How do video content creation goals impact which concepts people prioritize Holly Huey^a, Mackenzie Leake^b, Deepali Aneja^b, Matthew Fisher^b, & Judith E. Fan^c

when generating B-roll imagery? ^aUniversity of California San Diego 🕡 ^bAdobe Research 🔼 ^cStanford 💃



Overview

Compelling videos often combine a main video narrative (A-Roll) and supplemental images (B-Roll) to convey impactful messaging to viewers.

But what makes great B-Roll content?

We developed a large-scale behavioral benchmark (N > 800 participants) of how people with different video content creation goals prioritize words in their transcripts to illustrate as B-Roll images.



Experimental task

Highlight words to make an entertaining video

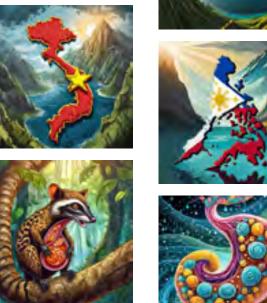
Example video transcript: Kopi Luwak Kopi Luwak, also known as civet coffee, is a unique and enigmatic coffee variety celebrated for its unusual production process. This coffee gains distinction from its association with the Asian palm civet, a small, cat-like mammal that plays a crucial role in its production. The unique coffee originates from Southeast Asia, particularly countries like Indonesia, Vietnam, and the Philippines.

Kopi Luwak coffee beans appear similar to traditional coffee beans, but their unique journey through the civet's digestive system imparts distinct qualities. The beans are often slightly larger and their flavor is altered by enzymes present in the civet's stomach. They are medium to dark brown in color and possess a smoother surface.

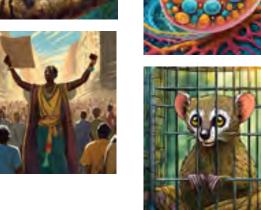
The rarity of Kopi Luwak stems from its labor-intensive and unconventional

production process. The limited number of civets in the wild, combined with their naturally selective feeding habits, makes the collection of the coffee beans a challenging and time-consuming task. Additionally, ethical concerns have arisen regarding the treatment of captive civets for commercial production, leading to a focus on responsible sourcing and sustainable practices but elevating the value of Kopi Luwak as a highly prized coffee variety.

Highlight words to make an informative video

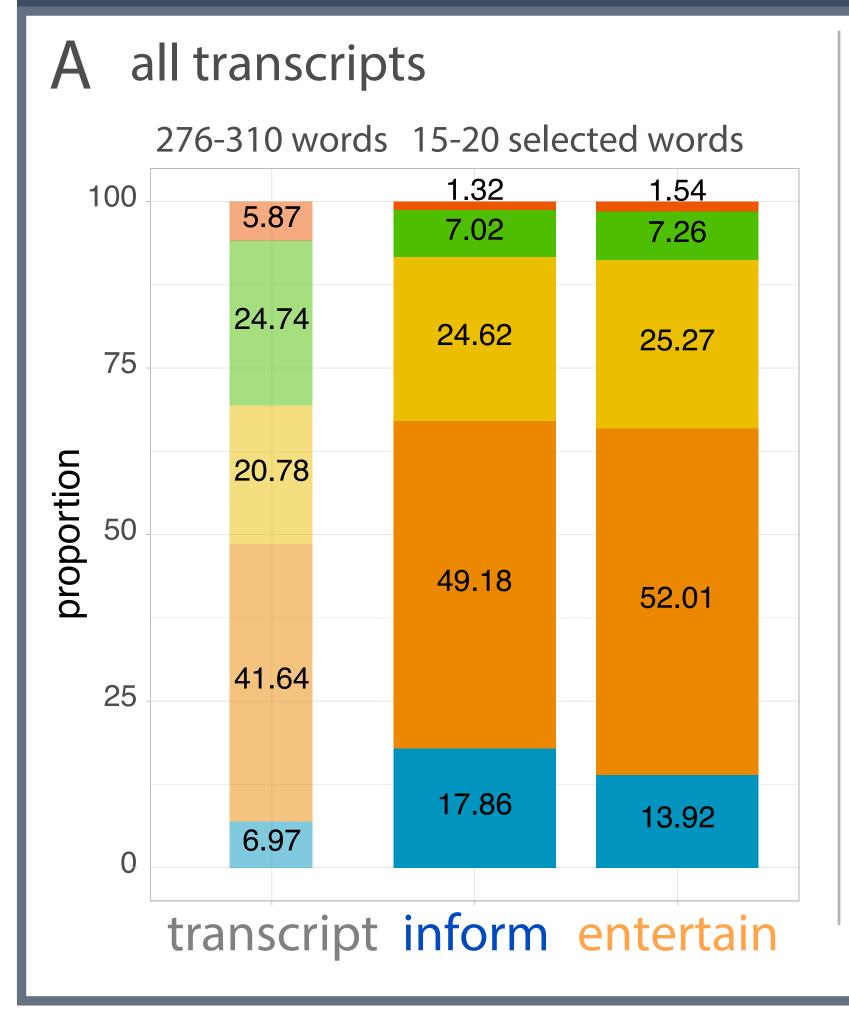


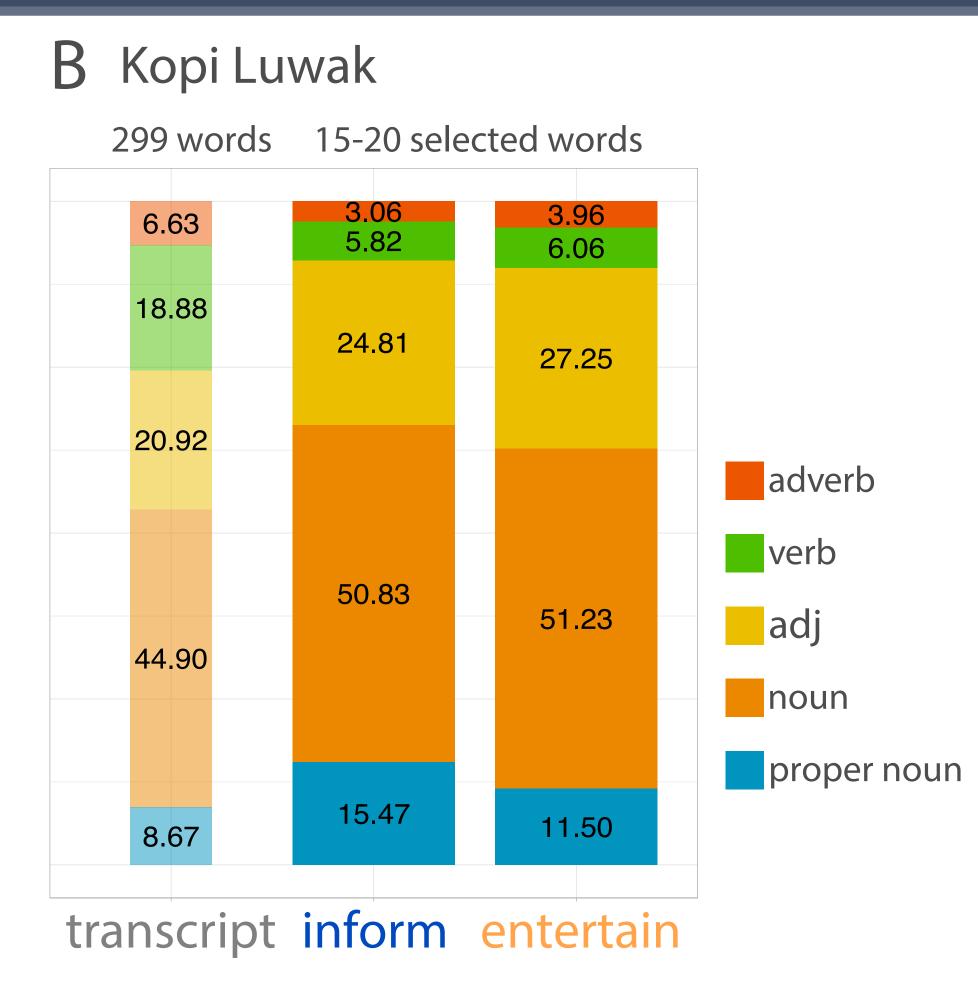


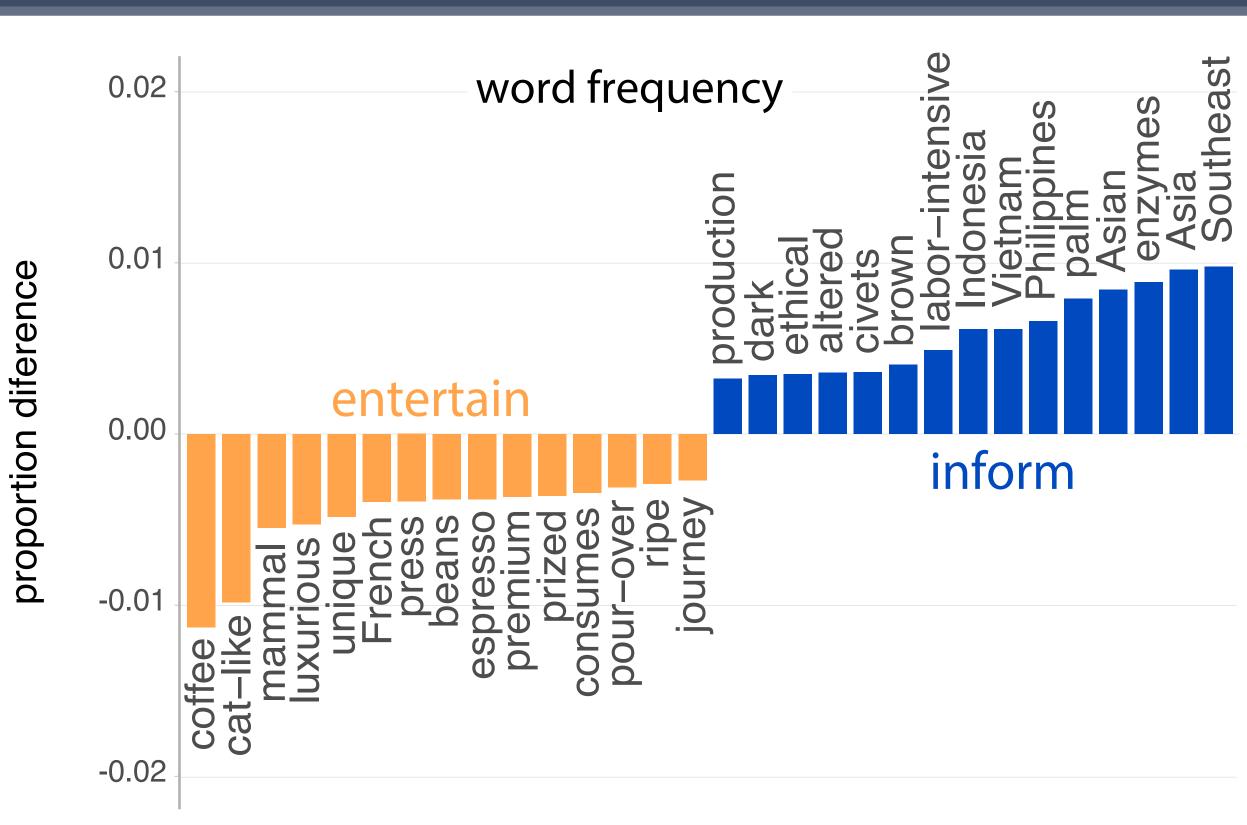


Participants annotated 12 transcripts spanning 4 popular video topics: food, fashion, city travel, animals Randomly assigned to highlight words to display as B-Roll images in videos either meant to entertain or inform viewers

Results: Measuring impact of video goals on human visual concept selection

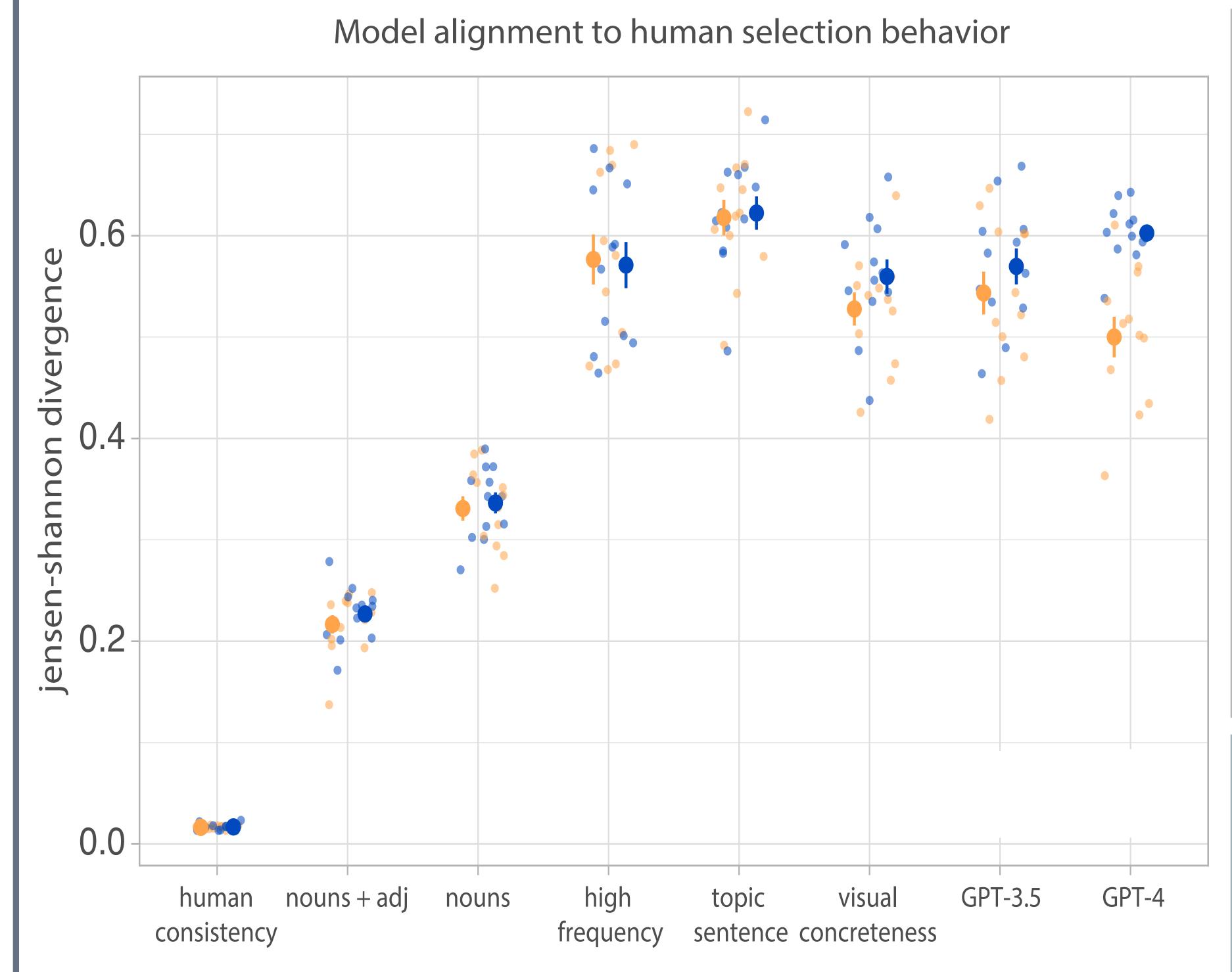






People prioritize nouns & adjectives but different words depending on goal

Model comparison: Predicting human visual concept selection using text selection models



Substantial gap remains between human behavior & models

Parts-of-speech model sampling noun & adjectives best approximates human selection behavior

To approximate human selection behavior, we randomly sampled words from transcripts using the following procedures:

Baseline models

Nouns

random keyword selection

Nouns & adjs

random keyword selection

Heuristic models

High frequency words

relevant keywords based on frequency of inclusion in transcripts

Topic sentence words

relevant keywords based on topic sentences capturing the main idea of each pararaph

Visually concrete words

relevant keywords based on more visually concrete

words evoke more visual imagery needed for B-Roll

LLMs

ChatGPT3.5 and 4 prompted with the same task instructions given to human participants in the different video conditions

Summary

People systematically prioritize different visual concepts to illustrate as B-Roll depending on their goals to make entertaining or informative videos.

Results can help guide improvements for text-to-image systems aimed at supporting different content creation goals.

email: hhuey@ucsd.edu