On the Challenges in Evaluating Visually Grounded Stories

Aditya K Surikuchi Raquel Fernández Sandro Pezzelle

University of Amsterdam
Institute for Logic, Language and Computation

Visually Grounded Story Generation

Visually Grounded Story Generation

Ql, ? k_l bsaiq dmp amtcp, ⇔cpc _pc `sjjcr fmjcq gl rfc u_jj lc_p fgk*_lb fc amtcpq fgq d_ac gl dc_p, Fc rpgcq rm ecr_u_w* `sr_lmrfcp ncpqml ep_`q fgk dpmk `cfglb, Fc fmjbq fgk gl nj_ac _q rfc k_l qrpseejcq rm ecr_u_w, Qsbbcljw* rfc `sjjcrq qr_pr ^wgle, @mrf kcl bsai mp bgtc dmp q_dcrw, Dmmb_lb `grq md nj_qrcp em ^wgle_q rfc `sjjcrq ´jj rfc pmmk, Rum kcl _pc fgr*_lb rfcw `mrf em ^wgle `_ai,

Visually Grounded Story Generation

Q*l*, *? k_ l bsaiq dmp amtcp,* ⇔ *cpc _pc `sjjcr fmjcq gl rfc u_jj lc_p fgk*_ lb fc amtcpq fgq d_ac gl dc_p, Fc rpgcq rm ecr _u_w*` sr _ lmrfcp ncpqml ep_`q fgk dpmk `cfglb, Fc fmjbq fgk gl nj_ac _q rfc k_ l qrpseejcq rm ecr _u_w, Qsbbcljw*rfc `sjjcrq qr_pr ^wgle, @mrf kcl bsai mp bgtc dmp q_dcrw, Dmmb _ lb `grq md nj_qrcp em ^wgle _q rfc `sjjcrq ´jj rfc pmmk, Rum kcl _pc fgr*_ lb rfcw `mrf em ^wgle `_ai,*

Q0, *Gl rfc mnsjclr pmmk*_ k_ l gl _ rsvcbm u_q nj_wgle rfc _aampbgml* m`jgtgmsq rm rfc af_mq sldmjbgle _pmslb fgk, Qsbbcljw*_ lmrfcp k_ l `spqr glrm rfc pmmk* fgq d_ac amlrmprcb ugrf _lecp _q fc jslecb _r rfc ´pqr k_ l, ⇔ c qcamlb k_ lęq _↓_ai u_q kcr ugrf _ qug← amslrcp dpmk rfc ´pqr* ufm k_ l_ecb rm nsqf fgk `_ai, ?q rfcw qrpseejcb*_ ajmsb md bsqr ´jjcb rfc _gp* m`qaspgle rfcgp kmtckclrq, Gl rfc kgbqr md rfc amk kmrgml*_ rfgpb k_ l _nnc_pcb* fgq cvnpcqqgml mlc md qfmai _lb dc_p _q fc u_rafcb rfc rum kcl ep_nnjc,*

Visually Grounded Story Generation

🗣 ? k_l bsaiq dmp amtcp, ⇔cpc _pc `sjjcr fmjcq gl rfc u_jj lc_p fgk*_lb fc amtcpq fgq d_ac gl dc_p, Fc rpgcq rm ecr _u_w* `sr _lmrfcp ncpqml ep_`q fgk dpmk `cfglb, Fc fmjbq fgk gl nj_ac _q rfc k_l qrpseejcq rm ecr _u_w, Qsbbcljw* rfc `sjjcrq qr_pr ^wgle, @mrf kcl bsai mp bgtc dmp q_dcrw, Dmmb _lb `grq md nj_qrcp em ^wgle _q rfc `sjjcrq ´jj rfc pmmk, Rum kcl _pc fgr*_lb rfcw `mrf em ^wgle `_ai,

Gl rfc mnsjclr pmmk*_ k_l gl _rsvcbm u_q nj_wgle rfc _aampbgml* m`jgtgmsq rm rfc af_mq sldmjbgle _pmslb fgk, Qsbbcljw*_lmrfcp k_l `spqr glrm rfc pmmk* fgq d_ac amlrmprcb ugrf _lecp _q fc jslecb _r rfc ´pqr k_l, ⇔c qcamlb k_lęq_↓_ai u_q kcr ugrf _qug← amslrcp dpmk rfc ´pqr* ufm k_l_ecb rm nsqf fgk `_ai, ?q rfcw qrpseejcb*_ ajmsb md bsqr ´jjcb rfc _gp* m`qaspgle rfcgp kmtckclrq, Gl rfc kgbqr md rfc amk kmrgml*_rfgpb k_l _nnc_pcb* fgq cvnpcqqgml mlc md qfmai _lb dc_p _q fc u_rafcb rfc rum kcl ep_nnjc,

Datasets

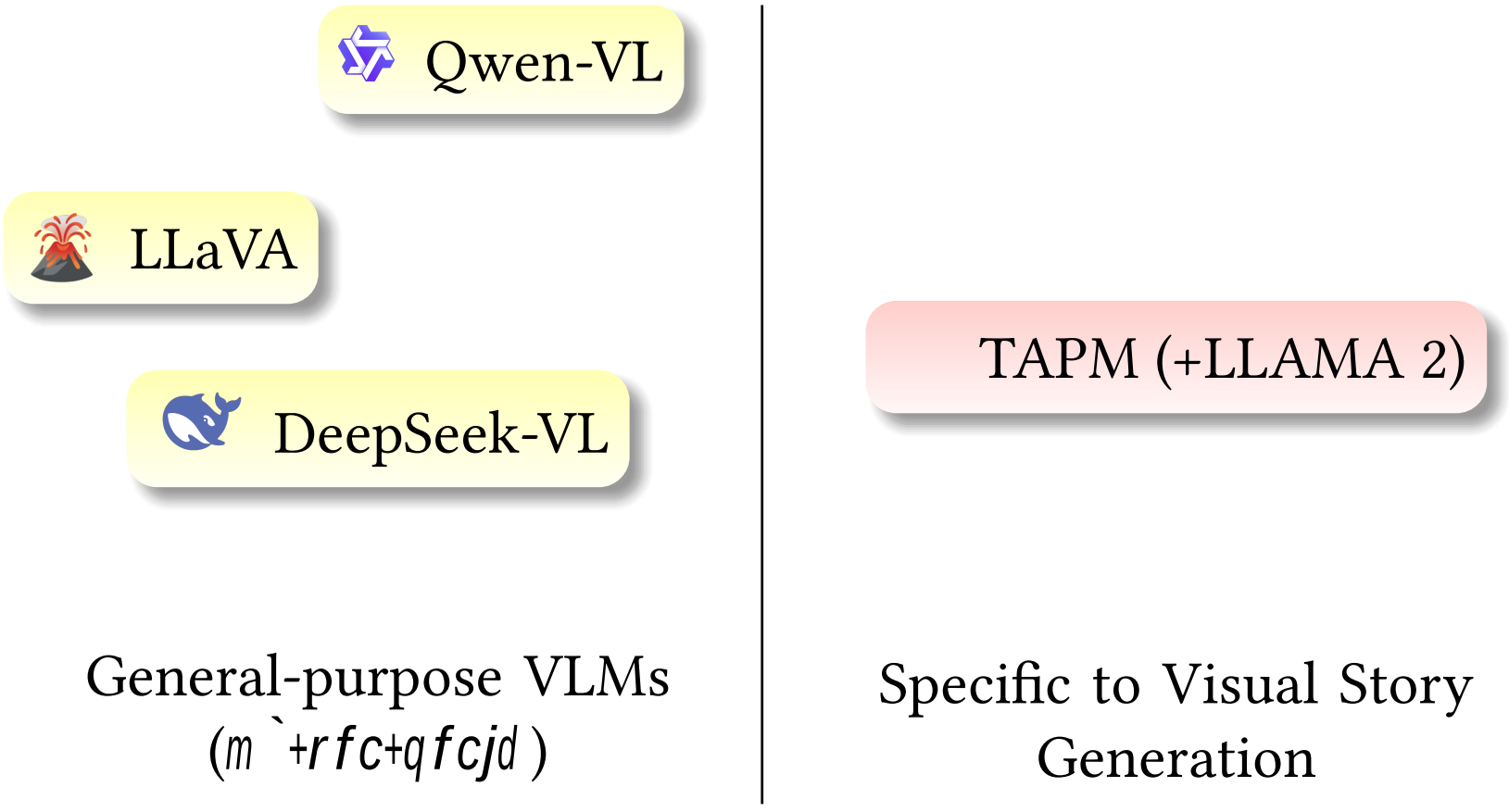
TGQR

- Sequences constructed using images from Flickr albums.
- Lacks consistency of entities (e.g., *af_p_arcpq*, *m`hcarq*)
- Corresponding stories are generally descriptive in nature

TUN

- Sequences constructed using scenes from movies
- Semantically well-connected with recurring characters
- Stories contain diverse entities, are longer, and coherent

Models


Qwen-VL
LLaVA
DeepSeek-VL
TAPM (+LLAMA 2)
General-purpose VLMs
(m `+rfc+qfcjd)
Specific to Visual Story
Generation

Evaluation

Human evaluation is challenging in terms of:

- Scalability and costs
- Selecting qualified annotators and reliability

Evaluation

Human evaluation is challenging in terms of:

- Scalability and costs
- Selecting qualified annotators and reliability

Automatic Metrics:

- BLEU, METEOR, CIDEr, SPICE, ROUGE

Evaluation

Human evaluation is challenging in terms of:

- Scalability and costs
- Selecting qualified annotators and reliability

Automatic Metrics:

- ~~BLEU, METEOR, CIDEr, SPICE, ROUGE~~
- Reference-free metrics that assess stories along different aspects—*Amfcpclac*, *Tgqs_jepmslbgle*, *Pcncrgrgmll*

Evaluation

Human evaluation is challenging in terms of:

- Scalability and costs
- Selecting qualified annotators and reliability

Automatic Metrics:

- ~~BLEU, METEOR, CIDEr, SPICE, ROUGE~~
- Reference-free metrics that assess stories along different aspects—*Amfcpclac*, *Tgqs_jepmslbgle*, *Pcncrgrgmnl*
- d_{FK} measure that computes distances between individual metric scores of model- and corresponding human stories

Results

What is overall best performing model?

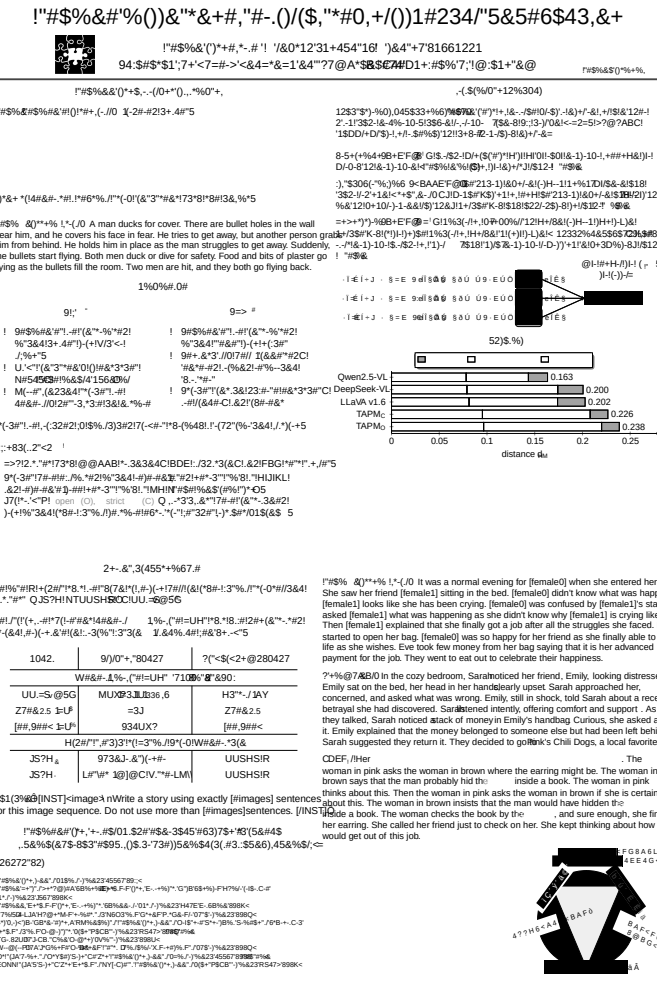
Results

What is overall best performing model?

How did the models fare along each of the 3 dimensions?

"SF UIFSF PUIFS EJNFOTJPOT SFMFW
WJTVBMMZ HSPVOEFE TUPSJFT

To learn more, please stop by our poster:



Thanks for your attention! ✨