How do video content creation goals impact which concepts people prioritize when generating B-roll imagery?

HOLLY HUEY, University of California San Diego, USA
MACKENZIE LEAKE, Adobe Research, USA
DEEPALI ANEJA, Adobe Research, USA
MATTHEW FISHER, Adobe Research, USA
JUDITH E. FAN, Stanford University, USA

B-roll is vital when producing high-quality videos, but finding the right images can be difficult and time-consuming. Moreover, what B-roll is most effective can depend on a video content creator's intent—is the goal to entertain, to inform, or something else? While new text-to-image generation models provide promising avenues for streamlining B-roll production, it remains unclear how these tools can provide support for content creators with different goals. To close this gap, we aimed to understand how video content creator's goals guide which visual concepts they prioritize for B-roll generation. Here we introduce a benchmark containing judgments from > 800 people as to which terms in 12 video transcripts should be assigned highest priority for B-roll imagery accompaniment. We verified that participants reliably prioritized different visual concepts depending on whether their goal was help produce *informative* or *entertaining* videos. We next explored how well several algorithms, including heuristic approaches and large language models (LLMs), could predict systematic patterns in human judgments. We found that none of these methods fully captured human judgments in either goal condition, with state-of-the-art LLMs (i.e., GPT-4) even underperforming a baseline that sampled only nouns or nouns and adjectives. Overall, our work identifies opportunities to develop improved algorithms to support video production workflows.

CCS Concepts: • **Human-centered computing** → **User models**; **Empirical studies in visualization**; • **Information systems** → **Multimedia content creation**.

Additional Key Words and Phrases: text-to-image; video B-roll; visual communication; human behavioral benchmarking

ACM Reference Format:
Holly Huey, Mackenzie Leake, Deepali Aneja, Matthew Fisher, and Judith E. Fan. 2024. How do video content creation goals impact which concepts people prioritize when generating B-roll imagery?. In *Creativity and Cognition (C&C '24), June 23–26, 2024, Chicago, IL, USA*. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3635636.3664252

1 INTRODUCTION

If you have ever felt moved by a documentary, social media clip, video lecture, or even advertisement, it is likely that you have been as influenced by the narrative of the main video (called *A-roll*) as its accompanying cutaway images and footage (called *B-roll*). Recently, text-to-image generation models like Firefly, Dall-E 2, Midjourney, and Stable Diffusion are providing new opportunities for content creators to quickly and flexibly generate B-roll. These models build on prior work leveraging B-roll recommendation algorithms using text from video transcripts [3, 10]. These B-roll generation techniques have also been generalized to non-photorealistic outputs, including illustrations [13] and animated graphics [17, 22, 25]. Together, such rapid pace of recent progress suggests the potential to fully automate

Manuscript submitted to ACM

text-to-visual generation in the near future [15, 16, 18, 26]. Nevertheless, many of these approaches do not account for the different goals that content creators might have (e.g., to entertain or inform) beyond the text information in a video transcript and how these goals have important consequences for what kind of B-roll may be most effective for those viewership goals.

Recent work in cognitive science investigating visual communication has found that people produce different kinds of images depending on their communicative goals [4, 11, 27], and these differences also impact viewers' interpretation of these images [8, 9, 11]. Such findings might have implications for understanding how different goals during video production could impact their content, and thus how engaging and memorable they are. Prior work has found that videos intended to be entertaining often use humor [7] and emotionally resonant themes [2] to hold a viewer's attention, even if it means including less information about the main subject of the video [24] or less relevance to viewers [14]. On the other hand, videos intended to be educational are more effective when emphasizing information relevant to viewers' learning objectives [23], rather than including entertaining but irrelevant information [12]. These findings suggest videos primarily intended to be engaging and those intended to support learning often use different strategies to accomplish those goals. However, these studies do not characterize in detail what kinds of image content people prefer to use under these different goals.

A first step toward characterizing how video goals impact content selection during video production is to evaluate what information creators prioritize deciding what B-roll content to include. Towards this end, we developed a benchmark dataset containing judgments from >800 people as to which visual concepts across 12 video transcripts they believed should be portrayed as B-roll in either *informative* or *entertaining* videos. We hope that public release of our dataset will help catalyze progress toward text-to-image generation algorithms with enhanced capacities that align with skilled human video content creators with diverse viewership goals. Progress toward answering these questions would not only help guide improvements for text-to-image systems aimed at supporting creators' different goals, but also help constrain the problem space for automated B-roll generation systems of different kinds of videos.

2 RESEARCH DESIGN

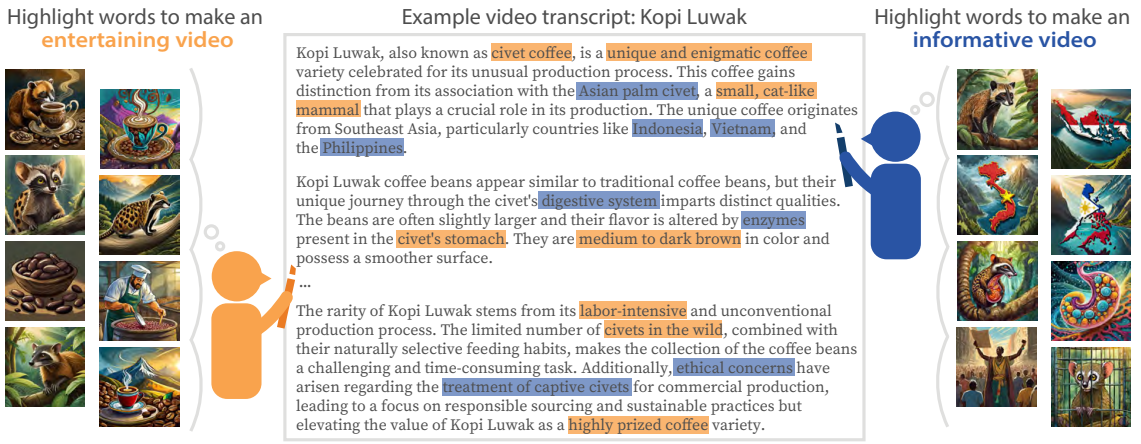

Fig. 1. Participants highlighted text they wanted portrayed as B-roll images in *informative* or *entertaining* videos. (Example B-roll images in this figure was generated using Adobe Firefly Image 2.)

2.1 Methods

2.1.1 Participants. 949 participants (608 female, mean age = 20.46 years) were recruited from our university's undergraduate study pool and completed the task. Data sessions were excluded from 131 participants (N=32 for technical errors, N=37 for self-reporting little to no ability to form mental images, N=62 for self-reporting little to no effort devoted to the task). Our final dataset contained 8,880 annotated transcripts from 818 participants. All participants provided informed consent in accordance with our university's IRB.

2.1.2 Transcript stimuli. 12 transcripts spanning 4 popular video topics were generated by ChatGPT: *food*: Almas caviar, Murnong, Kopi Luwak coffee; *fashion*: African kanzu, Indonesian kebaya, Mongolian deel; *city travel*: Vilnius, Lichinga, Cuenca; *animals*: vaquita, saola, dugong. We selected rare subtopics (e.g., endangered animals) that video content creators might choose to educate naive viewers about. For each transcript, ChatGPT generated a description spanning 5 topically focused paragraphs, 3 sentences each. Transcripts were approximately matched in word count (range = 276-310 words) and audio recorded (range = 1:54-2:08 minutes).

2.1.3 Task procedure. We designed a web-based experiment in which participants viewed the 12 video transcripts. Participants were told the transcripts would later be converted into talking head videos (i.e., in which a speaker talks directly to the camera), but were currently missing B-roll. During each trial, participants viewed one transcript and listened to its audio recording. They were instructed to highlight segments of consecutive text with their cursor (e.g., "long pointed horns") that they believed would be best portrayed as B-roll. Participants were told a video editor would receive their highlighted transcripts and use their highlights to create B-roll for the final videos. Participants were instructed to highlight 15-20 words and told each highlighted text segment corresponded to one B-roll image.

Half of participants helped produce *informative* videos intended to help viewers remember the transcript content for general life knowledge or school test (Fig 1). The other half of participants helped produce *entertaining* videos intended to motivate viewers to "like", write comments, or subscribe to a video channel. Before test trials, they watched an example of a talking head video that included B-roll images and an example of a talking head video that did not include B-roll images. They then completed a practice trial to familiarize themselves with the transcript highlighting platform.

3 RESULTS

3.1 Estimating reliability of human visual concept selection behavior

Prior work has shown that people flexibly adapt their visual production behavior depending on their communicative goals [6, 9, 11, 19, 20, 27]. Building on this work, we predicted that different content creation goals to help produce *informative* or *entertaining* videos would shift which visual concepts people believe should be portrayed as B-roll. However, if visual concept selection is primarily driven by individual aesthetic preferences, we would predict no systematic difference in text selections even if provided different content creation goals. To evaluate this, we conducted a chi-square goodness-of-fit test to evaluate the difference in word selections between goal conditions. Participants prioritizing informativity generated word selections that were significantly different from those prioritizing entertainment ($\chi^2(1742) = 5532$, $p = 0.0$; Fig 2B *right*, see supp. Fig4) and word selections from both conditions were significantly different from random word selections (informative: $\chi^2(2151) = 67570$, $p = 0.0$; entertaining: ($\chi^2(2151) = 65129$, $p = 0.0$).

We also conducted a split-half reliability test to evaluate how internally consistent participants' word selections were within goal conditions. We randomized participants within transcript and goal and then randomly assigned participants to two groups. Consistent with our prior finding, we found significant differences *between* randomized informative

($\chi^2(1742) = 3336$, $p = 0.0$) and entertaining ($\chi^2(1742) = 3570$, $p = 0.0$) groups and found that randomized groups *within* informative ($\chi^2(1742) = 1561$, $p = 0.999$) and entertaining conditions ($\chi^2(1742) = 1511$, $p = 0.999$) were internally similar. These analyses provide evidence that participants' word selection behavior was reliably consistent when given goals to prioritize video informativity or entertainment.

3.2 Measuring impact of content creation goals on human visual concept selection

We next sought to measure how different content creation goals might shift how many text segments people highlighted, as well as how many words were included in each highlight. Insofar as informative videos may include more informationally dense B-roll, we predicted participants with this goal might highlight fewer text segments but that those segments might include more words per highlight. We hypothesized that entertaining videos may include more visually diverse images and predicted participants with this goal might highlight more text segments but with less words included per highlight. To evaluate these hypotheses, we fit a linear mixed-effects model to predict the number of highlighted text segments from goal, with random intercepts for each participant and video transcript and a second linear mixed-effects model with the same random intercepts to predict the number of words included per highlight. Contrary to our predictions, we found that across all transcripts participants highlighted a similar number of text segments (informative: 7.82, entertaining: 8.069; $b = 0.302$, $t = 0.995$, $p = 0.32$) and included a similar number of words per highlight (informative: 1.86, entertaining: 1.905; $b = -0.157$, $t = -1.19$, $p = 0.235$).

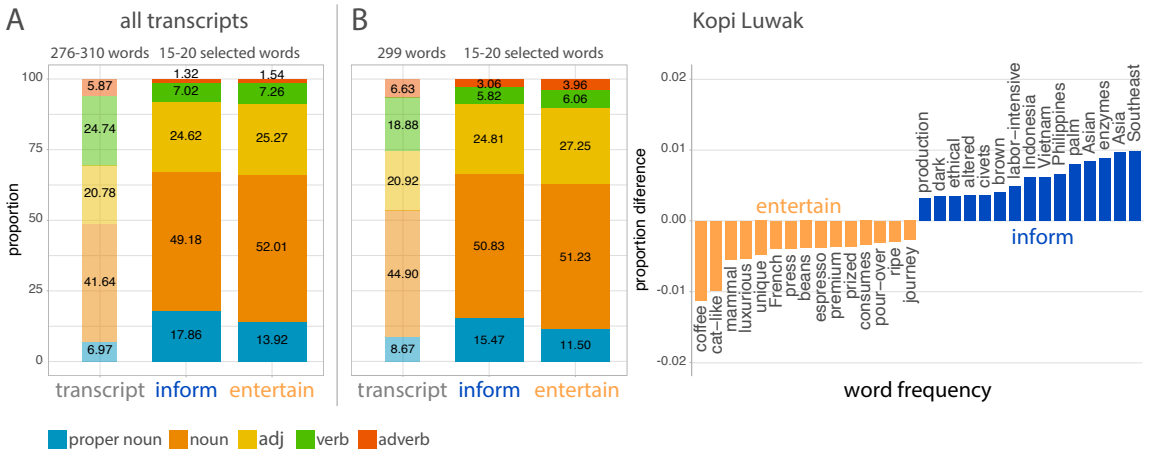

Fig. 2. (A) Proportion of proper nouns, nouns, adjectives, verbs, and adverbs within video transcripts: *left*, transcript base rate; *middle*, informative condition; *right*, entertaining condition. (B, *left*) Kopi Luwak transcript: proportion of proper nouns, nouns, adjectives, verbs, and adverbs. See Supplemental Fig. 5 for proportions across all transcripts. (B, *right*) Kopi Luwak transcript: Proportion difference of 15 most frequently selected words between conditions: *informative* condition represented in the positive direction, *entertaining* condition represented in the negative direction. See Supplemental Fig 4 for word frequencies across all transcripts.

We then measured the semantic differences between visual concepts prioritized within each goal condition. We applied parts-of-speech tagging to selected words using *NLTK* tokenization and removed stopwords. We hypothesized that if informative videos might focus on more specific information like geographical locations, historical figures, or time periods or information like processes and actions, we predicted participants with this goal might prioritize proper nouns and verbs. By contrast, we hypothesized that if entertaining videos focus on more visually interesting

information, we predicted participants with this goal might prioritize nouns, adjectives, and adverbs. To evaluate these hypotheses, we fit a linear mixed-effects model to predict the proportion of proper nouns from goal, with random intercepts for each participant and video transcript. We predicted the proportion of nouns, adjectives, verbs, and adverbs using models with the same structure. Consistent with our hypotheses, participants prioritizing informativity highlighted more proper nouns ($b = 0.49$, $t = 6.15$, $p = 1.23e{-9}$) and fewer nouns ($b = -0.253$, $t = -2.081$, $p = 3.77e{-2}$) and adverbs ($b = -0.0429$, $t = -2.103$, $p = 3.56e{-2}$), relative to those prioritizing entertainment (Fig 2A, see supp. Fig5). We did not find significant differences in the frequency of verbs ($b = -9.88e{-3}$, $t = -0.311$, $p = 0.756$) or adjectives ($b = 1.059e{-2}$, $t = 0.135$, $p = 0.892$) between goal conditions. However, we compared the proportion of proper nouns, nouns, adjectives, verbs, and adverbs in participants' highlighted text segments against the base rate of proper nouns, nouns, adjectives, verbs, and adverbs in the transcripts. We again fit linear mixed-effects models to predict the proportion of each parts-of-speech type from goal, with random intercepts for each transcript. Across both conditions, participants highlighted more adjectives (informative: $b = 3.84$, $t = 3.44$, $p = 5.54e{-3}$; entertaining: $b = 4.49$, $t = 5.49$, $p = 1.87e{-4}$) but fewer verbs (informative: $b = -17.7$, $t = -21.5$, $p = 2.47e{-10}$; entertaining: $b = -17.5$, $t = -26.3$, $p = 12.74e{-11}$), relative to the transcript base rate of adjectives and verbs. Additionally, participants in both conditions highlighted more proper nouns (informative: $b = 10.9$, $t = 5.87$, $p = 1.08e{-4}$; entertaining: $b = 6.95$, $t = 4.18$, $p = 1.54e{-3}$) and nouns (informative: $b = 7.54$, $t = 6.27$, $p = 6.1e{-5}$; entertaining: $b = 10.4$, $t = 8.67$, $p = 3.01e{-6}$) and less adverbs (informative: $b = -4.55$, $t = -12.2$, $p = 1.02e{-7}$; entertaining: $b = -4.32$, $t = -11.4$, $p = 2.0e{-7}$), relative to the transcript base rate of proper nouns and adverbs. These results indicate that video content creation goals systematically shifted which words participants prioritized within the video transcripts.

3.3 Predicting human visual concept selection using text selection models

What might explain how people select visual concepts for B-roll in informative or entertaining videos? To explore this, we evaluated several heuristic approaches for text selection and compared their alignment to human judgments (Fig 3). For each model, we generated simulated datasets consisting of the same number of word selections as participants within each goal condition. We then employed Jensen–Shannon divergence (JSD) to measure the difference between the probability of a word being selected by a model against the observed frequency of that same word being selected by human participants (0 = identical alignment). We tested three heuristic models:

(1) Word frequency: Insofar as people prioritize visual concepts highly relevant to a video topic, we predicted people would be biased to choose words appearing at a higher frequency within transcripts. To generate word selections, we randomly sampled the top 20 most frequently included words from each transcript.

(2) Topic sentence selection: To the extent that topic sentence expresses the main idea of a paragraph, we evaluated whether people are biased to select words from the first sentence of each transcript paragraph. To generate word selections, we randomly sampled words from the first sentence of each transcript paragraph, excluding stopwords.

(3) Visual concreteness: Because B-roll is inherently visual, we evaluated whether participants prioritized words more strongly evoking visual imagery. Using visual concreteness ratings [1] (1 = abstract, 5 = concrete), we generated concreteness scores for each word in the transcripts. Building on work leveraging these scores to auto-generate slideshow imagery [15], we randomly sampled words scoring > 4.5 and filled in remaining word selections, if needed, with words scoring > 3. We predicted people would prioritize more visually concrete words in both goal conditions. Specifically, we predicted participants prioritizing entertainment would prefer more visual-based imagery and select words with higher concreteness scores. By contrast, word concreteness has been shown to increase memorability [5, 21] and so we predicted participants might favor portraying more visually concrete words to increase viewer memory retention.

We also evaluated how large language models, like GPT-3.5 and GPT-4, performed on the same task. To test this, we prompted models with instructions mirroring, as much as possible, the task instructions participants received: *"I am trying to make two kinds of videos. For one video, I am trying to make a video that is as entertaining as possible to my viewers. My goal is to get viewers to like the video, subscribe to my channel, and comment on my video. I need to add pictures to the video. Please select 20 single words that would be best to convert into accompanying entertaining pictures. For the second video, I am trying to make a video that is as informative as possible to my viewers. My goal is to get viewers remember as much content, whether they may need it for a school test or general life knowledge. Please select another 20 single words that would be best to convert into accompanying informative pictures."*

As a baseline for human consistency, we calculated the JSD between the random samples of goal conditions from our split-half reliability test. We also developed two baseline models by randomly sampling nouns, including proper nouns, and adjectives from the transcripts.

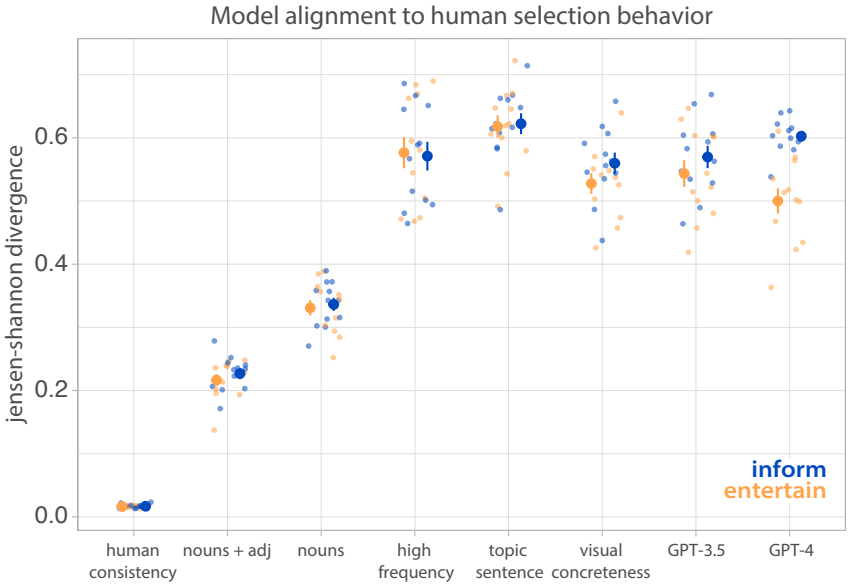

Fig. 3. Model comparison against human selection behavior, using Jensen-Shannon Divergence (0 = identical alignment). Smaller translucent dots represent the JSD score for each video transcript. Error bars represent standard error.

3.3.1 How well do models approximate content creation goal differences? To evaluate each model's ability to capture differences in word selections between goals, we fit a linear mixed-effects model to predict the JSD score from goal, with random intercepts for transcript. Models sampling high frequency words ($b = -5.505e{-3}$, $t = -1.33$, $p = 0.21$), topic sentence words ($b = 45.02e{-4}$, $t = 0.66$, $p = 0.52$), and nouns ($b = 55.03e{-4}$, $t = 1.35$, $p = 0.204$) were not significantly different in their approximation of content creation goals. However, models sampling words scoring high in concreteness ($b = 32.02e{-3}$, $t = 4.22$, $p = 1.43e{-3}$) and nouns and adjectives ($b = 10.24e{-3}$, $t = 2.37$, $p = 0.037$) more closely captured word selections in the entertainment condition, relative to the informative condition. Although GPT-3.5 was not sensitive to goal manipulations ($b = 26.2e{-3}$, $t = 1.14$, $p = 0.28$), GPT-4 more closely approximated word selections in the entertainment condition ($b = 10.26e{-2}$, $t = 6.51$, $p = 4.36e{-5}$), relative to informative condition.

3.3.2 Which models better align with human judgments? Although some models captured differences in word selections between goals, all models fell short of aligning with human judgments (*high frequency*: informative: JSD = 0.571; bootstrapped 95% CI = [0.57, 0.572], entertaining: JSD = 0.577; bootstrapped 95% CI = [0.576, 0.578]; *topic sentence*: informative: JSD = 0.622; bootstrapped 95% CI = [0.621, 0.623], entertaining: JSD = 0.618; bootstrapped 95% CI = [0.618, 0.62]; *visual concreteness*: informative: JSD = 0.56; bootstrapped 95% CI = [0.559, 0.561]; entertaining: JSD = 0.528; bootstrapped 95% CI = [0.526, 0.528]). While all models were significantly different from human word selections (all p<0.05), our baseline models sampling only nouns and/or adjectives more closely approximated human behavior (*nouns*: informative: JSD = 0.336, bootstrapped 95% CI = [0.335, 0.338], entertaining: JSD = 0.331; bootstrapped 95% CI = [0.331, 0.334]; *nouns + adjectives*: informative: JSD = 0.227; bootstrapped 95% CI = [0.224, 0.228], entertaining: JSD = 0.217; bootstrapped 95% CI = [0.212, 0.216]).

Specifically, our baseline model randomly sampling nouns (including proper nouns) and adjectives outperformed all other models, including sampling only nouns (informative: $b = -10.95e{-}2$, $t = -15.7$, $p = 7.22e{-}9$; entertaining: $b = -11.4e{-}2$, $t = -13.2$, $p = 4.32e{-}8$), high frequency words (informative: $b = -34.4e{-}2$, $t = -14.2$, $p = 1.38e{-}12$; entertaining: $b = -35.9e{-}2$, $t = -13.7$, $p = 3.07e{-}12$), topic sentence words (informative: $b = -39.6e{-}2$, $t = 21.5$, $p = 2.91e{-}16$; entertaining: $b = -40.12e{-}2$, $t = -20.22$, $p = 1.06e{-}15$), and visually concrete words (informative: $b = -33.3e{-}2$, $t = -19.8$, $p = 5.87e{-}10$; entertaining: $b = -31.11e{-}2$, $t = -16.6$, $p = 8.43e{-}11$). This model also outperformed GPT-3.5 (informative: $b = -34.3e{-}2$, $t = -18.71$, $p = 1.09e{-}9$; entertaining: $b = -32.7e{-}2$, $t = -14.4$, $p = 1.4e{-}12$) and GPT-4 (informative: $b = -37.6e{-}2$, $t = -32.9$, $p < 2e{-}16$; entertaining: $b = -28.3e{-}2$, $t = -12.9$, $p = 9.83e{-}12$). These results suggest that when participants selected visual concepts, they likely prioritized adjective and noun or adjective and proper noun pairings, consistent with our prior finding that participants included approximately 1.88 words per highlighted text segment. Moreover, these analyses indicate that participants selected words beyond topic sentences, although they also did not merely prioritize frequently occurring words throughout transcripts. Additionally, these data demonstrate that participants tended to select words lower in visual concreteness. This may be explained by the fact that participants prioritizing informativity appeared to select numerical words (e.g., time, dimensions) and more abstract words (e.g., "conservation", "endangered", "craftsmanship") that are likely low in visual concreteness. Participants prioritizing entertainment also appeared to select words with high emotional valence but, again, likely low in visual concreteness (e.g., "elusive", "grace", "gentle"). Overall, these findings suggest that more nuanced models are needed to approximate more fine-grain human-like content creation goals and their impact on B-roll visual concept selection.

4 CONCLUSION

Video content creators aim to captivate, amuse, and excite, as well as to impart critical expertise and information to their viewers. How might creators' different viewership goals impact what content they choose to visualize as B-roll? We developed a novel highlighting paradigm to benchmark which visual concepts people prioritize for B-roll generation in informative vs. entertaining videos and found that people systematically prioritize different visual concepts in video transcripts depending on their goals.

Our findings raise key research opportunities: (1) our findings demonstrate that people prioritize different concepts to portray as B-roll, but does not directly capture how people imagined the B-roll images (e.g., a "civet" within the Kopi Luwak transcript could be portrayed as a naturalist illustration or cartoon dancing civet); (2) in addition to quantitative estimates of creator preferences for B-roll content, exploring how these systematic biases impact downstream viewership entertainment and information retention is equally important. Building on our current work,

results from such opportunities would contribute critical insights for improving video production tools and workflows, as well as deeper cognitive understanding of how people use imagery to augment their communication with others.

ACKNOWLEDGMENTS

We thank Zoe Tait, Rio Aguina-kang, and Sade Oyekenu for assisting with the development of study materials. We also thank members of the Cognitive Tools Lab at UC San Diego and Stanford University. J.E.F. is supported by NSF CAREER award #2047191 and an ONR Science of Autonomy award.

All code and materials available at https://github.com/cogtoolslab/video-broll-public2024

REFERENCES

[1] Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior research methods* 46 (2014), 904–911.

[2] Colin Campbell, Frauke Mattison Thompson, Pamela E Grimm, and Karen Robson. 2017. Understanding why consumers don't skip pre-roll video ads. *Journal of Advertising* 46, 3 (2017), 411–423.

[3] Pei-Yu Chi and Henry Lieberman. 2011. Intelligent assistance for conversational storytelling using story patterns. In *Proceedings of the 16th international conference on Intelligent user interfaces*. 217–226.

[4] Judith E Fan, Robert D Hawkins, Mike Wu, and Noah D Goodman. 2020. Pragmatic inference and visual abstraction enable contextual flexibility during visual communication. *Computational Brain & Behavior* 3, 1 (2020), 86–101.

[5] Klaus Fliessbach, Susanne Weis, Peter Klaver, Christian Erich Elger, and Bernd Weber. 2006. The effect of word concreteness on recognition memory. *NeuroImage* 32, 3 (2006), 1413–1421.

[6] Simon Garrod, Nicolas Fay, John Lee, Jon Oberlander, and Tracy MacLeod. 2007. Foundations of representation: Where might graphical symbol systems come from? *Cognitive science* 31, 6 (2007), 961–987.

[7] Kendall Goodrich, Shu Z Schiller, and Dennis Galletta. 2015. Consumer reactions to intrusiveness of online-video advertisements: do length, informativeness, and humor help (or hinder) marketing outcomes? *Journal of advertising research* 55, 1 (2015), 37–50.

[8] Robert D Hawkins, Megumi Sano, Noah D Goodman, and Judith E Fan. 2023. Visual resemblance and interaction history jointly constrain pictorial meaning. *Nature Communications* 14, 1 (2023), 2199.

[9] Sebastian Holt, Judith E Fan, and David Barner. 2024. Creating ad hoc graphical representations of number. *Cognition* 242 (2024), 105665.

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

[11] Holly Huey, Xuanchen Lu, Caren M Walker, and Judith E Fan. 2023. Visual explanations prioritize functional properties at the expense of visual fidelity. *Cognition* 236 (2023), 105414.

[12] Mohamed Ibrahim, Pavlo D Antonenko, Carmen M Greenwood, and Denna Wheeler. 2012. Effects of segmenting, signalling, and weeding on learning from educational video. *Learning, media and technology* 37, 3 (2012), 220–235.

[13] Yu Jiang, Jing Liu, Zechao Li, Changsheng Xu, and Hanqing Lu. 2012. Chat with illustration: a chat system with visual aids. In *Proceedings of the 4th International Conference on Internet Multimedia Computing and Service*. 96–99.

[14] Claire Youngnyo Joa, Kisun Kim, and Louisa Ha. 2018. What makes people watch online in-stream video advertisements? *Journal of Interactive Advertising* 18, 1 (2018), 1–14.

[15] Mackenzie Leake, Hijung Valentina Shin, Joy O Kim, and Maneesh Agrawala. 2020. Generating Audio-Visual Slideshows from Text Articles Using Word Concreteness.. In *CHI*, Vol. 20. 25–30.

[16] Jihyeon Janel Lee, Mitchell Gordon, and Maneesh Agrawala. 2017. Automatically Visualizing Audio Travel Podcasts. In *Adjunct Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*. 165–167.

[17] Jian Liao, Adnan Karim, Shivesh Singh Jadon, Rubaiat Habib Kazi, and Ryo Suzuki. 2022. RealityTalk: Real-time speech-driven augmented presentation for AR live storytelling. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–12.

[18] Xingyu" Bruce" Liu, Vladimir Kirilyuk, Xiuxiu Yuan, Alex Olwal, Peggy Chi, Xiang" Anthony" Chen, and Ruofei Du. 2023. Visual captions: Augmenting verbal communication with on-the-fly visuals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–20.

[19] K. Mukherjee, R. X. D. Hawkins, and J. E. Fan. 2019. Communicating semantic part information in drawings. In *Proceedings of the 41th Annual Meeting of the Cognitive Science Society*. 2413–2419.

[20] Kushin Mukherjee, Holly Huey, Xuanchen Lu, Yael Vinker, Rio Aguina-Kang, Ariel Shamir, and Judith Fan. 2023. SEVA: Leveraging sketches to evaluate alignment between human and machine visual abstraction. In *Advances in Neural Information Processing Systems*.

[21] Allan Paivio, Mary Walsh, and Trudy Bons. 1994. Concreteness effects on memory: When and why? *Journal of Experimental Psychology: Learning, Memory, and Cognition* 20, 5 (1994), 1196.

[22] Takaaki Shiratori, Moshe Mahler, Warren Trezevant, and Jessica K Hodgins. 2013. Expressing animated performances through puppeteering. In *2013 IEEE Symposium on 3D User Interfaces (3DUI)*. IEEE, 59–66.

[23] Abdulhadi Shoufan. 2019. Estimating the cognitive value of YouTube's educational videos: A learning analytics approach. *Computers in Human Behavior* 92 (2019), 450–458.

[24] Douglas West and John Ford. 2001. Advertising agency philosophies and employee risk taking. *Journal of Advertising* 30, 1 (2001), 77–91.

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

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

[27] Justin Yang and Judith E Fan. 2021. Visual communication of object concepts at different levels of abstraction. *arXiv preprint arXiv:2106.02775* (2021).

A SUPPLEMENTAL FIGURES

A.1 Proportion difference in word frequency across all video transcripts

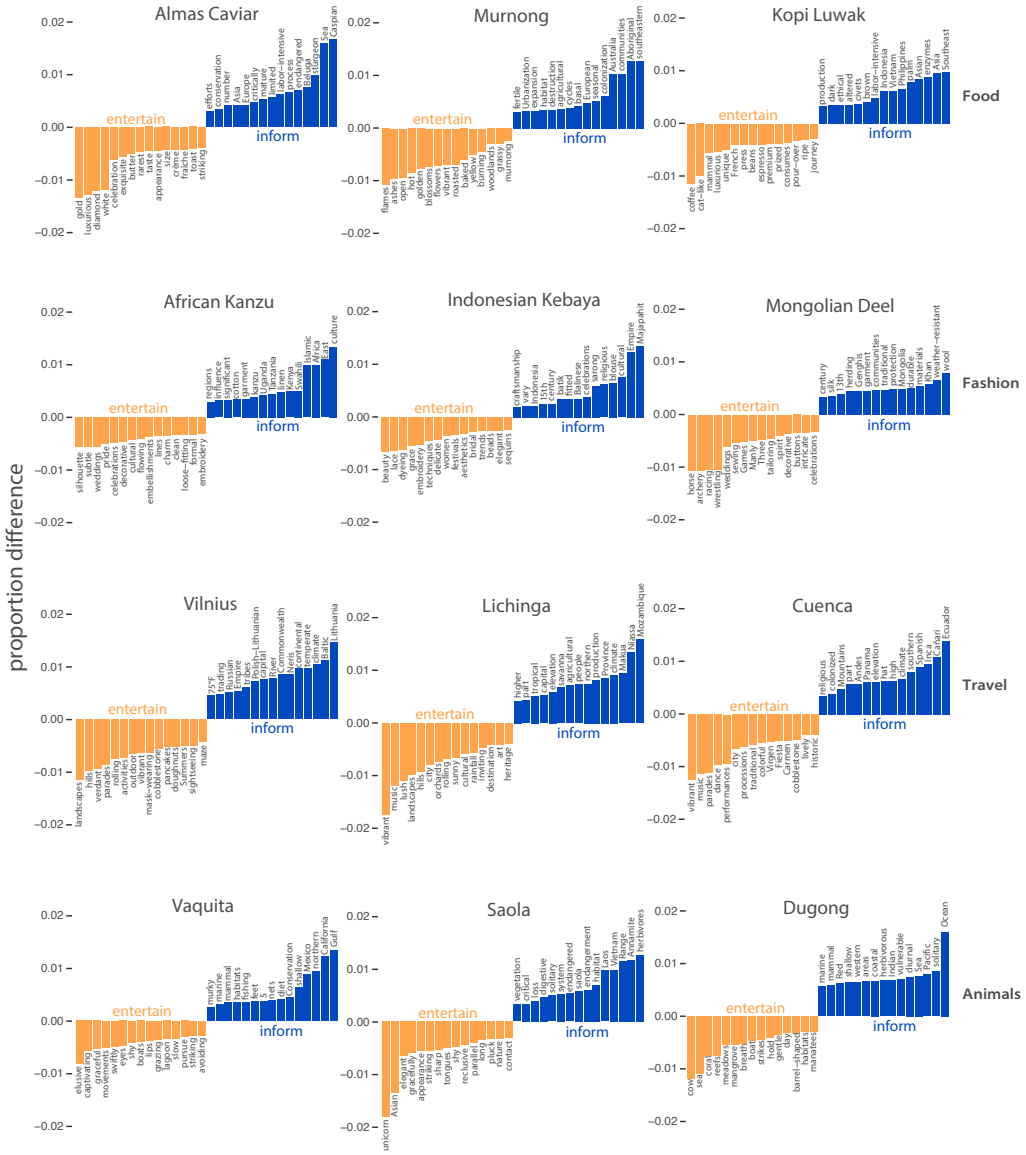

Fig. 4. Top 15 most frequently selected words with the largest difference in prioritization across conditions: *informative* represented in the positive direction; *entertaining* represented in the negative direction

A.2 Proportion of proper nouns, nouns, adjectives, verbs, and adverbs across all video transcripts

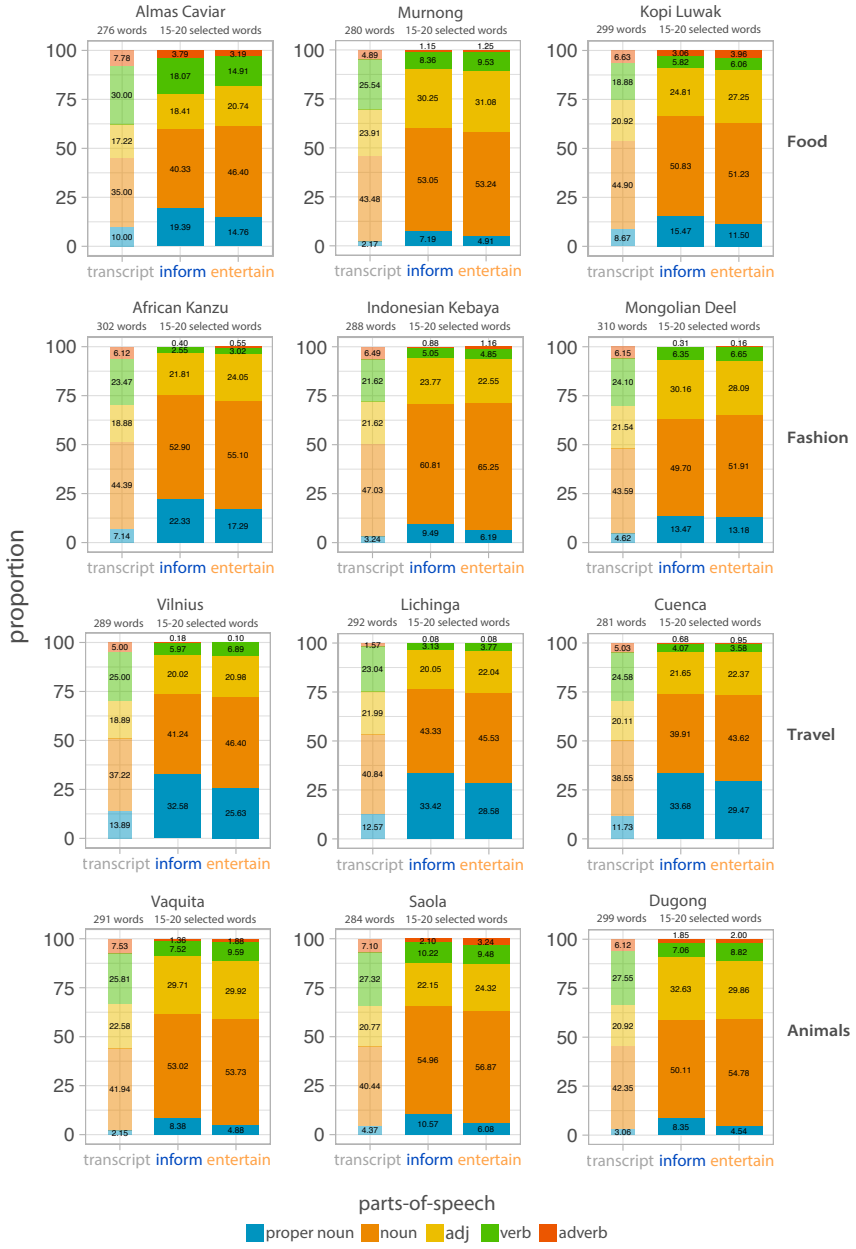

Fig. 5. Proportion of parts of speech within transcripts and within word selections across goals

Received 8 April 2024; revised 12 March 2009; accepted 5 June 2009