

Videogenic: Identifying Highlight Moments in Videos with Professional Photographs as a Prior

David Chuan-En Lin
Carnegie Mellon University
Pittsburgh, PA, USA
chuanenl@cs.cmu.edu

Fabian Caba Heilbron
Adobe Research
San Jose, CA, USA
caba@adobe.com

Joon-Young Lee
Adobe Research
San Jose, CA, USA
jolee@adobe.com

Oliver Wang
Adobe Research
Seattle, WA, USA
owang@adobe.com

Nikolas Martelaro
Carnegie Mellon University
Pittsburgh, PA, USA
nikmart@cmu.edu

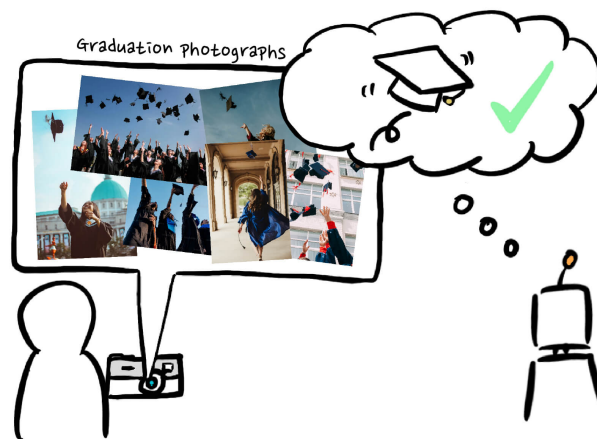


Figure 1: Videogenic leverages high-quality photographs as a prior to identify domain-specific highlights. For example, graduation photographs may frequently include people tossing their graduation caps.

ABSTRACT

This paper investigates the challenge of extracting highlight moments from videos. To perform this task, we need to understand what constitutes a highlight for arbitrary video domains while at the same time being able to scale across different domains. Our key insight is that photographs taken by photographers tend to capture the most remarkable or *photogenic* moments of an activity. Drawing on this insight, we present Videogenic, a technique capable of creating domain-specific highlight videos for a diverse range of domains. In a human evaluation study ($N=50$), we show that a high-quality photograph collection combined with CLIP-based retrieval (which uses a neural network with semantic knowledge of images) can

serve as an excellent prior for finding video highlights. In a within-subjects expert study ($N=12$), we demonstrate the usefulness of Videogenic in helping video editors create highlight videos with lighter workload, shorter task completion time, and better usability.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; • **Human-centered computing** → **Human computer interaction (HCI)**.

KEYWORDS

video highlight detection, video moment retrieval, video summarization

ACM Reference Format:

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 *Creativity and Cognition (C&C '24)*, June 23–26, 2024, Chicago, IL, USA. ACM, New York, NY, USA, 19 pages. <https://doi.org/10.1145/3635636.3656186>

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

C&C '24, June 23–26, 2024, Chicago, IL, USA

© 2024 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-0485-7/24/06.

<https://doi.org/10.1145/3635636.3656186>

1 INTRODUCTION

Video highlight generation is the task of creating a short video clip that captures the highlight moments of a longer video or video collection. Such highlight videos can be useful for a variety of purposes. For example, people may wish to create short highlight clips of an activity (e.g., the moment of a great skateboard trick) or event (e.g., the main ceremony of a wedding) to share on social media. Video creators may wish to find good moments among large amounts of raw video footage to use for their videos. Video creators also may wish to upload short snippets of their longer videos on increasingly popular short-form video platforms such as TikTok, Instagram Reels, and YouTube Shorts to advertise their works to a larger audience [1]. Video sharing platforms may wish to let users see short previews of videos before watching them (e.g., YouTube plays a 3-second preview when the user hovers over a video’s thumbnail [9]).

Many video highlight generation approaches have been proposed to support the demand for highlight videos. Since the definition of what constitutes a “highlight” is highly dependent on the domain of interest (e.g., a skateboard trick or a cool dance move), a key challenge of video highlight generation is to find some way of encoding domain knowledge about what a good highlight is within the system. Several works make use of domain-specific features to identify highlights, such as detecting the presence of people [24] or identifying when a goal is scored in sports videos [43]. However, such systems only work for the specific domain that they are designed for. More recent works make use of neural networks to learn a model of video highlights from data, such as from pairs of highlight videos and their source videos [36] or videos with particular segments labeled as highlights [42]. Nonetheless, such systems require resource-intensive training of models on large amounts of data.

“Photography is the simultaneous recognition, in a fraction of a second, of the significance of an event.”

— Henri Cartier-Bresson, Photographer

In this research, we take a different approach to creating highlight clips from longer videos by leveraging the *domain knowledge of photographers*. We posit that pictures of events or activities taken by photographers capture the most remarkable or *photogenic* moments of an activity. Given an arbitrary domain of interest, we search for a small collection of professional photographs depicting the activity and create an average representation of these photographs. We then compare the average representation against each frame of the source video to compute similarities. High similarity between the video frame and the average photograph representation corresponds to a high “highlight score”. We show that a high-quality photograph collection combined with the use of a semantic model to encode its representation (CLIP [31]) can serve as a strong prior for finding video highlights in arbitrary domains without any further training. In addition, our method also *implicitly encodes useful photography knowledge*, such as good composition and framing.

In this paper, we present Videogenic, a technique for identifying highlight moments in videos by leveraging a small sample of

photographs that represent an arbitrary activity or event. We first introduce a set of principles to ground the task of domain-agnostic highlight generation. We build an interactive proof-of-concept system for creating highlight videos to validate our technique. We test the performance of our technique against a baseline by generating highlight videos for a variety of source videos covering various domains and asking people to pick their preferences. In an expert study, we further evaluate the usefulness of Videogenic for video editors against a baseline, demonstrating improvements in workload, usability, and task completion time. We asked external raters to evaluate the videos created by the editors, demonstrating strong “highlightness” and production quality for the highlight videos created with Videogenic.

This research thus makes the following contributions:

- **Videogenic, a simple yet effective technique for identifying highlight moments in videos.** By leveraging professional photographs as a prior and leveraging CLIP’s representation power, our technique scales to *arbitrary* video domains *out-of-the-box*. We built a proof-of-concept system to validate our technique with professional video creators.
- **A human evaluation** ($N=50$ participants) of the performance of Videogenic against a baseline. Participants preferred Videogenic’s highlights on average 80% of the time.
- **An expert study** ($N=12$ video editors) evaluating the usefulness of Videogenic against a manual editing baseline. Participants experienced lighter workload, shorter task completion time, and better usability when using Videogenic. External raters rated strong “highlightness” and production quality for the highlight videos created with Videogenic.

2 RELATED WORK

Our work is situated among extensive literature on video highlight generation. We discuss works that adopt heuristics-based approaches, data-driven approaches, as well as HCI approaches.

2.1 Heuristic Approaches

As the content within videos can vary considerably across different domains, many works focus on a single video domain to define a set of domain-specific heuristics. A large body of work focuses on sports videos. A sports game usually has a well-defined structure and consists of various stages. Among the various stages, only a small selection contains highlight moments. For example, the highlights of a soccer video are generally the stages in which the goals are scored. Since the relevant highlight stages can also vary across different types of sports, a variety of approaches have been proposed across various sports categories, such as for soccer [43], baseball [32], basketball [30], and cricket [22]. In addition, a growing body of work addresses video highlight generation for egocentric videos, possibly due to a rise in popularity of personal action cameras [3, 4]. These works often make use of various pre-defined cues, such as the detection of people, faces, and objects [24, 27]. Various commercial software also rely on pre-defined heuristics. For example, Insta360’s FlashCut feature [2] uses image recognition models to detect hands and faces. Song et al. [34] uses a variety of aesthetic heuristics to select thumbnails for videos. In this research, we avoid

using pre-defined heuristics and make use of the latent knowledge encoded within professional photographs.

2.2 Data-Driven Approaches

Following recent advancements in machine learning research, recent works investigate data-driven methods for video highlight generation. For example, Yao et al. [42] learn from a dataset consisting of long videos segmented into various highlight and non-highlight segments. Nonetheless, video datasets with labeled highlights are difficult to collect as highlights are subjective and require human annotation. Moreover, the annotation task can be ambiguous since what constitutes a highlight is dependent on the video domain of interest. Looking at sports videos as an example, the human annotator would need to understand the rules of the particular sport to be able to make the annotations. As a result, the annotated datasets may contain noisy labels. Thus, several researchers have looked into exploiting *proxy* priors. Examples include using pairs of edited videos and their raw videos [36], pairs of GIFs and their video sources [18], short user-generated videos [40], web-images [20, 21], titles [35], and detecting shared visual events across multiple videos [16]. Our work is closely related to this thread of work. However, prior works involve expensive model training on (noisy) large-scale video or image data. In contrast, we show how only a small set of high-quality photographs from photographers combined with the semantic encodings of CLIP [31] can act as an excellent prior for creating highlight videos. This allows our technique to work *out-of-the-box* on arbitrary videos with no additional training necessary.

2.3 HCI Approaches

The HCI community has explored methods of obtaining domain knowledge information from people to support finding highlights in videos. Several past works adopt a crowdsourcing approach. For example, Wu et al. [39] leverage crowd wisdom for summarizing videos, capable of adapting to various video domains and summary abstraction levels. Bernstein et al. [11] use synchronous crowds to crowdsourcing highlight moments within videos in real-time. Instead of recruiting crowd workers, several works leverage existing users within social communities of video content. San et al. [33] identify recurring scenes within a video domain on video-sharing sites as a form of “social summarization.” Sun et al. [37] convert video and user comments into tree-like visual summaries to help people identify content highlights. Yang et al. [41] provide real-time summaries of livestreams based on streaming content and user interaction data. In this research, we follow a similar spirit, by tapping into the domain knowledge of professional photographers through their photography as a way of finding highlights.

HCI researchers have also built interfaces to better support browsing highlight moments in videos. Matejka et al. [28] use a grid of thumbnails designed to support users with easier scrubbing and selection of video moments. Matejka et al. [29] also help users quickly find relevant video sections through large collections of videos using text-based timelines and associated metadata about what is in the video. In this research, we support users by scrubbing through an interface with predicted highlight score data and allow people to use a brush feature across video segments to create their final highlight video.

3 PRINCIPLES OF VIDEO HIGHLIGHT GENERATION

We define three key design principles for the task of generating video highlights based on prior work to ground the development of Videogenic.

3.1 Principle 1: Domain-Agnostic Highlights

Our first principle is to create a system that is able to scale to a diverse range of domains. To achieve this, our system should be established on a *domain-agnostic prior* (i.e., does not limit the system to the specific domain it was designed for). In this research, we use the domain-agnostic prior of professional photographs and demonstrate flexibility across sports videos, events videos, nature videos, and more.

3.2 Principle 2: Domain-Specific Knowledge

Our second principle is to design the system such that is able to understand what constitutes a highlight for the specific domain of interest [36]. For example, the highlights of a skateboarding video (e.g., the impressive skateboard trick) can be very different from the highlights of a wedding video (e.g., when the officiant addresses the marriage partners). We show how photographs taken by photographers can encode rich domain-specific knowledge in addition to encouraging good composition and framing.

3.3 Principle 3: Flexible User Control

Our third principle is to give users flexible controllability. For example, users may wish to interpret the highlight predictions [23], correct errors [10], select their preference among multiple feasible highlights, or add individual touches [26]. To support this, we should give users multiple modes to pick from at various stages of the system. By selecting the automatic mode in all stages, the user may create highlight videos in a completely automatic fashion. Alternatively, the user may opt for an interactive mode to have greater fine-grained control over the final output.

4 IMPLEMENTATION

Our three principles are manifested in Videogenic and guide its implementation. In this section, we detail the implementation of our system, including (1) selecting the topic, (2) computing highlight scores, and (3) generating the highlight video.

4.1 Classifying the Activity

Our first step is determining the primary activity of the video. We offer two methods for the user: (1) by providing a keyword and (2) by using an automatic classifier.

4.1.1 User-Specified Keyword. The user may specify a keyword (e.g., skydiving) as the primary activity. This gives the user the flexibility to experiment with different keywords (Principle 3), and customize the highlight video (Principle 3). For example, given video footage of skydiving, the user may give the keyword of skydiving landing to create a highlight video for skydiving touchdowns (Figure 6).

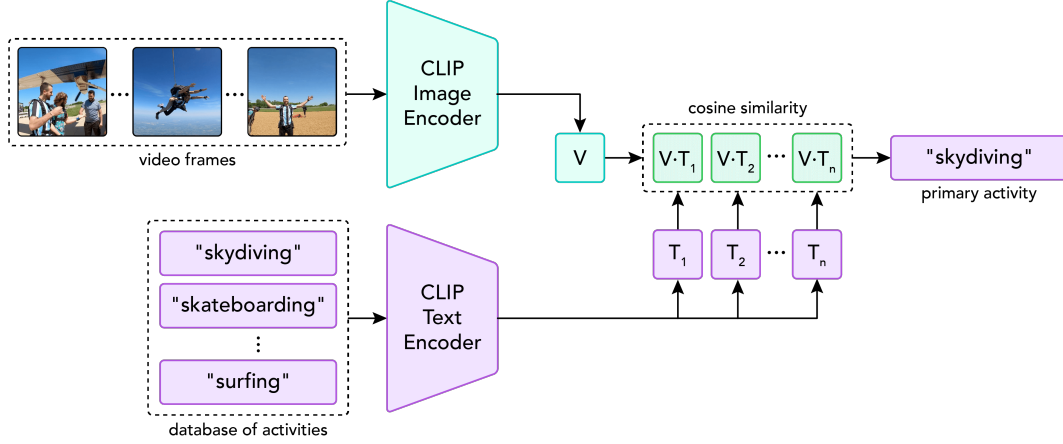


Figure 2: Automatic classifier. Given the frames of a video and a database of activity labels, Videogenic performs pairwise comparisons to predict the primary activity of the video.

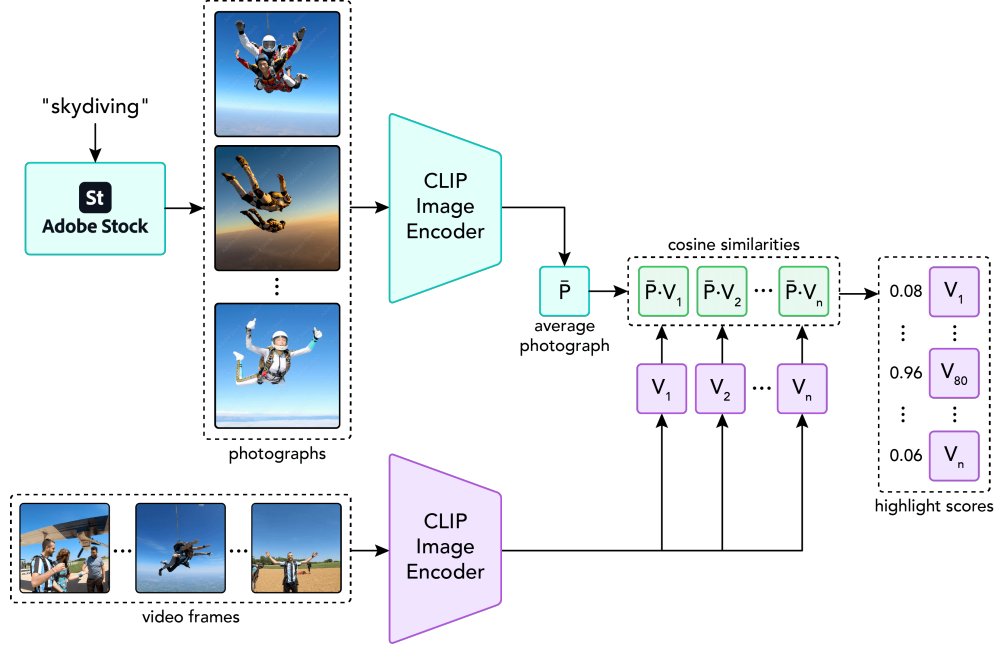


Figure 3: Computing highlight scores. Given an activity label (e.g., skydiving), Videogenic retrieves 10 stock photographs and computes the average photograph representation. Given each frame of a video and the average photograph, Videogenic performs pairwise comparisons to predict a highlight score for each frame.

4.1.2 Automatic Classifier. To support an automatic mode, the user may allow our system to automatically determine the primary activity of the video (Figure 2). Given a video, we encode each video frame through the CLIP [31] image encoder to produce its semantic representation. We then concatenate the frame-wise representations into the representation for the video. Given a database of various activity categories (e.g., skydiving, skateboarding, surfing), we encode each activity label through the CLIP text encoder. We then compare each encoded activity label representation with the

encoded video representation via cosine similarities, and select the top-ranking activity label as the primary activity.

4.2 Computing Highlight Scores

The second step is computing frame-wise highlight scores for the video (Figure 3). Given the primary activity (e.g., skydiving), Videogenic automatically retrieves 10 professional photographs depicting the activity from a database of professional photography (e.g., top 10 photos from Adobe Stock search results). We encode

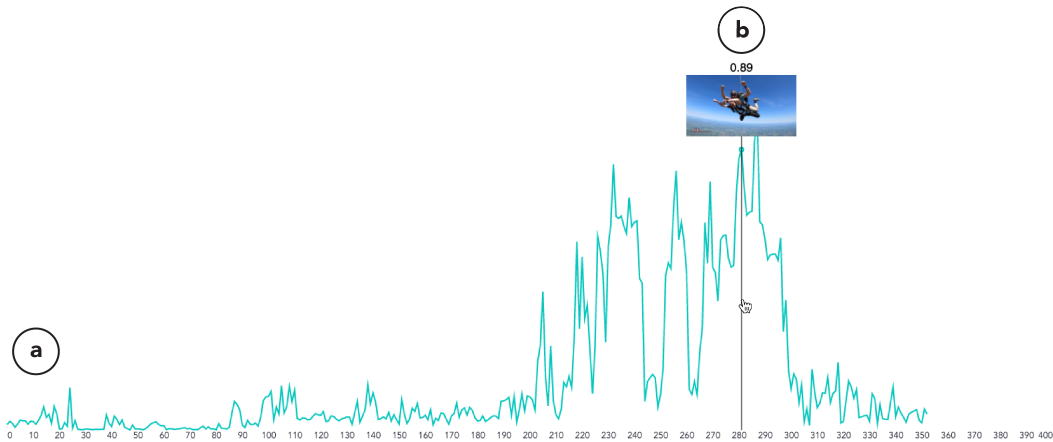


Figure 4: Highlight graph. The highlight graph visualizes the distribution of predicted highlight scores across the video (a). The user may scrub through the graph to inspect a corresponding video frame and its highlight score (b).

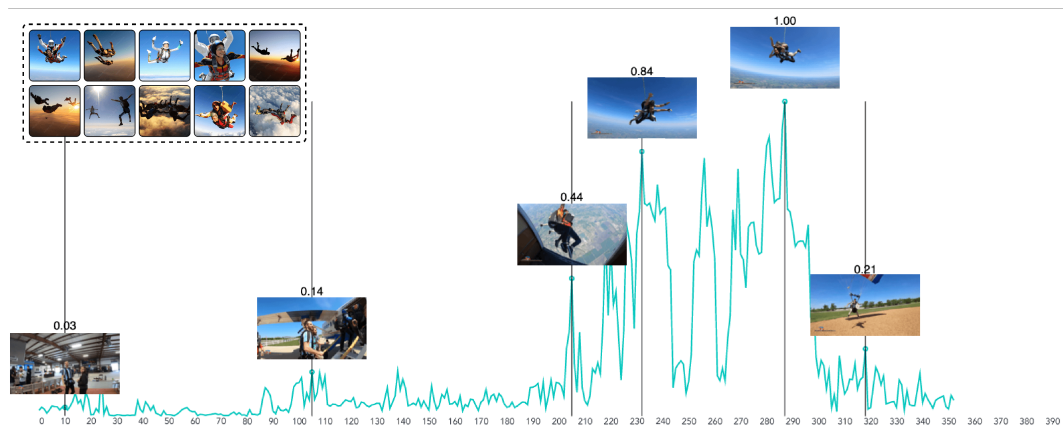


Figure 5: Example video frames and their corresponding highlight scores within a long skydiving video, using the keyword skydiving. The top-left corner displays the photograph collection used by Videogenic.

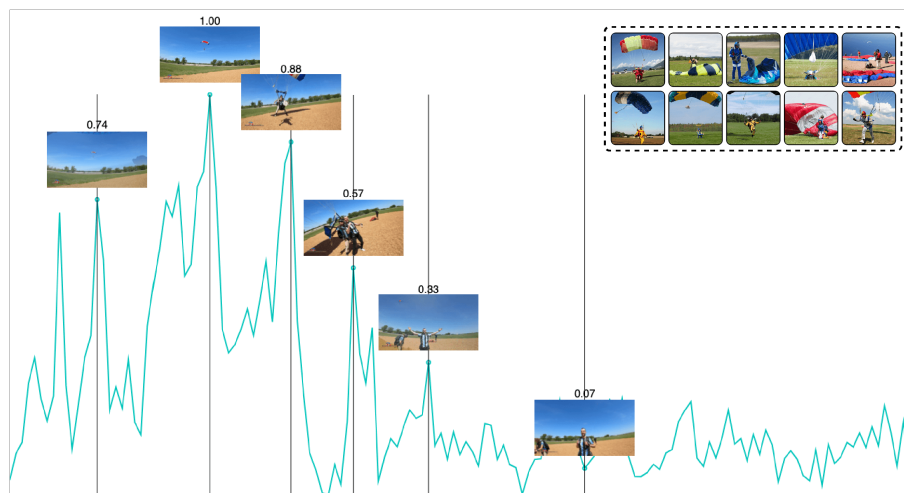


Figure 6: Example video frames and their corresponding highlight scores within a long skydiving video, using the keyword skydiving landing. The top-right corner displays the photograph collection used by Videogenic.

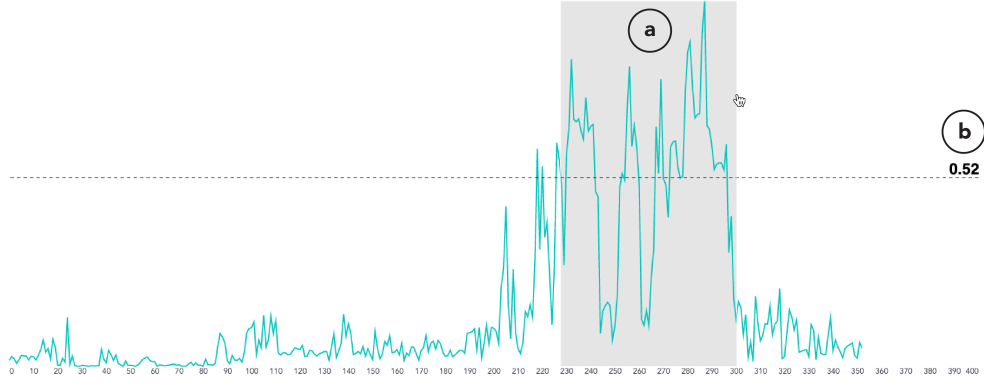


Figure 7: The user may brush through the highlight graph to select an interval of the video to use for the highlight video (a). The interface displays a dashed line and a text label to indicate the average highlight value of the selected interval (b).

each photograph through the CLIP image encoder (P). We then average all the photographs’ representations $P_1, \dots, P_{10}$ to create a representation for the average photograph ($\bar{P}$), which we use as the prior for judging the highlight scores of each video frame. Our intuition is that professional photographs capture the most highlight-worthy moments of an activity with skillful composition and framing. By exploiting the domain knowledge of photographers (Principle 2), we are not using any domain-specific priors (e.g., detecting people or faces), making our system scalable across diverse domains (Principle 1) through the use of new sets of photographs to compute the average photograph for different domains (e.g., nature photographs or wedding photographs). We empirically experimented with using different numbers of photographs. We found that using a single photograph sometimes introduces irrelevant attributes to the highlight. For example, the photograph is taken during a particular time of day, includes a particular background, or depicts a subject of a particular gender. We found that creating an *average photograph* from ten photographs successfully removes the effects of irrelevant attributes. Next, we encode the video frames through the CLIP image encoder (V). We then compare distance the average photograph representation with each encoded video frame via cosine similarities (Eq. 1). This gives us a vector of highlight scores (H) for each video frame. Finally, we normalize the highlight scores across the video on a scale of $[0, 1]$.

$$H = \bar{P} \cdot V^T \quad (1)$$

4.2.1 Highlight Graph. We provide users with an interface to visualize the distribution of highlight scores across the video (Principle 3) (Figure 4). This allows users to easily identify potential highlights in the video at a glance. We plot the highlight scores of each frame, where the y-axis represents the normalized highlight score and the x-axis represents the ID of each video frame in chronological order (Figure 4a). Analogous to the playhead of a video player, the user may scrub through the visualization to inspect the corresponding video frame thumbnail and highlight score (Figure 4b). In Figure 5, we show several example moments in a skydiving video and its highlight scores (with skydiving as the primary activity). We see that the moments of freefall have the highest scores, the moments

of jumping out of the plane and landing have moderate scores, and the moments of preparation and boarding the plane have low scores. By changing the keyword to skydiving landing, we show several example highlight moments based on a new set of images that were sampled to create the average image (Figure 6). Allowing simple changes of the keyword can allow users to explore different kinds of highlights within the same video.

4.3 Generating the Highlight Video

Our final step is generating the highlight video. We offer two methods for the user: (1) by a user-selected interval in the highlight visualization and (2) by automatically identifying an interval with high scores.

4.3.1 User Selection. The user may select an interval of the source video as their highlight video by brushing through the highlight visualization (Figure 7a). As the user brushes through the visualization, we display the average highlight score of the selected interval (Figure 7b). We then output the user-selected interval as the final highlight video.

4.3.2 Automatic Selection. We search for a continuous interval of length N (e.g., 10 seconds) within the video for an interval that has the maximum sum of highlight scores by sliding a window across the frame-wise highlight scores to find the maximum subarray.

5 HUMAN EVALUATION

To test the functionality of Videogenic, we run a human evaluation study on video highlights generated with Videogenic and a strong baseline method of CLIP similarity between text and video frames. The following outlines our experimental setup, procedure, and results.

5.1 Setup

5.1.1 Source Video Collection. We first collect our set of source videos that we would like to generate highlight videos for. To test the flexibility of Videogenic, we collect 16 diverse source videos from YouTube of (1) various *lengths* ranging from 30 seconds to 4 hours (mean=1,823 seconds, SD=3,584 seconds), (2) various *formats*

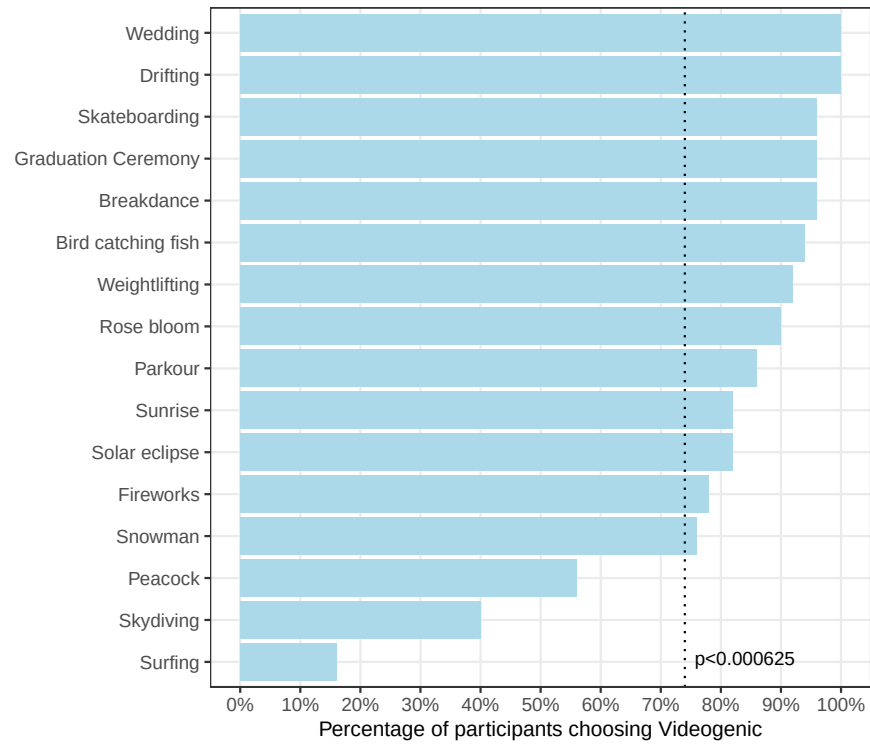


Figure 8: Human evaluation results ($N = 50$). The y-axis lists the videos in the evaluation study. The x-axis shows the percentage of participants who preferred Videogenic’s highlight over the baseline’s. The dashed line marks the point of statistical significance ($p < 0.000625$). The baseline is based on CLIP [31] text-video similarity.

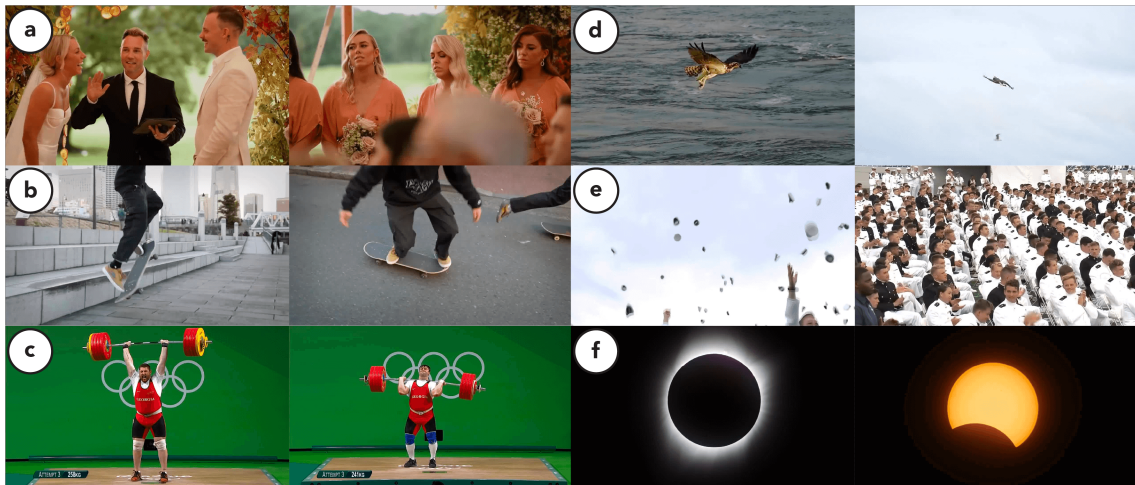


Figure 9: Example qualitative human evaluation results for wedding (a), skateboarding (b), weightlifting (c), bird hunting fish (d), graduation ceremony (e), and solar eclipse (f). For each pair, the left shows the highlight by Videogenic and the right shows the highlight by the baseline. Videogenic identifies the most remarkable moments with good composition and framing.

such as live broadcasts (e.g., drift racing live broadcast), timelapses (e.g., sunrise timelapse), vlogs (e.g., a day in the life of a surfer vlog), documentaries (e.g., peacock documentary), and unedited videos (e.g., unedited fireworks show) and (3) various *categories* such

as sports (e.g., skateboarding, weightlifting, parkour), events (e.g., wedding, graduation ceremony, building snowman), nature (e.g., sunrise, solar eclipse, rose bloom), and animals (e.g., bird hunting fish, peacock courtship).

5.1.2 Videogenic and Baseline Setup. For each collected video, we generate a highlight video for it with two systems: Videogenic and a baseline system.

Videogenic Setup. We automatically generate a highlight video for each collected video with Videogenic as detailed in Section 4.

Baseline Setup. We adopt CLIP similarity between *text* and video frames [8] as our baseline system. Specifically, rather than computing an average photograph representation to compare against video frames, the method compares the keyword against video frames directly via CLIP cosine similarities. Given a keyword, the method is to find semantically relevant video frames with strong performance. We make use of this method as our baseline for two reasons. First, like Videogenic, the method is able to perform in a zero-shot manner with no training required. This matches our key principle of creating a system that can be used *out-of-the-box* to generate video highlights for arbitrary videos (Principle 1). Second, a comparison between Videogenic and the method can help examine the usefulness of using *professional photographs* as our prior. To create our highlight video, we determine the primary activity keyword as shown in Figure 2. We then use the baseline method to compute scores for each video frame based on the keyword and take an interval with the maximum sum of scores as the highlight video.

5.2 Procedure

We run a human evaluation study on Prolific [6] to evaluate the highlight videos generated by Videogenic and the baseline system. We recruit 50 US-based participants with standard sampling and prescreen participants such that they must have experience in using TikTok, a short-form video platform, to ensure that participants are familiar with the concept of highlight videos. After receiving participants' consent, we ask each participant to compare the highlight videos generated with the two conditions for each of the 16 sample videos in a paired comparison, two-alternative forced choice manner [17]. The study takes approximately 3 minutes to complete and we compensate participants \$2 USD for their time.

5.3 Results

Overall, participants prefer the highlights by Videogenic over the baseline system for 14 out of 16 videos (preference determined by the majority) (Figure 8). We further analyze the results for statistical significance through a binomial test with Bonferroni correction (16 tests, significance level at $\alpha < \frac{0.01}{16} = 0.000625$). Participants significantly prefer the highlight videos generated with Videogenic (mean=80.00%, SD=9.45%, $p < 0.000625$) compared to the baseline.

Figure 9 shows qualitative examples of highlights identified by the Videogenic (left) versus the baseline (right). We see that, in general, the baseline method is able to correctly identify relevant content within diverse videos. For example, the skateboarding source video includes various irrelevant content such as the skateboarders eating, shopping, playing arcade games, and walking around the city. However, we see that Videogenic is able to more accurately identify the most remarkable highlight moments, given professional photographs as a prior. In the figure, Videogenic identifies the officiant address of the wedding, the skateboard kickflip, the weightlifter completing the clean and jerk, the bird carrying its

prey, the graduation hat toss, and the total solar eclipse. In addition, Videogenic inherits knowledge on good composition and framing from professional photographs (e.g., low-angle shot for skateboarding (Figure 9b) and close-up shot of the hunting bird (Figure 9d)).

6 EXPERT STUDY

We conduct a within-subjects expert study to evaluate the usefulness of Videogenic in helping video editors create highlight videos. The following outlines our study design, participants, procedure, and results. Our main research questions are:

- RQ1. How would participants' workload level be affected with the use of Videogenic?
- RQ2. How do participants' find the usability of Videogenic?
- RQ3. How would the use of Videogenic affect the participants' task completion time?
- RQ4. Qualitatively, what would participants see as the pros and cons of Videogenic?

6.1 Study Design

6.1.1 Independent Variable. The independent variable of the study is the system: Videogenic versus a baseline of manual editing. In the experimental condition, we ask participants to create a highlight video using Videogenic. In the baseline condition, we ask participants to create a highlight video using Adobe Premiere Pro, a standard video editing software that editors use to create highlight videos.

6.1.2 Dependent Variable. The dependent variables of the study are workload (RQ1) measured by the mental, temporal, effort, and frustration components of the NASA TLX questionnaire [19], usability (RQ2) measured by the System Usability Scale (SUS) questionnaire [13], and task completion time (RQ3) reported by the participant (in seconds). All questionnaire questions are represented on a 7-point Likert scale.

6.2 Participants

We recruit 12 professional video editors (4 female, 8 male) aged 18 to 48 (mean=28.75, SD=9.77) from Upwork [7], a platform for hiring freelancers. We conduct a background survey with the participants before each study to assess their video editing experience. Overall, participants have high self-rated familiarity with video editing (mean=6.25, SD=0.75) (7-point Likert scale) and have several years of experience (mean=6.71, SD=3.93). All participants regularly use Adobe Premiere Pro for video editing.

6.3 Procedure

We conduct the expert study remotely. After receiving the participant's consent, we collect information about individual backgrounds. We then ask the participant to create a highlight video from a source video with Videogenic and to create a highlight video from another source video by manually editing in Adobe Premiere Pro. We counterbalance both the order of the conditions and the order of the source videos. The source videos are of comparable difficulty, both being videos containing a large variety of complex scenes that depict a full-day vlog of an activity (i.e., skydiving and

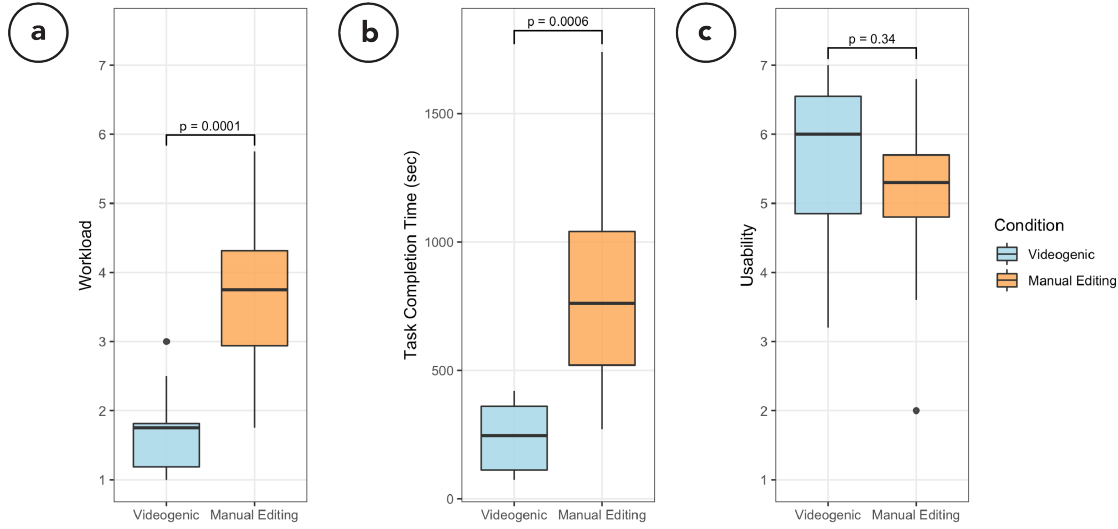


Figure 10: Expert study results ($N=12$). Boxplots from left to right: workload measured with NASA TLX [19] (7-point Likert scale, lower is better) (a), task completion time (seconds, lower is better) (b), and usability measured with SUS [13] (7-point Likert scale, higher is better) (c). The baseline is manual editing with Adobe Premiere Pro.

surfing). We also ask participants to record the time they spend in each condition by starting a timer after opening the application and stopping the timer after completing the highlight video. After each condition, we ask participants to complete the NASA TLX, SUS, and task completion time questionnaires. After the participant completes both conditions, we ask the participant to answer open-ended questions regarding the overall experience of using Videogenic. The study lasts for approximately 40 minutes. We compensate participants \$30 USD for their time.

6.4 Results and Discussion

For quantitative analysis, we analyze the scores for workload, usability, and task completion through paired t-tests. In addition, we recruit external raters to evaluate whether the final highlight videos created by the participants capture the highlight moments of activities and have good production quality. We analyze these external ratings through unpaired t-tests. Figure 10 shows an overview of the quantitative results comparing Videogenic against the manual editing baseline. For qualitative analysis, we analyze the participants' open-ended responses with deductive thematic analysis [12] according to the dimensions of the quantitative measurements (i.e., workload, task completion time, and usability).

6.4.1 Workload. The differences in workload per participant across conditions pass the Shapiro-Wilk test of normality ($W=0.95$, $p=0.60$). We thus compare the differences in workload through a parametric paired t-test. Participants report a significantly lower workload when using Videogenic (mean=1.69, $SD=0.61$) compared to the baseline (mean=3.67, $SD=1.20$) ($t(11)=5.84$, $p=0.0001$, $r=0.87$, $d_s=1.68$) (7-point Likert scale, lower is better) (Figure 10a). Participants state that “[Videogenic] found the most interesting moments for the clip (P11)” and that it “does away a lot of the bland and monotonous

editing ‘chores’ like having to scrub through a lot of fluff (P2)”. Participants enjoy the flexibility of multiple modes: “In auto mode, it is literally zero effort. In user selection mode, the fact that [Videogenic] gets you 90% there is pretty cool too. (P2)” Overall, participants feel that Videogenic helps reduce the tedious components of creating highlight videos. For example, in action footage such as skateboarding or surfing, the camera has to be constantly rolling to capture unexpected moments. However, going through the raw footage can be a tiresome chore. Videogenic could change the nature of editing work by allowing editors to dedicate more mental energy to the creative aspects of editing. For content creators, this could motivate them to create and share more video content.

6.4.2 Task Completion Time. The differences in task completion time per participant across conditions pass the Shapiro-Wilk test of normality ($W=0.93$, $p=0.36$). We thus compare the differences in task completion time through a parametric paired t-test. Participants report a significantly lower task completion time in seconds when using Videogenic (mean=247, $SD=132$) compared to the baseline (mean=856, $SD=476$) ($t(11)=4.73$, $p=0.0006$, $r=0.82$, $d_s=1.37$) (time taken in seconds, lower is better) (Figure 10b). Participants state that Videogenic helps “cut the time when it comes to searching for the right clips to use (P1)”: “With manual editing, I have to sort through a 5-minute video just for that 5 seconds of good footage. (P1)” Several participants mention how Videogenic is especially suitable for editing short videos: “Premiere and other software can be too clunky to make short videos (P8)” and “no matter how short the project is, [Premiere] always takes more time than I would like it to (P6)”. One participant also suggests a combination of the two systems: “I’d use [Videogenic] to detect highlight moments for further editing in Premiere (P10)”. Overall, participants feel that Videogenic significantly shortens the time it takes to create a highlight video, notably by

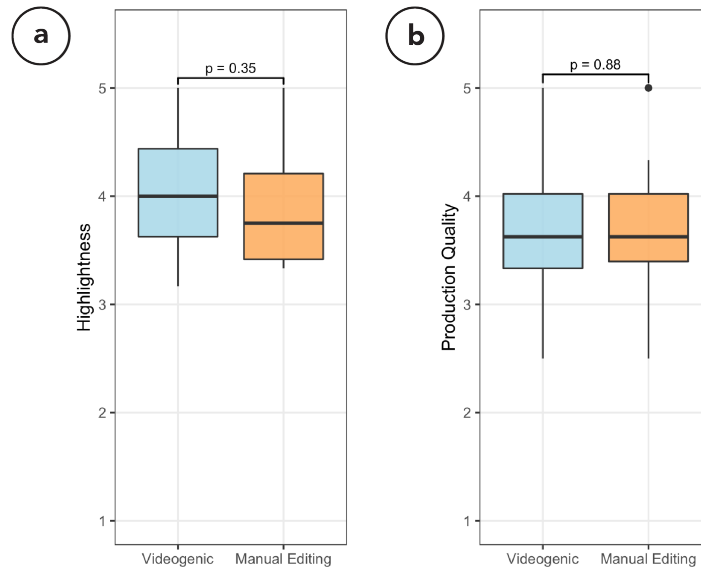


Figure 11: Highlightness and quality results ($N=20$). Boxplots from left to right: highlightness (5-point Likert scale, higher is better) (a), production quality (5-point Likert scale, higher is better) (b). The baseline is manual editing with Adobe Premiere Pro.

reducing the time spent on searching for highlight moments within large amounts of footage.

6.4.3 Usability. The differences in usability per participant across conditions pass the Shapiro-Wilk test of normality ($W=0.95$, $p=0.64$). We thus compare the differences in usability through a parametric paired t-test. Participants report a higher but not significantly higher usability when using Videogenic (mean=5.62, SD=1.34) compared to the baseline (mean=5.05, SD=1.29) ($t(11)=-0.99$, $p=0.34$) (7-point Likert scale, higher is better) (Figure 10c). The absence of statistical significance is unsurprising as the baseline (Adobe Premiere Pro) is a polished editing software that the participants are familiar with. Participants found Videogenic to be “a very simple feature that works well and is easy to navigate and use (P6)” and is “pretty fool-proof (P7)”: “I’ve never used the program before, but was still able to create the video I needed in under two minutes. That’s incredible. (P6)”.

Participants also comment on specific components of Videogenic: **Highlight Graph.** Participants enjoy scrubbing through the highlight graph “to see spots where there could be highlights in graph form” (P8): “It shows me the high and low points of the video and where to cut. (P1)” Participants also appreciate being able to see the data: “it made [Videogenic] feel sophisticated because of the data being shown (P7)”.

Overall, participants express that Videogenic is easy to use to create highlight videos. Given that many editors enjoy interacting with the highlight graph, it could potentially be useful for the highlight graph to be integrated within video editing programs.

6.4.4 Quality. We recruit external raters to evaluate the quality of the final highlight videos created by the editors using Videogenic and using the baseline.

Procedure. We recruit 20 US-based raters on Prolific [6] with standard sampling and prescreen participants such that they must have experience in using TikTok so that they are familiar with the concept of highlight videos. After receiving participants’ consent, we ask each participant to rate the quality of 24 highlight videos (12 created with Videogenic, 12 created with baseline) by answering how they feel about the following two statements on a scale of 1 to 5 (strongly disagree, disagree, neutral, agree, strongly agree):

- The video captures the highlight moments of <activity>.
- This is a well-made highlight video.

The study takes approximately 10 minutes to complete and we compensate participants \$2 USD for their time.

Results. Figure 11 shows an overview of the results. The differences per participant across conditions pass the Shapiro-Wilk test of normality for both “highlightness” ($W=0.98$, $p=0.93$) and production quality ($W=0.93$, $p=0.16$). We thus analyze the results for statistical significance through parametric unpaired t-tests. There is no significant statistical difference between the highlight videos created using Videogenic compared to the baseline (manual editing with Premiere) for both highlightness and production quality. Participants rate a slightly higher highlightness for the videos created using Videogenic (mean=4.06, SD=0.59) compared to the baseline (mean=3.89, SD=0.55) ($t(37.9)=0.95$, $p=0.35$) (5-point Likert scale, higher is better) (Figure 11a). Participants rate a similar production quality for the videos created using Videogenic (mean=3.70, SD=0.66) compared to the baseline (mean=3.73, SD=0.60) ($t(37.7)=-0.15$, $p=0.88$) (5-point Likert scale, higher is better) (Figure 11b). Overall, raters feel that both videos created by Videogenic and by human editors capture the highlight moments of activities and are well-made as rated on our 5-point scale.

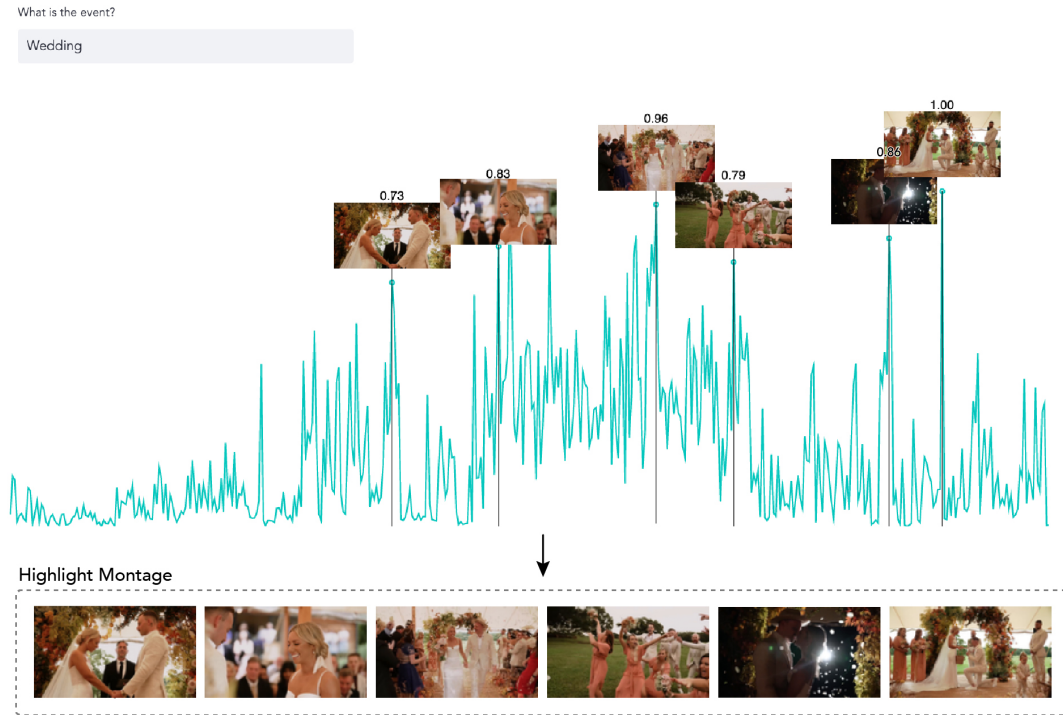


Figure 12: Videogenic can be extended to create a highlight moments montage by combining multiple local maxima highlight moments.

7 EXTENDED APPLICATIONS

We implemented two extended applications to explore video applications that could use Videogenic as a building block.

7.1 Highlight Moments Montage

A highlight moments montage features a sequence of many highlight moments from an event. Instead of identifying one key highlight moment, we may extend Videogenic to identify multiple highlight moments by computing several local maxima. Figure 12 shows how several different local highlight moments from a wedding video can be combined to create a wedding highlight moments montage.

7.2 Personalized Highlight Video

Videogenic can be easily adapted to support personalized highlight videos. The user may upload their own photographs to create a custom photography database, powering a personalized Videogenic. Figure 13 shows how a photography database containing a particular skateboarding trick can help create a personalized highlight video featuring this trick.

8 LIMITATIONS AND FUTURE WORK

While Videogenic was positively received in our user studies, there are several avenues for improvement that we plan to address for future work. First, Videogenic determines highlights from the visual domain. Thus, Videogenic is not designed for specific categories of audio-oriented videos with few visual concepts, such as recordings

of podcasts or interview videos. In addition, further work on gaining a deeper understanding of the target domain may be required to extend Videogenic for videos involving complex subtasks, such as how-to videos [14, 15, 38]. Second, as Videogenic uses photographs as its prior, it currently does not factor in motion information from videos. We hypothesize this to be the main reason why the highlight videos automatically generated with Videogenic for surfing and skydiving received lower preference (Figure 8), as they were more static shots. While this may be addressed through human-in-the-loop selection of highlight segments (Section 4.3), we may extend Videogenic to additionally leverage *professional stock videos* as a prior. We could represent motion information using video motion extraction methods such as optical flow [25]. Third, Videogenic uses a small collection of stock photographs to compute highlight scores. There may be cases where novel activities do not have readily available professional photographs. As the number of photographs required by Videogenic is small (10 photographs), we could allow users to create their own personalized examples (Section 7.2). A personalized example set could be an interesting form of control for users to steer the direction of the highlight moments from a handful of personal examples [26]. Finally, Videogenic currently takes in an average photo representation from a collection of photographs. While we found this to generally perform well through extensive human evaluations, additional experiments could be done, such as ablating different numbers of photographs for different video domains or having multiple clustered averages instead of a single average (e.g., photograph clusters of handshakes, crowd shots, and

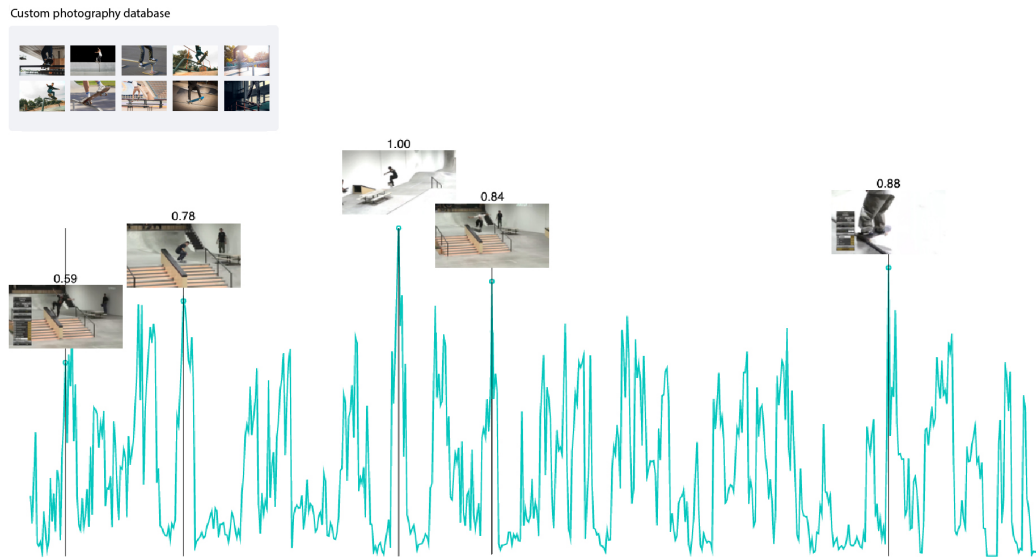


Figure 13: Videogenic can be extended to create a personalized highlight video by creating a custom photography database.

thrown hats for a “graduation” highlight video). This could also enable greater a level of customization of highlight videos.

9 CONCLUSION

In this paper, we present Videogenic, a simple yet very effective technique for creating highlight videos. Our key insight is that professional photographs tend to capture the most remarkable moments of a given activity. We conduct a human evaluation study ($N=50$), showing that a set of high-quality photographs combined with encodings of CLIP can act as a strong prior for extracting domain-specific highlights for videos encompassing a diverse range of domains. We further evaluate the usefulness of Videogenic for video editors through a within-subjects expert study ($N=12$) comparing Videogenic to a baseline (Adobe Premiere Pro), demonstrating decreased workload, decreased task completion time, and increased usability. External raters rated high “highlightness” and production quality for the highlight videos created by editors with Videogenic. This work takes a step towards *out-of-the-box*, domain-agnostic highlight video generation by building on the domain knowledge of photographers. In recent years, we see growth in long-form video content (e.g., livestreaming [5]) as well as a proliferation of video capturing devices (e.g., smartphones and action cameras [4]). On the other hand, we see a rapid surge in popularity in short-form video consumption (e.g., TikTok, Instagram Reels, and YouTube Shorts). We hope Videogenic can help to bridge this gap by lowering the barrier required to convert long-form videos into engaging short-form highlights.

ACKNOWLEDGMENTS

We would like to thank colleagues in the Augmented Design Capability Studio for providing valuable feedback on our system.

REFERENCES

- [1] 2021. *Apple announces the 2021 App Store Award winners and most downloaded apps of the year*. Retrieved August 15, 2022 from <https://techcrunch.com/2021/12/02/apple-announces-the-2021-app-store-award-winners-and-most-downloaded-apps-of-the-year/>
- [2] 2022. *FlashCut Auto Editing - ONE R Support*. Retrieved August 15, 2022 from <https://onlinemanual.insta360.com/oner/en-us/editing/flashcut>
- [3] 2022. *HERO10 Black 5.3K Video 23MP Action Camera Bundle | GoPro*. Retrieved August 15, 2022 from <https://gopro.com/en/us/shop/cameras/hero10-black/CHDX-101-master.html>
- [4] 2022. *Insta360 ONE RS – Waterproof Action Camera + 360 Camera in One*. Retrieved August 15, 2022 from <https://www.insta360.com/product/insta360-oners>
- [5] 2022. *Live Streaming Market Worth \$4.26 Billion by 2028*. Retrieved August 15, 2022 from <https://www.bloomberg.com/press-releases/2022-05-05/live-streaming-market-worth-4-26-billion-by-2028-market-size-share-forecasts-trends-analysis-report-with-covid-19-impact>
- [6] 2022. *Prolific*. Retrieved August 15, 2022 from <https://www.prolific.co/>
- [7] 2022. *Upwork*. Retrieved August 15, 2022 from <https://www.upwork.com/>
- [8] 2022. *Which Frame?* Retrieved August 15, 2022 from <http://whichframe/>
- [9] 2022. *YouTube Help - Video previews*. Retrieved August 15, 2022 from <https://support.google.com/youtube/answer/7074781?hl=en>
- [10] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–13.
- [11] Michael S Bernstein, Joel Brandt, Robert C Miller, and David R Karger. 2011. Crowds in two seconds: Enabling realtime crowd-powered interfaces. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*. 33–42.
- [12] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.
- [13] John Brooke et al. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.
- [14] 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.
- [15] Pei-Yu Chi, Joyce Liu, Jason Linder, Mira Dontcheva, Wilmo Li, and Bjoern Hartmann. 2013. Democut: generating concise instructional videos for physical demonstrations. In *Proceedings of the 26th annual ACM symposium on User interface software and technology*. 141–150.
- [16] Wen-Sheng Chu, Yale Song, and Alejandro Jaimes. 2015. Video co-summarization: Video summarization by visual co-occurrence. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3584–3592.
- [17] Gustav Theodor Fechner. 1860. *Elemente der psychophysik*. Vol. 2. Breitkopf u. Härtel.

- [18] Michael Gygli, Yale Song, and Liangliang Cao. 2016. Video2gif: Automatic generation of animated gifs from video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1001–1009.
- [19] 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.
- [20] Aditya Khosla, Raffay Hamid, Chih-Jen Lin, and Neel Sundaresan. 2013. Large-scale video summarization using web-image priors. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2698–2705.
- [21] Hoseong Kim, Tao Mei, Hyeran Byun, and Ting Yao. 2018. Exploiting web images for video highlight detection with triplet deep ranking. *IEEE Transactions on Multimedia* 20, 9 (2018), 2415–2426.
- [22] Maheshkumar H Kolekar and Somnath Sengupta. 2006. Event-importance based customized and automatic cricket highlight generation. In *2006 IEEE International Conference on Multimedia and Expo*. IEEE, 1617–1620.
- [23] Josua Krause, Adam Perer, and Kenney Ng. 2016. Interacting with predictions: Visual inspection of black-box machine learning models. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 5686–5697.
- [24] Yong Jae Lee, Joydeep Ghosh, and Kristen Grauman. 2012. Discovering important people and objects for egocentric video summarization. In *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 1346–1353.
- [25] Chuan-en Lin. 2020. Introduction to motion estimation with optical flow. *can be found under <https://nanonets.com/blog/optical-flow>* (2020).
- [26] David Chuan-En Lin and Nikolas Martelaro. 2021. Learning Personal Style from Few Examples. In *Designing Interactive Systems Conference 2021*. 1566–1578.
- [27] Zheng Lu and Kristen Grauman. 2013. Story-driven summarization for egocentric video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2714–2721.
- [28] 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.
- [29] 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.
- [30] Surya Nepal, Uma Srinivasan, and Graham Reynolds. 2001. Automatic detection of 'Goal' segments in basketball videos. In *Proceedings of the ninth ACM international conference on Multimedia*. 261–269.
- [31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. PMLR, 8748–8763.
- [32] Yong Rui, Anoop Gupta, and Alex Acero. 2000. Automatically extracting highlights for TV baseball programs. In *Proceedings of the eighth ACM international conference on Multimedia*. 105–115.
- [33] Jose San Pedro, Vaiva Kalnikaite, and Steve Whittaker. 2009. You can play that again: exploring social redundancy to derive highlight regions in videos. In *Proceedings of the 14th international conference on Intelligent user interfaces*. 469–474.
- [34] Yale Song, Miriam Redi, Jordi Vallmitjana, and Alejandro Jaimes. 2016. To click or not to click: Automatic selection of beautiful thumbnails from videos. In *Proceedings of the 25th ACM international on conference on information and knowledge management*. 659–668.
- [35] Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. 2015. Tvsum: Summarizing web videos using titles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5179–5187.
- [36] Min Sun, Ali Farhadi, and Steve Seitz. 2014. Ranking domain-specific highlights by analyzing edited videos. In *European conference on computer vision*. Springer, 787–802.
- [37] Zhida Sun, Mingfei Sun, Nan Cao, and Xiaojuan Ma. 2016. VideoForest: interactive visual summarization of video streams based on danmu data. In *SIGGRAPH ASIA 2016 symposium on visualization*. 1–8.
- [38] Anh Truong, Peggy Chi, David Salesin, Irfan Essa, and Maneesh Agrawala. 2021. Automatic generation of two-level hierarchical tutorials from instructional makeup videos. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.
- [39] Shao-Yu Wu, Ruck Thawonmas, and Kuan-Ta Chen. 2011. Video summarization via crowdsourcing. In *CHI '11 Extended Abstracts on Human Factors in Computing Systems*. 1531–1536.
- [40] Bo Xiong, Yannis Kalantidis, Deepti Ghadiyaram, and Kristen Grauman. 2019. Less is more: Learning highlight detection from video duration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 1258–1267.
- [41] Saelyne Yang, Jisu Yim, Juho Kim, and Hujung Valentina Shin. 2022. CatchLive: real-time summarization of live streams with stream content and interaction data. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–20.
- [42] Ting Yao, Tao Mei, and Yong Rui. 2016. Highlight detection with pairwise deep ranking for first-person video summarization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 982–990.
- [43] Dennis Yow, Boon-Lock Yeo, Minerva Yeung, and Bede Liu. 1995. Analysis and presentation of soccer highlights from digital video. In *proc. ACCV*, Vol. 95. Citeseer, 11–20.
- [44] JD Zamfirescu-Pereira, Richmond Y Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny can't prompt: how non-AI experts try (and fail) to design LLM prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–21.

A EXAMPLE HIGHLIGHT GRAPHS

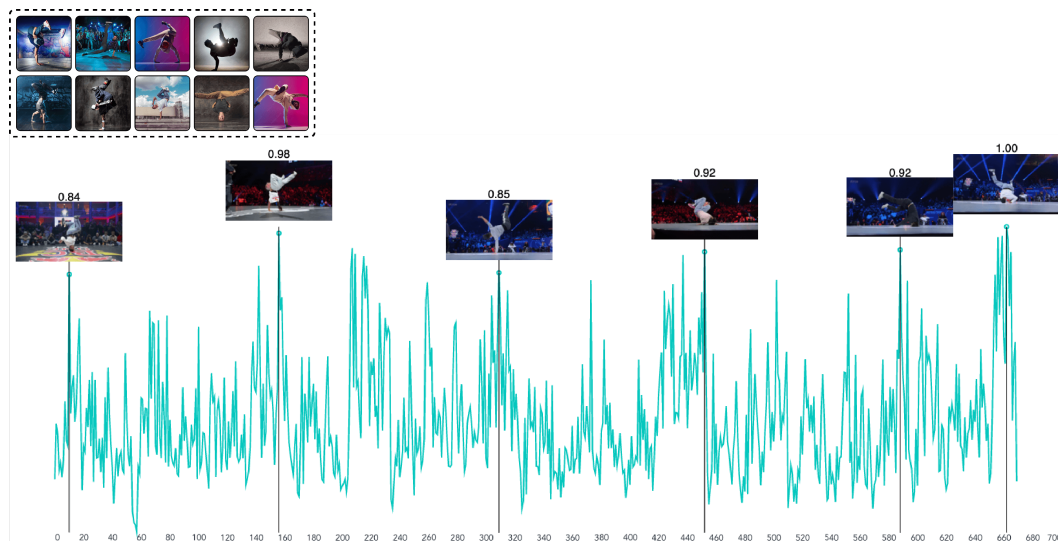


Figure 14: Example highlights from a breakdance competition video. The keyword is breakdance. The photo collection used by Videogenic is shown on the top-left. Videogenic identifies the iconic power moves.

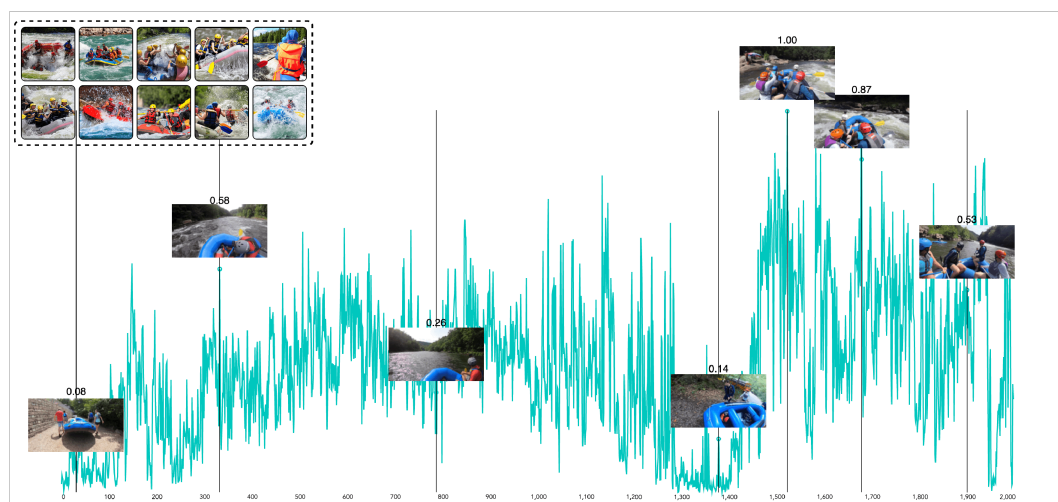


Figure 15: Example video frames and highlight scores within around 30 minutes video footage from a rafting trip. The video clips are recorded by one of the authors using an action camera. The keyword is rafting. The photo collection used by Videogenic is shown on the top-left. We see that Videogenic scores the whitewater moments (i.e., raft going through the river rapids) more highly.

B COMPARISON AGAINST CLIP WITH DESCRIPTIVE TEXT PROMPTS

In Section 5, we compare Videogenic against a CLIP text-image baseline prompted with the topic of the highlight video (e.g., “skateboarding”). In this section, we examine prompting the baseline with a variety of more descriptive prompts, even including prompts that contain domain-specific knowledge (e.g., “kickflip”). Table 1 shows several prompts and retrieved highlight moments. We show more prompts and retrieved highlight moments in Tables 2 and 3. We observe that Videogenic still produces higher quality results than CLIP prompted with descriptive text prompts.

We discuss the findings using results shown in Table 2. First, we test prompts that explicitly describe the task of finding highlights (e.g., “skateboarding highlight” and “photogenic skateboarding”). We also test prompts that implicitly leverage highlight-related cues (e.g., “skateboarding tiktok” and “professional photograph of skateboarding”). We see that these prompts generally fail to retrieve high-quality highlight moments. Second, assuming that the user has domain knowledge of the activity, we test prompts that include domain-specific keywords (e.g., “skateboarding focus on the highlight trick”, “skateboarding flip in the air”, and “kickflip”). We see that while these prompts help identify impressive skateboarding moments (i.e., the trick), they are often aesthetically unpleasing in terms of composition and framing. Third, we prompt for an aesthetic shot (e.g., “aesthetic skateboarding highlight shot”). We take a step further, assuming that the user has domain knowledge of what shots are most aesthetically pleasing (e.g., “skateboarding close up shot” and “skateboarding low angle shot”). We see that these prompts capture aesthetic but less interesting shots. Finally, we test prompts that include domain knowledge of moments and shot composition (e.g., “close up shot of skateboarding flip in the air” and “low angle shot of a kickflip”). We see that these prompts still produce subpar highlight moments compared to Videogenic. Overall, we can conclude that prompting for good highlight moments is challenging. This aligns with prior research findings of how people generally struggle with crafting good prompts [44]. We see that even descriptive text prompts containing rich domain knowledge such as impressive actions and shot specifications are far less capable of identifying highlight moments, compared with leveraging the visual priors encoded within professional photographs.

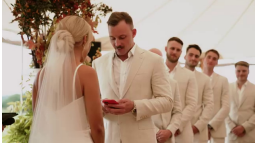
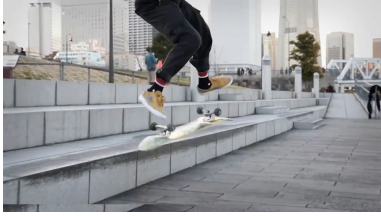
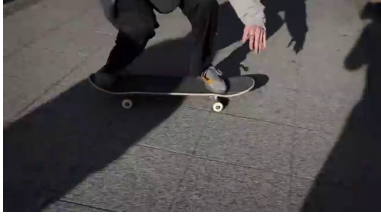
Videogenic	“professional photograph of ...”	Descriptive prompt	Domain knowledge prompt
<professional photographs> 	“professional photograph of skateboarding” 	“skateboarding flip in the air” 	“kickflip” 
<professional photographs> 	“professional photograph of a wedding” 	“wedding focusing on the couple” 	“officiant address” 

Table 1: Highlight moments identified using Videogenic versus CLIP prompted with a descriptive prompt, a domain knowledge prompt (e.g., the name of a skateboard trick), and “professional photograph of <domain>”.

Query	Highlight moment
Videogenic	
skateboarding	
skateboarding highlight	
photogenic skateboarding	
skateboarding tiktok	
professional photograph of skateboarding	
skateboarding focusing on the highlight trick	

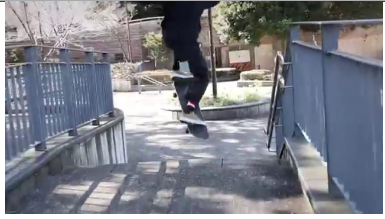
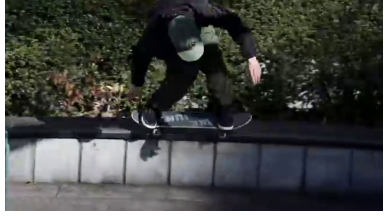
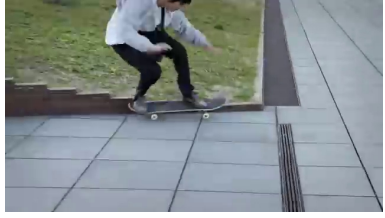
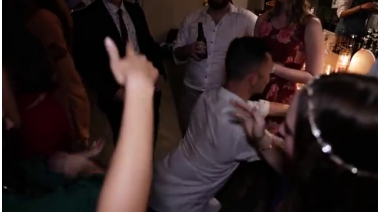
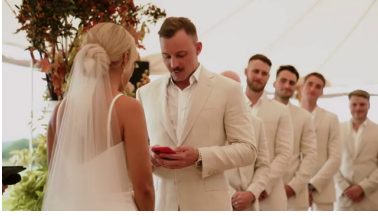
skateboarding flip in the air	
kickflip	
aesthetic skateboarding highlight shot	
skateboarding close up shot	
skateboarding low angle shot	
close up shot of skateboarding flip in the air	
low angle shot of a kickflip	

Table 2: Skateboarding highlight moments identified using Videogenic versus CLIP prompted with various text prompts.

Query	Highlight moment
Videogenic	
wedding	
wedding highlight	
photogenic wedding	
wedding highlight reel	
professional photograph of a wedding	
wedding focusing on the couple	

officiant address	
wedding couple vow	
aesthetic wedding shot	
wedding close up shot	
wedding symmetric shot	
close up shot of couple vow	
symmetric shot of the couple	

Table 3: Wedding highlight moments identified using Videogenic versus CLIP prompted with various text prompts.