by filmmakers, we examine how to support scenes in which creators need to control how a subject moves through a real-world environment and how the camera relates to that motion. Such scenarios arise in applications such as previsualizing scenes in specific locations before filming, prototyping scenes in public spaces that are difficult or expensive to access directly, and creating walking or driving shots in real-world locations.

To depict real-world locations in AI video generation, creators often rely on reference images [1], such as on-site photography or online search results (e.g., street view imagery) [2]. While such references provide visual grounding, existing AI video generation interfaces offer limited support for specifying where the subject should appear, how it should move through the scene, and how the camera should relate to that motion. Meanwhile, maps and street view imagery provide spatial information tied to real-world locations, making them a promising foundation for specifying subject trajectories and planning camera behavior in grounded scenes.

To further investigate this opportunity, we used experience prototyping to examine a map- and street view imagery-driven workflow

[1] Runway Gen-4 Image References Guidance
[2] Curious Refuge blog post on using street view imagery for AI video generation

for grounded video creation in real-world locations. We created a lightweight proof-of-concept pipeline that combined a minimal map-based Street View Imagery (SVI) [5] retrieval interface with Runway, a popular AI video generation tool, and conducted a formative study with five filmmakers using this pipeline. Participants found the workflow promising for filmmaking, but they also identified four limitations: (1) difficulty specifying shot composition and subject placement via text prompts, (2) inconsistency in backgrounds caused by hallucination, (3) limited control over subject motion within the scene, and (4) lack of fine-grained control over camera movement.

Building on these insights, we developed Map2Video, a map- and street view imagery-driven AI video generation tool for grounded scenes involving subject movement through real-world locations. Map2Video supports this workflow through map-based annotation of masks and trajectories, direct manipulation of virtual camera framing on 360° street view imagery, and video inpainting. Inspired by traditional filmmaking practices such as location scouting and on-location rehearsal, Map2Video lets users refine subject placement, motion, and camera behavior in context before generation, supporting greater spatial control and continuity across shots. Our system supports a six-step workflow: 1) *Location Scouting*, where users select scenes and backgrounds directly on a map; 2) *Mask Positioning*, where proxy elements are placed into the street view scene and embedded in 3D space for later inpainting; 3) *Movement Sketching*, where users draw trajectories to choreograph how actors or objects move across the scene, with synchronized updates in the camera view; 4) *Camera Walkthrough*, where users refine camera placement, panning, zoom, and rotation; 5) *Prompting Scene*, where users provide natural language descriptions of characters' appearances and actions; and 6) *Video Inpainting*, where the system combines these inputs to generate the final video. We implemented Map2Video using Unity and ComfyUI with the VACE video generation model, along with OpenStreetMap and Mapillary for map and street view imagery.

We evaluated Map2Video with 12 filmmakers across replication and open-ended tasks. In the replication task, Map2Video enabled faster task completion and required fewer fine-tuning iterations (1.2×) while achieving higher perceived controllability than an image-to-video baseline. Across tasks, participants reported lower cognitive effort, higher usability, greater creative controllability, and higher perceived spatial consistency. These results suggest that Map2Video better supports controllable, location-grounded video authoring for subject motion and camera planning in real-world scenes.

In summary, our main contributions are:

(1) An experience prototyping study ($N = 5$), informed by a survey of AI filmmaking practices ($N = 34$), that identified key challenges and interaction needs for grounded video creation in real-world scenes, particularly around subject movement and camera control.
(2) The concept of map- and street view imagery-driven AI video generation for controllable subject movement and camera planning in grounded scenes.
(3) Map2Video, an interactive authoring tool that supports a six-step workflow by integrating video inpainting with map-based annotation of subject trajectories and masks, and virtual camera framing.
(4) A user evaluation ($N = 12$) comparing our system to an image-to-video baseline, demonstrating higher controllability, lower cognitive effort, higher perceived spatial consistency across both tasks, and fewer iterations in the replication task.

2 RELATED WORKS

2.1 Generative Video and Grounding

Recent text-to-video models [7, 21, 66, 70] have lowered the barrier to video production and influenced emerging filmmaking practices [1, 27, 91]. Prior work has sought to improve controllability through structured conditions such as motion trajectories, camera parameters, and scene layouts. *Direct-a-Video* [88] and *MotionCtrl* [83] provide more explicit control over motion and camera behavior. Model- and pipeline-level approaches such as *Consisti2V* [64], *SynCamMaster* [2], and *VideoCrafter2* [13] improve visual consistency, including temporal stability and identity preservation. Approaches such as *AnimateAnyone* [32] and *VACE* [38] further support fine-grained control through pose sequences and segmentation masks. Commercial tools such as *Higgsfield Cinema Studio* [31] extend this direction by enabling camera rig and lens-level control in 3D environments. However, these approaches mainly focus on control within generated or virtual environments rather than on authoring videos grounded in real-world locations.

Meanwhile, recent work has explored using real-world spatial data for grounded video generation. *Streetscapes* [22] uses map layouts to generate long-range street view videos, *StreetCrafter* [86] leverages LiDAR-based geometric signals for camera control, and *DreamDrive* [58] generates 4D driving scenes from street view imagery. *Seoul World Model* [68] grounds video world simulation in a city through retrieval-augmented conditioning on nearby street view imagery. These works show that real-world spatial data can serve as useful foundations for grounded video generation with controllable motion and camera behavior.

Our work examines how user interaction with real-world spatial data can guide grounded video authoring, particularly for subject movement and camera planning. We connect maps and street view imagery with video diffusion models such as *VACE* [38] to support iterative authoring through subject placement, trajectory specification, and virtual camera control.

2.2 Video Authoring Interfaces

A large body of HCI research supports video editing and authoring through analysis, search, and, more recently, LLM-based interaction [33, 45, 59, 71, 76]. Transcript- and signal-driven editors, such as *B-Script* [33], *AVScript* [35], *GAZED* [59], *RubySlippers* [12], and *Computational Video Editing* [44], support trimming, B-roll insertion, and navigation of existing footage. *QuickCut* [72] further uses time-aligned transcripts and annotations to match narration with relevant footage segments. Recent systems incorporate LLMs into editing workflows. *ExpressEdit* [71], *LAVE* [76], *EditDuet* [67], *From Shots to Stories* [52], and *VideoDiff* [34] support planning, editing, and refinement. These systems improve work on existing footage, but do not focus on authoring generated video content.

Another line of work supports video authoring from existing media, text, and other structured sources [3, 42, 53, 73, 75]. *Videogenic* [53], *Lotus* [3], *ReelFramer* [80], and *PodReels* [81] transform source media into highlight, summary, news, and podcast teaser videos. *VidSTR* [87], *VideoMap* [54], and *VIVA* [65] support retargeting, exploration, and annotation, while *Surch* [41], *SceneSkim* [61], and *SwapVid* [60] support structural search and navigation. Other work also explores authoring from text and other structured sources: *Write-A-Video* [79], *Crosscast* [85], and *Crosspower* [84] use text, transcripts, or linguistic structure as high-level inputs, while other systems generate videos from scripts [46], documents [17], web pages [19], or other media [18]. These approaches support video authoring, but mainly rely on retrieval, transformation, predefined mappings, or composition from source material.

Recent work also explores human–AI co-creation and generative tools for video authoring. *VideOrigami* [10] and *Doki* [57] support structured authoring workflows through representations such as timelines, canvases, narrative structures, and text-based authoring views. Other systems also explore generative media creation in adjacent domains, including music video production (e.g., *Generative Disco* [56], *MVPrompt* [48]) and animation (e.g., *Katika* [36], *LogoMotion* [55], *MapStory* [26]). In parallel, *DreamCinema* [14] and *CineVerse* [62] explore cinematic generation and multi-shot coherence as generative authoring systems.

Our work also treats generated video itself as an authoring target, rather than focusing on editing, repurposing, or assembling existing media. We contribute an interactive workflow for iteratively shaping grounded scenes before video generation.

2.3 Previsualization and Map-Based Storytelling

To support filmmaking, HCI research has explored previsualization (previs) tools for composition and storytelling. For instance, *CollageVis* [39] prototypes rapid preview through video collages; *CinemAssist* [29, 30] provides real-time composition guidance; *CineAI* [25] generates director-style cutscenes; and *CinePreGen* [15] adapts image generation for previs on virtual 3D stages. Other pipelines, such as *MovieFactory* [94], *Script2Screen* [82], and *ASAP* [40] advance script-to-previs workflows with controllable cameras and virtual actors. These systems support cinematic planning and preview, but primarily rely on previews, virtual stages, or pre-authored assets rather than real-world visual data.

Our system also draws inspiration from location-grounded and map-centric storytelling tools. Mobile and AR systems such as *Story-Driven* [4], *Dynamic Theater* [43], and *Location-Aware Adaptation* [49] synchronize narratives with locations, while other tools choreograph camera movement within geographic visualizations [49, 51, 69]. For animated cartography and spatial explanation, *MapStory* [26] demonstrates LLM-agent-driven map animations. Related work also spans embodied data narratives [37] and spatial navigation of 360°media [47]. Engine-based geospatial workflows such as Cesium [11] also make 3D data accessible in creative tools.

Map2Video supports location-based video authoring with maps and street view imagery. Rather than using these media only for planning or navigation, our system uses them directly as authoring materials in AI filmmaking workflows.

3 FORMATIVE STUDY

3.1 SVI-Driven AI Video Generation

Building on trends identified in the literature [91], we conducted a preliminary survey with 34 filmmakers to understand challenges in AI-assisted video creation (Appendix Section 7). The survey revealed recurring difficulties in controlling generated videos, including challenges related to composition and camera movement, prompt adherence, continuity across shots, and spatial context.

These findings motivated us to explore street view imagery-driven AI video generation as a possible direction for more controllable, spatially grounded video authoring. In this approach, street-level panorama images (e.g., Google Street View, Mapillary) and map annotations serve as *spatial references* for video diffusion models (Figure 1). This direction offers three potential benefits. First, it may enable more controllable subject movement and camera planning through a familiar street view imagery interface, reducing prompt tinkering and aligning with directors' mental models of location scouting, blocking, and rehearsal. While similar to previsualization, it can directly generate video outputs. Second, it may support shot-to-shot spatial continuity by preserving scene geometry (e.g., streets, facades, and vanishing lines) across sequential shots. Third, it aligns with current filmmaking practices. In our survey, several participants described relying on real locations, reference imagery, or virtual scouting materials during pre-production, suggesting a lower learning curve.

This approach also has several limitations. One is *indoor coverage*. Because street view imagery typically does not cover building interiors, our current approach is mainly limited to outdoor scenes. For location-dependent indoor shots, creators could instead manually capture and reconstruct 3D environments using techniques such as 4D Gaussian splatting or NeRF, then use those captures as spatial references in a similar way. Another limitation is *smooth lateral motion between panorama viewpoints*. Since current street-level panorama imagery provides only discrete viewpoints, it is less suited for generating smooth sideways parallax. Although our system does not address this, prior work such as *Streetscapes* [22] suggests a possible direction by generating long-range, consistent street-view videos with autoregressive video diffusion, which could enable smoother transitions across viewpoints.

3.2 Experience Prototyping

To explore how a SVI-driven workflow could support more controllable and spatially grounded video authoring, we used experience prototyping [8] to examine how practitioners imagine, adapt to, or resist such a workflow. We built a lightweight proof-of-concept pipeline for SVI-driven AI video generation and placed it directly in filmmakers' hands [8, 9]. By simulating the end-to-end experience of location scouting, selecting panoramas, and generating AI videos, we observed the kinds of interactions they desired.

3.2.1 *Proof-of-concept Pipeline.*

(1) **Map and Street View Imagery Exploration:** a map-based street view imagery search tool [3] (OpenStreetMap + Mapillary) allows users to browse 360° panoramas, set yaw and pitch, and save viewpoints as reference backgrounds.

[3] https://map2streetview.vercel.app/

(2) **Video Generation with Reference Images:** participants upload the saved backgrounds to a popular AI video tool (Runway [4]), add a character reference image, and write prompts for actions and camera motions such as pan, tilt, and dolly.

We intentionally kept the prototype minimal and manually translated between tools, similar to how practitioners currently work.

3.2.2 Participants. We recruited 5 of 34 survey participants (1 female, 4 male) who had mentioned challenges related to *"spatial," "environments,"* or *"geography"* in their responses, to gather deeper insights into their experiences and strategies for managing spatial context in AI-assisted filmmaking. All participants had prior experience in both traditional filmmaking ($M = 10.8$, $SD = 11.5$ years) and AI-assisted filmmaking ($M = 10.2$, $SD = 2.5$ months).

Figure 2: Proof-of-concept pipeline of SVI-driven AI video generation: ⓐ SVI search, ⓑ AI video generation (Runway)

3.2.3 Protocol. The study included two parts: 1) a semi-structured interview about challenges in spatially grounded AI video creation, and 2) an exploratory experience prototyping session using the proof-of-concept pipeline. During prototyping, participants shared their screens and thought aloud while creating 3–5 videos within a single location. They searched for candidate locations, selected 360° panoramas, saved viewpoints as reference backgrounds, and used them in Runway to generate videos (Figure 2).

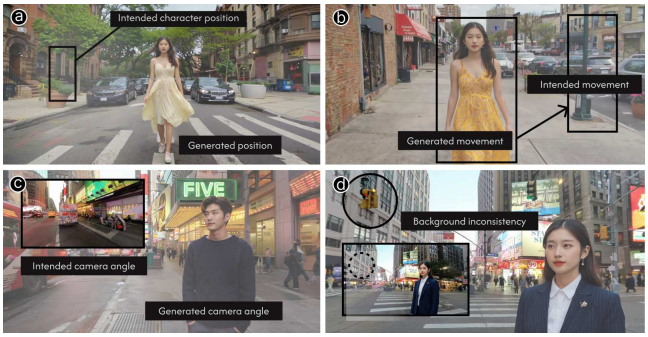

Figure 3: Challenges in spatially grounded video generation.

3.2.4 Challenges and Desired Interactions. Overall, participants responded positively to the SVI-driven workflow, noting that virtual exploration of candidate locations helped them reason more concretely about scene layout and camera setup (see Appendix Figure 17 for an example participant-created sequence). Despite these benefits, participants described recurring challenges in controlling generated videos within this workflow (Figure 3).

[4] https://app.runwayml.com/

Placing Characters in Space. Specifying where a character should appear via text prompts proved unreliable across models (Figure 3.ⓐ). Participants tried screen-direction terms (e.g., "screen left," "foreground"; P5) or alignment with reference footage (P3), yet outcomes varied, forcing trial and error and compromising framing (P1, P4, P5). They preferred to place characters directly on a map or within the frame, visually anchoring position and depth so that intent was unambiguous and more consistent across models (P5).

Directing Character Motion. Controlling character motion was also limited (Figure 3.ⓑ). Participants aligned the last frame of one clip with the first of the next (P1) or trimmed away erratic movement and physical violations (P3, P5). They envisioned path-based control, such as sketching trajectories and key poses on the map or in the frame, so the system could generate transitions that preserved intended spatial relationships across shots (P1, P2, P5).

Exploring Camera Movement. Camera motion often collapsed to generic angles despite reference images and detailed prompts (P1, P5) (Figure 3.ⓒ). Participants generated many takes and curated usable clips. Some maintained prompt libraries (P3), while others preferred tools that parsed cinematographic terms more effectively (P5). They wanted free navigation within 360° panoramas and the ability to set multiple camera positions on a map to preview pans, tilts, and dollies before committing to generation (P2, P3, P5).

Preserving Background Continuity. While the SVI-driven workflow helped maintain background continuity, participants still observed inconsistencies across multi-shot sequences, such as hallucinated or shifting props (P5), uneven environmental effects (P5), and disappearing or altered details (P2, P3) (Figure 3.ⓓ). Workarounds included choosing simple backgrounds, using shallow depth of field, and hiding cuts with close-ups (P2, P4, P5). They envisioned selective editing or automatic continuity auditing that could flag inter-shot background differences for correction (P2, P5).

Key Takeaway. Across these themes, participants converged on the need for more directly controllable interfaces for grounded video authoring, particularly for subject placement, motion path sketching, and camera preview in map and panorama views.

4 MAP2VIDEO SYSTEM

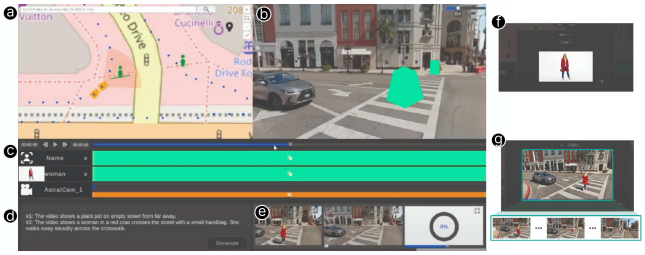

Figure 4: Map2Video interface: ⓐ Map Panel, ⓑ SVI Panel, ⓒ Timeline Panel, ⓓ Prompt Panel, ⓔ Render Queue Panel, ⓕ Reference Image Window, and ⓖ Video Player Window.

Based on the formative study, we developed Map2Video, a map-based authoring tool for SVI-grounded video generation. Map2Video

uses street view imagery to support direct control over subject placement, subject motion, and camera behavior before rendering.

4.1 System Walkthrough

As shown in Figure 5, Map2Video combines map-based planning and video inpainting through six steps.

Step 1. Location Scouting: Users first choose a location in the Map Panel (Figure 4ⓐ) and select a street view image as the scene background. The SVI Panel (Figure 4ⓑ) previews the selected view and allows users to adjust framing, such as view angle and field of view, while keeping the shot grounded in the chosen location.

Step 2. Mask Positioning: Users place a green mask, which serves as an subject proxy, in the Map Panel (Figure 4ⓐ). The system projects this proxy into the SVI Panel (Figure 4ⓑ) as a 2D card anchored to the scene. This lets users determine where a subject should appear in the grounded environment before generation.

Step 3. Movement Sketching: Users sketch the movement path of the mask in the Map Panel (Figure 4ⓐ), and the corresponding motion is synchronized in the SVI Panel (Figure 4ⓑ). This allows them to specify how the subject moves through the scene while previewing the motion in context.

Step 4. Camera Walkthrough: Users define placement and motion of the camera in the SVI Panel (Figure 4ⓑ), including translation, rotation, and zoom. The Timeline Panel (Figure 4ⓒ) synchronizes subject motion and camera actions over time, allowing users to preview and refine shot structure in relation to the scene.

Step 5. Scene Prompt: Once users are satisfied with the spatial setup, motion, and camera plan, they describe the target subject or event in the Prompt Panel (Figure 4ⓓ) and submit the shot to the Render Queue (Figure 4ⓔ). Users can optionally provide a reference image in the Reference Image Window (Figure 4ⓕ) to guide subject appearance across iterations or shots. Otherwise, the system generates the subject from the prompt.

Step 6. Video Inpainting: Finally, the system generates the video by inpainting only the masked region based on the specified scene, motion, camera, and prompt. This supports iterative refinement of the subject while preserving the surrounding scene. Outputs can be reviewed in the Video Player (Figure 4ⓖ).

Users can adjust mask scale to create a range of grounded scene elements, including realistic characters, environmental events such as fire, surreal elements such as flying jellyfish, and multiple entities such as a crowd (Figure 6). The system also supports multiple masks, which are processed sequentially and increase generation time (Figure 7).

4.2 Implementation

As shown in Figure 8, Map2Video is implemented with Unity as the frontend and ComfyUI, an open-source node-based interface for diffusion models, as the backend. The ComfyUI workflow runs on a remote server equipped with an NVIDIA H100 NVL GPU.

Based on user interaction in Unity, ComfyUI takes four inputs to generate the video: (1) a background video, consisting of a street view imagery sequence with camera movement; (2) a mask video, defining a geographically fixed mask position and movement aligned with the camera work; (3) text prompts describing the intended scene; and (4) an optional reference image, which users can upload or generate to guide the masked content. ComfyUI then upscale the background video and inpaints only the masked region. Each sequence is limited to five seconds by the Wan2.1 model [74], and longer videos can be created by stitching multiple clips together [50].

Map Interaction and Mask Placement. We use OpenStreetMap for the map and Mapillary for street view imagery. To place a subject in the panoramic view, we project the subject's map coordinates into the camera view. We approximate the local area with a tangent plane (East–North–Up) centered at the camera, compute the subject's relative azimuth and elevation, and map these angles to pixel coordinates with a pinhole camera model. This projection keeps masks spatially anchored to the subject and the user-selected camera. Details are provided in Appendix Table 4 for reproducibility.

4.2.1 Video Inpainting Workflow. Our video inpainting workflow builds on the default VACE pipeline in the ComfyUI documentation [5] and adds nodes for street-view-based video authoring. VACE is a unified video foundation model for text-to-video, image-to-video, and masked video-to-video tasks. It builts on the Wan2.1 and integrates reference and mask inputs through a video condition unit and context adapter, enabling flexible video generation [38].

Pre-processing. To address the low resolution of 360° Mapillary imagery, we use SeedVR [78] to upscale low-resolution 360° Mapillary imagery. The enhanced background is then combined with an inverted mask and fed into the inpainting module.

Encoder. We use `umt5-xxl fp8 e4m3fn scaled (wan type)` to encode text prompts into conditioning vectors for diffusion. We then use the `Wan 2.1 VAE` to encode frames into latent space and decode them back into RGB frames, reducing computation.

Video Inpainting. Inpainting is handled by the `Wan VACE To Video` node. The upscaled background with an inverted alpha mask constrains edits to masked regions while preserving the grounded background and camera motion. We use `Wan2.1 VACE 14B fp16 (weights dtype: fp8 e4m3fn fast)` for diffusion model.

We also use two LoRA modules: CausVid [89] to accelerate convergence and LightX2V [20] to improve temporal stability. In our environment, the workflow takes about 50 seconds to load the model, around 3 minutes to sample a 5-second 1280×720 video at 16 fps, and another 40 seconds to render and download the video from the server.

Although optional, a reference image helps maintain subject appearance. Users can upload or generate one in the reference image window (Figure 4.ⓕ). If a user chooses to generate a character instead of uploading a subject image, we run a lightweight workflow using the Unet Loader with a GGUF-formatted model (`Wan2.1 14B VACE Q4 K M.gguf`). We also apply the `Wan14B Realism Boost T2V LoRA` so the generated character better matches real-world street view imagery.

[5] https://docs.comfy.org/tutorials/video/wan/vace/

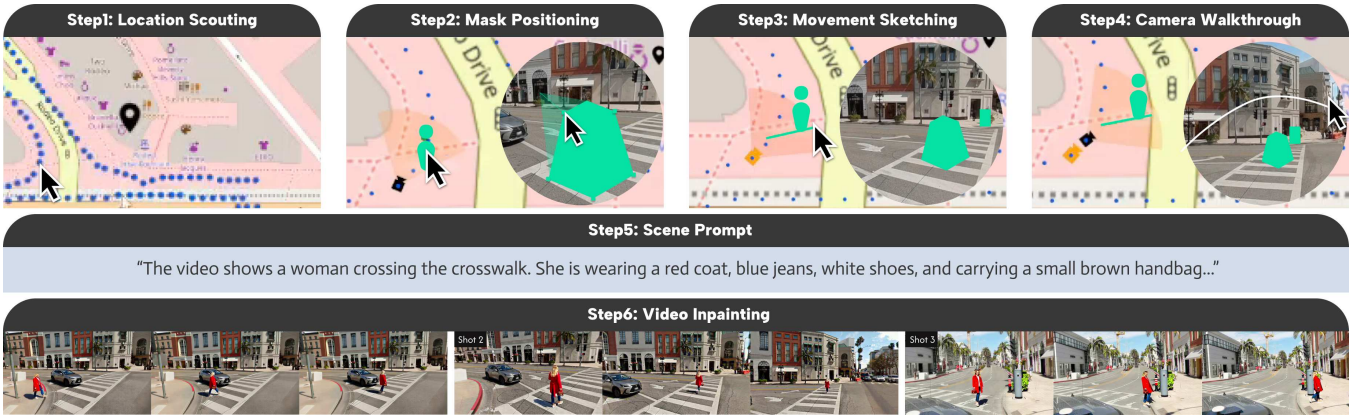

Figure 5: Map2Video system walkthrough.

Sampling and Decoding Strategy. For inference, we use the `Model Sampling SD3` [23] scheduler followed by `KSampler` in ComfyUI. The `SD3` shift parameter (set to 5) balances coarse structure and detail. `KSampler` runs six steps (guidance scale 1, `uni_pc`, `simple` scheduler, strength 1), prioritizing fast generation while maintaining background fidelity.

5 USER EVALUATION

We evaluated Map2Video with 12 filmmakers to examine whether it provides greater control than an image-to-video baseline over subject placement, subject motion, and camera behavior in SVI-grounded video generation.

5.1 Method

We used a within-subjects design in which participants completed two video authoring tasks using both Map2Video and an image-to-video baseline.

5.1.1 Participants. We recruited 12 participants (ages 20–46; $M = 28.00$, $SD = 7.47$; 5 female, 7 male), all with experience in traditional filmmaking. Their filmmaking experience ranged from 1–12 years ($M = 5.38$, $SD = 3.18$). All had prior experience with AI video tools.

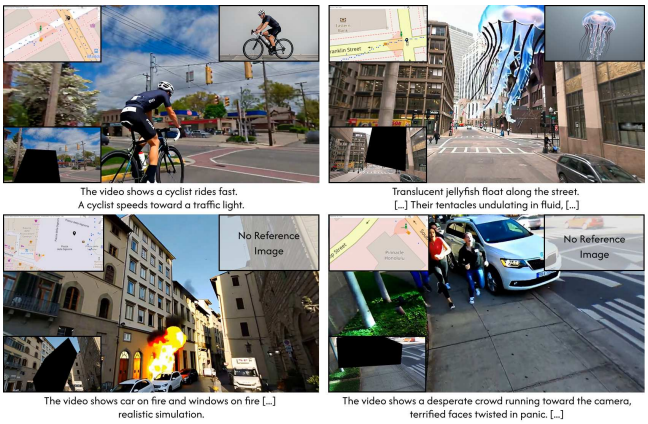

Figure 6: Various inpainted subjects created with or without reference images, such as a cyclist, jellyfish, fire, and a crowd.

Five reported using them actively in their own work, while the others had used them more casually for experimentation.

5.1.2 Conditions. We compared Map2Video with an image-to-video baseline to evaluate our interaction designs. To control for model differences, we implemented the baseline as a standard image-to-video workflow using the same underlying models.

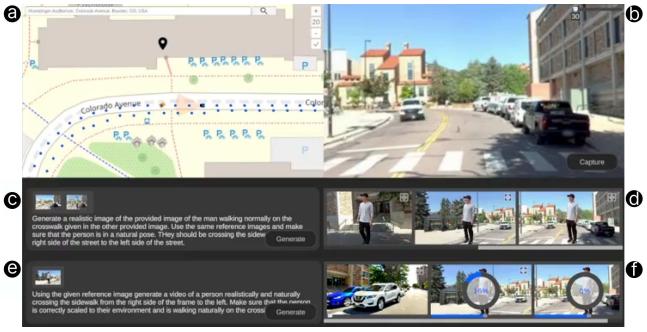

Figure 9: Baseline interface: ⓐ Map Panel, ⓑ Street View Imagery Panel, ⓒ Image Prompt Panel, ⓓ Video Prompt Panel, ⓔ Image Render Queue Panel, and ⓕ Video Render Queue Panel

(1) **Baseline** (Figure 9): An image-to-video (I2V) tool with map-based street view retrieval. Participants selected a background and subject image to create a composite reference image for video generation. We added automatic retrieval to avoid manual download and upload.
(2) **Map2Video** (Figure 4): A map-to-video (M2V) tool with editable masks and camera control. Participants placed masks on the map, animated them by drawing trajectories, and specified camera motion with timeline keyframes.

This comparison isolates the effects of mask placement, motion authoring, and camera control. Both conditions used the same Unity frontend, ComfyUI backend, Flux1 Kontext [24] for reference image synthesis, and Wan2.1 VACE [38] for video generation.

5.1.3 Task. Participants completed two video authoring tasks in both conditions:

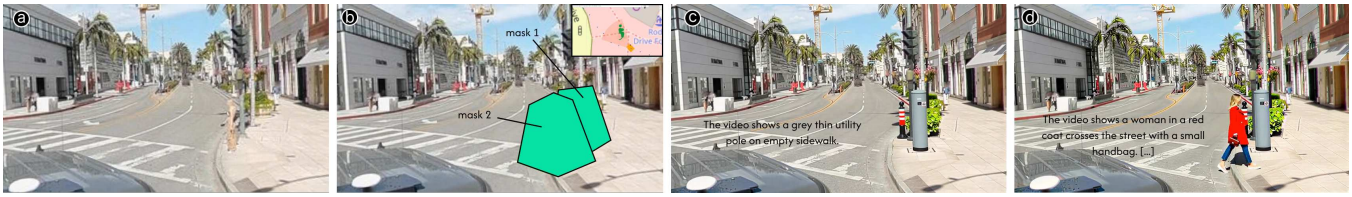

Figure 7: Multi-mask inpainting process. ⓐ original image, ⓑ masked image, ⓒ pedestrian removal and background upscaling, ⓓ character composition.

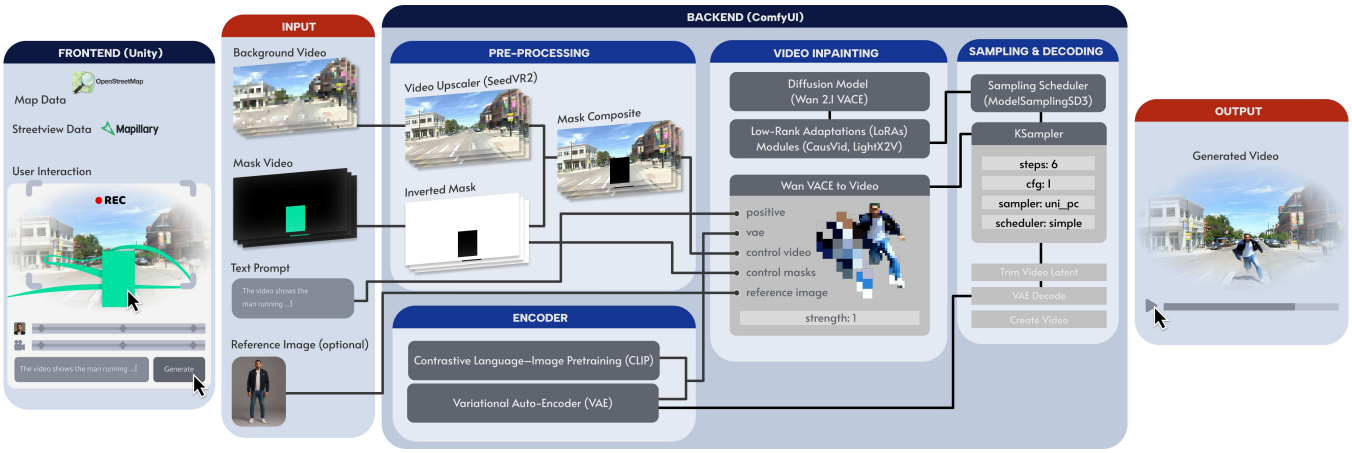

Figure 8: Map2Video system architecture.

(1) **Replication Task**: We filmed three target street shots: (1) a car arriving with a static camera, (2) a character crossing the street with a right-to-left pan, and (3) a character walking away as the camera zoomed out and tilted upward. The shots were inspired by scenes from *Breakfast at Tiffany's*, *Vanilla Sky*, and *Baby Driver*, but recreated due to limited Mapillary coverage at the original locations.

(2) **Open-Ended Task**: Participants created three clips in one location selected from ten candidates with high-quality Mapillary imagery. They could choose any theme, but were asked to maintain spatial continuity across shots. To limit study time, we allowed only one mask.

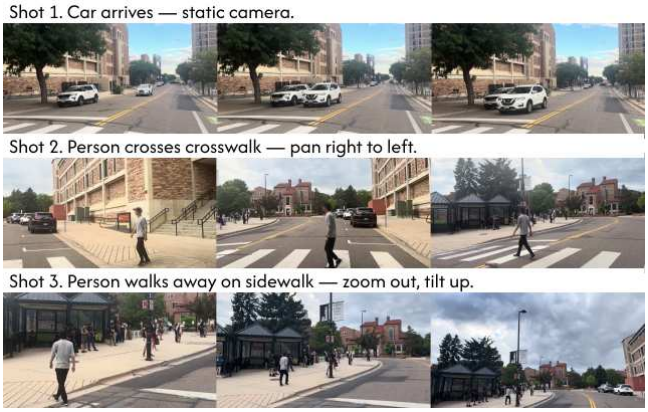

Figure 10: Replication task material

5.1.4 Procedure. After an introduction and demographics, participants received a brief walkthrough of each system and completed both tasks in both conditions, with system order counterbalanced. After each condition, they completed questionnaires and a short semi-structured interview. Sessions lasted 90–120 minutes, and participants were compensated 25 USD/hour (75–100 USD total).

5.1.5 Measurement. For each condition, we recorded task completion time and number of iterations, and collected workload (NASA-TLX) [28], usability (SUS) [6], creativity support (CSI) [16], and custom questionnaires. Participants self-rated task performance in the replication and open-ended tasks, and evaluated outputs on 7-point Likert scales for (1) spatial consistency, (2) intent alignment, (3) creative controllability, (4) ease of iteration, (5) visual quality, and (6) overall experience (Appendix Table 5). They also rated Map2Video-specific features on 7-point Likert scales: (1) map-based blocking, (2) trajectory design, (3) map-based camera setup, (4) framing preview, and (5) workflow adoption (Appendix Table 6).

5.2 Results

Overall, Map2Video gave participants greater control over subject placement, motion, and camera behavior in SVI-grounded video generation than the baseline. Appendix Figure 18 shows that participants also produced diverse videos with Map2Video in the open-ended task.

Factor	Map2Video	Baseline	Count
Exploration	16.08 (2.61)	9.25 (6.08)	2.41
Enjoyment	16.17 (2.41)	8.33 (5.35)	1.75
Expressiveness	15.08 (3.26)	10.33 (6.24)	2.00
Results Worth Effort	14.92 (2.61)	8.67 (6.31)	2.25
Immersion	14.17 (3.93)	7.42 (4.78)	1.58
Weighted Sum (CSI)	**76.65**	**44.42**	–

Table 1: Creative Support Index (CSI) Questionnaire Result. CSI factors (N=12) by interface with averages (SD), factor counts, and weighted sum at the bottom.

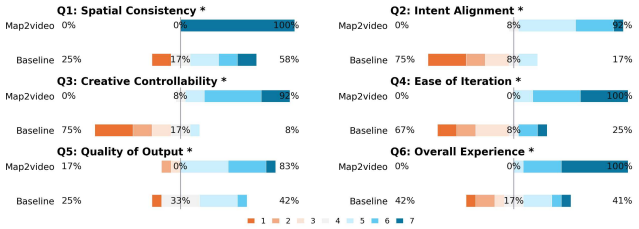

Figure 11: Custom Questionnaire Result

5.2.1 *Controllability and Authoring Outcomes.*

Map2Video made grounded subject and camera control more reliable. Creative controllability was significantly higher for Map2Video ($M = 5.92$, $SD = 0.90$) than for the baseline ($M = 2.50$, $SD = 1.38$), $t(11) = 6.46$, $p < .001^{***}$ (Figure 11). This pattern was also reflected in participants' self-evaluations. In the replication task, ratings were significantly higher for Map2Video ($M = 6.00$, $SD = 1.13$) than for the baseline ($M = 2.91$, $SD = 1.78$; Welch's $t(18.6) = -5.08$, $p < .001^{***}$). In the open-ended task, ratings were again higher for Map2Video ($M = 4.64$, $SD = 1.23$ vs. $M = 3.46$, $SD = 2.06$), although the difference was not significant ($t(17.9) = -1.71$, $p = .105$).

The baseline was easy to use, but hard to direct. In the replication task, many participants failed to reproduce intended compositions within the time limit and proceeded with imperfect references. They often had to regenerate images multiple times, yet still struggled to match subject scale, position, and motion (Figure 12). Even with detailed prompts specifying framing and camerawork, such as zoom or tilt, outputs frequently defaulted to centered subjects with minimal motion. This was consistent with our formative study (Section 3.2), in which filmmakers similarly described difficulty controlling composition and motion through prompt-based workflows.

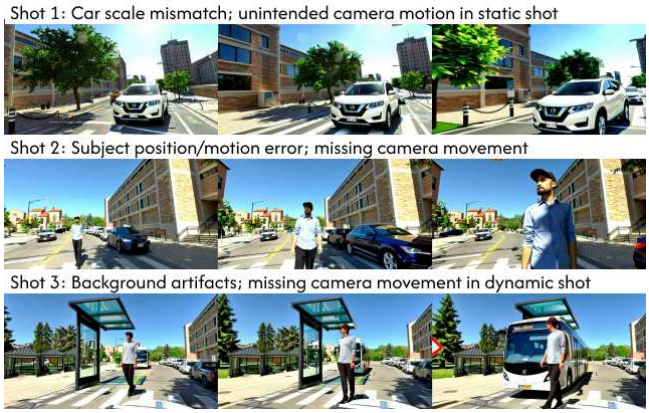

Figure 12: Baseline failure cases in the replication task

Fine-grained motion and mask control remained difficult. In the open-ended task, Map2Video improved overall composition control, but limitations remained in finer-grained motion and mask specification. Compounded masked motions, such as "a squid rolls and explodes" or "a man does a somersault," as well as vertical motion such as jumping, were difficult to generate reliably. Two participants (P10, P12) also produced incomplete or oversized subjects because of mask scaling issues, such as a bicycle without a person or a giant human (Figure 13). Control limitations also persisted in the baseline during Task 2, where grounded composition and motion remained difficult to direct precisely. These results suggest that Map2Video shifted the main challenge from coarse scene specification to finer-grained motion and mask control.

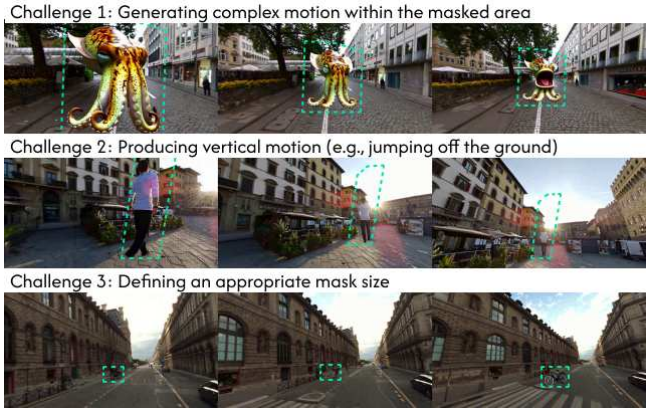

Figure 13: Map2Video failure cases in the open-ended task. Green-highlighted parts are the user-defined mask.

5.2.2 *Efficiency, Workload, and Usability.*

Map2Video reduced trial and error. In Task 1, Map2Video was significantly faster than the baseline (20.42±7.48 min vs. 24.65±8.14 min; Table 2). In the baseline, participants repeatedly generated images to obtain an acceptable reference, whereas Map2Video shifted effort toward layout planning and camera framing. Because video iteration counts were similar in Task 1 (baseline: 3.66±1.06; Map2Video: 3.67±1.50), the efficiency gain appears to come mainly

Task	Item	Baseline	Map2Video
Task 1	Duration (min)	24.65 (±8.14)	20.42 (±7.48)*
	Iteration (image)	8.93 (±3.82)	NA
	Iteration (video)	3.66 (±1.06)	3.67 (±1.50)
Task 2	Duration (min)	23.33 (±8.20)	23.60 (±6.79)
	Iteration (image)	5.62 (±2.27)	NA
	Iteration (video)	3.68 (±1.33)	5.75 (±4.13)*

Table 2: Task duration and iterations (paired t-test, p<.05).

from reducing image-level trial and error rather than video refinement. In Task 2, completion time was similar across conditions, but video iterations were higher for Map2Video (5.75±4.13) than for the baseline (3.68±1.33)*. This pattern suggests that, even after scene layout and camera setup were established, fine-grained motion within the masked region still required additional iteration.

Controllability mattered more than interface simplicity. On NASA-TLX, the baseline imposed significantly higher mental demand ($t(11) = -4.01$, $p = .002^{**}$), effort ($t(11) = -4.24$, $p = .001^{**}$), and frustration ($t(11) = -4.89$, $p < .001^{***}$) (Figure 14). They also reported better performance with Map2Video ($t(11) = -5.27$, $p < .001^{***}$). Map2Video received a SUS score of 83.12 ($SD = 10.43$, *Excellent*), while the baseline received 51.25 ($SD = 22.09$, *Poor*) (Figure 15). Despite the baseline's simpler interface, participants rated it lower because its outputs were inconsistent and unpredictable ($W = 0$, $p = .008^{**}$).

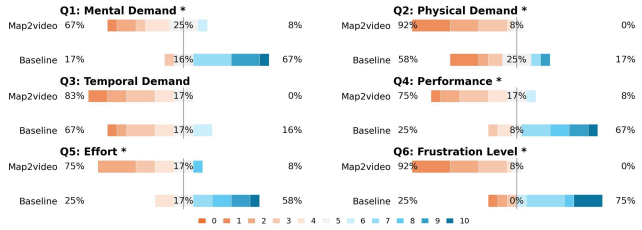

Figure 14: Task Load Questionnaire Result (NASA-TLX). The performance scale is opposite.

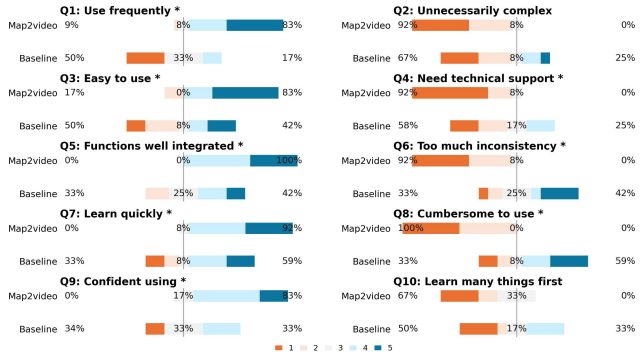

Figure 15: System Usability Questionnaire Result (SUS)

5.2.3 *Creative Support and Output Perception.*

Map2Video made creative exploration feel more worthwhile. Following prior work [16], we computed a 5-factor weighted CSI by omitting the collaboration factor for our single-user workflow. As shown in Table 1, Map2Video received a higher weighted CSI score than the baseline (76.65 vs. 44.42) and was rated higher across all factors. In particular, higher scores in Exploration and Results-Worth-Effort indicate that participants found Map2Video more useful for trying different ideas and producing rewarding outcomes.

Greater control improved satisfaction and spatial consistency. Overall satisfaction was higher for Map2Video than for the baseline ($M = 6.50$, $SD = 0.67$ vs. $M = 3.92$, $SD = 1.78$), $t(11) = 4.24$, $p < .001^{***}$ (Figure 11). Participants also rated Map2Video as better at maintaining spatial consistency across sequential clips. As shown in Figure 12, baseline outputs often regenerated entire backgrounds when users described camera changes or referred to SVI objects such as bus stops, buildings, cars, or trees, whereas Map2Video modified only masked regions. Accordingly, all participants gave Map2Video the maximum spatial consistency rating, while the baseline received a lower mean rating ($M = 4.42$, $SD = 2.11$), $t(11) = 4.24$, $p < .001^{***}$.

Low-quality street view images and sparse viewpoints limited output quality and camera setup. Although Map2Video received significantly higher output quality ratings than the baseline (Figure 11), its mean rating remained modest ($M = 5.08$, $SD = 1.38$). Even after upscaling, Mapillary backgrounds still contained degraded pixels, affecting subsequent video quality. The same source limitations appeared in feature-level ratings; trajectory drawing received the highest score ($M = 6.67$, $SD = 0.49$), while map-based camera setup received the lowest ($M = 6.42$, $SD = 0.51$) (Figure 16). This lower rating likely reflects Mapillary's limited 360° viewpoints, which restricted precise street view selection.

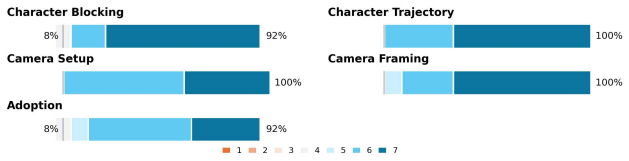

Figure 16: Map2Video Feature Evaluation.

5.3 Qualitative Insights

Street View Imagery as a Grounded Starting Point. Participants described street view imagery as a grounded starting point for ideation, rather than merely a reference source. Unlike text prompting, SVI provided spatial context that helped them make creative decisions about plot, action, and shot design. For example, P2 noted that maps helped "channel creativity into the plot and action," while P11 said that beginning with an existing environment was "better than a blank canvas." Others used it to think through continuity and pre-production details such as building shadows (P8). At the same time, several participants noted that fixed real-world backgrounds could constrain more surreal or imaginative directions, suggesting a tradeoff between grounding and abstraction.

Balancing Control and Randomness. Participants valued both deliberate control and occasional randomness, suggesting that they supported different creative modes. Map2Video made trajectories and camera motion more intentional and predictable, reducing guesswork during authoring. At the same time, several participants saw value in the baseline's unpredictability. P9 noted that it "made me try more variations," suggesting that randomness could support improvisational exploration. These reflections suggest an opportunity to combine grounded control with room for surprise and divergence.

Agency Through Preview and Feedback. Participants often described Map2Video as closer to "directing" than "generating." They valued being able to place masks directly within the camera view, preview trajectories and camera paths, and decide whether to proceed before full generation. This preview-feedback loop contributed to a stronger sense of agency. P1 noted that manually setting trajectories meant "less time spent waiting," while P10 emphasized that previewing masks reduced unnecessary iterations. Similarly, P4 described "giving up faster" with the baseline because too much effort was spent on uncertain generations.

Expanding Camera and Motion Control. Participants pointed to several directions for extending grounded control beyond the current prototype. Some requested more flexible camera control beyond fixed 360° image points, for example by sketching camera trajectories directly on the map or SVI to achieve bird's-eye views or dolly-like shots (P4, P11). Others wanted more expressive subject motion, including vertical trajectories and shape changes for actions such as flying, exploding, or transforming. They also proposed tighter integration with familiar workflows, such as Premiere-style shortcuts (P1, P3) and Blender export for timeline editing (P5), along with depth recognition (P5, P12) and object interaction (P10) through 3D reconstruction of SVI environments and characters.

Overall, these insights suggest that Map2Video supported grounded video generation through greater control, more concrete planning, and iterative feedback. They also highlight opportunities for future tools to support more expressive motion, flexible camera control, and occasional departures from real-world constraints.

6 LIMITATIONS AND FUTURE WORK

Addressing Street View Source Limitations. Some limitations stemmed not from the interface, but from the underlying street view source. Fixed eye-level viewpoints constrained camera control, while low-quality Mapillary imagery reduced output quality even after upscaling. Although street view imagery served as a useful grounded scaffold, future systems could better separate spatial grounding from source appearance, support smoother viewpoint changes beyond sparse, discrete views, and allow changes in appearance such as atmosphere or time of day while preserving scene geometry. Possible directions include panoramic environment reconstruction [77], autoregressive video diffusion over street-view sequences [22], and layout-conditioned generation approaches that preserve scene geometry while allowing controllable changes in appearance. More broadly, while our system focuses on street view imagery and maps, the underlying idea of controllable grounded authoring could extend to other structured scene priors, such as 3D models [63] and point clouds [93].

Supporting Fine-Grained Motion Control. Fine-grained motion remained difficult to specify, especially for compounded actions, vertical movement, and mask scaling. Planning motion on the map also did not always translate intuitively to street view-based editing. Future work could therefore focus on improving motion specification and preview across both map and SVI views. One direction is to support multi-scale authoring, where users move fluidly between high-level geographic planning and fine-grained street-level adjustment. Another is to provide motion preview through real-time text-to-motion models such as DART [92]. For example, text annotations attached to character trajectories could allow users to specify actions or movement styles that update interactively during preview.

Ethical Considerations. In developing Map2Video, we recognize ethical implications in using generative AI together with street view imagery for filmmaking. Because the system relies on real-world visual data, it raises concerns around privacy, data provenance, and geographic coverage bias, which may affect what locations can be represented faithfully or controllably. Our goal is not to automate storytelling, but to support creators in authoring spatially grounded scenes with greater control. To mitigate these risks, users remain responsible for selecting locations, specifying prompts, and defining motion and camera behavior, keeping AI in an assistive rather than autonomous role.

7 CONCLUSION

We introduced Map2Video, a street view imagery-driven system for more controllable SVI-grounded video authoring. By combining map-based annotation with video inpainting, Map2Video helped users author location-specific scenes with more direct control over subject placement, motion, and camera behavior. Our studies showed that street view imagery served as a grounded starting point, while the rapid preview-feedback loop supported iterative refinement. These findings suggest opportunities for future systems to expand camera and motion control, improve coordination between map and street view representations, and reduce limitations of current street view imagery sources.

REFERENCES

[1] Torin Anderson and Shuo Niu. 2025. Making AI-Enhanced Videos: Analyzing Generative AI Use Cases in YouTube Content Creation. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–7.

[2] Jianhong Bai, Menghan Xia, Xintao Wang, Ziyang Yuan, Xiao Fu, Zuozhu Liu, Haoji Hu, Pengfei Wan, and Di Zhang. 2024. SynCamMaster: Synchronizing Multi-Camera Video Generation from Diverse Viewpoints. *arXiv preprint arXiv:2412.07760* (2024).

[3] Aadit Barua, Karim Benharrak, Meng Chen, Mina Huh, and Amy Pavel. 2025. Lotus: Creating short videos from long videos with abstractive and extractive summarization. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*. 967–981.

[4] Jan Henry Belz, Lina Madlin Weilke, Anton Winter, Philipp Hallgarten, Enrico Rukzio, and Tobias Grosse-Puppendahl. 2024. Story-driven: exploring the impact of providing real-time context information on automated storytelling. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. 1–15.

[5] Filip Biljecki and Koichi Ito. 2021. Street view imagery in urban analytics and GIS: A review. *Landscape and Urban Planning* 215 (2021), 104217.

[6] John Brooke et al. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.

[7] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. 2024. Video generation models as world simulators. *OpenAI Blog* 1, 8 (2024), 1.

[8] Marion Buchenau and Jane Fulton Suri. 2000. Experience prototyping. In *Proceedings of the 3rd conference on Designing interactive systems: processes, practices, methods, and techniques*. 424–433.

[9] Bradley Camburn, Vimal Viswanathan, Julie Linsey, David Anderson, Daniel Jensen, Richard Crawford, Kevin Otto, and Kristin Wood. 2017. Design prototyping methods: state of the art in strategies, techniques, and guidelines. *Design Science* 3 (2017), e13.

[10] Yining Cao, Yiyi Huang, Anh Truong, Hijung Valentina Shin, and Haijun Xia. 2025. Compositional structures as substrates for human-ai co-creation environment: A design approach and a case study. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–25.

[11] Cesium GS, Inc. 2023. Cesium. Photorealistic 3D Tiles from Google Maps Platform in Cesium for Unity. Retrieved Feb 24, 2026 from https://cesium.com/learn/unity/unity-photorealistic-3d-tiles/

[12] Minsuk Chang, Mina Huh, and Juho Kim. 2021. Rubyslippers: Supporting content-based voice navigation for how-to videos. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–14.

[13] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. 2024. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7310–7320.

[14] Weiliang Chen, Fangfu Liu, Diankun Wu, Haowen Sun, Jiwen Lu, and Yueqi Duan. 2024. Dreamcinema: Cinematic transfer with free camera and 3d character. *arXiv preprint arXiv:2408.12601* (2024).

[15] Yiran Chen, Anyi Rao, Xuekun Jiang, Shishi Xiao, Ruiqing Ma, Zeyu Wang, Hui Xiong, and Bo Dai. 2024. Cinepregen: Camera controllable video previsualization via engine-powered diffusion. *arXiv preprint arXiv:2408.17424* (2024).

[16] Erin Cherry and Celine Latulipe. 2014. Quantifying the creativity support of digital tools through the creativity support index. *ACM Transactions on Computer-Human Interaction (TOCHI)* 21, 4 (2014), 1–25.

[17] Peggy Chi, Tao Dong, Christian Frueh, Brian Colonna, Vivek Kwatra, and Irfan Essa. 2022. Synthesis-assisted video prototyping from a document. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–10.

[18] Peggy Chi, Nathan Frey, Katrina Panovich, and Irfan Essa. 2021. Automatic instructional video creation from a markdown-formatted tutorial. In *The 34th Annual ACM Symposium on User Interface Software and Technology*. 677–690.

[19] Peggy Chi, Zheng Sun, Katrina Panovich, and Irfan Essa. 2020. Automatic video creation from a web page. In *Proceedings of the 33rd annual ACM symposium on user interface software and technology*. 279–292.

[20] LightX2V Contributors. 2025. LightX2V: Light Video Generation Inference Framework. https://github.com/ModelTC/lightx2v.

[21] Google DeepMind. [n.d.]. Veo. https://deepmind.google/models/veo/

[22] Boyang Deng, Richard Tucker, Zhengqi Li, Leonidas Guibas, Noah Snavely, and Gordon Wetzstein. 2024. Streetscapes: Large-scale consistent street view generation using autoregressive video diffusion. In *ACM SIGGRAPH 2024 Conference Papers*. 1–11.

[23] Patrick Esser, Robin Rombach, Andreas Blattmann, Nikolay Savinov, et al. 2024. Stable Diffusion 3: Advancing Text-to-Image Generation with Multi-Modal Diffusion Transformers. *arXiv preprint arXiv:2403.03206* (2024). https://arxiv.org/abs/2403.03206

[24] Black Forest Labs et al. 2025. FLUX.1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space. arXiv:2506.15742 [cs.GR] https://arxiv.org/abs/2506.15742

[25] Inan Evin, Perttu Hämäläinen, and Christian Guckelsberger. 2022. Cine-ai: Generating video game cutscenes in the style of human directors. *Proceedings of the ACM on Human-Computer Interaction* 6, CHI PLAY (2022), 1–23.

[26] Aditya Gunturu, Ben Pearman, Keiichi Ihara, Morteza Faraji, Bryan Wang, Rubaiat Habib Kazi, and Ryo Suzuki. 2025. MapStory: Prototyping Editable Map Animations with LLM Agents. *arXiv preprint arXiv:2505.21966* (2025).

[27] Brett A Halperin, Diana Flores Ruíz, and Daniela K Rosner. 2025. Underground AI? Critical approaches to generative cinema through amateur filmmaking. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–18.

[28] Sandra G Hart. 1986. NASA Task Load Index (TLX): Paper and Pencil Package-Volume 1.0. (1986).

[29] Rui He, Ying Cao, Johan F Hoorn, and Huaxin Wei. 2024. Cinemassist: An Intelligent Interactive System for Real-Time Cinematic Composition Design. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–7.

[30] Rui He, Huaxin Wei, and Ying Cao. 2024. An Interactive System for Supporting Creative Exploration of Cinematic Composition Designs. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. 1–15.

[31] Higgsfield. 2026. Cinema Studio 2.0. The first AI Cinema Studio with real optical physics for cinematic video generation. Retrieved Feb 24, 2025 from https://higgsfield.ai/blog/cinema-studio-guide

[32] Li Hu. 2024. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8153–8163.

[33] Bernd Huber, Hijung Valentina Shin, Bryan Russell, Oliver Wang, and Gautham J Mysore. 2019. B-script: Transcript-based b-roll video editing with recommendations. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–11.

[34] Mina Huh, Ding Li, Kim Pimmel, Hijung Valentina Shin, Amy Pavel, and Mira Dontcheva. 2025. VideoDiff: Human-AI Video Co-Creation with Alternatives. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–19.

[35] Mina Huh, Saelyne Yang, Yi-Hao Peng, Xiang'Anthony' Chen, Young-Ho Kim, and Amy Pavel. 2023. Avscript: Accessible video editing with audio-visual scripts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17.

[36] Amir Jahanlou and Parmit K Chilana. 2022. Katika: An end-to-end system for authoring amateur explainer motion graphics videos. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–14.

[37] Radhika Pankaj Jain, Adam Drogemuller, Kadek Ananta Satriadi, Ross Smith, and Andrew Cunningham. 2025. Strollytelling: Coupling Animation with Physical Locomotion to Explore Immersive Data Stories. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–17.

[38] Zeyinzi Jiang, Zhen Han, Chaojie Mao, Jingfeng Zhang, Yulin Pan, and Yu Liu. 2025. Vace: All-in-one video creation and editing. *arXiv preprint arXiv:2503.07598* (2025).

[39] Hye-Young Jo, Ryo Suzuki, and Yoonji Kim. 2024. CollageVis: Rapid Previsualization Tool for Indie Filmmaking using Video Collages. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–16.

[40] Hanseob Kim, Ghazanfar Ali, and Jae-In Hwang. 2021. Asap: Auto-generating storyboard and previz with virtual humans. In *2021 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. IEEE, 316–320.

[41] Jeongyeon Kim, Daeun Choi, Nicole Lee, Matt Beane, and Juho Kim. 2023. Surch: Enabling structural search and comparison for surgical videos. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17.

[42] Minsun Kim, Dawon Lee, and Junyong Noh. 2025. Generating highlight videos of a user-specified length using most replayed data. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–13.

[43] You-Jin Kim, Joshua Lu, and Tobias Höllerer. 2023. Dynamic theater: Location-based immersive dance theater, investigating user guidance and experience. In *Proceedings of the 29th ACM Symposium on Virtual Reality Software and Technology*. 1–11.

[44] Mackenzie Leake, Abe Davis, Anh Truong, and Maneesh Agrawala. 2017. Computational video editing for dialogue-driven scenes. *ACM Trans. Graph.* 36, 4 (2017), 130–1.

[45] Mackenzie Leake and Wilmot Li. 2024. ChunkyEdit: Text-first video interview editing via chunking. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–16.

[46] Mackenzie Leake, Hijung Valentina Shin, Joy O Kim, and Maneesh Agrawala. 2020. Generating audio-visual slideshows from text articles using word concreteness. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–11.

[47] Chaeeun Lee, Jinwook Kim, Hyeonbeom Yi, and Woohun Lee. 2024. Viewer2Explorer: Designing a Map Interface for Spatial Navigation in Linear 360 Museum Exhibition Video. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–15.

[48] ChungHa Lee, DaeHo Lee, and Jin-Hyuk Hong. 2025. MVPrompt: Building Music-Visual Prompts for AI Artists to Craft Music Video Mise-en-scène. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–21.

[49] Wanwan Li, Changyang Li, Minyoung Kim, Haikun Huang, and Lap-Fai Yu. 2023. Location-aware adaptation of augmented reality narratives. In *Proceedings of the 2023 CHI conference on human factors in computing systems*. 1–15.

[50] Wuyang Li, Wentao Pan, Po-Chien Luan, Yang Gao, and Alexandre Alahi. 2025. Stable video infinity: Infinite-length video generation with error recycling. *arXiv preprint arXiv:2510.09212* (2025).

[51] Wenchao Li, Zhan Wang, Yun Wang, Di Weng, Liwenhan Xie, Siming Chen, Haidong Zhang, and Huamin Qu. 2023. GeoCamera: Telling stories in geographic visualizations with camera movements. In *Proceedings of the 2023 CHI conference on human factors in computing systems*. 1–15.

[52] Yuzhi Li, Haojun Xu, and Feng Tian. 2025. From Shots to Stories: LLM-Assisted Video Editing with Unified Language Representations. *arXiv preprint arXiv:2505.12237* (2025).

[53] David Chuan-En Lin, Fabian Caba Heilbron, Joon-Young Lee, Oliver Wang, and Nikolas Martelaro. 2024. Videogenic: Identifying Highlight Moments in Videos with Professional Photographs as a Prior. In *Proceedings of the 16th Conference*

on Creativity & Cognition. 328–346.

[54] David Chuan-En Lin, Fabian Caba Heilbron, Joon-Young Lee, Oliver Wang, and Nikolas Martelaro. 2022. VideoMap: Supporting Video Editing Exploration, Brainstorming, and Prototyping in the Latent Space. *arXiv preprint arXiv:2211.12492* (2022).

[55] Vivian Liu, Rubaiat Habib Kazi, Li-Yi Wei, Matthew Fisher, Timothy Langlois, Seth Walker, and Lydia Chilton. 2025. Logomotion: Visually-grounded code synthesis for creating and editing animation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–16.

[56] Vivian Liu, Tao Long, Nathan Raw, and Lydia Chilton. 2023. Generative disco: Text-to-video generation for music visualization. *arXiv preprint arXiv:2304.08551* (2023).

[57] Xingyu Bruce Liu, Mira Dontcheva, and Dingzeyu Li. 2026. A Text-Native Interface for Generative Video Authoring. *arXiv preprint arXiv:2603.09072* (2026).

[58] Jiageng Mao, Boyi Li, Boris Ivanovic, Yuxiao Chen, Yan Wang, Yurong You, Chaowei Xiao, Danfei Xu, Marco Pavone, and Yue Wang. 2025. Dreamdrive: Generative 4d scene modeling from street view images. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 367–374.

[59] KL Bhanu Moorthy, Moneish Kumar, Ramanathan Subramanian, and Vineet Gandhi. 2020. Gazed–gaze-guided cinematic editing of wide-angle monocular video recordings. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–11.

[60] Taichi Murakami, Kazuyuki Fujita, Kotaro Hara, Kazuki Takashima, and Yoshifumi Kitamura. 2024. SwapVid: Integrating Video Viewing and Document Exploration with Direct Manipulation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–13.

[61] Amy Pavel, Dan B Goldman, Björn Hartmann, and Maneesh Agrawala. 2015. Sceneskim: Searching and browsing movies using synchronized captions, scripts and plot summaries. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*. 181–190.

[62] Quynh Phung, Long Mai, Fabian David Caba Heilbron, Feng Liu, Jia-Bin Huang, and Cusuh Ham. 2025. CineVerse: Consistent Keyframe Synthesis for Cinematic Scene Composition. *arXiv preprint arXiv:2504.19894* (2025).

[63] PolySphere Studio. 2025. City Core – Paris. 3D models. Retrieved Feb 24, 2026 from https://polysphere-studio.gitbook.io/city-core-paris/

[64] Weiming Ren, Huan Yang, Ge Zhang, Cong Wei, Xinrun Du, Wenhao Huang, and Wenhu Chen. 2024. Consisti2v: Enhancing visual consistency for image-to-video generation. *arXiv preprint arXiv:2402.04324* (2024).

[65] Anita Ruangrotsakun, Dayeon Oh, Thuy-Vy Nguyen, Kristina Lee, Mark Ser, Arthur Hiew, Rogers Ngo, Zeyad Shureih, Roli Khanna, and Minsuk Kahng. 2023. VIVA: Visual exploration and analysis of videos with interactive annotation. In *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces*. 162–165.

[66] Inc. Runway. [n.d.]. Runway. https://runwayml.com/

[67] Marcelo Sandoval-Castañeda, Bryan Russell, Josef Sivic, Gregory Shakhnarovich, and Fabian Caba Heilbron. 2025. EditDuet: A Multi-Agent System for Video Non-Linear Editing. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*. 1–11.

[68] Junyoung Seo, Hyunwook Choi, Minkyung Kwon, Jinhyeok Choi, Siyoon Jin, Gayoung Lee, Junho Kim, JoungBin Lee, Geonmo Gu, Dongyoon Han, et al. 2026. Grounding World Simulation Models in a Real-World Metropolis. *arXiv preprint arXiv:2603.15583* (2026).

[69] Jae-Eun Shin and Woontack Woo. 2023. How space is told: linking trajectory, narrative, and intent in augmented reality storytelling for cultural heritage sites. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–14.

[70] Rui Sun, Yumin Zhang, Tejal Shah, Jiahao Sun, Shuoying Zhang, Wenqi Li, Haoran Duan, Bo Wei, and Rajiv Ranjan. 2024. From sora what we can see: A survey of text-to-video generation. *arXiv preprint arXiv:2405.10674* (2024).

[71] Bekzat Tilekbay, Saelyne Yang, Michal Adam Lewkowicz, Alex Suryapranata, and Juho Kim. 2024. Expressedit: Video editing with natural language and sketching. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*. 515–536.

[72] Anh Truong, Floraine Berthouzoz, Wilmot Li, and Maneesh Agrawala. 2016. Quickcut: An interactive tool for editing narrated video. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. 497–507.

[73] Tess Van Daele, Akhil Iyer, Yuning Zhang, Jalyn C Derry, Mina Huh, and Amy Pavel. 2024. Making short-form videos accessible with hierarchical video summaries. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–17.

[74] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. 2025. Wan: Open and Advanced Large-Scale Video Generative Models. *arXiv preprint arXiv:2503.20314* (2025).

[75] Bryan Wang, Zeyu Jin, and Gautham Mysore. 2022. Record Once, Post Everywhere: Automatic Shortening of Audio Stories for Social Media. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–11.

[76] Bryan Wang, Yuliang Li, Zhaoyang Lv, Haijun Xia, Yan Xu, and Raj Sodhi. 2024. Lave: Llm-powered agent assistance and language augmentation for video editing. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*. 699–714.

[77] Guangcong Wang, Peng Wang, Zhaoxi Chen, Wenping Wang, Chen Change Loy, and Ziwei Liu. 2024. Perf: Panoramic neural radiance field from a single panorama. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46, 10 (2024), 6905–6918.

[78] Jianyi Wang, Shanchuan Lin, Zhijie Lin, Yuxi Ren, Meng Wei, Zongsheng Yue, Shangchen Zhou, Hao Chen, Yang Zhao, Ceyuan Yang, Xuefeng Xiao, Chen Change Loy, and Lu Jiang. 2025. SeedVR2: One-Step Video Restoration via Diffusion Adversarial Post-Training. (2025).

[79] Miao Wang, Guo-Wei Yang, Shi-Min Hu, Shing-Tung Yau, Ariel Shamir, et al. 2019. Write-a-video: computational video montage from themed text. *ACM Trans. Graph.* 38, 6 (2019), 177–1.

[80] Sitong Wang, Samia Menon, Tao Long, Keren Henderson, Dingzeyu Li, Kevin Crowston, Mark Hansen, Jeffrey V Nickerson, and Lydia B Chilton. 2024. ReelFramer: Human-AI co-creation for news-to-video translation. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–20.

[81] Sitong Wang, Zheng Ning, Anh Truong, Mira Dontcheva, Dingzeyu Li, and Lydia B Chilton. 2024. PodReels: Human-AI Co-Creation of Video Podcast Teasers. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference*. 958–974.

[82] Zhecheng Wang, Jiaju Ma, Eitan Grinspun, Bryan Wang, and Tovi Grossman. 2025. Script2Screen: Supporting Dialogue Scriptwriting with Interactive Audio-visual Generation. *arXiv preprint arXiv:2504.14776* (2025).

[83] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. 2024. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*. 1–11.

[84] Haijun Xia. 2020. Crosspower: Bridging graphics and linguistics. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. 722–734.

[85] Haijun Xia, Jennifer Jacobs, and Maneesh Agrawala. 2020. Crosscast: adding visuals to audio travel podcasts. In *Proceedings of the 33rd annual ACM symposium on user interface software and technology*. 735–746.

[86] Yunzhi Yan, Zhen Xu, Haotong Lin, Haian Jin, Haoyu Guo, Yida Wang, Kun Zhan, Xianpeng Lang, Hujun Bao, Xiaowei Zhou, et al. 2025. Streetcrafter: Street view synthesis with controllable video diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 822–832.

[87] Joshua Kong Yang, Mackenzie Leake, Jeff Huang, and Stephen DiVerdi. 2025. VidSTR: Automatic Spatiotemporal Retargeting of Speech-Driven Video Compositions. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–17.

[88] Shiyuan Yang, Liang Hou, Haibin Huang, Chongyang Ma, Pengfei Wan, Di Zhang, Xiaodong Chen, and Jing Liao. 2024. Direct-a-video: Customized video generation with user-directed camera movement and object motion. In *ACM SIGGRAPH 2024 Conference Papers*. 1–12.

[89] Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. 2025. From Slow Bidirectional to Fast Autoregressive Video Diffusion Models. In *CVPR*.

[90] Zhengqing Yuan, Yixin Liu, Yihan Cao, Weixiang Sun, Haolong Jia, Ruoxi Chen, Zhaoxu Li, Bin Lin, Li Yuan, Lifang He, et al. 2024. Mora: Enabling generalist video generation via a multi-agent framework. *arXiv preprint arXiv:2403.13248* (2024).

[91] Ruihan Zhang, Borou Yu, Jiajian Min, Yetong Xin, Zheng Wei, Juncheng Nemo Shi, Mingzhen Huang, Xianghao Kong, Nix Liu Xin, Shanshan Jiang, et al. 2025. Generative AI for Film Creation: A Survey of Recent Advances. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 6267–6279.

[92] Kaifeng Zhao, Gen Li, and Siyu Tang. 2024. DartControl: A diffusion-based autoregressive motion model for real-time text-driven motion control. *arXiv preprint arXiv:2410.05260* (2024).

[93] Sixiao Zheng, Minghao Yin, Wenbo Hu, Xiaoyu Li, Ying Shan, and Yanwei Fu. 2026. VerseCrafter: Dynamic Realistic Video World Model with 4D Geometric Control. *arXiv preprint arXiv:2601.05138* (2026).

[94] Junchen Zhu, Huan Yang, Huiguo He, Wenjing Wang, Zixi Tuo, Wen-Huang Cheng, Lianli Gao, Jingkuan Song, and Jianlong Fu. 2023. Moviefactory: Automatic movie creation from text using large generative models for language and images. In *Proceedings of the 31st ACM International Conference on Multimedia*. 9313–9319.

APPENDIX

Survey on AI Filmmaking

As generative AI video tools enter filmmaking practice, their real-world use and control challenges remain underexplored. To inform our system design, we conducted a formative survey with AI-assisted filmmaking practitioners.

Method. The survey examined usage patterns, workflows, and challenges. It was distributed through online AI filmmaking communities (Creative Refuge, Facebook groups, and Discord). The survey included questions about the AI video tools participants used, their use cases, and the challenges they encountered. Participation was voluntary and uncompensated, though respondents were informed about the possibility of a compensated follow-up formative study.

In total, 34 filmmakers (7 female, 27 male) completed the survey. All had experience with AI-assisted filmmaking, and 30 of 32 respondents also had traditional filmmaking experience. Participants held roles such as director, screenwriter, producer, editor, and cinematographer. Many worked across the full AI production pipeline, including prompt design, creative direction, and editing.

Most participants (29/34) used multiple AI tools, with self-assessed proficiency skewing advanced (18 advanced, 5 intermediate, 2 beginner, 9 unspecified). Table 3 lists the reported tools.

Tool Name	Mention	Tool Name	Mention
Runway	29	Luma / Lumalabs	2
Kling	18	Seedance	2
Veo / Veo 3	12	Minimax	2
ComfyUI	11	Reve	1
Pika	6	Pixverse	1
Sora	5	Kaiber	1
Flux / Flux Kontext	3	Stable Video Diffusion	1
Vidu	1	Moonvally	1
Dreamina	1	RunningHub	1

Table 3: AI video generation tools reported by participants.

Current Practices. Filmmakers reported using generative AI video tools across all stages of the filmmaking process: pre-production, production, and post-production, with varying degrees of integration into traditional workflows. In pre-production, many participants used tools like *Midjourney* and *Runway* to create mood boards, storyboards, and previsualizations, allowing for rapid iteration on visual concepts before committing to full production. During production, several participants used tools such as *Veo*, *Kling*, and *ComfyUI* to generate video sequences, especially for stylized or speculative scenes that would be difficult or costly to produce conventionally. Some used AI-generated clips as B-roll or animation to complement live-action footage. In post-production, AI tools were commonly applied to editing, compositing, background replacement, stylization, and visual effects. While a few participants (6) used AI tools to support traditional workflows, the majority reported *"producing fully AI-generated films"*, handling scripting, scene design, and video generation with AI tools, and refining the output using conventional editing software such as Premiere Pro or DaVinci Resolve. In these cases, AI filmmaking either replaced conventional production processes entirely or filled specific gaps such as location footage, animation, or VFX. Participants reported *"creative exploration"*, *"faster turnaround"*, and the ability to *"visualize ideas that would be too expensive or impossible to shoot"* as key motivations for adopting generative AI tools. Participants implied the use of diverse strategies to prepare reference images for video generation, including sourcing real-world images from stock platforms (e.g., Getty), generating visuals through image-creation tools, and constructing virtual environments with 3D/virtual set software such as Unreal Engine. Among the portfolio examples shared by participants, the majority focused on relatively static single-subject shots, often with shallow depth of field or blurred backgrounds. We also observed examples centered on vehicle movement through urban environments.

Challenges. One of the most common challenges reported by participants was ***"maintaining visual and spatial consistency"*** (mentioned by 17 respondents). These issues included inconsistencies in character appearance, lighting, backgrounds, and overall scene geography, especially across longer sequences. Nine participants noted difficulties with ***"controlling composition and camera movement"***, such as aligning shots with intended framing or achieving smooth transitions while the character is moving. ***"Prompt adherence"*** was another frequent concern, with 12 participants describing frustration when generated outputs did not match their intended visual or narrative details. This included problems with ***"character placement"*** (11) and specifying ***"character actions"*** via text (3). ***"Quality degradation over time,"*** such as hallucinated elements and inconsistent lip-sync and gesture timing, was mentioned by 6 participants. ***"Multi-character coordination"*** was also noted as a difficulty by 6 participants. Together, these responses point to broader difficulties in controlling generated videos to match filmmakers' intentions. In addition to generation issues, participants pointed to broader obstacles: 8 mentioned the ***"high cost"*** of using certain AI tools, 5 reported the ***"steep learning curve"*** in tools like *ComfyUI*, and 4 noted the ***"difficulty of keeping up with rapidly evolving technologies"***. Although some participants observed improvements in character consistency and tool capabilities, most agreed that achieving coherent, high-quality outputs still requires extensive iteration, manual correction, and post-production work. These responses, together with the submitted portfolio examples, suggested a promising opportunity for more directly controlling subject placement, motion, and camera behavior in location-grounded scenes.

Figure 17: An example of three sequential user-created shots and corresponding input street view image and user prompt.

Step	Description	Formula
1. Local ENU projection	Convert subject geodetic coordinates (φ_a, λ_a) into local East–North–Up (ENU) offsets relative to the camera (φ_c, λ_c).	$\Delta x \approx R\cos\varphi_c(\lambda_a - \lambda_c), \quad \Delta z \approx R(\varphi_a - \varphi_c)$
2. Range and bearing	Compute ground distance d and bearing angle ψ_{ca} between subject and camera.	$d = \sqrt{\Delta x^2 + \Delta z^2}, \quad \psi_{ca} = \text{atan2}(\Delta x, \Delta z)$
3. Camera-relative angles	Subtract camera heading ψ_c and pitch θ_c to obtain relative azimuth and elevation.	$\Delta\psi = \psi_{ca} - \psi_c, \quad \Delta\theta = -\text{atan2}(h_c, d) - \theta_c$
4. Pinhole projection	Map relative angles into normalized screen coordinates using a pinhole camera model.	$s_x = \frac{\tan(\Delta\psi)}{\tan(\alpha_h/2)}, \quad s_y = -\frac{\tan(\Delta\theta)}{\tan(\alpha_v/2)}$
5. Screen scaling	Convert normalized coordinates to pixel coordinates (u, v) given screen resolution (W, H).	$u = \frac{s_x+1}{2}W, \quad v = \frac{s_y+1}{2}H$

Table 4: Subject placement from geodetic coordinates to panoramic screen coordinates.

Dimension	Question Item
Task 1 Self-Evaluation	I think I did well in the replication task, and the quality of the output meets my expectations.
Task 2 Self-Evaluation	I think I did well in the open-ended task, and the quality of the output meets my expectations.
Spatial Consistency	The three clips appeared consistent as if filmed in the same location.
Intent Alignment	The output reflected my intended direction well.
Creative Controllability	It was easy to control the system to produce the changes I wanted.
Ease of Iteration	I could quickly try, modify, and refine my ideas.
Quality of Output	The generated results met my expectations for visual quality.
Overall Experience	Overall, my experience with the system was positive.

Table 5: Custom questionnaire items used for self-evaluation and output assessment. All items were rated on a 7-point Likert scale (1 = Strongly Disagree, 7 = Strongly Agree).

Feature	Question Item
Character Blocking	Placing characters on the map helped me maintain spatial continuity.
Character Trajectory	Designing character trajectories felt intuitive and useful.
Camera Setup	Setting up cameras on the map helped me visualize spatial relationships between shots.
Camera Framing	Previewing framing helped me plan the shot effectively.
Adoption	I would incorporate these features into my creative workflow.

Table 6: Feature evaluation questionnaire items. All items were rated on a 7-point Likert scale (1 = Strongly Disagree, 7 = Strongly Agree).

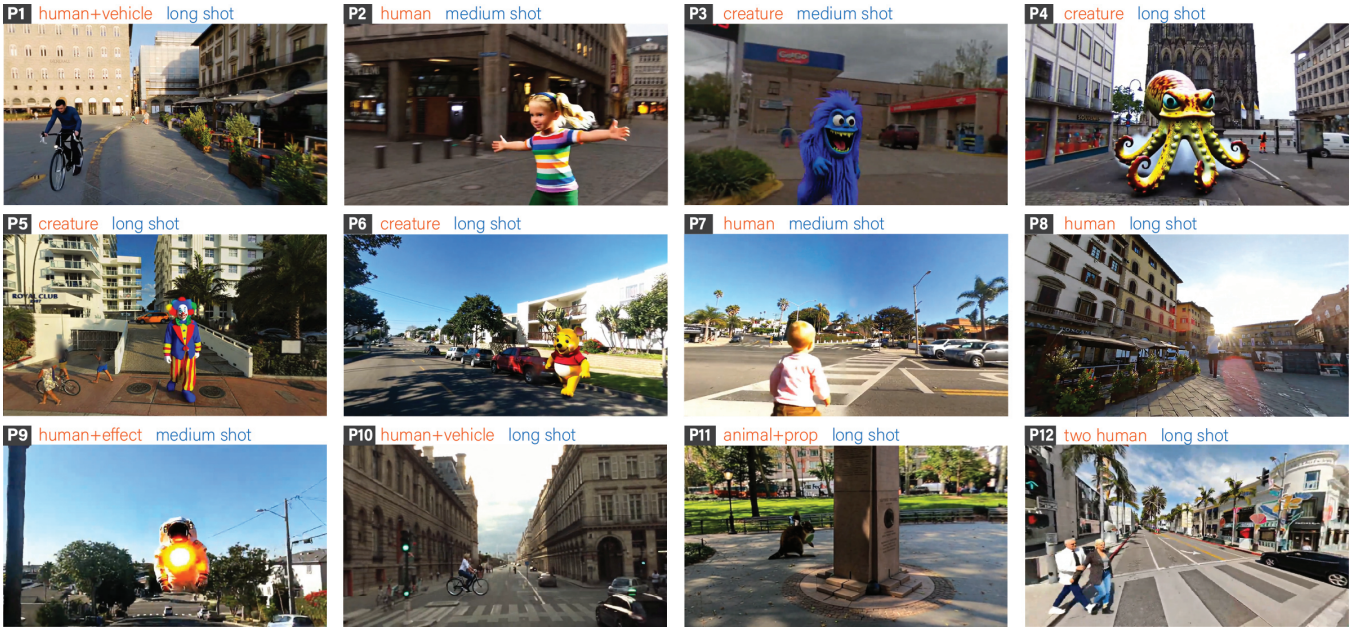

Figure 18: Examples of video artifacts created by users with the Map2Video system, illustrating diverse variations in subjects, shot styles, geographic places, and subject positioning within the frame.