Many-Videos Browsing: Interactive Visualization of Generated Video Collections

JOTARO SAKAMIYA, The University of Tokyo, Japan
GEORGE XU, The University of Tokyo, Japan
ANRAN QI, The University of Tokyo, Japan
I-CHAO SHEN, The University of Tokyo, Japan
NADAV Z. COHEN, The University of Tokyo, Japan and Reichman University, Israel
YUKI KOYAMA, The University of Tokyo, Japan
TAKEO IGARASHI, The University of Tokyo, Japan

Fig. 1. We present *Many-Videos Browsing* (MVB), an interface for interactively exploring (a) diverse object motions that emerge across a large collection of videos stochastically generated by image-to-video models. (b) MVB represents these motions as colored trajectories and speed, allowing users to compare and examine how objects move across videos, and to progressively narrow their focus to subsets of videos that match their interests through stroke-based (red thick stroke) and region-based (green box) filtering. (c) illustrates the best-matched video. Using MVB, users can select desired motions of objects (green circles in (d), red circles in (e)) from different videos. (f) A new motion montage video can be synthesized by recombining these motions.

Image-to-video models can generate many plausible object motion variations from the same input image and text prompt due to their stochastic nature, creating a rich but difficult-to-navigate motion space. We present *Many-Videos Browsing* (MVB), an exploration interface that enables users to understand, compare, and progressively narrow motion variations across such a generated video collection. Our approach is technically orthogonal to control-based video generation: rather than steering a single generation given a predefined motion, MVB focuses on motion *exploration* of a video collection. In practical workflows, the two paradigms are complementary, supporting different phases of creative video production. MVB exposes motion distributions through trajectory and speed visualization and enables users to interactively explore motions by filtering, adjusting the scale of focus over trajectories and speed. Additionally, MVB supports browsing of newly appearing objects beyond the first frame, allowing users to understand scene evolution without requiring video playback. The interface operates in a model-agnostic manner and does not require access to model internals. We provide a three-fold validation of our approach: (1) a quantitative analysis of effectiveness in motion understanding and search tasks, (2) a qualitative validation of our interface design through interviews with professional animation artists, and (3) multiple application scenarios that demonstrate practical use.

CCS Concepts: • **Computing methodologies → Graphics systems and interfaces**.

Additional Key Words and Phrases: Video generative model, video exploration, exploration interface

ACM Reference Format:
Jotaro Sakamiya, George Xu, Anran Qi, I-Chao Shen, Nadav Z. Cohen, Yuki Koyama, and Takeo Igarashi. 2021. Many-Videos Browsing: Interactive Visualization of Generated Video Collections. *ACM Trans. Graph.* 38, 4, Article 39 (July 2021), 12 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Image-to-video (I2V) generation is a recent and rapidly emerging technology that is reshaping how videos are created. Unlike traditional video creation workflows, which typically produce a single video for a given specification, I2V models are inherently stochastic: a single input can produce many distinct videos through probabilistic sampling. In such a generated video collection, the visual appearance of the scene is largely fixed, and diversity manifests

Authors' Contact Information: Jotaro Sakamiya, The University of Tokyo, Japan; George Xu, The University of Tokyo, Japan; Anran Qi, The University of Tokyo, Japan; I-Chao Shen, The University of Tokyo, Japan; Nadav Z. Cohen, nadavc220@gmail.com, The University of Tokyo, Japan and Reichman University, Israel; Yuki Koyama, koyama@pe.t.u-tokyo.ac.jp, The University of Tokyo, Japan; Takeo Igarashi, takeo@acm.org, The University of Tokyo, Japan.

2021. ACM 1557-7368/2021/7-ART39
https://doi.org/10.1145/nnnnnnn.nnnnnnn

through differences in object motion over time. Motion—rather than appearance—thus becomes the dominant axis of variation across generated videos. As a result, users are confronted with large, unstructured motion spaces that are difficult to understand and navigate. Therefore, in this work, we aim to support the **interactive exploration** of stochastic I2V outputs generated from the *same* input, enabling users to understand, compare, and progressively narrow *motion variations* across large collections of generated videos.

Traditional video browsing techniques mostly rely on visual appearance or metadata in the video for summarization and navigation [Barthel et al. 2015, 2016; Matejka et al. 2014]. While effective for general video collections, such visualizations become ineffective in our setting, where videos differ primarily in object motion. As a result, existing browsing interfaces are ill-suited to understanding, comparing, or organizing motion variation across videos. We argue that effective exploration of I2V outputs requires an interface that enables users to quickly survey the range of plausible motion outcomes, compare alternatives, and progressively refine their preferences. Our approach is designed to support this exploratory workflow by keeping users in the loop as they navigate motion uncertainty. We advocate Bill Buxton's opinion: there is no such thing as a best interface—only interfaces that are appropriate for a given task [Buxton 2010].

Workflow-wise, our interface is **complementary** to control-based motion generation methods (e.g., trajectories [Geng et al. 2025; Wang et al. 2024; Zhang et al. 2025b], skeletal [Tu et al. 2025], camera parameters [He et al. 2025], and keyframes [Feng et al. 2025; Wang et al. 2025b]) rather than a replacement. Control-based methods presuppose that users already know the desired motion and how to encode it as a control signal. This is an assumption rarely held in early-stage creative exploration. In practice, creative video creation workflows often begin with under-specified or absent motion intent [Brown et al. 2008; Pandey et al. 2023; Schön 2017]. Users may not know what motion to aim for and must first understand what motion is possible before deciding how motion might unfold. This exploratory phase is fundamentally different from the goal-directed synthesis that control-based methods support. Note that, even with control signals in place, there can still be value in keeping the user in the exploration loop, since controlled generation can still produce a wide range of plausible motion variations (refer to Section 2.1).

To this end, we present Many-Videos Browsing (MVB), a model-agnostic interface for **interactive exploration** of motion variation across large collections of I2V-generated videos (Figure 1). MVB supports exploration by helping users: (a) **Understand** what motions are possible: MVB provides an overview of motion distributions by visualizing object trajectories and speed from all generated videos directly on the shared initial frame, allowing users to quickly perceive dominant patterns as well as rare but valid motions. To reduce visual clutter, MVB also supports users in previewing a subset of representative trajectories that capture distinct motion patterns. As objects may enter the scene after the first frame, MVB extends its exploration support to display all *newly appearing objects* across the video set, allowing users to anticipate what will happen in the scene without requiring video playback. (b) **Compare** motion patterns across many videos: Users can select individual trajectory to immediately preview the corresponding video and speed, supporting rapid back-and-forth comparison without leaving the spatial context of the motion. (c) Progressively **narrow** the space of desired motion candidates: First, users can perform sketch-based filtering by drawing strokes that express motion patterns of interest, allowing them to retrieve videos whose object trajectories or speed best match the sketch. Second, users can perform region-based filtering, including or excluding motions that pass through specified spatial regions to quickly narrow the search space. Third, the interface enables users to adjust the scale of focus over objects, allowing them to shift their attention between coarse motion trends and localized regions or sub-parts of object motions. These functions support an exploratory workflow that moves from comprehension to comparison and ultimately to selection, enabling users to understand what motions are possible, evaluate alternatives, and progressively converge on desired outcomes.

We evaluated MVB through a controlled user study by comparing it with a text-based interface on video exploration tasks. The results indicate that MVB enables more efficient exploration, particularly for understanding motion distributions and searching for target motions. In addition, through interviews with professional animation artists focused on inbetweening tasks in 2D hand-drawn animation, we empirically validate our interface design and identify the potential of MVB as a complementary tool in real-world animation workflows. Finally, we demonstrate its creative use through multiple application scenarios with I2V models, including video motion montage, motion scout, and interactive long video generation. Our contributions are threefold:

(1) Many-Videos Browsing (MVB), a model-agnostic interactive exploration interface for browsing diverse object motions across large collections of I2V-generated videos, with a set of interaction functions.
(2) Validating the efficiency and effectiveness of MVB through a controlled user study in comparison with a text-query interface, and empirically grounding its design goals through interviews with professional animation artists.
(3) Multiple applications showcasing the creative usage of MVB.

2 Related Work

2.1 Video Generation

In recent years, a large number of studies have investigated video generation models capable of synthesizing temporally coherent and visually plausible videos. Contemporary approaches transform noise into videos using diffusion models [Blattmann et al. 2023; Ho et al. 2022a,b] or flow matching [Jin et al. 2025; Lipman et al. 2023; Liu et al. 2022]. While these models achieve high visual quality, the stochasticity of the generation process leads to large variation in object motion across samples.

To address this motion variability, a large body of prior work has focused on motion *control*—guiding video generation toward user-specified motion patterns through explicit conditioning signals. Beyond text [Singer et al. 2022] and image prompts [Ni et al. 2023], researchers have explored motion-centric controls such as skeletal pose [Tu et al. 2025], camera parameters [He et al. 2025], keyframes [Feng et al. 2025; Wang et al. 2025b], and object motion trajectories [Geng et al. 2025; Wang et al. 2023, 2024; Yin et al.

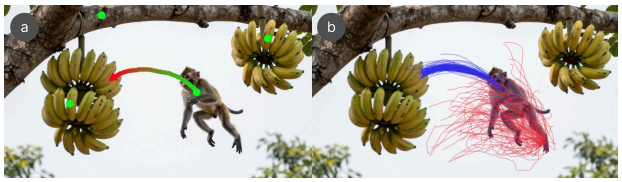


Fig. 2. **Motion variations under trajectory-controlled video generation.** Despite applying (a) trajectory control of the monkey from the start (green) to the end (red) of the stroke, (b) videos generated by ATI [Wang et al. 2025a], a trajectory conditioned video generation method, continue to show significant motion variation at the right leg (red trajectories).

2023; Zhang et al. 2025b]. When the desired motion is known in advance and can be precisely specified, these approaches provide effective and direct control over generation outcomes. In practical video-creation workflows, however, designers often begin with little or no motion intent, making control-based methods unsuitable in this situation. Moreover, control-based methods frequently introduce trade-offs between controllability, visual realism, and generality [Bahmani et al. 2025, 2024; Ma et al. 2025], further motivating the need for video *exploration* methods.

Even with control signals in place, due to their sparsity, not all pixels are constrained. Thus, the videos generated by these control-based methods still exhibit substantial motion variations, especially in regions weakly constrained by control signals. Figure 2 shows an example where, despite a trajectory control on the monkey, the motions of the right leg vary significantly across generated videos. These considerations motivate us to study this underexplored problem of enabling effective exploration of object motion across large collections of generated videos.

2.2 Design Exploration

Design exploration is an iterative process of navigating a design space through repeated cycles of divergence and convergence [Liu et al. 2003; Shneiderman 2007]. During this process, designers progressively narrow down alternatives by comparing options within the design space [Dow et al. 2010].

A major line of research supports design exploration by enabling browsing over existing data collections, including images [Zhu et al. 2014], shapes [Matejka et al. 2018; Pandey et al. 2023], or human motions [Bernard et al. 2013; Choi et al. 2012]. Another line of work focuses on exploring parameterized design spaces by exposing a small number of interpretable parameters or low-dimensional projections of high-dimensional spaces by sliders [Chiu et al. 2020; Gandikota et al. 2025], gallery-based interfaces [Koyama et al. 2020], sometimes augmented with suggestions [Koyama and Goto 2022]. Other approaches enable exploration by pre-sampling a large number of designs and supporting browsing over the resulting samples, including interactive evolutionary systems [Sims 1991], browsing pre-sampled physics simulations based on motion or spatial properties [Goel and James 2022; Twigg and James 2007], and exploring pre-generated UI design suggestions [O'Donovan et al. 2015].

Our work is inspired by Many-Worlds Browsing (MWB) [Twigg and James 2007] and adapts its pre-sampling-and-selection paradigm to video collections. In contrast to prior MWB settings, we tailor the interface to I2V outputs by any recent diffusion or flow-based video models such as [Wan et al. 2025; Wang et al. 2025a; Zhang et al. 2025a], where videos share the same initial frame and differ primarily in object motion (see next section for more details).

2.3 Video Exploration

Existing video exploration work can be broadly categorized into single video or video collection exploration.

For the exploration of single videos, several approaches enable coarse-to-fine navigation by presenting overviews composed of keyframes or video segments extracted from the original video [Barnes et al. 2010; Boreczky et al. 2000; Jackson et al. 2013]. Other work proposes efficient slider-based interfaces that allow rapid and continuous navigation along the temporal axis [Matejka et al. 2013; Ramos and Balakrishnan 2003]. In addition, prior research has explored video exploration through direct manipulation in pixel space [Dragicevic et al. 2008; Goldman et al. 2008; Tang et al. 2008], as well as exploration based on textual information associated with videos, such as transcripts or scene descriptions [Pavel et al. 2015, 2014].

For exploring video collections, Tompkin et al. [2012] enables exploration of video collections captured in the same scene by aligning videos in scene space and supporting visually smooth transitions between clips. Matejka et al. [2014] supports browsing video collections through a wide range of attributes associated with videos. In addition, Barthel et al. [2015, 2016] propose graph-based browsing systems that connect related frames with edges, enabling navigation through video collections based on visual similarity. Interactive video retrieval systems from the Video Browser Showdown (VBS) community focus on searching large video collections. Representative systems include SOMHunter [Veselỳ et al. 2021] and Vibro [Schall et al. 2023], which support efficient retrieval through multimodal search interfaces.

These approaches are effective when videos differ in appearance, content, or scene structure. Ours belongs to this category but under a different setting: I2V-generated videos share the same initial frame and nearly identical visual appearance; variation lies mostly in object motion. This setting poses several challenges for existing collection browsing techniques. First, appearance-based similarity metrics become ineffective, as visual content provides little discriminative signal. Second, thumbnail grids and keyframe- or metadate-based summaries fail to convey motion variation, forcing users to compare subtle motion differences across many videos without understanding the motion distribution beforehand. We argue that effective exploration in this setting requires interfaces that make motion itself explicit, comparable, and navigable.

3 Design Goals

While we are not aware of existing users for this task due to the very recent emergence of the I2V generation, we derived our design goals from prior exploration literature [Buxton 2010] and system [Goel and James 2022; Pandey et al. 2023; Twigg and James 2007].

- **D1: Rapid understanding of motion distribution.** Help users quickly understand what motions are possible under a given I2V model (or a set of models). The interface should

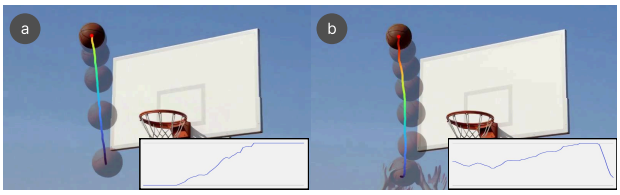


Fig. 3. **Objects with similar trajectories but different speed curves.** (a) The basketball follows a natural free-fall trajectory, shown as a color-coded path, alongside its speed curve (bottom right) reflecting consistent acceleration due to gravity. (b) A visually similar trajectory whose speed curve (bottom right) differs notably, suggesting the ball was decelerated by an external force (hand). Despite their near-identical trajectories, the two motions are clearly distinguishable through their speed curves.

allow users to perceive both dominant and rare but valid motion patterns without inspecting videos one by one.

- **D2: Rapid narrowing of the motion space.** While understanding the overall motion distribution provides valuable context, effective exploration also requires efficient convergence. The interface should enable users to quickly focus on a promising subset of videos based on their design intent.
- **D3: Multi-scale object focus.** The interface should support exploration of motion at multiple semantic scales, from coarse object-level patterns to fine-grained part-level details, depending on users' design needs (e.g., both a full body motion and an arm motion).
- **D4: Model-agnostic exploration.** The interface should not require access to model parameters or latent representations and must work with videos generated by various I2V models.

4 Many-Videos Browsing (MVB)

At the centre of the design of MVB is our motion representation of the video. Guided by these design goals, we represent each object's (or part's) motion as a polyline trajectory over time centered around the object (or part) center and overlay trajectories of all videos on the shared initial frame (D1, D2, D3). We anchor motion patterns to a common spatial reference, allowing users to quickly perceive the overall distribution of possible motions across all frames and across all videos, which is fundamentally different from representing each video as a thumbnail or as a point in an abstract feature space. As noted in Williams [2012], speed is a critical feature of motion perception. In practice, objects can share nearly identical trajectories while exhibiting very different speed profiles. As shown in Figure 3, two basketball trajectories appear almost identical yet differ substantially in their speed curves. We therefore augment each trajectory with its corresponding speed curve, visualized in a separate panel to avoid visual clutter. Importantly, because the representation is derived solely from generated videos, it treats the underlying I2V model as a black box and remains independent of model architectures, parameters (D4). This representation is lightweight, enabling the caching of many trajectories and speed curves for fine-grained motion exploration (D3).

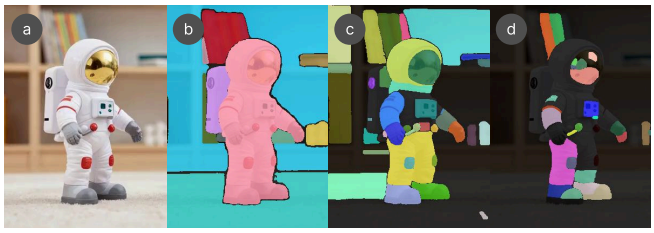


Fig. 4. **Hierarchical object detection.** We apply SAM to (a) the input image to extract objects at multiple semantic scales, from (b–d) coarse instances to finer part-level regions. This hierarchical decomposition allows users to adjust the scale of focus over objects.

4.1 System Overview

Our system consists of two stages: offline preprocessing, which generates a video collection from an image-text prompt, extracts motion trajectories and speed, and visualizes them on the shared initial frame; and interactive exploration, in which users manipulate this precomputed motion representation to explore motion variations.

4.2 Preprocessing

Object Detection. We first identify a set of objects O_{init} from the shared initial frame I_{init}. We apply SAM-based multi-scale segmentation to extract region masks at different scales (Figure 4), enabling exploration from coarse object motion to fine-grained part motion. This multi-scale segmentation takes approximately 20–30 seconds per image on an NVIDIA RTX 4080 GPU. Because all videos share the same initial frame, these objects provide a consistent reference for analyzing motion variations across the video collection.

Motion Analysis. Given a generated video collection $\mathcal{V}$, we analyze the motion of each detected object $o_i \in O_{\text{init}}$, where i denotes the object index. Specifically, for each object and each video $v \in \mathcal{V}$, we estimate a 2D motion trajectory X_i^v over time. We select a representative point p_i for each object o_i by computing the medial axis of its mask and choosing the skeleton point with the largest inscribed radius. We then track this point across frames using a pretrained point-tracking model [Karaev et al. 2025]. Collectively, this yields a trajectory set $T_i = \{X_i^v \mid v \in \mathcal{V}\}$ for each object o_i, with all trajectories originating from the same spatial location in I_{init}. In terms of computational cost, tracking 34 points in a 60-frame video takes 1.13 seconds on an NVIDIA RTX 4080 GPU; detailed runtime statistics are provided in the supplementary material.

Representative Trajectory Selection. To reduce visual clutter, the interface can display a subset of representative motion trajectories. The set can be progressively expanded by iteratively adding trajectories that are most dissimilar to those currently shown.

However, the selection is a combinatorial problem. The diversity contribution of a video cannot be evaluated independently, as it depends on which videos have already been selected, and finding a globally optimal ordering would require considering an exponential number of possible orderings. To make this problem tractable, we compute the ordering via a greedy max–min strategy inspired by

Fig. 5. **Interface overview.** Users explore the generated video collection using (a) the input image as a canvas. (b) First, users select an object (butterfly) by double-clicking, and MVB visualizes the associated motion trajectories. (c) Clicking a trajectory (highlighted on the left) displays the corresponding generated video (right), while the speed panel highlights the associated speed curve (bottom left). (d) Users can adjust the number of representative trajectories using the mouse wheel. In addition, videos can be filtered by drawing strokes (thick lime-green strokes) on either the (e) trajectories or the (f) speed curves. (g) Holding the 'Shift' key while dragging the mouse allows users to specify positive (green box) and negative (red box) regions for spatial trajectory filtering.

Maximal Marginal Relevance (MMR) [Carbonell and Goldstein 1998]. Since all videos are generated from the same prompt, relevance is treated as constant, and we use the diversity-only variant of MMR. Unlike clustering-based approaches, this strategy does not require specifying the number of clusters in advance and naturally produces an ordering for progressive exploration. Let $\mathcal{S}_{j-1}$ denote the set of videos selected after $j-1$ iterations, and let $\mathcal{U}_{j-1} = \mathcal{V} \setminus \mathcal{S}_{j-1}$ denote the remaining unselected videos. At iteration j of the representative trajectory selection process, we select the video that maximizes its minimum distance to the current selected set:

$$v_j = \arg \max_{v \in \mathcal{U}_{j-1}} \min_{u \in \mathcal{S}_{j-1}} d(v, u).$$

The distance between two videos $v, u \in \mathcal{V}$ is defined as the maximum trajectory dissimilarity over all detected objects, measured using Dynamic Time Warping (DTW) [Itakura 1975]:

$$d(v, u) = \max_{o_i \in \mathcal{O}_{\text{init}}} \text{DTW}(X_i^v, X_i^u).$$

This definition measures diversity at the scene level rather than focusing on object-specific variation, encouraging the selection of videos with globally distinct motion patterns during broad exploration. In practice, we randomly select the first video. Additionally, we simplify each trajectory by Ramer–Douglas–Peucker algorithm [Douglas and Peucker 1973] to a fixed-length polyline to remove high-frequency noise while preserving overall motion shape. This strategy produces an ordering in which videos exhibiting more diverse object motions appear earlier in the sequence. We provide implementation details and computational analysis in the Supplementary Material.

Visualization. Trajectories are rendered on the first frame semi-transparently, such that overlap and transparency indicate the trajectories' density. Their colors are evenly sampled hues in the HSV color space. Speed profiles are displayed in a separate panel below the first frame.

4.3 Exploration

Based on the above design goals, we consider the following interaction scenarios that support exploratory workflows:

- **Overview.** Users double-click around an object or part center to view its trajectory and speed across all the videos (Figure 5(b)). Each trajectory serves as an interactive handle: users can click to view the object's speed curve and the corresponding video (Figure 5(c)).
- **Visualize representative trajectories.** Users can view and interactively adjust the number of representative trajectories using the mouse wheel, e.g., progressively reducing the displayed trajectories to retain only salient and novel motion patterns and vice versa (Figure 5(d)).
- **Filter by drawing strokes.** Users can sketch the desired motion trajectory on the initial frame to query videos whose object trajectories match the drawn strokes (Figure 5(e)). Strokes can be incrementally added, modified, or removed, allowing users to iteratively refine their search as their motion intent evolves. The same interaction applies to speed curves: users can sketch a desired speed curve to further filter results (Figure 5(g)).
- **Filter by positive or negative regions.** To express spatial constraints, users can drag the mouse while holding the 'Shift' key to specify the positive or negative regions on the frame to include or exclude trajectories that pass through designated areas (Figure 5(f)). This supports focused exploration when users have partial spatial requirements, e.g., trajectories avoiding obstacles.
- **Adjust the scale of focus.** Users can adjust the scale of focus over objects, shifting between coarse object-level motion and fine-grained part-level motion using the arrow keys (←, →) (see Figure 6).
- **Browse newly appearing objects.** The object discovery panel allows users to discover and explore objects that first appear after the initial frame (Figure 7 left). Users can click on any such object to view its trajectory in the corresponding video (Figure 7 right).

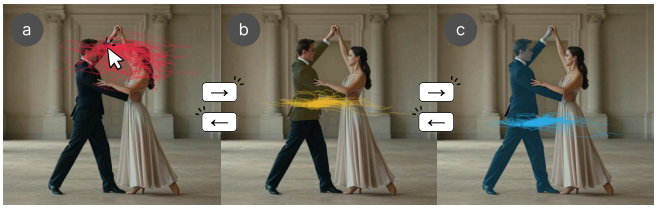


Fig. 6. **Adjust the scale of focus.** Users can use arrow keys to traverse object-part hierarchies at the selected location, enabling rapid shifts in focus from fine-grained parts to coarse object-level motion and vice versa (e.g., arm ↔ upper body ↔ full body of the male dancer).

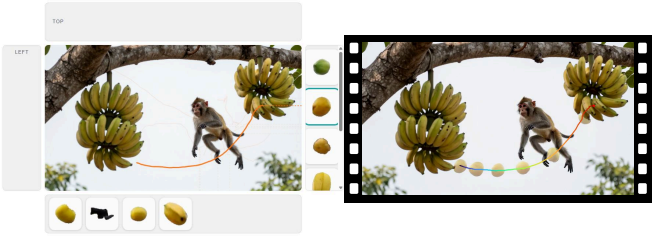


Fig. 7. **The object discovery panel.** Left: Newly appearing objects are displayed around the border of the initial frame, with each object's position reflecting the direction from which it first enters the scene. Right: Users can click on any such object to view its trajectory in the corresponding video.

5 User Study

5.1 Procedure

To evaluate the effectiveness of MVB, we conducted a user study comparing it with a text-based baseline interface (Figure 8). Following the study design of previous work [Pandey et al. 2023], we asked participants to explore pre-generated video collections using each interface and performed (a) an *understanding* task to comprehend motions, (b) a *comparison* task to compare motions, and (c) a *search* task to locate specific motions.

In the understanding and comparison tasks, we asked participants to answer questions about the motions in a video collection. An example understanding question was: "Is there any video where the basketball goes through the hoop?" An example comparison question was: "Which is more frequent: videos where the basketball hits the hoop or those where it does not?" For the search task, participants first viewed a target video and then searched for it within the video collection. We assess the interface based on how quickly and accurately users can complete these tasks.

We implemented our proposed interface and the baseline interface within the same layout (Figure 8). The left panel shows either our trajectory-based interface or the text-based baseline, while the right panel displays a video gallery with automatic looping playback. The baseline interface enables text-based exploration using an existing video-text encoder [Zhao et al. 2024]. When users enter a text query, the gallery displays videos sorted by similarity to the query.

We prepared two collections of single-object videos and two collections of multi-object videos for our user study. For each collection, we generated an initial frame using a text-to-image model [Cai et al. 2025]. We then generate 60-frame videos using an I2V model [Zhang et al. 2025a], with the initial frame and a text prompt as inputs. To focus on object motion, we filtered out videos that contained camera motion until each collection reached 100 videos. Please refer to the supplemental material for details of the video collections.

Study protocol. We conducted a within-subject user study with 12 participants (P1–P12) aged 20–30, recruited through convenience sampling. The study was conducted in person using a desktop PC equipped with Intel i9-14900KF, 64GB RAM, and NVIDIA GeForce RTX 4090. A GPU is required only for the baseline interface to compute text-video similarities. We began by introducing the overview of the study and procedure, and then participants performed video exploration tasks using both interfaces.

We organized the exploration tasks into two blocks: one for testing one interface and another for the text-based interface. In each block, participants explored a single-object collection and a multi-object collection, with the order counterbalanced across the two interfaces. Each block began with a tutorial on the interface and four minutes of practice. For each video collection, participants were asked to complete one understanding task, one comparison task, and two search tasks that focused on typical and rare motions. Each task was limited to three minutes; if a task was not completed within the time limit, it would be terminated. Once all exploration tasks were completed, participants filled out questionnaires. The entire session took around one hour. Please refer to the supplemental material for the list of tasks and the questionnaire used in the user study.

5.2 Results

We evaluated the effectiveness of each interface using both objective metrics and subjective questionnaires. Specifically, we analyzed task accuracy and completion time as objective metrics (Figure 9). For each task, we recorded whether participants answered each question correctly and measured task completion time. In addition, we collected questionnaires on mental workload based on NASA-TLX [Hart and Staveland 1988]. We used paired t-tests to assess statistical significance for each evaluation metric, with the significance level set at $\alpha = 0.05$. We observed significant differences between our interface and the baseline interface in overall task accuracy, completion time, and mental workload. These results indicate that MVB enables more efficient and accurate video exploration compared to the baseline.

For the *understanding task*, the results indicate that participants can develop an overall understanding of the motions contained in a video collection faster using MVB. Using MVB, all visualized trajectories associated with a user-selected object enable participants to quickly grasp an overview of the motions present, which likely contributed to more efficient completion of the understanding task.

For the *comparison task*, participants were more accurate using MVB compared to the text-based baseline. The limited precision of text queries likely made it difficult to compare motions accurately. In contrast, MVB allows users to explicitly specify object motions

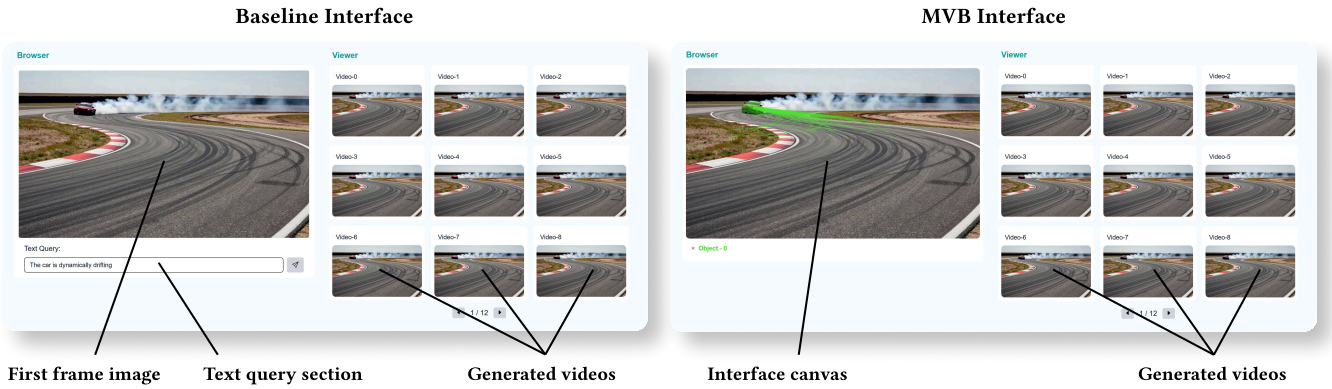


Fig. 8. **Interface designs used in the user study**. The baseline interface using text queries for retrieval (Left), and our MVB interface (Right).

through trajectory-based queries, enabling participants to categorize and distinguish motions effectively.

For the *search task*, we analyze two sub-tasks: searching for typical and rare motions. For the typical motion task, no statistically significant differences were observed between the two interfaces in terms of accuracy and task completion time. This can be attributed to the fact that, regardless of whether trajectory-based or text-based queries were used, participants needed to inspect individual videos to confirm visual similarity to the target video. In contrast, participants can search for rare motions faster and more accurately using MVB. The unique trajectory visualization in MVB makes it easier for participants to identify rare motions. Even when these motions look similar to other trajectories, participants can efficiently narrow down candidate videos using MVB's stroke-based filtering function. These capabilities of MVB were likely a factor in the improved performance observed in the rare motion task.

Additionally, we assessed participants' mental workload under each interface condition using NASA-TLX (Figure 10). Compared with the baseline, our interface yielded significantly lower ratings for Mental Demand ($t(12) = 2.727, p = .020$), Temporal Demand ($t(12) = 2.569, p = .026$), Effort ($t(12) = 2.454, p = .032$), and Frustration ($t(12) = 4.468, p = .001$). We found no significant differences in Physical Demand ($t(12) = 1.393, p = .191$) or Performance ($t(12) = 1.970, p = .075$). The overall workload score also differed significantly between the two conditions ($t(12) = 2.945, p = .013$).

6 Interface Component Evaluation

The user study in Section 5 evaluates the effectiveness of MVB as a complete interface. Here, we isolate two interface components unique to video browsing—the speed curve visualization and the object discovery panel—to evaluate their individual contributions to video exploration. We conducted a between-subjects study with six participants (three per condition). Participants completed two exploration tasks using either the full interface or a version with one interface component removed.

Speed curve visualization. Participants searched for "the basketball video in which the ball slows down" using either the full interface or a version without the speed curve visualization. All participants successfully completed the task under both conditions; however, the full interface reduced the average completion time from 50.0 to 32.3 seconds, showing that speed curves improve motion discrimination.

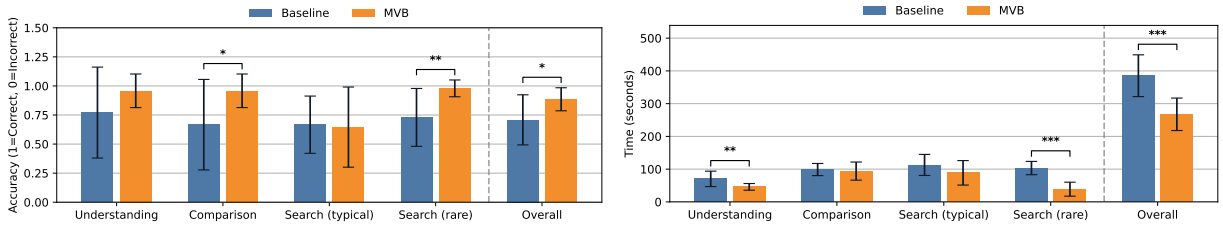
Object discovery panel. Participants searched for "the video in which a green fruit newly appears" using either the full interface or a version without the object discovery panel. All participants using the full interface successfully found the target video, whereas none using the reduced interface succeeded within the three-minute time limit, demonstrating the importance of the object discovery panel.

Although limited in scale, these results provide preliminary evidence that both interface components improve the efficiency of motion exploration.

7 Professional Artists Feedback: Animation Inbetweening

We conducted artist interviews to validate our interface design and discuss its potential use and challenges in real-world production environments. We chose *animation inbetweening* (the inbetweening task in 2D hand-drawn animation) as the focus of our artist interviews because it is one of the potential applications of MVB and also well-suited for the validation of MVB. First, in this task, visual appearance and scene content are largely fixed, making object motion the primary source of variation. This allows us to evaluate MVB in a setting where motion exploration is central. Second, this task highlights MVB's complementarity with control-based methods: not only the first but also the last frame serves as an explicit motion constraint, yet various plausible in-between motions remain, requiring users to explore and compare alternatives rather than specifying a single target trajectory. Finally, this task is well established in animation production, allowing us to recruit experienced artists and ground our evaluation in a familiar, practical workflow. We recruited three animation video artists with 10 years (PA1), 20 years (PA2), and 17 years (PA3) of professional experience. We prepared two animation inbetweening cases and asked the artists to use MVB to select the best one in their mind. Figure 11 illustrates the task setup. The videos are generated by [Zhang et al. 2025a].

The artists were generally interested in the concept of MVB and appreciated its potential practical value in everyday animation workflows. PA2 noted that *"even with the current functionality of MVB, it could potentially be used in practice for animation production if*



(a) Task accuracy.

(b) Task completion time.

Fig. 9. **User study results.** The p-values indicate statistical significance: $p < 0.05$ (*), $p < 0.01$ (**), and $p < 0.001$ (***).

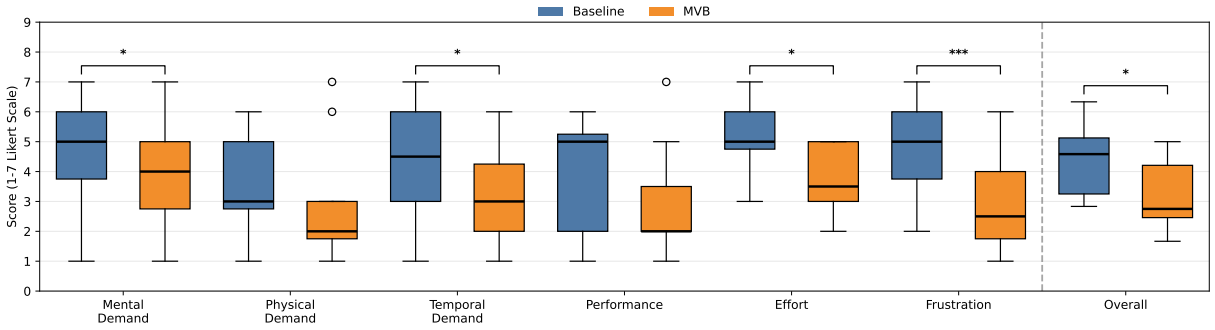


Fig. 10. RAW-TLX. The p-values indicate statistical significance: $p < 0.05$ (*), $p < 0.01$ (**), and $p < 0.001$ (***).

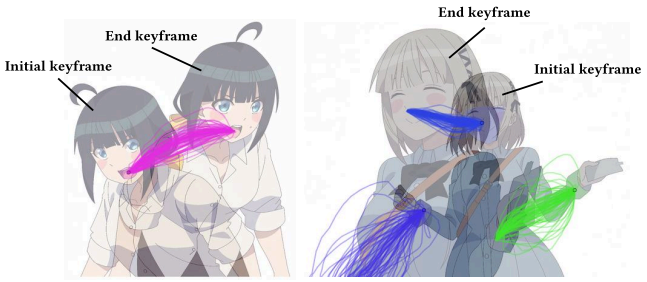


Fig. 11. **Animation Inbetweening.** Given an initial keyframe and an end keyframe as inputs, we generate intermediate frames that interpolate between the keyframes using [Zhang et al. 2025a]. As illustrated by the motion trajectories via MVB, the model generates a diverse set of motions while satisfying the keyframe constraints.

the accuracy of video generation models improves.". PA3 similarly remarked that *"exploring motion variations with this interface could be useful, particularly in cases where motion trajectories do not need to be precisely fixed in advance, such as simple object movements."*

Artists viewed our interface as a good contemporary tool for control-based video generation: PA2 proposed enabling users to *"specify three or more keyframes when generating videos, so that motions generated under more detailed conditions could be explored."* PA3 noted that *"even when having a preferred motion in mind, it is still beneficial to examine variations before deciding".*

Besides, the artists appreciate the functions supported in MVB. For motion overview, PA3 positively noted that *"compared to manually drawing animation frame by frame, the interface enabled users to preview a variety of motion candidates instantly. This capability has the potential to reduce the time required for exploring and refining motions."* As a concrete usage scenario, PA3 also suggested that *"young animators could use MVB to examine a wide range of motion examples to deepen their understanding of motion design."* Artists also valued adjusting the scale of focus and the filtering function. All designers noted that animators often care more about the motion of specific parts (e.g., a character's eyes or chin) than overall body motion. During the interviews, artists repeatedly used the filtering function to locate videos exhibiting the desired part-level motions.

While overall feedback was positive, the artists identified several directions for improvement. PA3 suggested a feature for exploring motion timing, proposing that *"not only motion trajectories but also variations in timing could be examined simultaneously".* Both PA2 and PA3 pointed out that the current videos generated from I2V models have not reached the professional animation production level, such as the occurrence of temporal blur (the model tries to enforce temporal consistency across frames and ends up averaging appearance over time), and the temporal inconsistency of motion. This limits the usability of our interface, but is expected to be addressed by the improvements of video generation models.

8 Creative Applications

Motivated by two of our initial design goals, i.e., rapid understanding of motion distribution (D1) and rapid narrowing of the motion space (D2), we present three applications illustrating the potential usage of MVB through its trajectory visualization and interaction design.

Fig. 12. **Video motion montage.** Users can use (a, e) MVB to select desired object motions from different videos (Top: (b) dog's motion from one video and (c) boy's motion from another video; Bottom: (f) astronaut's motion from one video and (g) robot's motion from another video). We generate (d, h) new motion montage videos by composing these motions with a video denoising approach [Singer et al. 2025]. This allows users to generate new videos that are not included in the existing video collections. To visualize the videos, we overlaid keyframes with semi-transparency and displayed representative trajectories. Please check the videos in the supplemental video.

8.1 Video Motion Montage

We present Video Motion Montage as an application of MVB that enables users to synthesize new videos by recombining motions of different objects selected from different video sources.

Using MVB, users first explore a large collection of generated videos to understand the motion space and identify object motions of interest. After exploration (Figure 12a), users can select desired object motions from different source videos—for example, selecting the dog's motion from one video (Figure 12b) and the boy's motion from another (Figure 12c)—and combine these selected motions to co-occur in a single generated video (Figure 12d).

In practice, we first combine the selected object motions in pixel space. However, this direct composition often introduces visual artifacts. We therefore use this video as conditioning signals for Time-to-Move [Singer et al. 2025], a control-based video denoising method, to generate a new video that lies on the model's natural manifold. As shown in Figure 12(d, h), this workflow allows users to flexibly mix object motions discovered through exploration and synthesize new, coherent videos.

8.2 MotionScout

Tactical decision-making in football often begins with under-specified intent, where a corner kick player must choose from a wide range of plausible ball motions. We present MotionScout as an application of MVB that enables users to scout and select valid attacking plays from a collection of corner kick videos generated by an I2V model. Given an input image of a corner kick scenario and a divergent prompt "A boy throws the ball to have a good corner," using MVB, users first explore the full distribution of ball trajectories overlaid on the shared initial frame (Figure 13(b)), gaining an immediate overview of what motions are possible. After exploration, users can identify and select distinct plays of interest—for example, a fake move (Figure 13(c)), a roll to the ground (Figure 13(d)), and a pass to the middle player (Figure 13(e))—each discovered by progressively narrowing the trajectory space using MVB's stroke-based and region-based filtering. This workflow demonstrates how MVB naturally supports tactical exploration in sports, enabling users to discover and compare possible motion outcomes without needing to browse all the videos individually.

8.3 Interactive Long Video Generation

When generating long-duration videos with I2V models, motion variations tend to accumulate over time, leading to an increasingly diverse motion space. Naively exploring this expanded motion space requires generating a prohibitively large number of videos. To address this challenge, we demonstrate how MVB can be used as an application for long-duration video generation by progressively exploring and extending generated videos in a multi-stage manner.

We begin by generating an initial video collection from an input image. Users explore this initial collection with MVB and select a video that best reflects their intent (Figure 14(a, e)). The selected video is then fed into a video extension model [Zhang et al. 2025a] to generate the next collection, which users again explore and select from using MVB (Figure 14(b, f)). This process repeats across multiple stages (Figure 14(c, g)), allowing users to efficiently explore the motion space of long-duration generated videos and design a video that matches their intent (Figure 14(d, h)).

9 Discussion, Limitations, and Conclusion

While our interface receives positive feedback from the user studies and professional animation artists, we believe there remains significant room for improvement. First, the pregeneration stage remains a practical bottleneck, as generating large video collections remains computationally expensive. E.g., with Wan2.2 [Wan et al. 2025], generating 100 81-frame videos at 832×480 resolution using four NVIDIA A100 40GB GPUs takes roughly 4.5 hours. This overhead can be mitigated by adopting faster video generation methods. At the same time, we note existing video authoring using generative models is itself an iterative process. Users repeatedly generate, inspect, and regenerate outputs, which can consume substantial

Fig. 13. **MotionScout.** Corner Kick Play Scouting. (a) The user starts with an input image of a corner kick scenario and a divergent prompt "A boy throws the ball to have a good corner." (b) MVB visualizes the full distribution of ball motion trajectories across the generated video collection. The user scouts three distinct attacking plays: (c) a fake move, (d) a roll to the ground, and (e) a pass to the middle player.

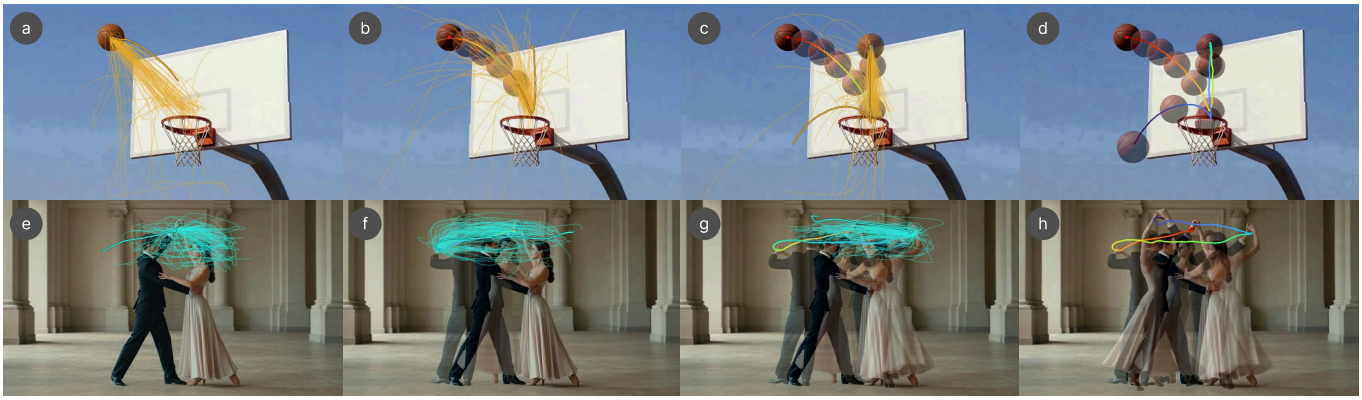


Fig. 14. **Interactive long video generation.** (a, e) The user starts by exploring videos generated from the input image and selects one video. (b, f) Using the selected video from the previous stage as input, the user generates a diverse set of extended videos, explores them, and selects one. (c, g) The user similarly generates and explores further extensions of the previously selected video and chooses one. (d, h) By iteratively repeating this stage-by-stage process of video extension and exploration, the user ultimately designs a long video that reflects their intent. To visualize the video, we overlaid keyframes with semi-transparency and displayed representative trajectories. Please check the videos in the supplemental video.

Fig. 15. **Examples of limited motion diversity in video collections.** (a) a toy mouse being chased by a cat and (b) two people crossing a bridge. In both examples, limited motion diversity within the video collections causes the visualized trajectories to largely overlap, reducing the usefulness of exploratory visualization.

interaction time. Our approach shifts this effort to an offline preprocessing stage, allowing users to explore candidates interactively and potentially reducing the overall interaction time.

Additionally, MVB assumes that generated collections contain sufficiently diverse motion variations. When videos exhibit limited variation, the benefits of exploration are reduced as shown in Figure 15.

Finally, integrating MVB with interactive video generation systems to support iterative synthesis and refinement within the user feedback loop remains an important future direction.

In summary, we presented Many-Videos Browsing (MVB), an interactive interface designed to support the exploration of large collections of videos generated by I2V models. By making motion explicit and navigable, MVB tackles a core challenge of this emerging technology: understanding and selecting desired videos from a rich yet unstructured space of stochastic motion outcomes. Through controlled studies and interviews with animation artists, we demonstrate that MVB is effective for exploration and can be naturally integrated into real-world creative workflows, complementing existing control-based generation methods.

Acknowledgments

We thank Forrester Cole for brainstorming the idea of this project. This work is based on results obtained from GENIAC (Generative AI Accelerator Challenge, a project to strengthen Japan's generative AI development capabilities), a project implemented by the Ministry of Economy, Trade and Industry (METI) and the New Energy and Industrial Technology Development Organization (NEDO). This work was also supported by JST, CRONOS JPMJCS25K1, JST ASPIRE JPMJAP2401, JSPS Grant-in-Aid JP23K16921, Japan, and UTokyo-Google AI Symbiotic Future Society Program.

References

Sherwin Bahmani, Ivan Skorokhodov, Guocheng Qian, Aliaksandr Siarohin, Willi Menapace, Andrea Tagliasacchi, David B Lindell, and Sergey Tulyakov. 2025. Ac3d: Analyzing and improving 3d camera control in video diffusion transformers. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 22875–22889.

Sherwin Bahmani, Ivan Skorokhodov, Aliaksandr Siarohin, Willi Menapace, Guocheng Qian, Michael Vasilkovsky, Hsin-Ying Lee, Chaoyang Wang, Jiaxu Zou, Andrea Tagliasacchi, et al. 2024. Vd3d: Taming large video diffusion transformers for 3d camera control. *arXiv preprint arXiv:2407.12781* (2024).

Connelly Barnes, Dan B Goldman, Eli Shechtman, and Adam Finkelstein. 2010. Video tapestries with continuous temporal zoom. In *ACM SIGGRAPH 2010 papers*. 1–9.

Kai Uwe Barthel, Nico Hezel, and Radek Mackowiak. 2015. Graph-based browsing for large video collections. In *International conference on multimedia modeling*. Springer, 237–242.

Kai Uwe Barthel, Nico Hezel, and Radek Mackowiak. 2016. Navigating a graph of scenes for exploring large video collections. In *International Conference on Multimedia Modeling*. Springer, 418–423.

Jürgen Bernard, Nils Wilhelm, Björn Krüger, Thorsten May, Tobias Schreck, and Jörn Kohlhammer. 2013. Motionexplorer: Exploratory search in human motion capture data based on hierarchical aggregation. *IEEE transactions on visualization and computer graphics* 19, 12 (2013), 2257–2266.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. 2023. Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets. arXiv:2311.15127 [cs] doi:10.48550/arXiv.2311.15127

John Boreczky, Andreas Girgensohn, Gene Golovchinsky, and Shingo Uchihashi. 2000. An interactive comic book presentation for exploring video. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 185–192.

Tim Brown et al. 2008. Design thinking. *Harvard business review* 86, 6 (2008), 84.

Bill Buxton. 2010. *Sketching user experiences: getting the design right and the right design*. Morgan kaufmann.

Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Steven Hoi, Shijie Huang, Zhaohui Hou, Dengyang Jiang, Xin Jin, Liangchen Li, et al. 2025. Z-Image: An Efficient Image Generation Foundation Model with Single-Stream Diffusion Transformer. *arXiv preprint arXiv:2511.22699* (2025).

Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. 335–336.

Chia-Hsing Chiu, Yuki Koyama, Yu-Chi Lai, Takeo Igarashi, and Yonghao Yue. 2020. Human-in-the-loop differential subspace search in high-dimensional latent space. *ACM Transactions on Graphics (TOG)* 39, 4 (2020), 85–1.

Myung Geol Choi, Kyungyong Yang, Takeo Igarashi, Jun Mitani, and Jehee Lee. 2012. Retrieval and visualization of human motion data via stick figures. In *Computer Graphics Forum*, Vol. 31. Wiley Online Library, 2057–2065.

David H Douglas and Thomas K Peucker. 1973. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Cartographica: the international journal for geographic information and geovisualization* 10, 2 (1973), 112–122.

Steven P Dow, Alana Glassco, Jonathan Kass, Melissa Schwarz, Daniel L Schwartz, and Scott R Klemmer. 2010. Parallel prototyping leads to better design results, more divergence, and increased self-efficacy. *ACM Transactions on Computer-Human Interaction (TOCHI)* 17, 4 (2010), 1–24.

Pierre Dragicevic, Gonzalo Ramos, Jacobo Bibliowitcz, Derek Nowrouzezahrai, Ravin Balakrishnan, and Karan Singh. 2008. Video browsing by direct manipulation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 237–246.

Haiwen Feng, Zheng Ding, Zhihao Xia, Simon Niklaus, Victoria Abrevaya, Michael J. Black, and Xuaner Zhang. 2025. Explorative Inbetweening of Time and Space. In *Computer Vision – ECCV 2024*, Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (Eds.). Vol. 15136. Springer Nature Switzerland, 378–395. doi:10.1007/978-3-031-73229-4_22

Rohit Gandikota, Zongze Wu, Richard Zhang, David Bau, Eli Shechtman, and Nick Kolkin. 2025. Sliderspace: Decomposing the visual capabilities of diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 15994–16003.

Daniel Geng, Charles Herrmann, Junhwa Hur, Forrester Cole, Serena Zhang, Tobias Pfaff, Tatiana Lopez-Guevara, Yusuf Aytar, Michael Rubinstein, and Chen Sun. 2025. Motion Prompting: Controlling Video Generation with Motion Trajectories. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 1–12.

Purvi Goel and Doug L. James. 2022. Unified many-worlds browsing of arbitrary physics-based animations. *ACM Trans. Graph.* 41, 4, Article 156 (July 2022), 15 pages. doi:10.1145/3528223.3530082

Dan B Goldman, Chris Gonterman, Brian Curless, David Salesin, and Steven M Seitz. 2008. Video object annotation, navigation, and composition. In *Proceedings of the 21st annual ACM symposium on User interface software and technology*. 3–12.

Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. 2025. CameraCtrl: Enabling Camera Control for Text-to-Video Generation. arXiv:2404.02101 [cs] doi:10.48550/arXiv.2404.02101

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. 2022a. Imagen Video: High Definition Video Generation with Diffusion Models. arXiv:2210.02303 [cs] doi:10.48550/arXiv.2210.02303

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. 2022b. Video Diffusion Models. *Advances in neural information processing systems* 35 (2022), 8633–8646.

Fumitada Itakura. 1975. Minimum prediction residual principle applied to speech recognition. *IEEE Transactions on acoustics, speech, and signal processing* 23, 1 (1975), 67–72.

Dan Jackson, James Nicholson, Gerrit Stoeckigt, Rebecca Wrobel, Anja Thieme, and Patrick Olivier. 2013. Panopticon: A parallel video overview system. In *proceedings of the 26th annual ACM symposium on User interface software and technology*. 123–130.

Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin. 2025. Pyramidal Flow Matching for Efficient Video Generative Modeling. arXiv:2410.05954 [cs] doi:10.48550/arXiv.2410.05954

Nikita Karaev, Yuri Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. 2025. Cotracker3: Simpler and better point tracking by pseudo-labelling real videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 6013–6022.

Yuki Koyama and Masataka Goto. 2022. Bo as assistant: Using bayesian optimization for asynchronously generating design suggestions. In *Proceedings of the 35th annual ACM symposium on user interface software and technology*. 1–14.

Yuki Koyama, Issei Sato, and Masataka Goto. 2020. Sequential gallery for interactive visual design optimization. *ACM Transactions on Graphics (TOG)* 39, 4 (2020), 88–1.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. 2023. Flow Matching for Generative Modeling. arXiv:2210.02747 [cs] doi:10.48550/arXiv.2210.02747

Xingchao Liu, Chengyue Gong, and Qiang Liu. 2022. Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow. arXiv:2209.03003 [cs] doi:10.48550/arXiv.2209.03003

Y-C Liu, Amaresh Chakrabarti, and Thomas Bligh. 2003. Towards an 'ideal'approach for concept generation. *Design studies* 24, 4 (2003), 341–355.

Wenping Ma, Xiaoting Yang, Licheng Jiao, Lingling Li, Xu Liu, Fang Liu, Puhua Chen, Yuting Yang, Mengru Ma, Long Sun, et al. 2025. Video diffusion generation: comprehensive review and open problems. *Artificial Intelligence Review* 58, 11 (2025), 338.

Justin Matejka, Michael Glueck, Erin Bradner, Ali Hashemi, Tovi Grossman, and George Fitzmaurice. 2018. Dream lens: Exploration and visualization of large-scale generative design datasets. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–12.

Justin Matejka, Tovi Grossman, and George Fitzmaurice. 2013. Swifter: improved online video scrubbing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1159–1168.

Justin Matejka, Tovi Grossman, and George Fitzmaurice. 2014. Video lens: rapid playback and exploration of large video collections and associated metadata. In *Proceedings of the 27th annual ACM symposium on User interface software and technology*. 541–550.

Haomiao Ni, Changhao Shi, Kai Li, Sharon X. Huang, and Martin Renqiang Min. 2023. Conditional Image-to-Video Generation with Latent Flow Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18444–18455.

Peter O'Donovan, Aseem Agarwala, and Aaron Hertzmann. 2015. Designscape: Design with interactive layout suggestions. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. 1221–1224.

Karran Pandey, Fanny Chevalier, and Karan Singh. 2023. Juxtaform: interactive visual summarization for exploratory shape design. *ACM Trans. Graph.* 42, 4, Article 52 (July 2023), 14 pages. doi:10.1145/3592436

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.

Amy Pavel, Colorado Reed, Björn Hartmann, and Maneesh Agrawala. 2014. Video digests: a browsable, skimmable format for informational lecture videos. In *Proceedings of the 27th annual ACM symposium on User interface software and technology*. 573–582.

Gonzalo Ramos and Ravin Balakrishnan. 2003. Fluid interaction techniques for the control and annotation of digital video. In *Proceedings of the 16th annual ACM symposium on User interface software and technology*. 105–114.

Konstantin Schall, Nico Hezel, Klaus Jung, and Kai Uwe Barthel. 2023. Vibro: video browsing with semantic and visual image embeddings. In *International Conference on Multimedia Modeling*. Springer, 665–670.

Donald A Schön. 2017. *The reflective practitioner: How professionals think in action*. Routledge.

Ben Shneiderman. 2007. Creativity support tools: accelerating discovery and innovation. *Commun. ACM* 50, 12 (2007), 20–32.

Karl Sims. 1991. Artificial evolution for computer graphics. In *Proceedings of the 18th annual conference on Computer graphics and interactive techniques*. 319–328.

Assaf Singer, Noam Rotstein, Amir Mann, Ron Kimmel, and Or Litany. 2025. Time-to-Move: Training-Free Motion Controlled Video Generation via Dual-Clock Denoising. *arXiv preprint arXiv:2511.08633* (2025).

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. 2022. Make-A-Video: Text-to-Video Generation without Text-Video Data. arXiv:2209.14792 [cs] doi:10.48550/arXiv.2209.14792

Anthony Tang, Saul Greenberg, and Sidney Fels. 2008. Exploring video streams using slit-tear visualizations. In *Proceedings of the working conference on Advanced visual interfaces*. 191–198.

James Tompkin, Kwang In Kim, Jan Kautz, and Christian Theobalt. 2012. Videoscapes: exploring sparse, unstructured video collections. *ACM Transactions on Graphics (TOG)* 31, 4 (2012), 1–12.

Shuyuan Tu, Zhen Xing, Xintong Han, Zhi-Qi Cheng, Qi Dai, Chong Luo, and Zuxuan Wu. 2025. Stableanimator: High-quality Identity-Preserving Human Image Animation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 21096–21106.

Christopher D. Twigg and Doug L. James. 2007. Many-worlds browsing for control of multibody dynamics. In *ACM SIGGRAPH 2007 Papers* (San Diego, California) *(SIGGRAPH '07)*. Association for Computing Machinery, New York, NY, USA, 14–es. doi:10.1145/1275808.1276395

Patrik Veselỳ, František Mejzlík, and Jakub Lokoč. 2021. Somhunter V2 at video browser showdown 2021. In *International Conference on Multimedia Modeling*. Springer, 461–466.

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).

Angtian Wang, Haibin Huang, Jacob Zhiyuan Fang, Yiding Yang, and Chongyang Ma. 2025a. ATI: Any Trajectory Instruction for Controllable Video Generation. *arXiv preprint arXiv:2505.22944* (2025).

Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. 2023. Videocomposer: Compositional video synthesis with motion controllability. *Advances in Neural Information Processing Systems* 36 (2023), 7594–7611.

Xiaojuan Wang, Boyang Zhou, Brian Curless, Ira Kemelmacher-Shlizerman, Aleksander Holynski, and Steven M. Seitz. 2025b. Generative Inbetweening: Adapting Image-to-Video Models for Keyframe Interpolation. arXiv:2408.15239 [cs] doi:10.48550/arXiv.2408.15239

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 *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers 24*. ACM, Denver CO USA, 1–11. doi:10.1145/3641519.3657518

Richard Williams. 2012. *The animator's survival kit: a manual of methods, principles and formulas for classical, computer, games, stop motion and internet animators*. Macmillan.

Shengming Yin, Chenfei Wu, Jian Liang, Jie Shi, Houqiang Li, Gong Ming, and Nan Duan. 2023. Dragnuwa: Fine-grained control in video generation by integrating text, image, and trajectory. *arXiv preprint arXiv:2308.08089* (2023).

Lvmin Zhang, Shengqu Cai, Muyang Li, Gordon Wetzstein, and Maneesh Agrawala. 2025a. Frame Context Packing and Drift Prevention in Next-Frame-Prediction Video Diffusion Models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Zhenghao Zhang, Junchao Liao, Menghao Li, Zuozhuo Dai, Bingxue Qiu, Siyu Zhu, Long Qin, and Weizhi Wang. 2025b. Tora: Trajectory-oriented Diffusion Transformer for Video Generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 2063–2073.

Long Zhao, Nitesh B Gundavarapu, Liangzhe Yuan, Hao Zhou, Shen Yan, Jennifer J Sun, Luke Friedman, Rui Qian, Tobias Weyand, Yue Zhao, et al. 2024. Videoprism: A foundational visual encoder for video understanding. *arXiv preprint arXiv:2402.13217* (2024).

Jun-Yan Zhu, Yong Jae Lee, and Alexei A Efros. 2014. Averageexplorer: Interactive exploration and alignment of visual data collections. *ACM Transactions on Graphics (TOG)* 33, 4 (2014), 1–11.