Videogenic: Video Highlight Generation via Photogenic Moments

David Chuan-En Lin[1], Fabian Caba Heilbron[2], Joon-Young Lee[2], Oliver Wang[2], Nikolas Martelaro[1]
[1]Carnegie Mellon University, [2]Adobe Research
[1]chuanenl@cs.cmu.edu, nikmart@cmu.edu,
[2]{caba, jolee, owang}@adobe.com

1 Introduction

> *"Photography is the simultaneous recognition, in a fraction of a second, of the significance of an event."*
>
> — Henri Cartier-Bresson, Photographer

This paper investigates the challenge of extracting highlight moments from videos. To perform this task, a system needs to understand what constitutes a "highlight" for a video domain (e.g. a skateboard trick) while at the same time being able to scale across different domains. Prior works have largely explored making use of domain-specific features (e.g. presence of people [7] or goal scores in sports [11]) or expensive training of models with large-scale data (e.g. videos with labeled highlight/non-highlight segments [10], pairs of highlight videos and source videos [8], or proxy labels such as short user-generated videos [9]). In this work, our key insight is that photographs taken by photographers tend to capture the most remarkable or *photogenic* moments of an activity with great composition and framing. Based on this insight, we present Videogenic, a system capable of creating domain-specific highlight videos for a wide range of domains. In a human evaluation study (N=50), we show that a high-quality photograph collection combined with semantic encodings of CLIP, a neural network with semantic knowledge of images, can serve as an excellent prior for finding video highlights in arbitrary video domains with no additional training necessary. In an expert study (N=10), we demonstrate Videogenic's usefulness in helping video editors create highlight videos with lighter workload, shorter task completion time, and better usability. We encourage you to have a look at example results of Videogenic at https://humanvideointeraction.github.io/videogenic.

2 Method

We first ask the user to type in a highlight prompt (e.g. `skydiving`) to specify the theme of the highlight video. Given a prompt, we search on a database of professional photography (e.g. Adobe Stock) for 10 professional photographs depicting the activity (Figure 2). We encode each photograph through the CLIP image encoder ($\mathbf{P}$). We then average all the photographs' representations $\mathbf{P_{1,\ldots,10}}$ to create a representation for the average photograph ($\bar{\mathbf{P}}$), 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. We found that creating an *average photograph* as opposed to using a single photograph reduces the effects of irrelevant attributes within individual photographs, such as having a particular background or depicting a particular gender. Next, we encode the video frames through the CLIP image encoder ($\mathbf{V}$). We then compare distances of the average photograph representation with each encoded frame via cosine similarities. This gives us a vector of highlight scores ($\mathbf{H} = \bar{\mathbf{P}} \cdot \mathbf{V}^\top$) for each video frame.

Preprint. Under review.

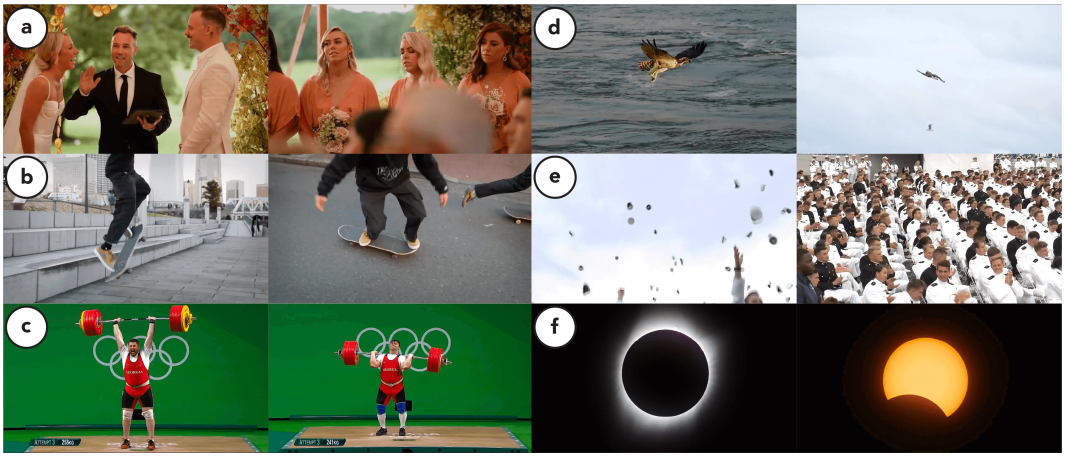

Figure 1: Qualitative highlight results by Videogenic (left) and the baseline (right) for wedding (a), skateboarding (b), weightlifting (c), bird hunting fish (d), graduation ceremony (e), and solar eclipse (f). Video sources: https://humanvideointeraction.github.io/videogenic#results.

We normalize the highlight scores across the video on a scale of $[0, 1]$. Finally, we slide a window across the frame-wise scores for a continuous N-length interval (e.g. 3 seconds) with maximum sum.

Interface. We provide users with an interface to visualize the distribution of highlight scores across the video (Figure 3a). The user may scrub through the interface to inspect the corresponding video frame thumbnail and highlight score (Figure 3b). In Figure 8, we show several example moments in a skydiving video (`skydiving` as the highlight prompt). 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 highlight prompt 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 9). Allowing simple changes of the highlight prompt can allow users to explore and personalize different kinds of highlights within the same video.

3 Evaluation

Human Evaluation Study. To evaluate the performance of Videogenic, we run a human evaluation study with 50 participants recruited from Prolific [3]. We ask participants to perform paired comparisons between highlight videos generated with Videogenic versus a baseline method of CLIP text-video similarity. We generate highlight videos using both conditions for 16 source videos collected from YouTube covering various *lengths* (e.g. 30 seconds to 4 hours), *formats* (e.g. live broadcasts, vlogs, unedited videos) and *categories* (e.g. sports, events, nature). Overall, participants prefer the highlight videos generated with Videogenic on average 80.00% of the time (Figure 10). Figure 1 shows qualitative examples of highlights by Videogenic (left) versus the baseline (right).

Expert Study. To evaluate the usefulness of Videogenic for video editors, we run an expert study with 12 professional video editors recruited from Upwork [4]. We ask participants to create highlight videos with Videogenic versus a baseline method of manual editing with Adobe Premiere Pro. We measure each participant's workload using the NASA Task Load Index [6], task completion time in seconds, and usability using the System Usability Scale [5] for both conditions. Participants report lower workload, lower task completion time, and higher usability when using Videogenic (Figure 11).

4 Conclusion

This paper takes a step towards general-purpose highlight video generation by building on the domain knowledge of photographers. In recent years, we see a growth in long-form video content (e.g. livestreaming [2]) as well as a proliferation of video capturing devices (e.g. smartphones to action cameras [1]). 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.

Ethical Implications

The introduction of Videogenic into the video editing process may also come with potential ethical implications. One example is bias in photographs. For example, a search of "wedding" photographs may depict predominately opposite-sex marriages. Such biases need to be surfaced to users.

References

[1] Insta360 ONE RS – Waterproof Action Camera + 360 Camera in One. URL https://www.insta360.com/product/insta360-oners.

[2] Live Streaming Market Worth $4.26 Billion by 2028. URL 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.

[3] Prolific. URL https://www.prolific.co/.

[4] Upwork. URL https://www.upwork.com/.

[5] John Brooke et al. Sus-a quick and dirty usability scale. *Usability evaluation in industry*, 189 (194):4–7, 1996.

[6] Sandra G Hart and Lowell E Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology*, volume 52, pages 139–183. Elsevier, 1988.

[7] Yong Jae Lee, Joydeep Ghosh, and Kristen Grauman. Discovering important people and objects for egocentric video summarization. In *2012 IEEE conference on computer vision and pattern recognition*, pages 1346–1353. IEEE, 2012.

[8] Min Sun, Ali Farhadi, and Steve Seitz. Ranking domain-specific highlights by analyzing edited videos. In *European conference on computer vision*, pages 787–802. Springer, 2014.

[9] Bo Xiong, Yannis Kalantidis, Deepti Ghadiyaram, and Kristen Grauman. Less is more: Learning highlight detection from video duration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1258–1267, 2019.

[10] Ting Yao, Tao Mei, and Yong Rui. Highlight detection with pairwise deep ranking for first-person video summarization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 982–990, 2016.

[11] Dennis Yow, Boon-Lock Yeo, Minerva Yeung, and Bede Liu. Analysis and presentation of soccer highlights from digital video. In *proc. ACCV*, volume 95, pages 11–20. Citeseer, 1995.

A System Figures

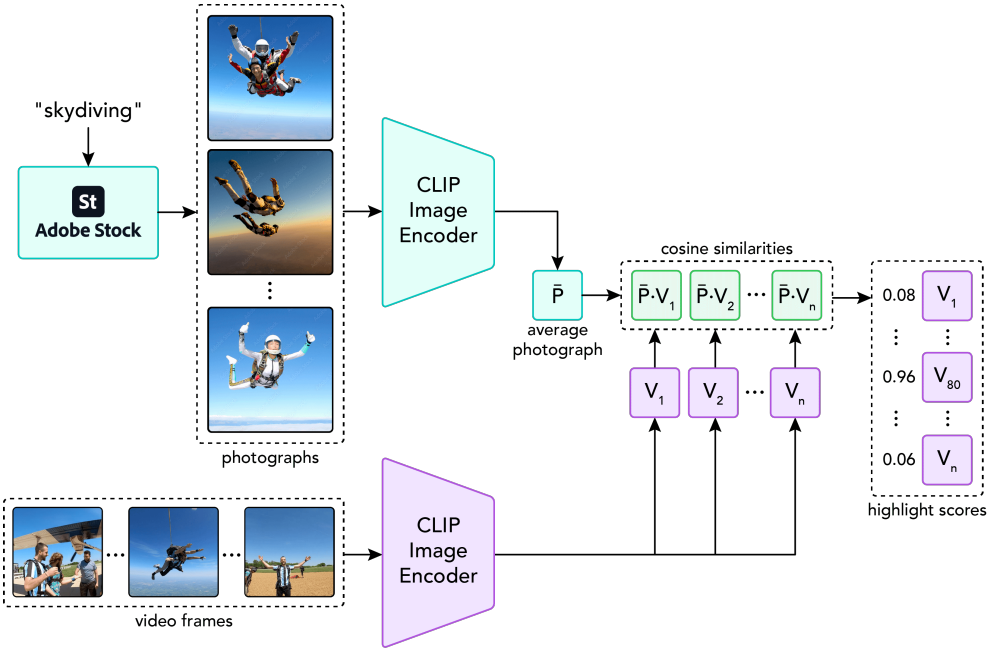

Figure 2: Given an activity label (e.g. "skydiving"), Videogenic retrieves 10 stock photographs and computes the average photograph representation. Given each frame of a video and the average photograph, Videogenic performs pairwise comparisons to predict a highlight score for each frame.

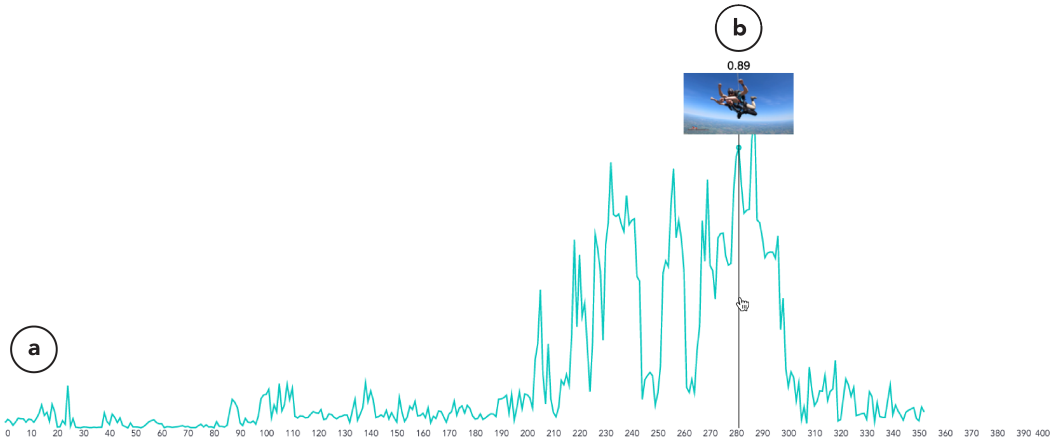

Figure 3: The highlight graph visualizes the distribution of predicted highlight scores across the video (a). The user may scrub through the graph to inspect a corresponding video frame and its highlight score (b).

B Qualitative Examples

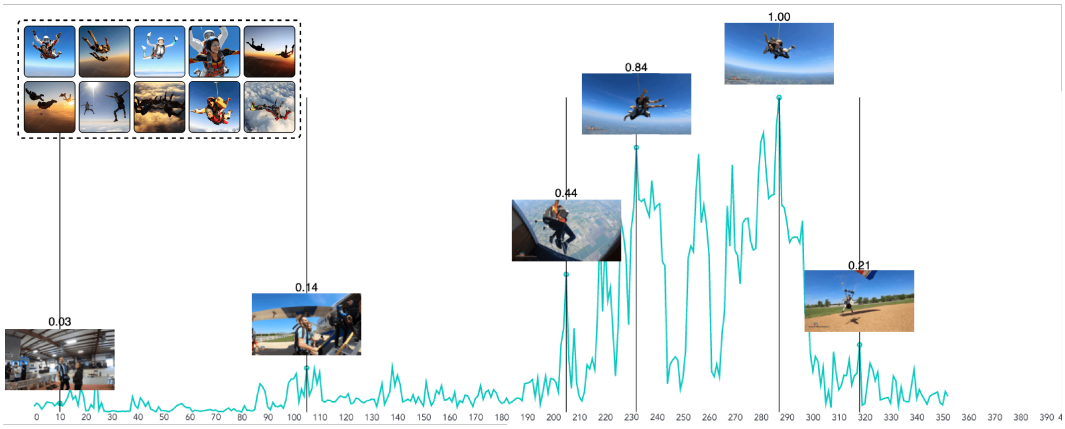

Figure 4: Example video frames and highlight scores from a skydiving vlog. The highlight prompt is `skydiving`. The photo collection used by Videogenic is shown on the top-left.

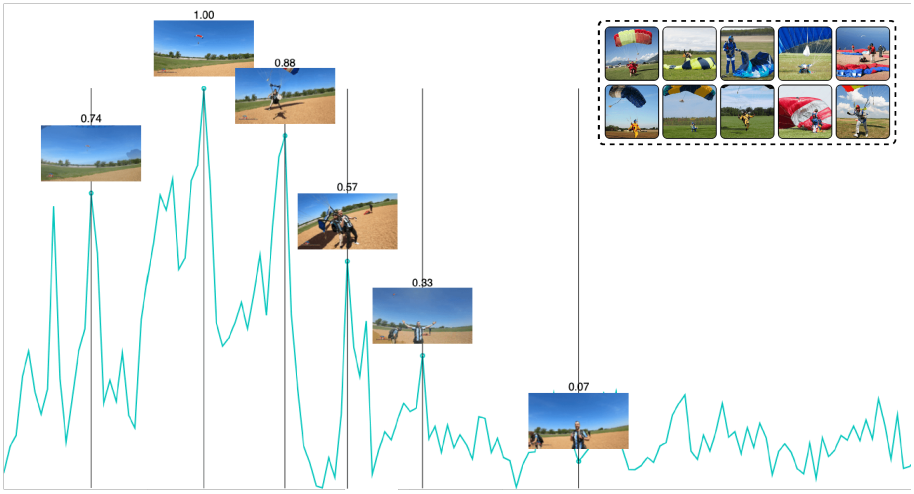

Figure 5: Example video frames and highlight scores from a skydiving vlog. The highlight prompt is `skydiving landing`. The photo collection used by Videogenic is shown on the top-right.

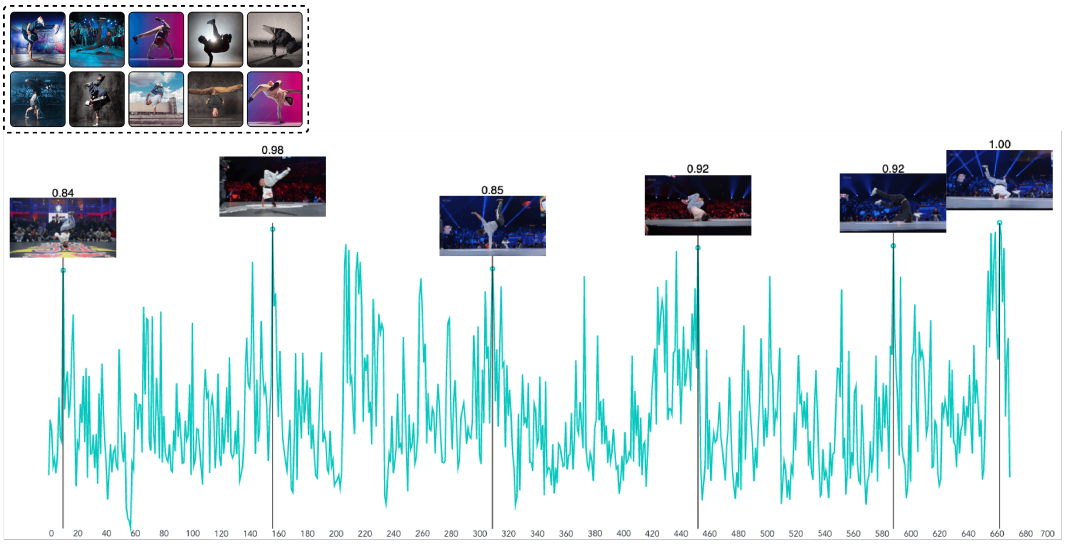

Figure 6: Example highlights from a breakdance competition video. The highlight prompt is `breakdance`. The photo collection used by Videogenic is shown on the top-left. Videogenic identifies the iconic power moves.

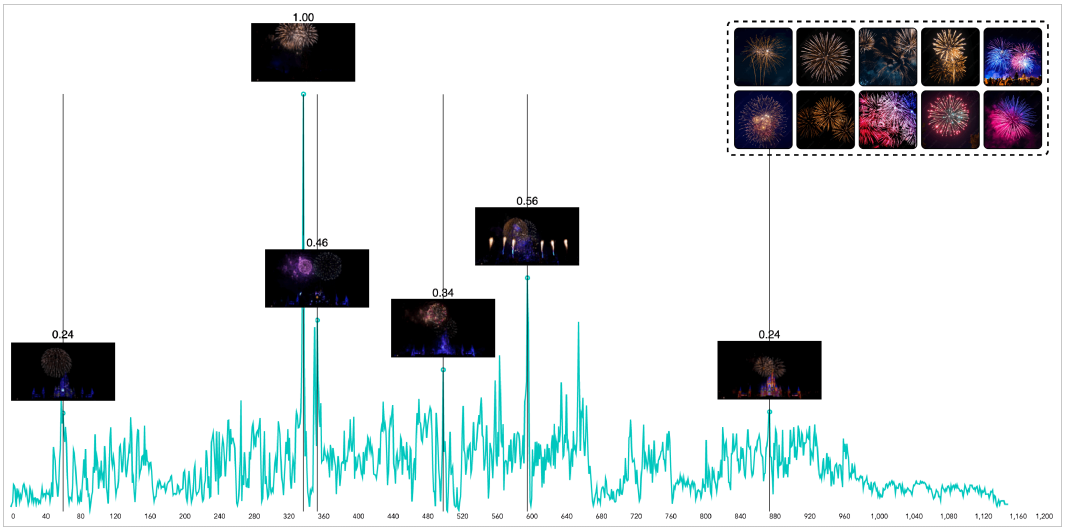

Figure 7: Example highlights from a recording of a fireworks show. The highlight prompt is `fireworks`. The photo collection used by Videogenic is shown on the top-right. Videogenic identifies when the fireworks are in full bloom.

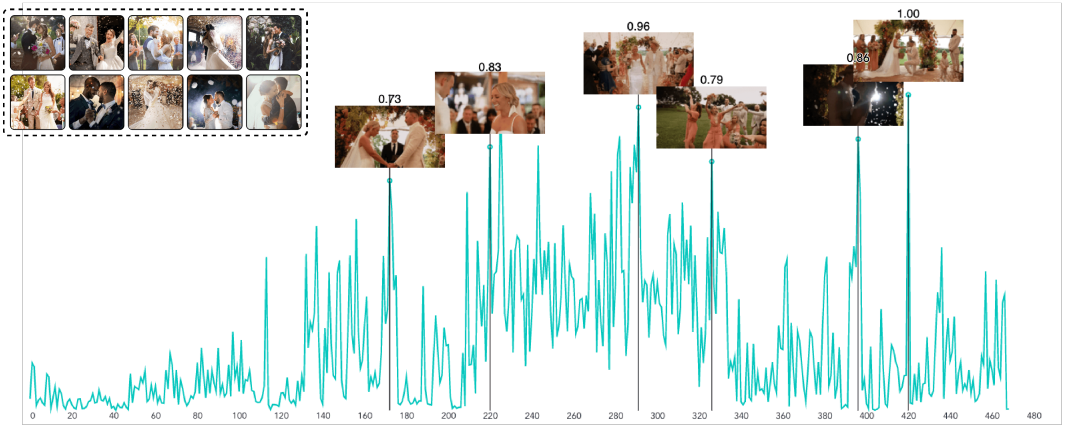

Figure 8: Example higlights from a wedding video. The highlight prompt is `wedding`. The photo collection used by Videogenic is shown on the top-left. Videogenic identifies the highlight moments of the couple.

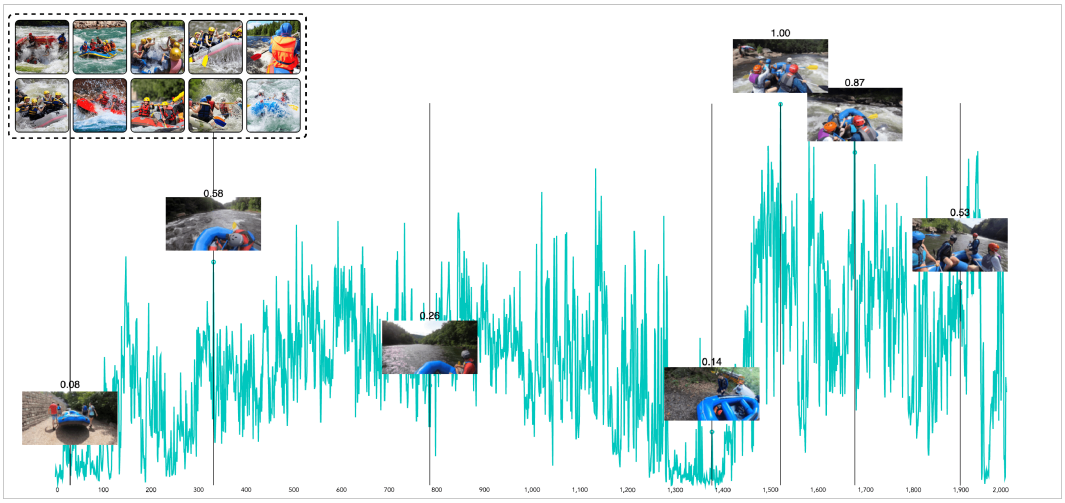

Figure 9: Example video frames and highlight scores within around 30 minutes video footage from a rafting trip. The video clips are recorded by one of the authors using an action camera. The highlight prompt is `rafting`. The photo collection used by Videogenic is shown on the top-left. We see that Videogenic scores the whitewater moments (i.e. raft going through the river rapids) more highly.

C Quantitative Results

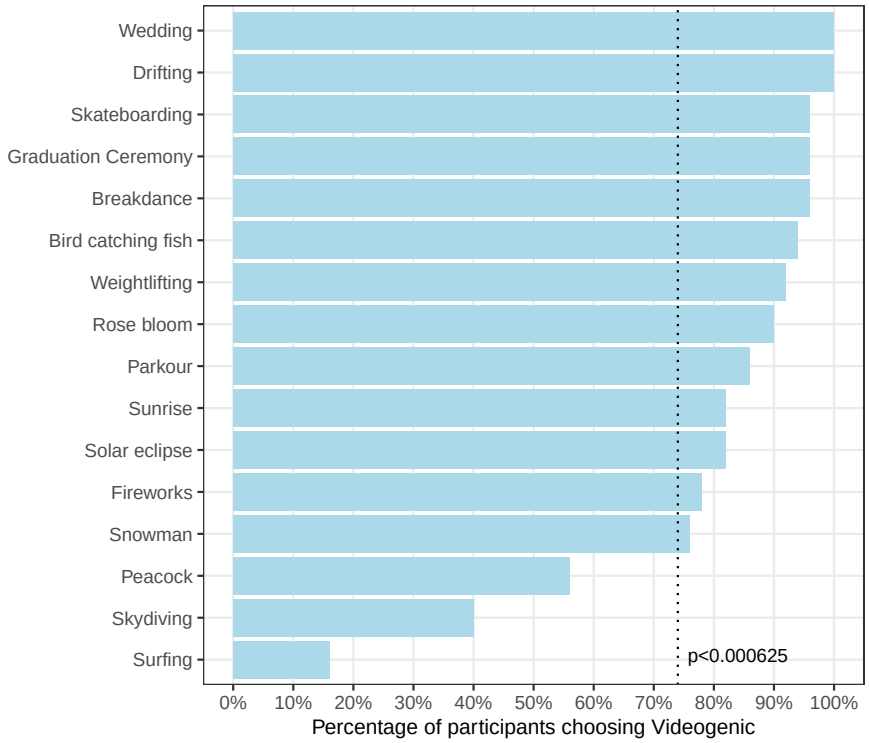

Figure 10: Human evaluation results ($N = 50$). The y-axis lists the videos in the evaluation study. The x-axis shows the percentage of participants who preferred Videogenic's highlight over the baseline's. The dashed line marks the point of statistical significance (p<0.000625).

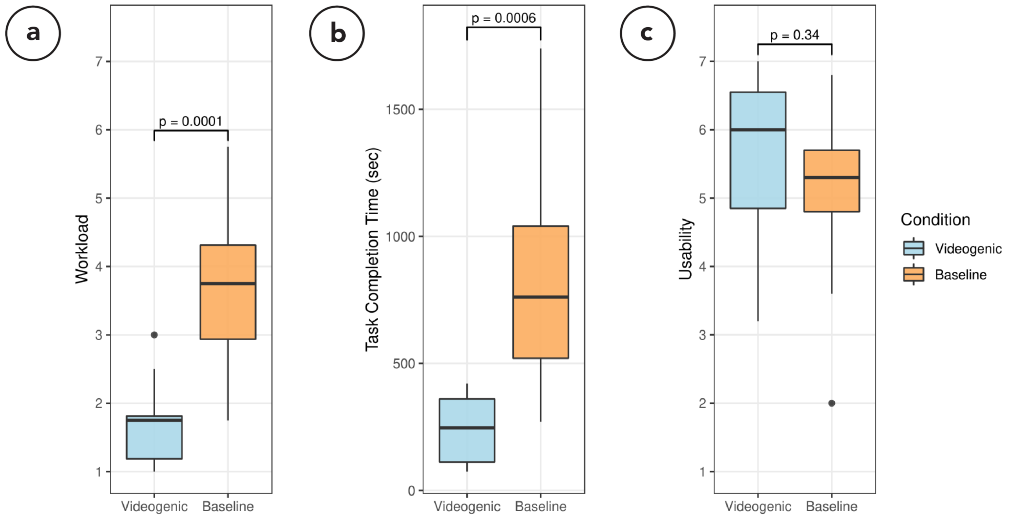

Figure 11: Expert study results (N=12). Boxplots from left to right: workload measured with NASA TLX [6] (7-point Likert scale, lower is better) (a), task completion time (seconds, lower is better) (b), and usability measured with SUS [5] (7-point Likert scale, higher is better) (c).