Not (yet) the whole story: Evaluating Visual Storytelling Requires More than Measuring Coherence, Grounding, and Repetition

Aditya K Surikuchi | Raquel Fernández | Sandro Pezzelle
Institute for Logic, Language and Computation, University of Amsterdam

EMNLP 2024

Visual Storytelling

Input: sequence of images

Task: to generate a textual story consistent with the input

Human-annotated story: *We invited lots of friends for a barbeque. The fire pit was very large. We roasted hot dogs right over the flame. Lots of people were happy. And there was a lot of beer too.*

Evaluation is challenging: plausibility of several creative stories for a single given image sequence, makes reference-based NLG metrics (e.g., METEOR) inappropriate.

Reference-free Evaluation Metrics

Coherence–RoViST-**C**[1]: average probability with which each sentence follows the preceding sentences (*entire prefix*) of the story; range ∈ [0, 1]

Visual grounding–GROOViST[2]: alignment scores between noun-phrases and image regions (*using CLIP*); penalization of low alignment scores and re-weighting using concreteness ratings; normalized and aggregated to range ∈ [−1, 1]

Repetition–RoViST-N**R**[1]: number of co-occurring words between two texts normalized by the total number of words in both texts (*Jaccard Similarity*); for every sentence average of inter- and intra-sentence repetition is computed; range ∈ [0, 1]

Q. Can we combine these metrics to determine how human-like a model-generated story is?

We take a human-centric approach and define the quality of model-generated stories in terms of their **closeness** to corresponding stories produced by humans, along the three different evaluation dimensions:

abs(**C**[human story] - **C**[model generated story]) = $\mathbf{d_C}$

abs(**G**[human story] - **G**[model generated story]) = $\mathbf{d_G}$

abs(**R**[human story] - **R**[model generated story]) = $\mathbf{d_R}$

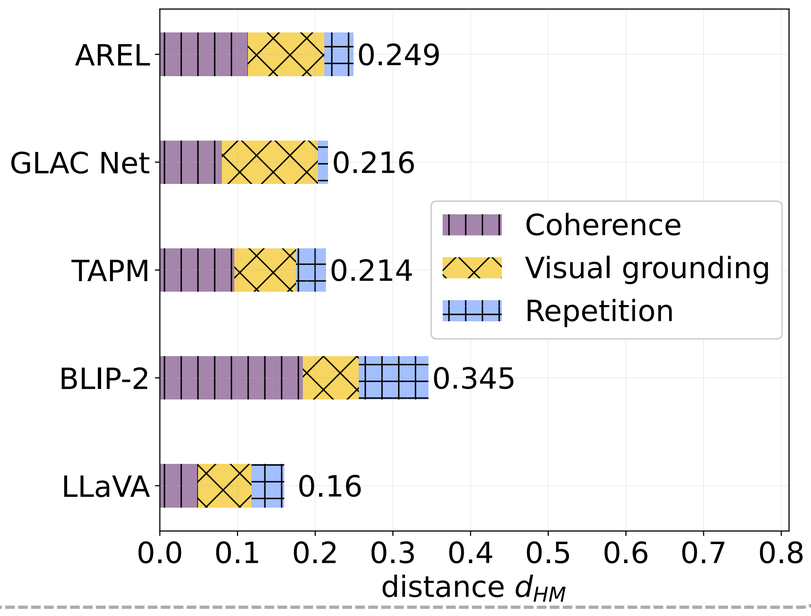
avg(...) = $\mathbf{d_{HM}}$

The lower the d_{HM}, the better.

Q. How do models perform on the d_{HM} measure?

On the VIST[3] test set, we evaluate three models designed and trained for visual storytelling: AREL[4], GLAC Net[5], TAPM[6]; and two general-purpose foundation models in a zero-shot manner: BLIP-2[7] and LLaVA[8].

Prompt: '*[INST]<image>\nWrite a story using exactly five sentences for this image sequence. Do not use more than five sentences. [/INST]*'



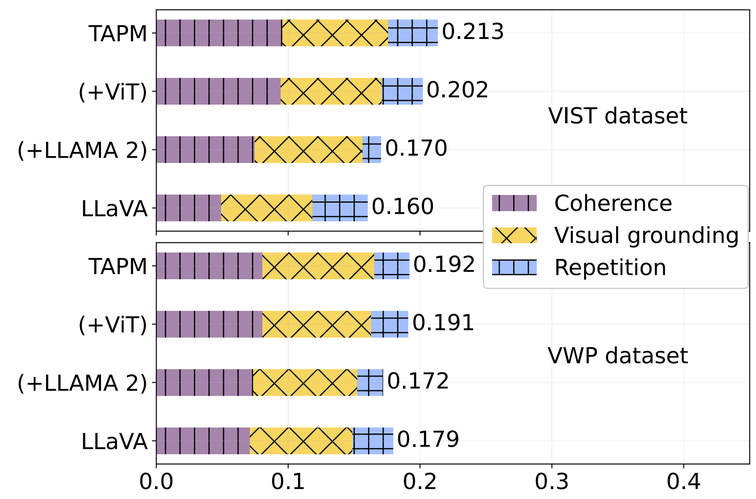
LLaVA obtains the best d_{HM} followed by TAPM.

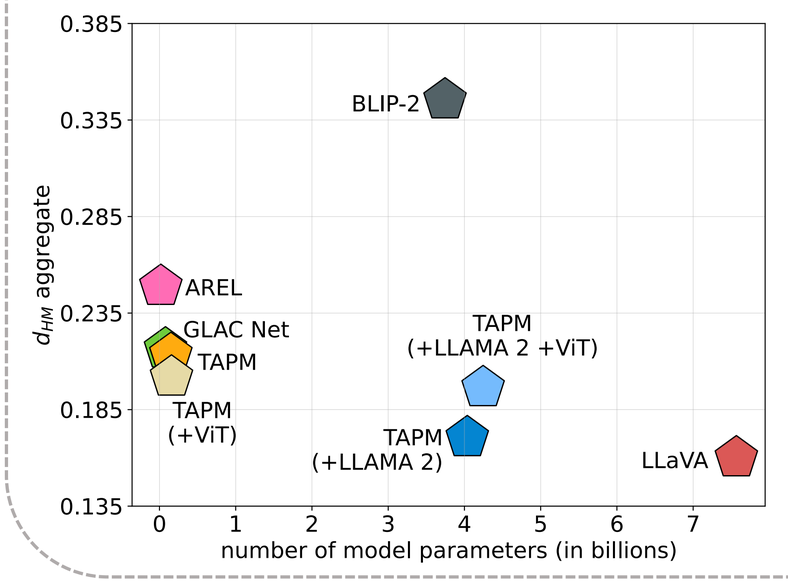
Despite being **50** times smaller than LLaVA, TAPM's d_{HM} is only slightly worse than LLaVA's.

Improvements to TAPM

LLaVA obtains better d_C and d_G compared to TAPM. So, we test whether we can obtain better results (lower distances), by **replacing TAPM's original language and vision components** with models comparable to those embedded in LLaVA, while **keeping the number of parameters significantly lower**.

With the updated language component, TAPM is on-par with LLaVA in terms of the overall d_{HM} value.



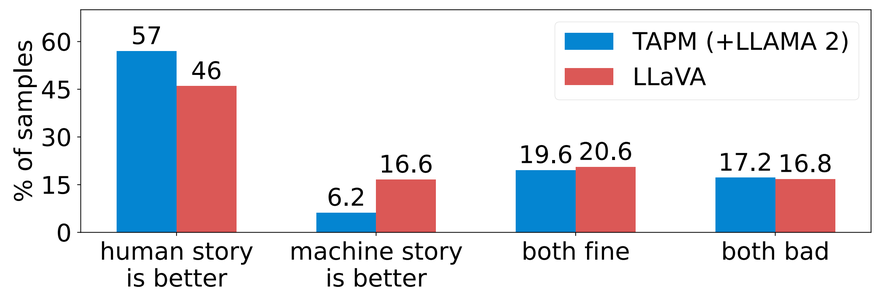


Improved TAPM is still significantly smaller in size compared to LLaVA.

Qualitative Analysis

Results using the d_{HM} measure suggest that the stories generated by the best-performing models—TAPM (+LLAMA 2)[9] and LLaVA—are very close to human-levels of visual grounding, coherence, and degree of repetition.

We ask human annotators (unrelated to the task) to assess the stories generated by these models against corresponding human-written stories.



Humans consistently preferred human-written stories over stories generated by the models!

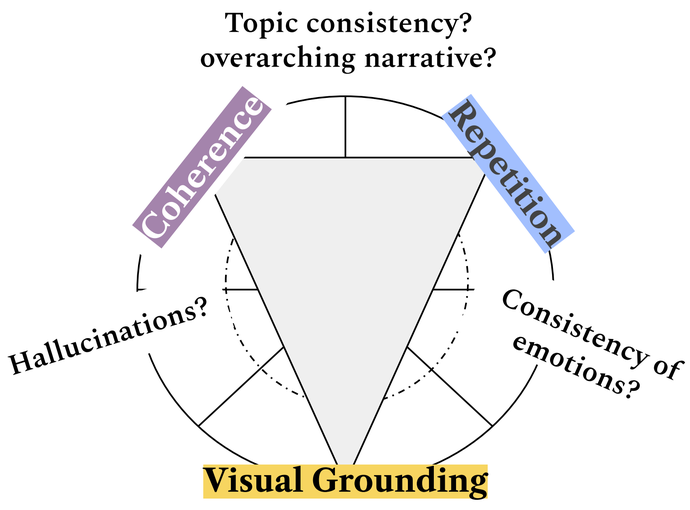
Human-annotated story: *Our business hosted its winter holiday party at a hotel. The food was served as a buffet. All the young workers sat at one end of the table. The older co-workers sat at the other end of the table. In the end, both young and old co-workers stayed until closing time for the hotel restaurant.*

Model-generated story: *The first image, a man stands in awe of a beautifully decorated Christmas tree. The second image captures a lively dinner party, with guests seated around a table laden with food and wine. The third image shows a bustling restaurant kitchen, where chefs are busy preparing meals. The fourth image reveals a cozy dining room, where a family enjoys a quiet meal together. The final image presents a grand banquet hall, filled with guests and adorned with elegant decorations.*

Takeaways

⚠ A 'good' story may require more than human-like level of visual grounding, coherence, and repetition.

🧐 There are several other dimensions of visual story evaluation that need to be considered and explored further.



References

[1]RoViST: Learning Robust Metrics for Visual Storytelling (Wang et al., NAACL Findings 2022)
[2]GROOViST: A Metric for Grounding Objects in Visual Storytelling (Surikuchi et al., EMNLP 2023)
[3]Visual Storytelling (Huang et al., NAACL 2016)
[4]No Metrics Are Perfect: Adversarial Reward Learning for Visual Storytelling (Wang et al., ACL 2018)
[5]GLAC Net: GLocal Attention Cascading Networks for Multi-image Cued Story Generation (Kim et al., 2018)
[6]Transitional Adaptation of Pretrained Models for Visual Storytelling (Yu et al., CVPR 2021)
[7]BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models (Li et al., PMLR 2023)
[8]LLaVA-NeXT: Improved reasoning, OCR, and world knowledge (Liu et al., 2024)
[9]Llama 2: Open Foundation and Fine-Tuned Chat Models (Touvron et al., 2023)