Supplementary Material for
Polysemous Visual-Semantic Embedding for Cross-Modal Retrieval

Yale Song
Microsoft Cloud & AI
yalesong@microsoft.com

Mohammad Soleymani
USC Institute for Creative Technologies
soleymani@ict.usc.edu

A. MRW Dataset

Our dataset consists of 50,107 video-sentence pairs collected from popular social media websites including reddit, Imgur, and Tumblr.[1] We crawled the data using the GIPHY API[2] with query terms `mrw`, `mfw`, `hifw`, `reaction`, and `reactiongif`; we crawled the data from August 2016 to March 2019. Table 1 shows the descriptive statistics of our dataset. We are continuously crawling the data, and plan to release updated versions in the future.

	Train	Validation	Test	Total
Number of video-sentence pairs	44,107	1,000	5,000	50,107
Average / median number of frames	104.91 / 72	209.04 / 179	209.55 / 178	117.43 / 79
Average / median number of words	11.36 / 10	15.02 / 14	14.79 / 13	11.78 / 11
Average / median word frequency	15.48 / 1	4.80 / 1	8.57 / 1	16.94 / 1
Vocabulary size	34,835	34,835	34,835	34,835

Table 1: Descriptive statistics of the MRW dataset.

A.1. Previous Work on Animated GIF

Note that most of the videos in our dataset have the animated GIF format. Technically speaking, animated GIFs and videos have different formats; the former is lossless, palette-based, and has no audio. In this paper, however, we use the two terms interchangeably because the distinction is unnecessary in our method. Below, to provide the context for our work, we briefly review previous work that focused on animated GIF.

There is increasing interest in conducting research around animated GIFs. Bakhshi *et al.* [2] studied what makes animated GIFs engaging on social networks and identified a number of factors that contribute to it: the animation, lack of sound, immediacy of consumption, low bandwidth and minimal time demands, the storytelling capabilities and utility for expressing emotions. Previous work in the computer vision and multimedia communities used animated GIFs for various tasks in video understanding. Jou *et al.* [11] propose a method to predict viewer perceived emotions for animated GIFs. Gygli *et al.* [9] propose the Video2GIF dataset for video highlighting, and further extended it to emotion recognition [8]. Chen *et al.* [3] propose the GIFGIF+ dataset for emotion recognition. Zhou *et al.* [18] propose the Image2GIF dataset for video prediction, along with a method to generate cinemagraphs from a single image by predicting future frames.

Recent work use animated GIFs to tackle the vision & language problems. Li *et al.* [13] propose the TGIF dataset for video captioning; Jang *et al.* [10] propose the TGIF-QA dataset for video visual question answering. Similar to the TGIF dataset [13], our dataset includes video-sentence pairs. However, our sentences are created by real users from Internet communities rather than study participants, thus posing real-world challenges. More importantly, our dataset has *implicit* concept association between videos and sentences (videos contain physical or emotional reactions to sentences), while the TGIF dataset has *explicit* concept association (sentences describe visual content in videos).

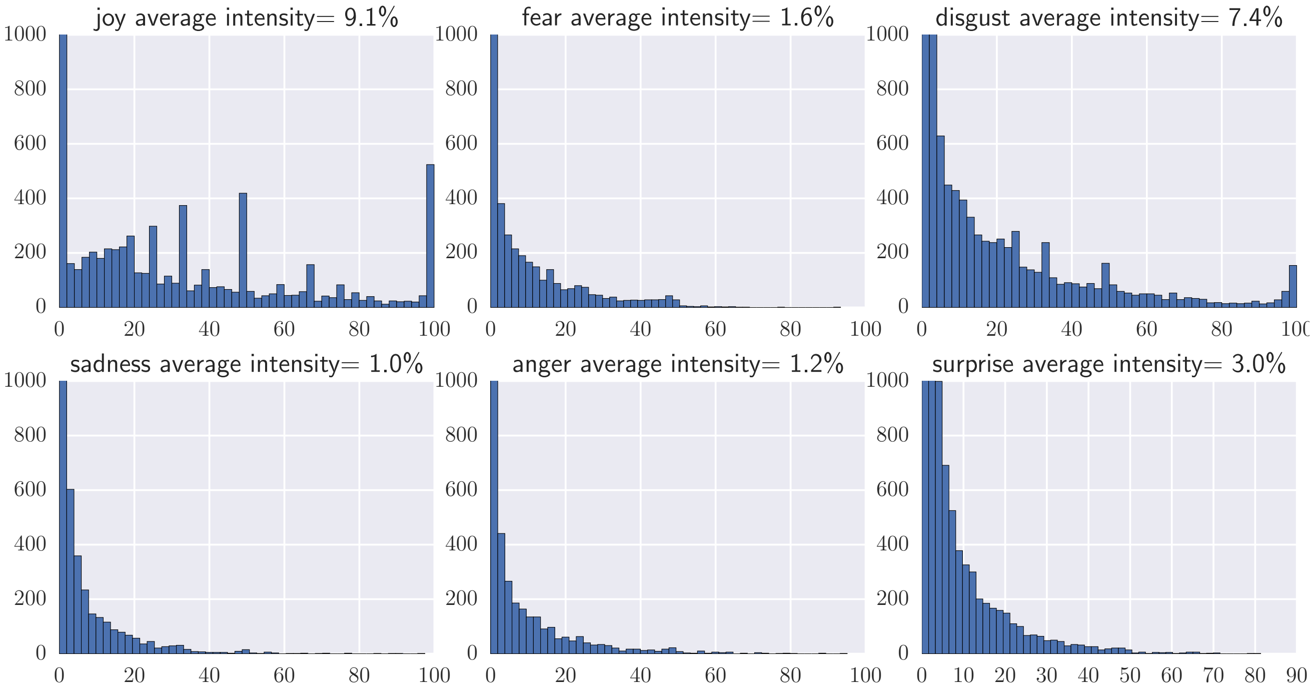


Figure 1. Histograms of the intensity of facial expressions. The horizontal axis represents the intensity of the detected expression, while the vertical axis is the sample count in frames with faces. We clip the y-axis at 1000 for visualization. Overall, joy, with average intensity of 9.1% and disgust (7.4%) are the most common facial expressions in our dataset.

A.2. Analysis of Facial Expressions

Facial expression plays an important role in our dataset: 6,380 samples contain the hashtag MFW (my face when), indicating that those GIFs contain emotional reactions manifested by facial expressions. To better understand the landscape of our dataset, we analyze the types of facial expressions contained in our dataset by leverage automatic tools.

First, we count the number of faces appearing in the animated GIFs. To do this, we applied the dlib CNN face detector [12] on five frames sampled from each animated GIF with an equal interval. The results show that there are, on average, 0.73 faces in a given frame of an animated GIF. Also, 34,052 animated GIFs contain at least one face. This means that 72% of our videos contain faces, which is quite significant. This suggests that employing techniques tailored specifically for face understanding could potentially improve performance on our dataset.

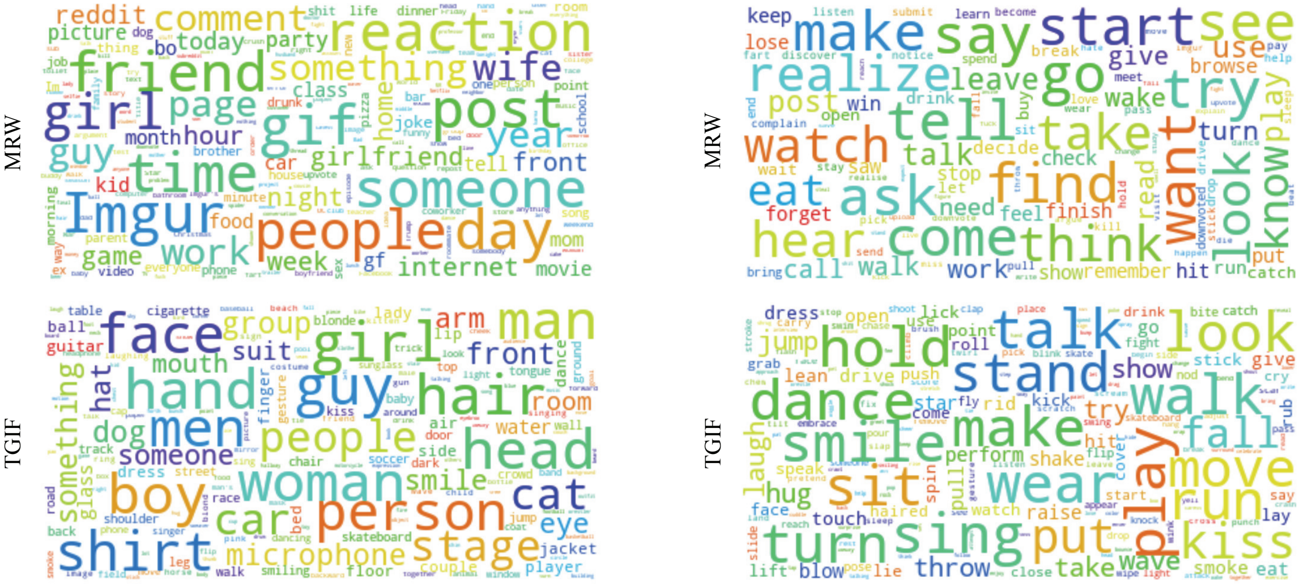
Next, we use the Affectiva Affdex [15] to analyze facial expressions depicted in the animated GIFs, detecting the intensity of expressions from two frames per second in each animated GIF. We looked at six expressions of basic emotions [4], namely, joy, fear, sadness, disgust, surprise and anger. We analyzed only the frames that contain a face with its bounding box region larger than 15% of the image. Figure 1 shows the results. Overall, joy with average intensity of 9.1% and disgust (7.4%) are the most common facial expressions in our dataset.

A.3. Comparison to the TGIF Dataset

Image and video captioning often involves describing objects and actions depicted explicitly in visual content [14, 13]. For reaction GIFs, however, visual-textual association is not always explicit. For example, as is the case in our dataset, objects and actions depicted in visual content might be a physical or emotional reaction to the scenario posed in the sentence.

In this section, we qualitatively compare our dataset with the TGIF dataset [13], which contains 120K video-sentence pairs for video captioning. We chose the dataset because both datasets contain videos that have the animated GIF format.

[1] https://www.reddit.com, https://imgur.com, https://www.tumblr.com
[2] https://developers.giphy.com



(a) Nouns

(b) Verbs

Figure 2. Distributions of nouns and verbs in our MRW and the TGIF [13] datasets. Compared to the TGIF dataset, words in our dataset depict more abstract concepts (e.g., post, time, day, start, realize, think, try), suggesting the ambiguous nature in our dataset.

We first compare words appearing in both datasets. Figure 2 shows word clouds of nouns and verbs extracted from our MRW dataset and the TGIF dataset [13]. Sentences in the TGIF dataset are constructed by crowdworkers to describe the visual content explicitly displayed in animated GIFs. Therefore, its nouns and verbs mainly describe physical objects, people and actions that can be visualized, e.g., cat, shirt, stand, dance. In contrast, MRW sentences are constructed by the Internet users, typically from subcommunities in social networks that focus on reaction GIFs. As can be seen from Figure 2, verbs and nouns in our MRW dataset additionally include abstract terms that cannot necessarily be visualized, e.g., time, day, realize, think. This shows that our dataset contains ambiguous terms and their associations, which pose significant challenges to cross-modal retrieval.

Next, we compare whether video-sentence associations are explicit/implicit in both datasets. To this end, we conducted a user study in which we asked six participants to verify the association between sentences and animated GIFs. We randomly sampled 100 animated GIFs from the test sets of both our dataset and TGIF dataset [13]. We paired each animated GIF with both its associated sentence and a randomly selected sentence from the corresponding dataset, resulting in 200 GIF-sentence pairs per dataset.

The results show that, in case of our dataset (MRW), 80.4% of the associated pairs are positively marked as being relevant, suggesting humans are able to distinguish the true vs. fake pairs despite implicit concept association. On the other hand, 50.7% of the randomly assigned sentences are also marked as matching sentences. The high false positive rate shows the ambiguous nature of GIF-sentence association in our dataset.

In contrast, for the TGIF dataset with clear explicit association, 95.2% of the positive pairs are correctly marked as relevant and only 2.6% of the irrelevant pairs are marked as being relevant. This human baseline demonstrates the challenging nature of GIF-sentence association in our dataset, due to their implicit rather than explicit association.

A.4. Application: Animated GIF Search For Social Media

Animated GIFs are becoming increasingly popular [2]; more people use them to tell stories, summarize events, express emotion, and enhance (or even replace) text-based communication. To reflect this trend, several social networks and messaging apps have recently incorporated GIF-related features into their systems, e.g., Facebook users can create posts and leave comments using GIFs, Instagram and Snapchat users can put “GIF stickers” into their personal videos, and Slack users can send messages using GIFs. This rapid increase in popularity and real-world demand necessitates more advanced and specialized systems for animated GIF search.

Current solutions to animated GIF search rely entirely on concept tags associated with animated GIFs and matching them with user queries. The tags are typically provided by users or produced by editors at companies like GIPHY. In the former case, noise becomes an issue; in the latter, it is expensive and would not scale well.

One of the motivations behind collecting our MRW dataset is to build a text-based animated GIF search engine, targeted for real-world scenarios mentioned above. Existing video captioning datasets, such as TGIF [13], are inappropriate for our purpose because of the explicit nature of visual-textual association, i.e., sentences simply describe what is being shown in videos. Rather, we need a dataset that captures various types of nuances used in social media, e.g., humor, irony, satire, sarcasm, incongruity, etc. Because our dataset provides video-text pairs with implicit visual-textual association, we believe that it has the potential to provide training data for building text-based animated GIF search engines targeted for social media.

To demonstrate the potential, we provide qualitative results on text-to-video retrieval using our dataset, shown in Figure 5. Each set of results show a query text and the top five retrieved videos, along with their ranks and cosine similarity scores. We would like the readers to take a close look at each set of results and decide which of the five retrieved videos depict the most likely visual response to the query sentence. The answers are provided below. For better viewing experience, we provide an HTML page with animated GIFs instead of static images. We strongly encourage the readers to check the HTML page to better appreciate the results. *(Answers: 3, 5, 2, 4, 1, 5, 4)*

B. Baseline Implementation Details

In the experiment section of the main paper, we provided baseline results for MS-COCO, TGIF, and MRW datasets. For MS-COCO, we provided previously reported results. For TGIF and MRW, on the other hand, we reported our own results because there has not been previous results on the datasets. Due to the space limit, we omitted implementation details of the baseline approaches; here we provide implementation details of the four baseline approaches: DeViSE [7], VSE++ [5], Order Embedding [17], and Corr-AE [6].

For fair comparison, all four baselines share the same video and sentence encoders as described in Section 3.1 of the main paper. The only difference is in the loss function we train the models with. Following the notation used in the main paper, we denote the output of the video and sentence encoders by $\phi(x)$ and $\phi(y)$, respectively. We employ the following loss functions for the baselines:

DeViSE [7]: We implement the conventional hinge loss in the triplet ranking setup, which penalizes the cases when the distance between positive pairs (i.e., the ground truth) is further away than negative pairs (e.g., randomly sampled) with a margin parameter ρ (we measure the cosine distance):

$$\mathcal{L}_{DeViSE} = \frac{1}{N} \sum_{i,j,k=1}^{N} \max\left(0, \rho - d(\phi(x_i), \phi(y_i)) + d(\phi(x_j), \phi(y_k))\right), \ \ \forall (i=j \vee i=k) \wedge j \neq k \tag{1}$$

VSE++ [5]: We implement the hard negative mining version of the conventional hinge loss triplet ranking loss, which is originally defined as:

$$\mathcal{L}_{VSE++} = \frac{1}{N} \sum_{i=1}^{N} \sum_{q=\{j,k\}} \max_{q} \max\left(0, \rho - d(\phi(x_i), \phi(y_i)) + d(\phi(x_j), \phi(y_k))\right), \ \ \forall (i=j \vee i=k) \wedge j \neq k \tag{2}$$

We have experimented with the original version and found that it fails to find a suitable solution to the objective, producing retrieval results that are almost identical to random guess. We suspect that the high noise present in both TGIF and MRW datasets makes the max function too strict as a constraint. We therefore replace the $\max_q$ function with a "filter" function that includes only highly-violating cases while ignoring others.

Intuitively, we implement the filter function to be an outlier detection function based on z-scores, where any z-score greater than 3 or less than -3 is considered to be an outlier. Specifically, we compute the z-scores for all of possible (i, j, k) combinations inside Equation 2 and discard instances if their absolute z-score is below 3.0. This way, we are considering multiple hard negatives instead of just one. We have empirically found this modification to be crucial to achieve reasonable performances on the TGIF and MRW datasets.

Order Embedding [17]: We used the original implementation provided by the authors of [17].

Corr-AE [6]: We implement the correspondence cross-modal autoencoder proposed by Feng *et al.* [6] (see Figure 4 in [6]). Given the encoder output $\phi(x)$ and $\phi(y)$, we build two autoencoders, one per modality, so that each autoencoder can reconstruct both $\phi(x)$ and $\phi(y)$. The autoencoders have four fully-connected layers with [512, 256, 256, 512] hidden units, respectively. Each of the fully connected layers is followed by a ReLU activation and a layer normalization [1].

Formally, a video autoencoder takes as input $\phi(x)$ and outputs $[\tilde{\phi}(x|x); \tilde{\phi}(y|x)]$, and a sentence autoencoder takes as input $\phi(y)$ and outputs $[\tilde{\phi}(x|y); \tilde{\phi}(y|y)]$. We then train the model with the following loss:

$$\mathcal{L}_{CorrAE} = \mathcal{L}_{DeViSE} + \frac{1}{N}\sum_{i=1}^{N} \sum_{c_i=\{x_i,y_i\}} \left(\|\phi(x_i) - \tilde{\phi}(x_i|c_i)\|_2^2 + \|\phi(y_i) - \tilde{\phi}(x_i|c_i)\|_2^2 \right) \quad (3)$$

We note that this loss is different from the original formulation of Corr-AE [6], where the first term in Equation 3 is replaced by a Euclidean loss, i.e., $\mathcal{L}_2 = \frac{1}{N}\sum_{i=1}^{N}\left(\|\phi(x_i) - \phi(y_i)\|_2^2\right)$. We found that using $\mathcal{L}_2$ instead of $\mathcal{L}_{DeViSE}$ makes the learning much harder, producing results that is almost identical to random guess.

C. Visualization of Multi-Head Self-Attention

C.1. Image-to-Text Retrieval Results on MS-COCO

Figure 3 shows examples of visual-textual attention maps on the MS-COCO dataset; the task is image-to-text retrieval. The first column shows query images with ground-truth sentences. Each of the other three columns shows visual (spatial) attention maps and their top-ranked text retrieval results, as well as their ranks and cosine similarity scores (green: correct, red: incorrect). We color-code words in the retrieved sentences according to their textual attention intensity values, normalized between [0, 1].

A glimpse at the results in each row shows that the three attention maps attend to different regions of the query image. Looking closely, we notice that salient regions are typically attended by multiple attention maps. For example, all three attention maps in Figure 3 highlight: (a) the photographer, (b) the bench, (c) the fruit stand, (e) the pink flowers, (f) the stop sign, (h) the woman, (j) the fire hydrant. However, this is not always the case: In Figure 3 (i), none of the attention maps highlights the most salient object, the black dog, and each attention map highlights different regions in the image. Even though all three attention maps do not "attend to" the dog, their top-ranked text retrieval results are still highly relevant to the query image; all three retrieved sentences have the word *dog* in them. This is possible because our PIE-Net computes embedding vectors by combining global context with locally-guided features. In this example, the global context provides information about the black dog, while each of the three locally-guided features contains region-specific information, specifically, (first map): the book shelf, (second map): the floor, (third map): the brown cushion.

The most interesting observation is that there are subtle variations in the retrieved sentences depending on where the visual attention is focused on. For example, in Figure 3 (a), the first result focuses on the photographer as a whole, the second focuses on the tiny camera (the visual attention is more narrowly focused on the photographer), and the third focuses on the pizza on the table (notice the visual attention on the table). In Figure 3 (d), the first result focuses on the ship, the second focuses on the building, and the third on an (imaginary) bird that could have been flying over the buildings. In Figure 3 (g), the first result focuses on the boat and the muddy water (notice visual attention on the muddy water region at the lower left corner), while the second focuses on the table of people (notice visual attention on the table region). In Figure 3 (j), the first results focuses on the fire hydrant and the yellow wall that is right behind the hydrant, while the second focuses on the hydrant as well as the building with two windows (notice now the visual attention is more widely spread out than the first result). We encourage the readers to look closely at Figure 3 to appreciate the subtle variations in the retrieved sentences depending on their corresponding visual attention.

C.2. Video-to-Text Retrieval Results on TGIF

Figure 4 shows examples of visual-textual attention maps on the TGIF dataset; the task is video-to-text retrieval. In each set of results, we show: (top) a query video and its ground-truth sentence, (bottom three rows): three visual (temporal) attention maps and their top-ranked text retrieval results, as well as their ranks and cosine similarity scores (green: correct, red: incorrect). We color-code words in the retrieval results according to their textual attention intensity values, normalized between [0, 1].

Similar to the results on MS-COCO, here we see that visual and textual attention maps tend to highlight salient video frames and words, respectively. Looking closely, we notice that the retrieved results tend to capture the concepts highlighted by their corresponding visual attention. For example, in **Figure 4 (a)**, the top ranked result contain "lady dressed in black" and "drinking a glass of wine", and the visual attention highlights both the early part of the video, where a woman is drinking from a bottle of whisky, and the latter part, where her black dress is shown. For the second ranked result, the visual attention no longer highlights the latter part, and the retrieved text focuses solely on drinking action (no mention of her black dress). In **Figure 4 (b)**, the top ranked result focuses on scoring a goal, while the second rank result also focus on the guy being hit in the face with the ball. Notice the difference of visual attention maps between the first and the second case.

C.3. Text-to-Video Retrieval Results on MRW

Figure 5 shows examples of text-to-video retrieval results on the MRW dataset. In each row, we show a query sentence and top five retrieved videos along with their ranks and cosine similarity scores. Unlike the previous two figures, here we do not directly show the ground-truth matches (but rather ask the readers to find them; we provide the answers above). The purpose of this is to emphasize the ambiguous and implicit nature of visual-textual association present in our dataset.

Most of the top five retrieved videos seem to be a good match to the query sentence. For example, Figure 5 (a) shows five videos that all contain a human face, each expressing subtly different emotions. Figure 5 (b) shows five videos that all contain an animal (squirrel, cat, etc), and most videos contain food. All five retrieved videos in Figure 5 show some form of awkward (dancing) moves.

We believe that the relatively poor retrieval performance reported in our main paper is partly explained by our qualitative results: visual-textual associations are highly ambiguous and there could be multiple correct matches. This calls for a different metric that measures the perceptual similarity between queries and retrieved results, rather than exact match. There has been some progress on perceptual metrics in the image synthesis literature (e.g., Inception Score [16]). We are not aware of a suitable perceptual metric for cross-modal retrieval, and this could be a promising direction for future research.

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[2] Saeideh Bakhshi, David Shamma, Lyndon Kennedy, Yale Song, Paloma de Juan, and Joseph'Jofish' Kaye. Fast, Cheap, and Good: Why Animated GIFs Engage Us. In *CHI*, 2016.

[3] Weixuan Chen, Ognjen Rudovic, and Rosalind W. Picard. GIFGIF+: Collecting Emotional Animated GIFs with Clustered Multi-Task Learning. In *ACII*, 2017.

[4] Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992.

[5] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: improved visual-semantic embeddings. In *BMVC*, 2017.

[6] Fangxiang Feng, Xiaojie Wang, and Ruifan Li. Cross-modal retrieval with correspondence autoencoder. In *ACM Multimedia*, 2014.

[7] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. Devise: A deep visual-semantic embedding model. In *NIPS*, 2013.

[8] Michael Gygli and Mohammad Soleymani. Analyzing and predicting GIF interestingness. In *ACM Multimedia*, 2016.

[9] Michael Gygli, Yale Song, and Liangliang Cao. Video2GIF: Automatic generation of animated GIFs from video. In *CVPR*, 2016.

[10] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. TGIF-QA: Toward spatio-temporal reasoning in visual question answering. In *CVPR*, 2017.

[11] Brendan Jou, Subhabrata Bhattacharya, and Shih-Fu Chang. Predicting viewer perceived emotions in animated GIFs. In *ACM Multimedia*, 2014.

[12] Davis E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009.

[13] Yuncheng Li, Yale Song, Liangliang Cao, Joel Tetreault, Larry Goldberg, Alejandro Jaimes, and Jiebo Luo. TGIF: A new dataset and benchmark on animated GIF description. In *CVPR*, 2016.

[14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[15] Daniel McDuff, Abdelrahman Mahmoud, Mohammad Mavadati, May Amr, Jay Turcot, and Rana el Kaliouby. Affdex sdk: A cross-platform real-time multi-face expression recognition toolkit. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '16, pages 3723–3726, New York, NY, USA, 2016. ACM.

[16] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *NIPS*, 2016.

[17] Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. Order-embeddings of images and language. In *ICLR*, 2016.

[18] Yipin Zhou, Yale Song, and Tamara L Berg. Image2GIF: Generating cinemagraphs using recurrent deep q-networks. In *WACV*, 2018.

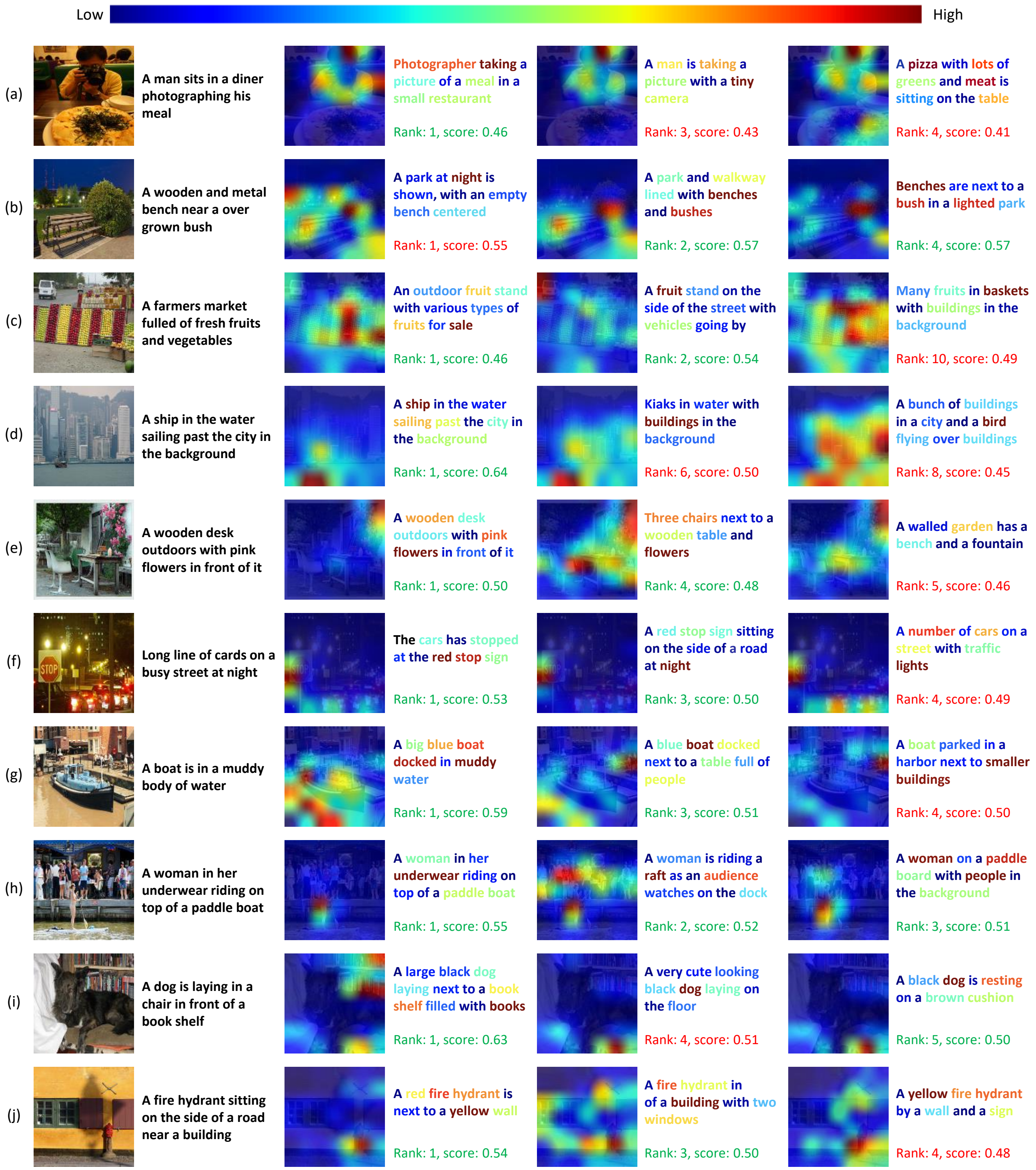

Low
High
(a)
A man sits in a diner photographing his meal
Photographer taking a picture of a meal in a small restaurant
Rank: 1, score: 0.46
A man is taking a picture with a tiny camera
Rank: 3, score: 0.43
A pizza with lots of greens and meat is sitting on the table
Rank: 4, score: 0.41
(b)
A wooden and metal bench near a over grown bush
A park at night is shown, with an empty bench centered
Rank: 1, score: 0.55
A park and walkway lined with benches and bushes
Rank: 2, score: 0.57
Benches are next to a bush in a lighted park
Rank: 4, score: 0.57
(c)
A farmers market fulled of fresh fruits and vegetables
An outdoor fruit stand with various types of fruits for sale
Rank: 1, score: 0.46
A fruit stand on the side of the street with vehicles going by
Rank: 2, score: 0.54
Many fruits in baskets with buildings in the background
Rank: 10, score: 0.49
(d)
A ship in the water sailing past the city in the background
A ship in the water sailing past the city in the background
Rank: 1, score: 0.64
Kiaks in water with buildings in the background
Rank: 6, score: 0.50
A bunch of buildings in a city and a bird flying over buildings
Rank: 8, score: 0.45
(e)
A wooden desk outdoors with pink flowers in front of it
A wooden desk outdoors with pink flowers in front of it
Rank: 1, score: 0.50
Three chairs next to a wooden table and flowers
Rank: 4, score: 0.48
A walled garden has a bench and a fountain
Rank: 5, score: 0.46
(f)
Long line of cards on a busy street at night
The cars has stopped at the red stop sign
Rank: 1, score: 0.53
A red stop sign sitting on the side of a road at night
Rank: 3, score: 0.50
A number of cars on a street with traffic lights
Rank: 4, score: 0.49
(g)
A boat is in a muddy body of water
A big blue boat docked in muddy water
Rank: 1, score: 0.59
A blue boat docked next to a table full of people
Rank: 3, score: 0.51
A boat parked in a harbor next to smaller buildings
Rank: 4, score: 0.50
(h)
A woman in her underwear riding on top of a paddle boat
A woman in her underwear riding on top of a paddle boat
Rank: 1, score: 0.55
A woman is riding a raft as an audience watches on the dock
Rank: 2, score: 0.52
A woman on a paddle board with people in the background
Rank: 3, score: 0.51
(i)
A dog is laying in a chair in front of a book shelf
A large black dog laying next to a book shelf filled with books
Rank: 1, score: 0.63
A very cute looking black dog laying on the floor
Rank: 4, score: 0.51
A black dog is resting on a brown cushion
Rank: 5, score: 0.50
(j)
A fire hydrant sitting on the side of a road near a building
A red fire hydrant is next to a yellow wall
Rank: 1, score: 0.54
A fire hydrant in of a building with two windows
Rank: 3, score: 0.50
A yellow fire hydrant by a wall and a sign
Rank: 4, score: 0.48

Figure 3. **Image-to-text retrieval results on MS-COCO.** For each query image we show three visual attention maps and their top-ranked text retrieval results, along with their ranks and cosine similarity scores (green: correct, red: incorrect). Words in each sentence is color-coded with textual attention intensity, using the color map shown at the top.

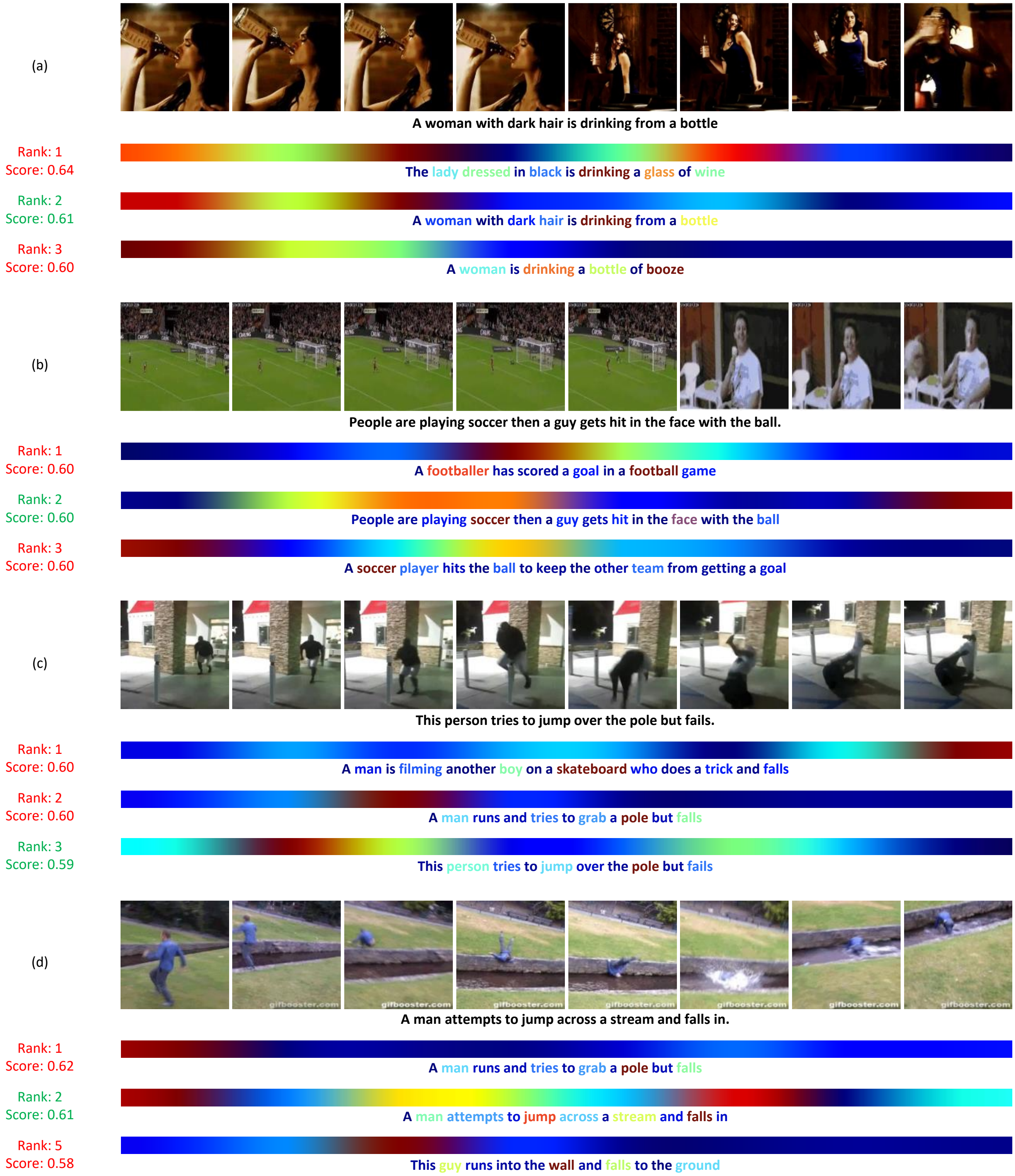

(a)
A woman with dark hair is drinking from a bottle
Rank: 1
Score: 0.64
The lady dressed in black is drinking a glass of wine
Rank: 2
Score: 0.61
A woman with dark hair is drinking from a bottle
Rank: 3
Score: 0.60
A woman is drinking a bottle of booze
(b)
People are playing soccer then a guy gets hit in the face with the ball.
Rank: 1
Score: 0.60
A footballer has scored a goal in a football game
Rank: 2
Score: 0.60
People are playing soccer then a guy gets hit in the face with the ball
Rank: 3
Score: 0.60
A soccer player hits the ball to keep the other team from getting a goal
(c)
This person tries to jump over the pole but fails.
Rank: 1
Score: 0.60
A man is filming another boy on a skateboard who does a trick and falls
Rank: 2
Score: 0.60
A man runs and tries to grab a pole but falls
Rank: 3
Score: 0.59
This person tries to jump over the pole but fails
(d)
A man attempts to jump across a stream and falls in.
Rank: 1
Score: 0.62
A man runs and tries to grab a pole but falls
Rank: 2
Score: 0.61
A man attempts to jump across a stream and falls in
Rank: 5
Score: 0.58
This guy runs into the wall and falls to the ground

Figure 4. **Video-to-text retrieval results on TGIF.** For each query video we show three visual attention maps and their top-ranked text retrieval results, along with their ranks and cosine similarity scores (green: correct, red: incorrect). Words in each sentence is color-coded with textual attention intensity.

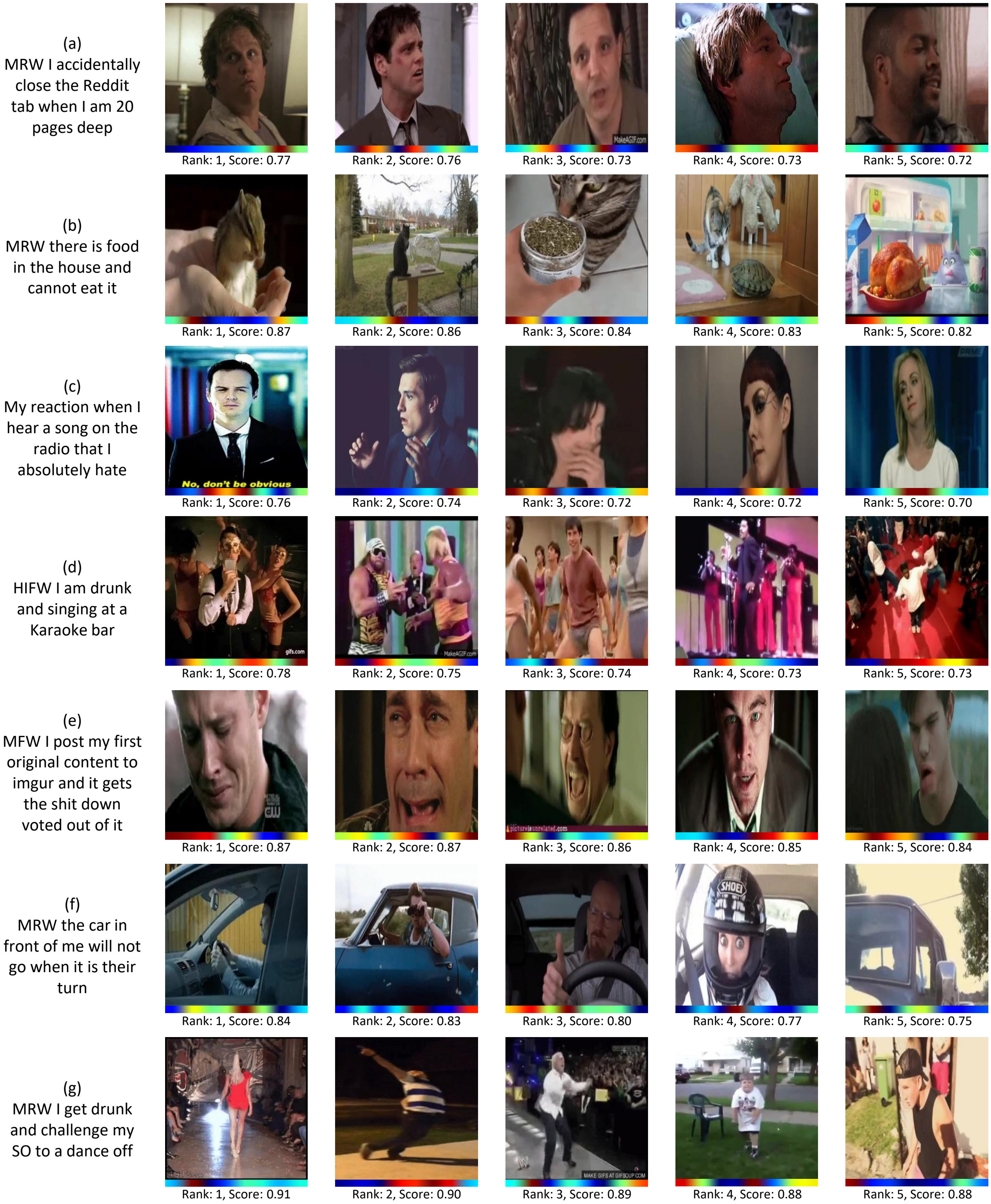


Figure 5. **Text-to-video retrieval results on MRW.** For each query sentence we show top five retrieved videos, along with their visual (temporal) attention maps, rank, and cosine similarity scores. For better viewing, we provide an HTML file with animated GIFs instead of static images. *Quiz:* We encourage the readers to find the best matching video in each set of results (see the text for answers).