An animated picture says at least a thousand words: Selecting Gif-based Replies in Multimodal Dialog

Xingyao Wang (`xingyaow@umich.edu`) & David Jurgens (`jurgens@umich.edu`)

1. What are we trying to do?

Online conversations include more than just text, people like using reaction gifs in their messages.

However, current NLP dialog systems (e.g., chatbots) are almost all text-based.

To fill the gap, we introduce a **new task:** Select a gif-based reply to a text message from a user.

@user: Ahhhh! The deadline is in 24 hours!

↳ **@model:**

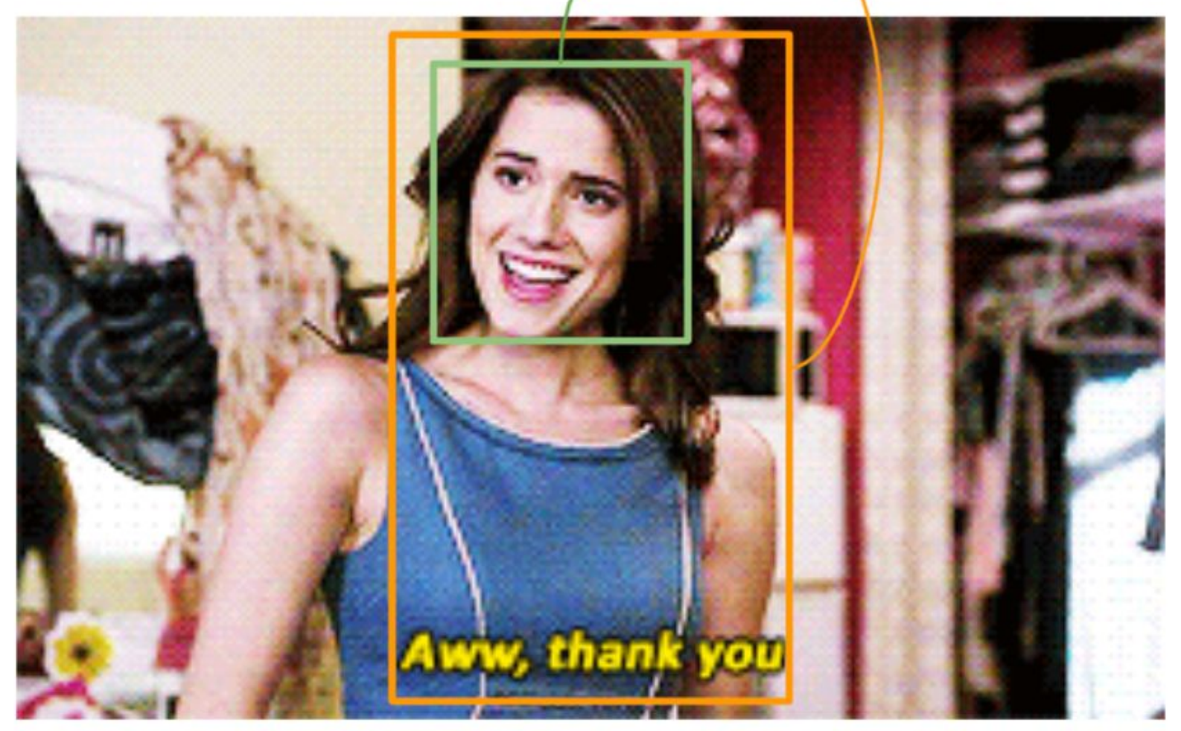
2. New dataset for multimodal text-gif dialog!

1.56M text-gif conversation turns & Metadata for 115k Gifs. Metadata includes annotated tags (~⅓ of gifs, e.g. happy, exciting), extracted captions. and machine-predicted object names and features (e.g., face, building)

Object names: face woman

Tweet: @USER is my hero

↳**Reply:**

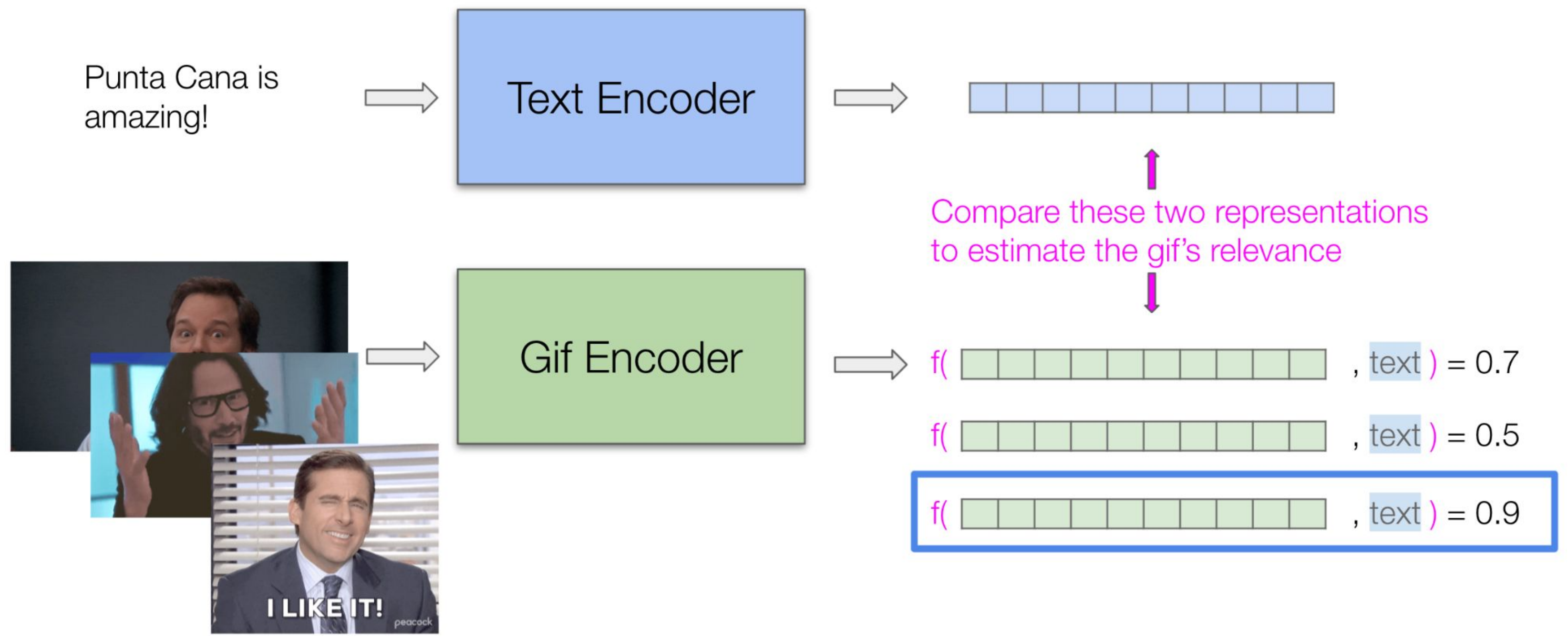


↳ **Annotated Tags:** [“`thank`”]

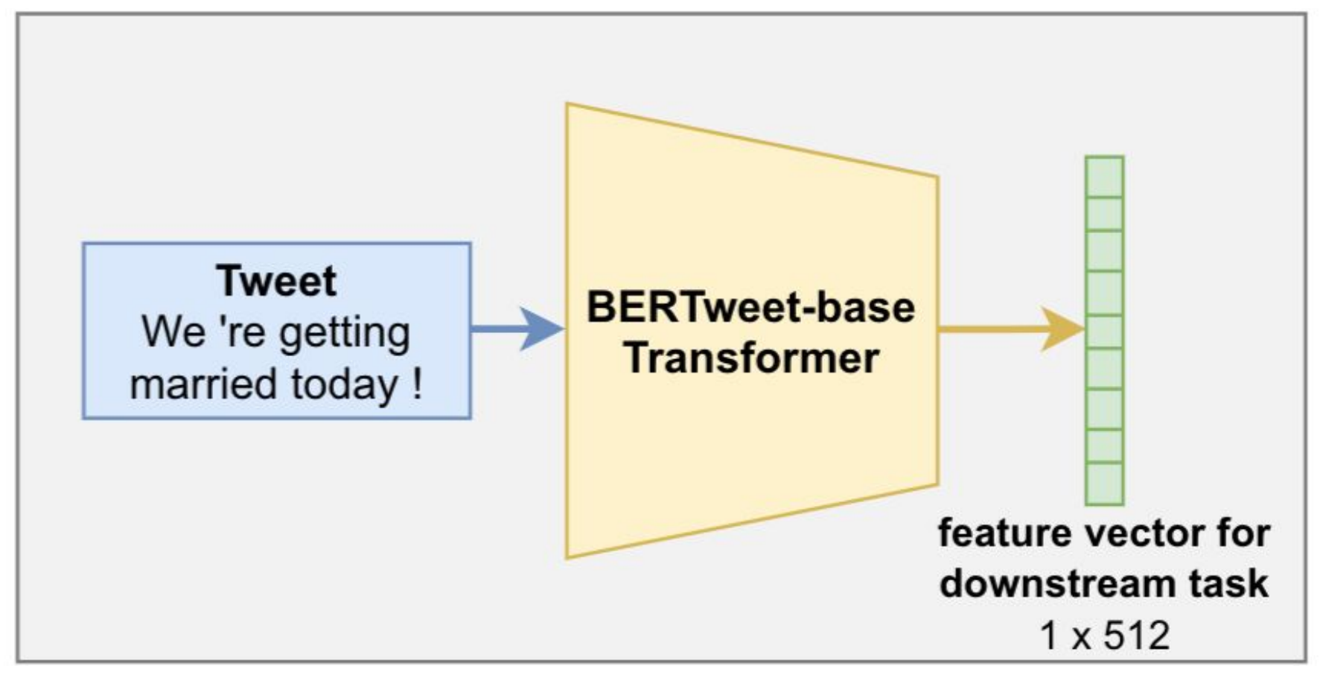
↳ **Captions:** Aww , thank you

3. How to select a gif reply? Ranking approach!

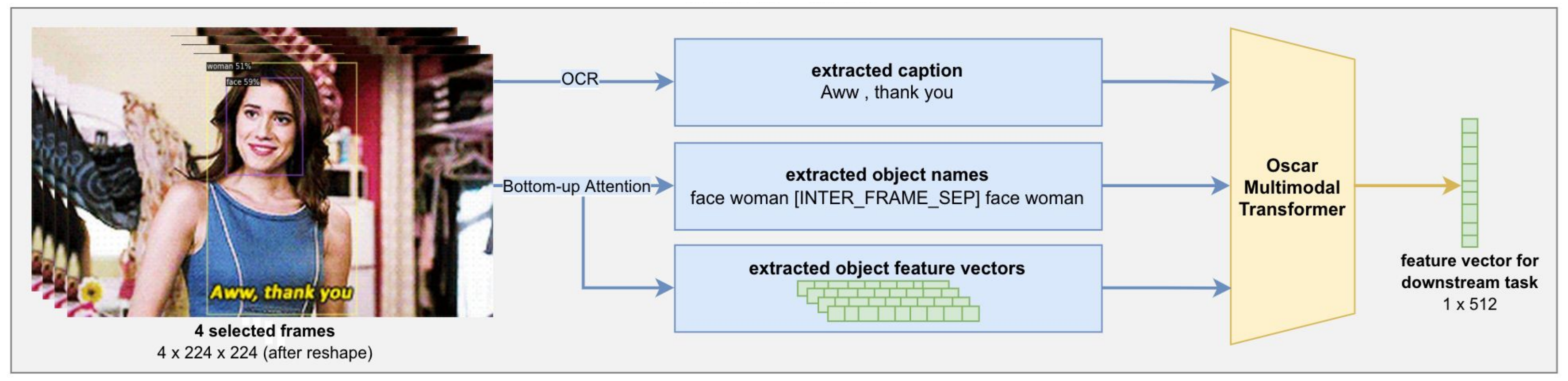
We compare **text** and **gif** representations to estimate the gif’s relevance. We pick the gif with the highest score as reply.



4. Our Model: Pepe the King Prawn[1]


Tweet
We 're getting married today !
BERTweet-base Transformer
feature vector for downstream task
1 x 512

Text encoder: *Roberta* (pre-trained on Twitter)


OCR
Bottom-up Attention
extracted caption
Aww , thank you
extracted object names
face woman [INTER_FRAME_SEP] face woman
extracted object feature vectors
Oscar Multimodal Transformer
feature vector for downstream task
1 x 512
4 selected frames
4 x 224 x 224 (after reshape)

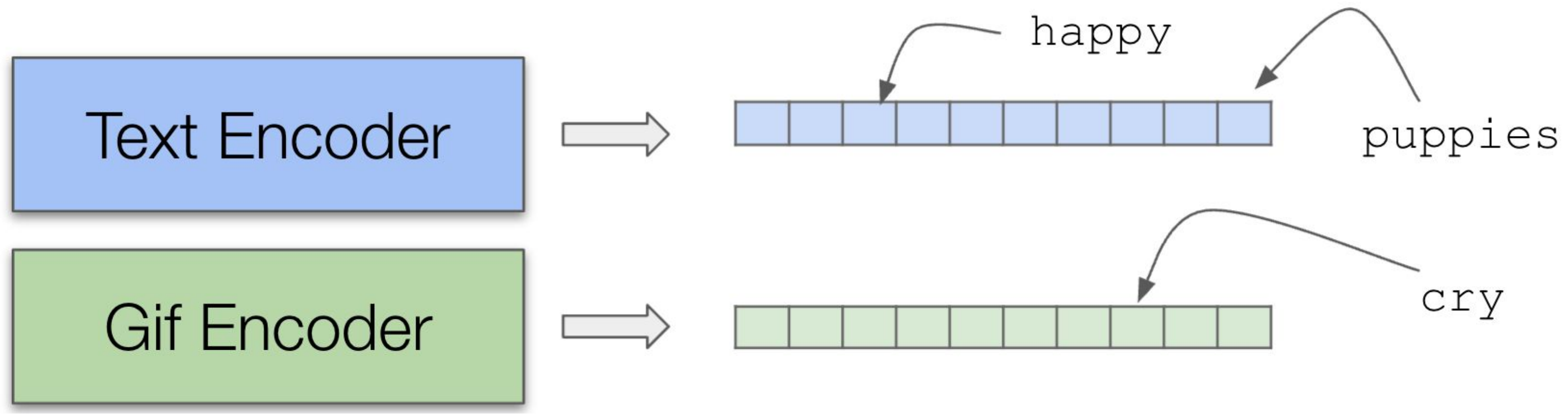
GIF Encoder: *Oscar Multimodal Transformer*

GIF encoder fuses information about the gif from text modality (captions), visual modality (object features), and bridge between the two modality (object names).

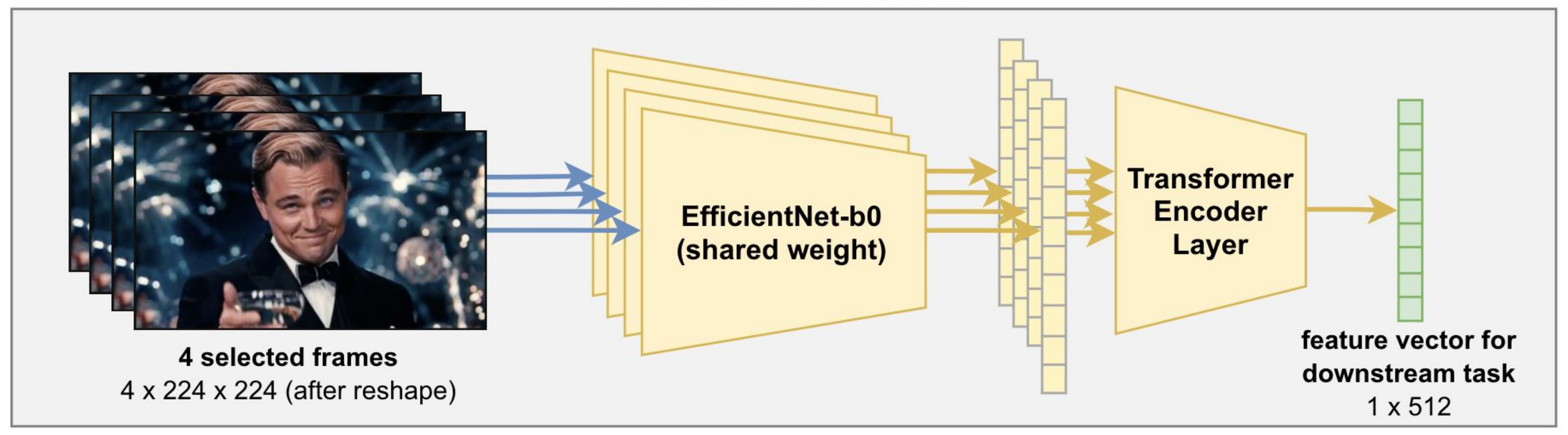
[1] “King Prawn” refers to “selecKting INteresting Gifs for Personal RespAWNses”

5. Can we rank gifs differently?

Idea 1 (Tag-based): Some gifs have describe content/intent, so use a **tag-based encoder** for both gif and text



Idea 2 (CLIP-based): Simplify the gif encoder to use a CLIP-like **CNN-based Image encoder**.

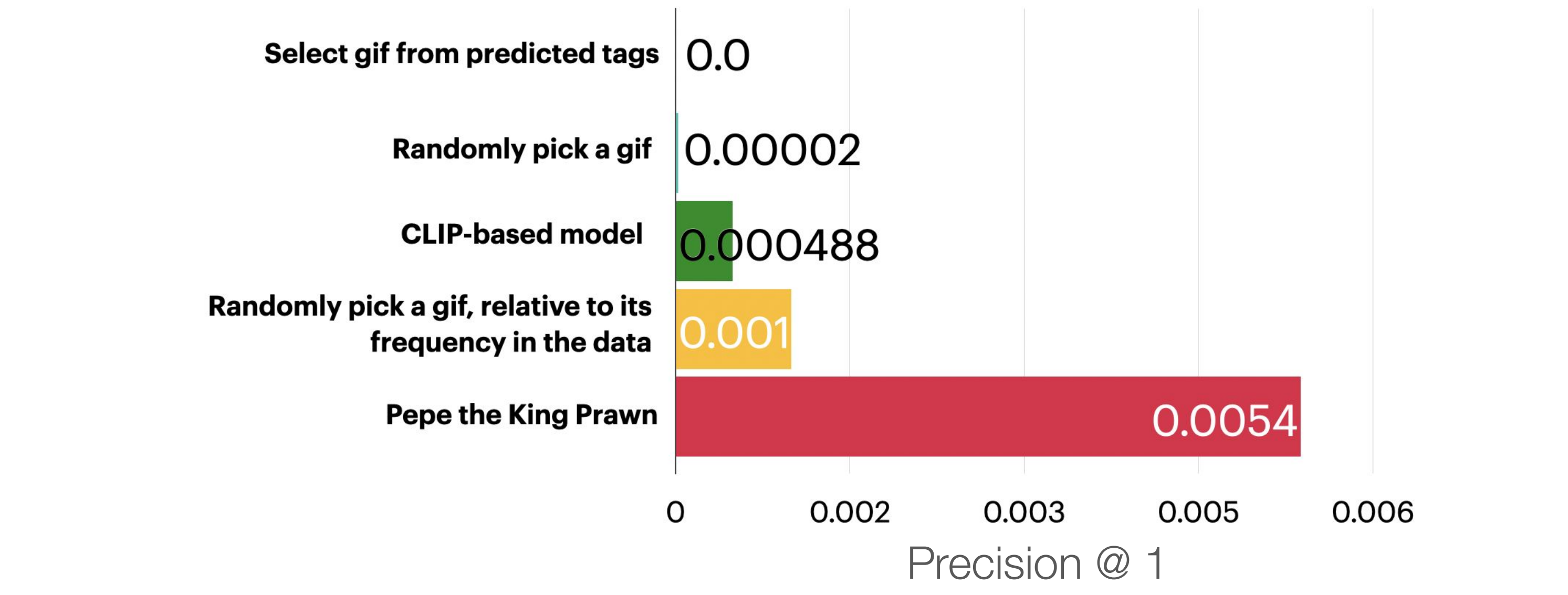

EfficientNet-b0 (shared weight)
Transformer Encoder Layer
feature vector for downstream task
1 x 512
4 selected frames
4 x 224 x 224 (after reshape)

Other baseline models:

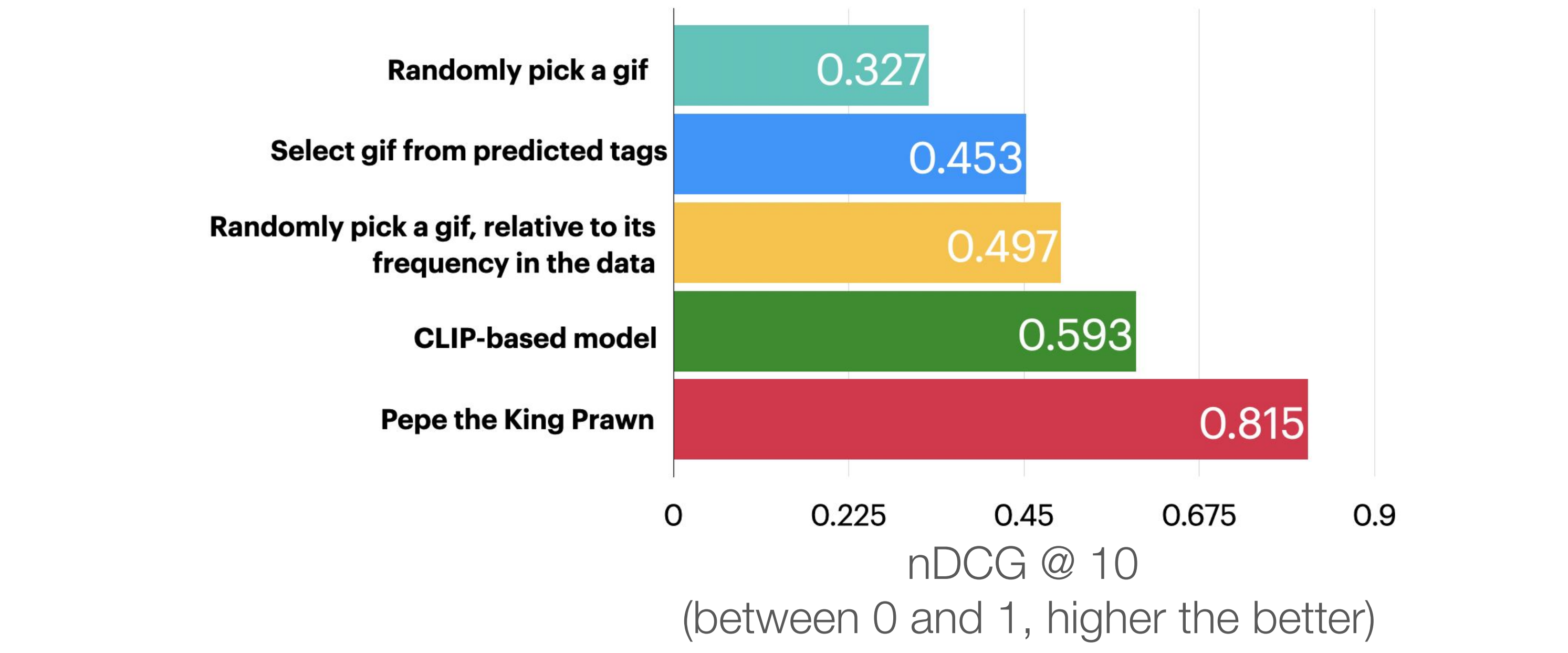
- **Random model**: randomly select a Gif ID
- **Distribution sampling**: sample Gif ID with probability proportional to its frequency (i.e. more frequently-used gif are sampled more often)

6. How good are our models?

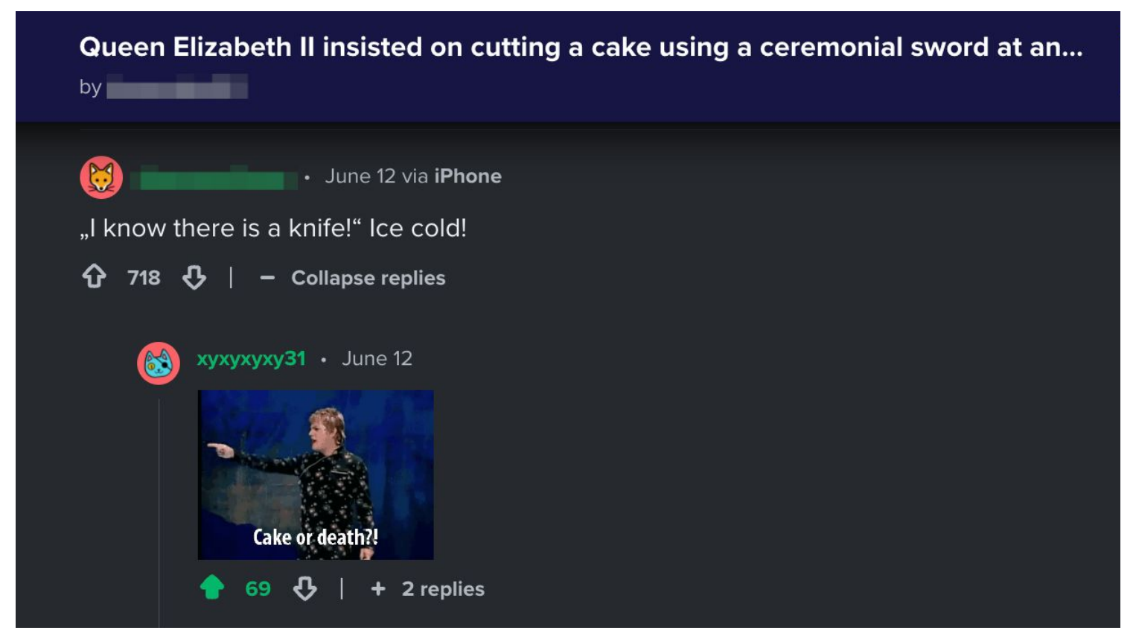
How accurate were the models in picking the **exact** gif someone used in the test set?



How good are the gifs used in the replies?
(better evaluation by annotating top-10 gifs of each model)



7. We deployed a Randomized Controlled Trial (RCT) in the real-world!



Queen Elizabeth II insisted on cutting a cake using a ceremonial sword at an...

by [redacted]

[redacted] • June 12 via iPhone

„I know there is a knife!“ Ice cold!

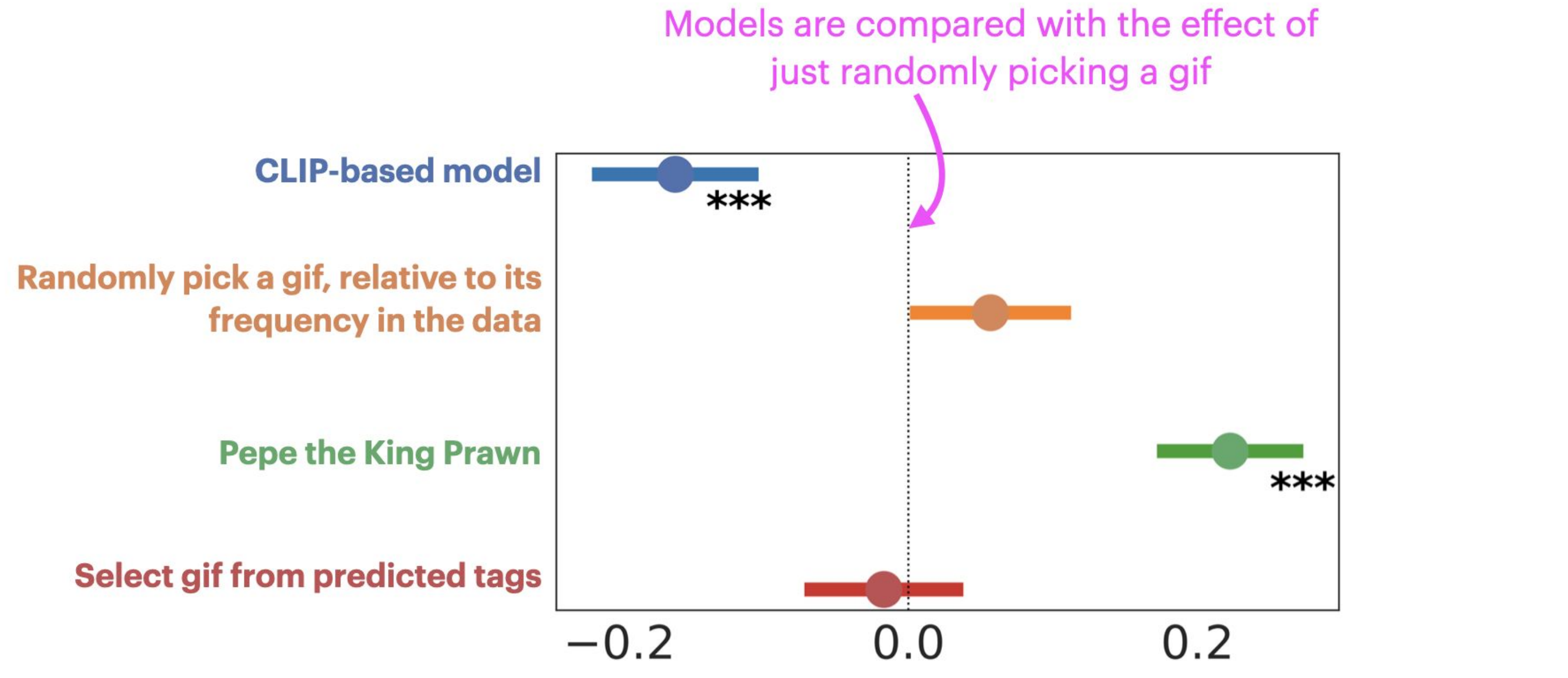
718 | − Collapse replies

xyxyxyxy31 • June 12

69 | + 2 replies

The RCT ran for ~5 months, and made 8,369 replies to users.

We run a regression on *score of the gif reply*, considering variables including model choice, parent comment topic, etc.



Negative Binomial regression coefficients (higher is better)
(bars show standard error and *** denotes significance at 0.01)

Code & Data: https://github.com/xingyaoww/gif-reply
Gif-Bot Slack App: https://github.com/xingyaoww/gif-reply-slack-bot