On the Challenges in Evaluating Visually Grounded Stories

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

Text2Story 2025

Visually Grounded Story Generation

Input: sequence of temporally-ordered images.

Task: to generate a textual story consistent with the input.

Human-written story: A man ducks for cover. There are bullet holes in the wall near him, and he covers his face in fear. He tries to get away, but another person grabs him from behind. He holds him in place as the man struggles to get away. Suddenly, the bullets start flying. Both men duck or dive for safety. Food and bits of plaster go flying as the bullets fill the room. Two men are hit, and they both go flying back.

Datasets

VIST[1]

- ◇ Sequences are constructed using images from Flickr albums.
- ◇ Lacks consistency of entities (e.g., *characters, objects*)
- ◇ Corresponding stories are generally descriptive in nature

VWP[2]

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

Stories are provided by qualified workers through a crowdsourcing platform.

VGSG challenge[3]

- VWP dataset with 11778 training, 849 validation, and 586 test samples.
- Stories were evaluated using reference-based metrics such as METEOR and reference-free metrics such as CM (*character matching*).
- Two tracks: `open (O), strict (C)`—participants were constrained from using other visual feature extractors besides *SwinTransformer*.

In this work, we evaluate various models on the VGSG challenge dataset, using the recently proposed d_{HM}[4] method.

Modeling Approaches

We use 2 models that are shown to perform well on other visual storytelling datasets—TAPM (+LLAMA 2)[4], LLaVAv1.6[5]

We also compare two recent general-purpose VLMs that have demonstrated strong performance on various vision-language benchmarks.

Model	Vision Encoder	Language Decoder
General-purpose VLMs *(off-the-shelf)*		
LLaVAv1.6	CLIP-ViT-L-336px	Mistral-7B
Qwen2.5-VL[6]	ViT	Qwen2.5
DeepSeek-VL[7]	SigLIP	DeepSeek
Models specific to Visual Story Generation		
TAPM$_C$	SwinTransformer	LLAMA 2
TAPM$_O$	ResNet-101, FasterRCNN	LLAMA 2

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

The general-purpose VLMs were not directly pre-trained using any of the visual storytelling datasets, including VWP.

References

[1] Visual Storytelling (Huang et al., NAACL 2016)
[2] Visual Writing Prompts: Character-Grounded Story Generation with Curated Image Sequences (Hong et al., TACL 2023)
[3] Visually Grounded Story Generation Challenge (Hong et al., INLG GenChal 2023)
[4] Not (yet) the whole story: Evaluating Visual Storytelling Requires More than Measuring Coherence, Grounding, and Repetition (Surikuchi et al., EMNLP 2024 *findings*)
[5] LLaVA-NeXT: Improved reasoning, OCR, and world knowledge (Liu et al., 2024)
[6] Qwen2.5-VL Technical Report (Bai et al., 2025)
[7] DeepSeek-VL: Towards Real-World Vision-Language Understanding (Lu et al., 2024)
[8] RoViST: Learning Robust Metrics for Visual Storytelling (Wang et al., NAACL 2022 *findings*)
[9] GROOViST: A Metric for Grounding Objects in Visual Storytelling (Surikuchi et al., EMNLP 2023)

Evaluation Methods

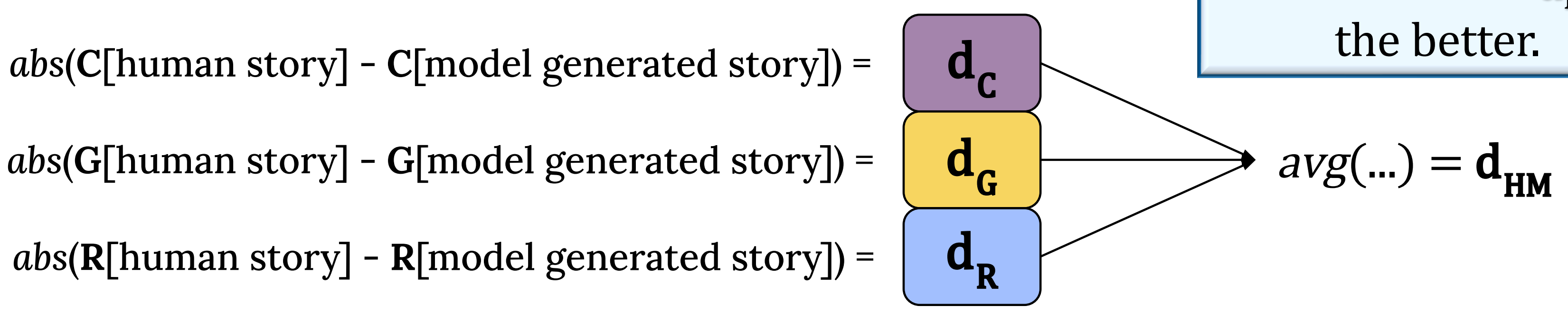
Evaluation is challenging. Plausibility of several creative stories for a single given image sequence, makes reference-based NLG metrics (e.g., METEOR) inappropriate for evaluating model-generated stories.

Coherence–RoViST-**C**[8]: average probability with which each sentence follows the preceding sentences (*entire prefix*) of the story; range ∈ [0, 1]

Visual grounding–GROOViST[9]: 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**[8]: 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]

The lower the d_{HM}, the better.



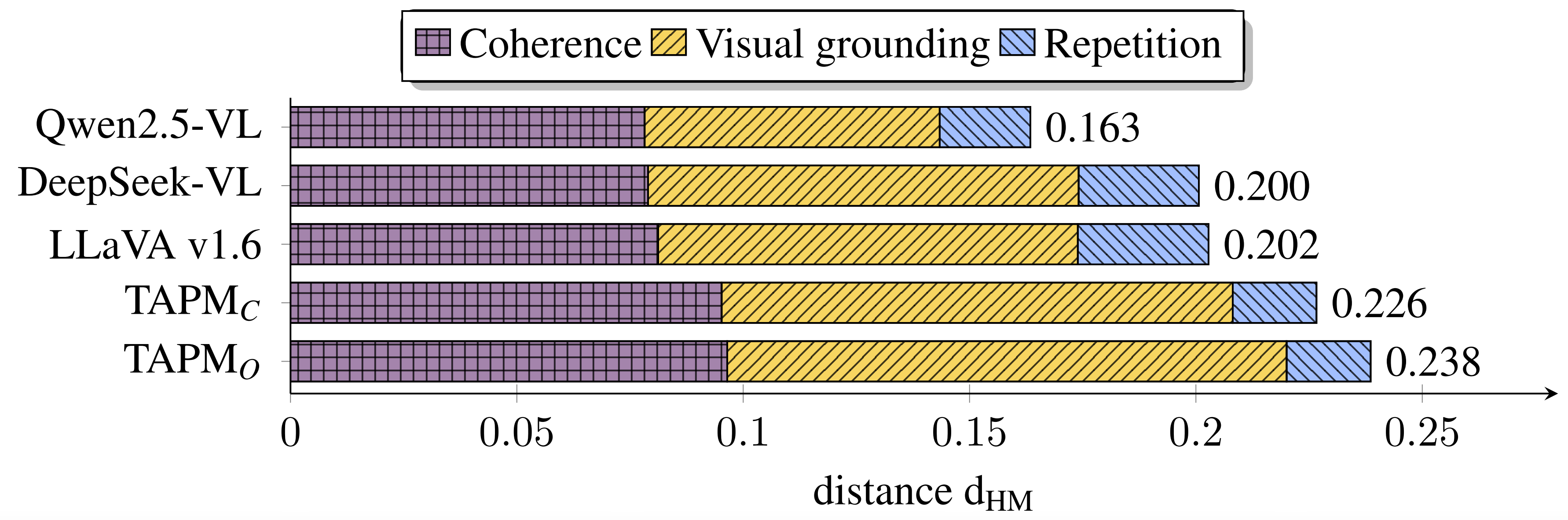
abs(**C**[human story] - **C**[model generated story]) = d_C

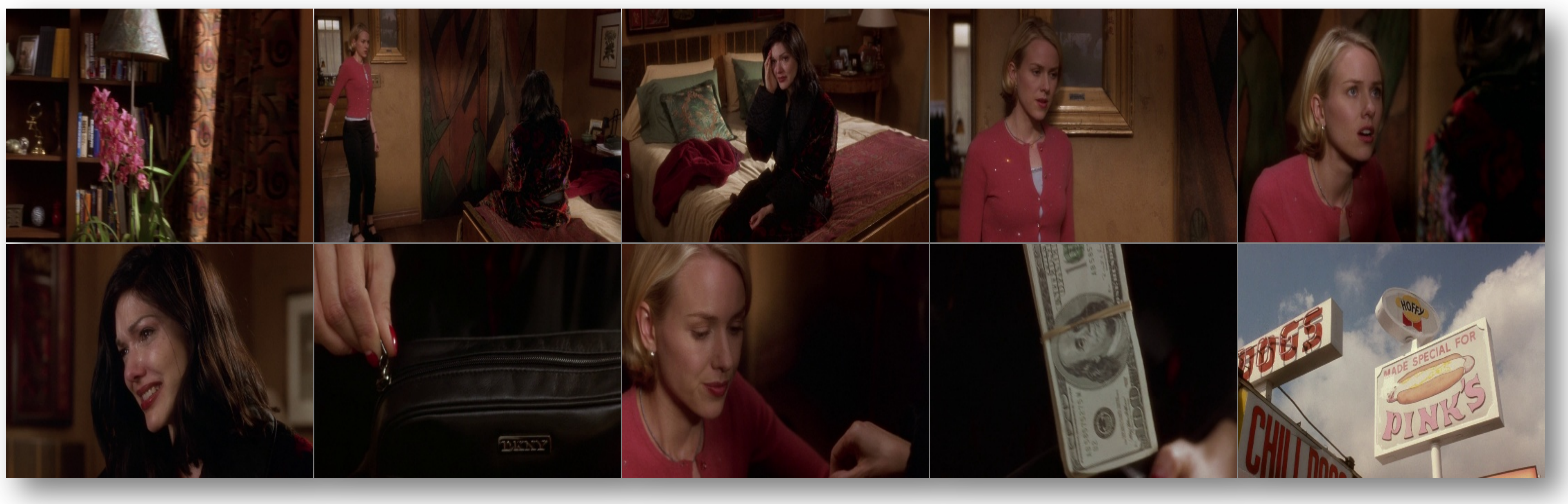
abs(**G**[human story] - **G**[model generated story]) = d_G

abs(**R**[human story] - **R**[model generated story]) = d_R

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

Results





Human-written story: It was a normal evening for [female0] when she entered her room. She saw her friend [female1] sitting in the bed. [female0] didn't know what was happening. [female1] looks like she has been crying. [female0] was confused by [female1]'s state. She asked [female1] what was happening as she didn't know why [female1] is crying like that. Then [female1] explained that she finally got a job after all the struggles she faced. Eve started to open her bag. [female0] was so happy for her friend as she finally able to live her life as she wishes. Eve took few money from her bag saying that it is her advanced payment for the job. They went to eat out to celebrate their happiness.

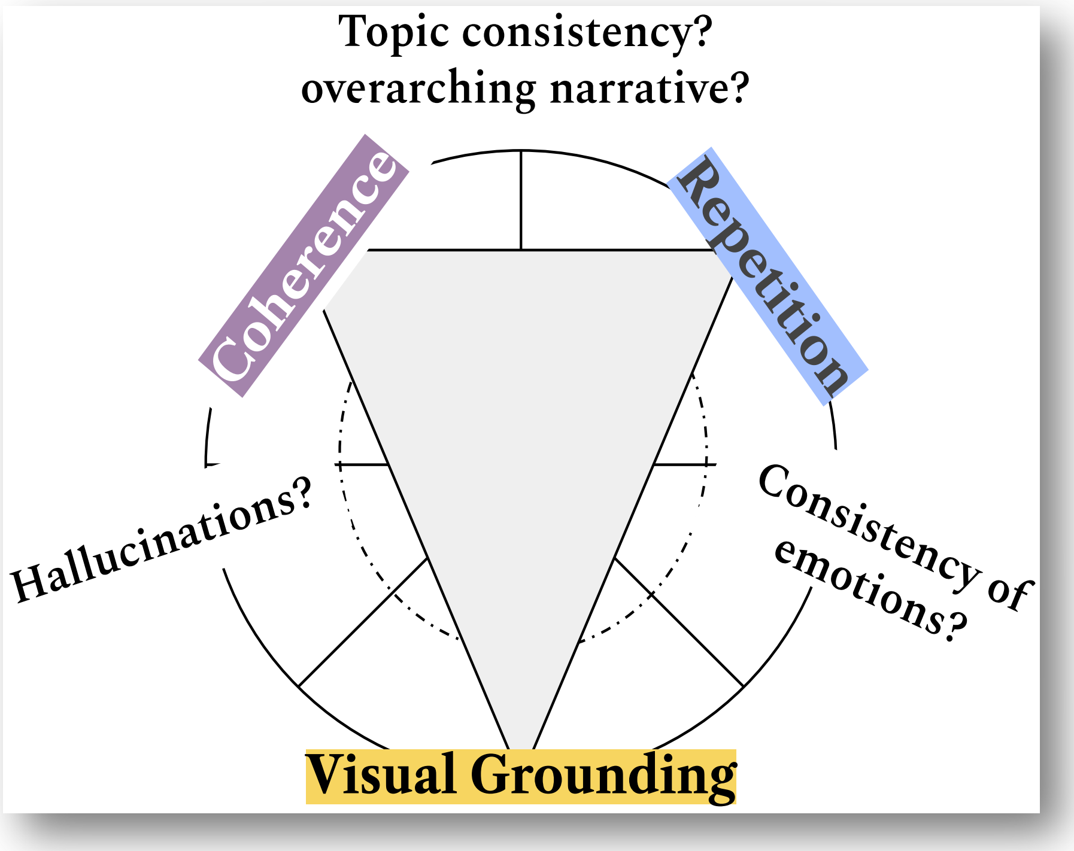
Qwen2.5-VL: In the cozy bedroom, Sarah **noticed her friend**, Emily, **looking distressed**. Emily sat on the bed, her head in her hands, **clearly upset**. Sarah approached her, concerned, and asked what was wrong. Emily, still in shock, told Sarah about a recent betrayal she had discovered. Sarah **listened intently, offering comfort and support**. As they talked, Sarah noticed a **stack of money** in Emily's **handbag**. Curious, she asked about it. Emily explained that the money belonged to someone else but had been left behind. Sarah suggested they return it. They decided to go to **Pink's Chili Dogs, a local favorite**.

TAPM$_C$: Her **husband [male0] was looking at her while she was on the floor**. The woman in pink asks the woman in brown where the earring might be. The woman in brown says that the man probably hid the **earring** inside a book. The woman in pink thinks about this. Then the woman in pink asks the woman in brown if she is certain about this. The woman in brown insists that the man would have hidden the **earring** inside a book. The woman checks the book by the **computer**, and sure enough, she finds her earring. She called her friend just to check on her. She kept thinking about how she would get out of this job. **It was a large and grand house with grand gardens.**

Takeaways

👍 The d_{HM} evaluation method is useful for differentiating 'good' stories in visually grounded story generation.

🧐 There are several other dimensions of a visual story that need to be considered for improving evaluation accuracy.


Topic consistency?
overarching narrative?
Coherence
Repetition
Hallucinations?
Consistency of emotions?
Visual Grounding