Coreference as an indicator of context scope in multimodal narrative

Nikolai Ilinykh[†], **Shalom Lappin**[§†], **Asad Sayeed**[†], and **Sharid Loáiciga**[†]
[†]Dept. of Philosophy, Linguistics, and Theory of Science, University of Gothenburg
[§]School of Electronic Engineering and Computer Science, Queen Mary University of London
[§]Dept. of Informatics, King's College London
nikolai.ilinykh@gu.se, s.lappin@qmul.ac.uk,
asad.sayeed@gu.se, sharid.loaiciga@gu.se

Abstract

We demonstrate that large multimodal language models differ substantially from humans in the distribution of coreferential expressions in a visual storytelling task. We introduce a number of metrics to quantify the characteristics of coreferential patterns in both human- and machine-written texts. Humans distribute coreferential expressions in a way that maintains consistency across texts and images, interleaving references to different entities in a highly varied way. Machines are less able to track mixed references, despite achieving perceived improvements in generation quality. Materials, metrics, and code for our study are available at https://github.com/GU-CLASP/coreference-context-scope.

1 Introduction

Generative models produce text that many perceive as increasingly human-like. However, machine-generated text conceals important distinctions to which people are sensitive (Russell et al., 2025). Work on visual narrative shows that there is still a gap between human and machine ability to generate coherent text (Xu et al., 2018).

A key difference between human and machine writing behaviour is the distribution of coreferential elements in English. We find that texts generated in a multimodal task setting have considerably different distributions of transitions between *coreference chains*, i.e., chains of referring expressions pointing to the same entity. In this work, we describe our methodology, using a set of metrics that we apply to state-of-the-art multimodal language models in a visual storytelling task. All tested models show substantial differences in measured behaviour with respect to human-generated reference texts. This points to a need for a quantitative approach to coherence that accounts for the characteristics of coreference in texts generated from images by machines. A proper handling of reference has implications for the ability of these models to perform multi-modal grounding, reasoning and inference tasks in the way humans expect (McKoon and Ratcliff, 1992). Our contributions are:

- We introduce a set of distributional metrics capturing coreference transition patterns.
- We evaluate recent multimodal models on visual storytelling with our metrics.
- We perform an analysis of multimodal alignment of character consistency in text.

Humans build stories in a certain way, focusing on visual events and the logical connections and participants involved in them. Event participants take the form of characters in the stories, with some reappearing as the story unfolds. This character introduction and reprisal is precisely what coreference resolution encodes. Coreference resolution identifies referring expressions or *mentions* in a long text and chains them into distinct entities or *coreference chains* (Ng, 2016). The type of mention is also influenced by the salience of a character – whether they are newly introduced or already known (Prince, 1992; Grosz et al., 1995).

Work on generation of visual narratives is often evaluated on automatic scores such as BLEU (Papineni et al., 2002) or ROUGE (Lin, 2004), but they correlate poorly with human judgements (Hsu et al., 2022). Other metrics that look at trigram repetitions (Goldfarb-Tarrant et al., 2020) or differences in distributions (Pillutla et al., 2021) have a higher correlation with fluency as perceived by humans. However, these metrics do not capture the *tellability* of a story. In this study we use surface-level reference patterns to capture key aspects of tellability, including continuity, salience, and character switching. Our main contribution is a diagnostic framework for evaluating narrative consistency, demonstrating how coreference-based features can distinguish between human and model-generated

Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM² 2025), pages 789–807
July 31 - August 1, 2025

stories in a multimodal setting.

1.1 Related work

Coherence A coherent text consists of logically connected utterances, maintained through cohesive elements like lexical similarity and discourse connectives (Halliday and Hasan, 1976). However, since coherence is not easily "extractable", its evaluation often relies on tasks like automatic summarisation (Barzilay and Lapata, 2008), sentence perturbation (Dini et al., 2025), or sentence intrusion detection (Shen et al., 2021).

On relation to Centering Theory Our work focuses on describing how character references in visual narratives behave, and how this can reflect coherence and tellability. While we do not directly build on formal discourse theories, we think that the concepts of information structure and the role of topic and focus together with insights from Centering Theory (Grosz et al., 1995) are relevant. However, we chose not to rely on Centering Theory directly, as it imposes constraints such as assumptions about anaphora resolution (Lappin and Leass, 1994) and requires parameter tuning, including concepts like "utterance" (Poesio et al., 2004) which it is non-trivial to define. As shown by Chai and Strube (2022), Centering Theory can complement but not replace modern neural coreference systems. Since our stories are grounded in image sequences, directly applying Centering Theory poses additional challenges that we hope to explore in future work.

Visual storytelling Visual storytelling is emerging as a key testbed for evaluating grounded language models. The point of departure has been series of static images with captions collected either from photo albums (Huang et al., 2016) or movie scenes (Hong et al., 2023b). Generating tellable stories requires identifying common elements across images and consistently referring to them. Previous work has focused on creation of character-centric stories (Liu et al., 2024; Liu and Keller, 2023) and on optimisation of the loss function for coherence without an explicit concept of coreference (Hong et al., 2023a). Visual storytelling poses unique challenges and provides a window into coherence, common sense reasoning, and discourse grounding. Because stories are structured around sequences of events and characters in images, they provide a rich but controlled environment for probing model capabilities – including reference tracking, focus management, and multimodal alignment.

Coreference There is a long tradition of text-only systems for coreference resolution (Liu et al., 2023). More recently, reference relationships beyond texts have been explored in simulated environments where agents or participants interact with the environment or each other (Lee et al., 2022). This type of setting elicits reference relations that are difficult to find in texts, such as changes in perspective and reference meaning negotiation between participants (Tang et al., 2024).

2 Methodology

Model	# sentences			# words		
	μ	↓	↑	μ	↓	↑
DeepSeek-VL2-4.5B	25.91	8	52	417.08	140	522
DeepSeek-VL2-1B	13.26	1	69	265.39	1	573
Gemini 2.0 Flash	8.18	1	24	104.76	3	456
GPT4o	12.51	6	21	212.99	114	321
InternVL2.5-78B	15.88	1	71	302.24	1	1021
Qwen2-VL-72B	14.33	4	47	230.01	60	516
Qwen2-VL-7B	11.89	4	32	223.83	66	511
Human	6.67	4	23	85.39	56	332

Table 1: General descriptive statistics for generated and human outputs: number of sentences and total word counts per output. Columns show mean (μ), minimum (↓), and maximum (↑) values.

We examine how large language models handle coreference by generating text from the same prompts used for human-written stories. Using a strong automatic coreference resolver, we annotate both model and human outputs and compute our metrics on them. This setup enables a direct comparison of coreferential behaviour, allowing us to identify differences between models and humans.

2.1 Generation

VWP (Hong et al., 2023b) is a collection of human-written narratives obtained by presenting participants with a curated sequence of up to 10 images from MovieNet (Huang et al., 2020), yielding sequences of images paired with one human-written story each. We choose to work with VWP because its stories are written and evaluated by humans to be tellable, diverse, and grounded. Hong et al. (2023b) also show that VWP offers better semantic cohesion and coherence than VIST (Huang et al., 2016) for the same image sequences.

We use the image sequences to generate text stories from several multimodal models. The prompts are provided in Appendix C and technical details are provided in Appendix B. We then employ Link-Append (Bohnet et al., 2023) to annotate both the machine- and human-generated texts. This means that for each sequence of images in our evaluation dataset, we have a story generated by a human, stories generated by multiple LLMs, and lists of coreference chains for all stories. Table 1 provides statistics for all the stories.

2.2 Data

VWP has a total of $12,627$ examples with pre-defined train/dev/test splits. To ensure a representative evaluation while limiting computational demands, we performed stratified sampling and selected 30% or $3,786$ examples for evaluation. The sampling is proportional to the distribution in the original splits and based on two factors: the number of stories per movie, and the number of images per story. Appendix D provides additional statistics.

2.3 Models

We employ two versions of DeepSeek-VL2 (1B and 4.5B of activated parameters, Wu et al. (2024)), two versions of Qwen2-VL (7B and 72B parameters, Wang et al. (2024)), InternVL2.5-78B (Chen et al., 2025), Gemini 2.0 Flash[1], and GPT4o[2] (version gpt-4o-2024-08-06). We also use the decoder-based Link-Append system (Bohnet et al., 2023) for automatic coreference resolution. This system is based on the 13B-parameter mT5 (Xue et al., 2021) and processes only text. Link-Append has a reported performance of 83.3 CoNLL score for English. While discriminative models may achieve slightly better scores (Martinelli et al., 2024), we chose Link-Append for its compatibility with our task, where end-to-end generation aligns naturally with sequence-based modelling.

2.4 Quantitative metrics

We identify character coreference chains by string-matching VWP character names with Link-Append mentions, labelling the entire chain as a character chain if at least one match is found. Our metrics are computed on the sentence level.

Character transition (CharTr) Let C_s be the set of character coreference chains in sentence s. For each consecutive sentence pair $(s, s+1)$, we define the indicator T_s as $T_s = 1$ if $C_s \cap C_{s+1} \neq \emptyset$, and $T_s = 0$ otherwise. A higher value indicates that consecutive sentences tend to share at least one character, implying character continuity.

Character drop (CharDr) For each sentence pair, if C_s is non-empty, the drop ratio is defined as $\text{CharDr} = \frac{|C_s \setminus C_{s+1}|}{|C_s|}$. This metric represents the proportion of characters that disappear from one sentence to the next, with higher values indicating less character continuity.

Character addition (CharAd) If C_{s+1} is non-empty, the addition ratio is $\text{CharAd} = \frac{|C_{s+1} \setminus C_s|}{|C_{s+1}|}$. A higher value indicates that many new characters are introduced in the next sentence.

Character change (CharCh) For pairs where both C_s and C_{s+1} are non-empty, we define $\text{CharCh}_s = 1$ if $C_s \cap C_{s+1} = \emptyset$, and 0 otherwise. The metric captures the proportion of sentence pairs with a complete change of characters.

Character reappearance (CharRe) For each character chain c, $s_{\min}(c)$ and $s_{\max}(c)$ are the first and last sentences in which it appears. Normalised by the maximum possible spread $(N-1)$, the reappearance metric is $\text{CharRe} = \frac{1}{|\mathcal{C}|}\sum_{c \in \mathcal{C}} \frac{s_{\max}(c) - s_{\min}(c)}{N-1}$. Higher values mean characters reappear in distant sentences.

Multimodal character continuity We introduce a metric to quantify how consistently each character is referenced across text and images. For each character C, we compute text continuity T_C as the fraction of sentences between the first ($s_{\min}$) and last ($s_{\max}$) mention of C that actually include C using coreference chains. Similarly, image continuity I_C is the fraction of images between the first ($j_{\min}$) and last ($j_{\max}$) appearance of C that include C based on bounding box annotations. While T_C and I_C are modality-specific, they are aligned over the same sequence length: each story has the same number of text descriptions and images. This alignment enables direct comparison, as both metrics are normalised over equivalent spans. We define continuity consistency as $1 - |T_C - I_C|$, reflecting the agreement between modalities for each character. The final multimodal character continuity (MCC) score for a story is the average continuity across all characters. We compute MCC for every story and compare distributions across sources (e.g., human vs. model) using two-sample t-tests and Cohen's d for effect size. Details on visual character detection are provided in Appendix A.

[1] https://deepmind.google/technologies/gemini/
[2] https://openai.com/index/hello-gpt-4o/

Model	# words-as-mentions			# chains			chain size			CCI		
	μ	↓	↑	μ	↓	↑	μ	↓	↑	μ	↓	↑
DeepSeek-VL2-4.5B	135.60	35	323	10.75	2	26	13.06	6	97.5	2.58	0	15.92
DeepSeek-VL2-1B	77.17	0	410	6.98	0	28	11.33	0	333	1.62	0	18.30
Gemini 2.0 Flash	27.48	0	147	3.88	0	16	6.47	0	27	0.83	0	6.60
GPT4o	52.84	11	131	6.16	1	15	8.89	3.2	27	1.62	0	7.64
InternVL2.5-78B	87.85	0	294	7.65	0	22	11.53	0	33.5	1.79	0	10.20
Qwen2-VL-72B	66.84	7	199	6.69	1	17	10.23	2.2	36	1.47	0	7.57
Qwen2-VL-7B	76.64	5	311	6.62	1	19	11.71	2.5	35.4	1.80	0	11.20
Human	26.09	2	176	3.85	1	14	6.87	2	23.5	0.89	0	6.43

Table 2: Descriptive statistics across models: number of words identified as mentions by LinkAppend, number of coreference chains, average chain size, and Chain Crossing Index (CCI). Columns show mean (μ), minimum (↓), and maximum (↑) values.

Model	CharTr	CharDr	CharAd	CharCh	CharRe	MCC		REC	ρ
						μ	↓		
DeepSeek-VL2-4.5B	0.06	0.90	0.88	0.57	0.57	$0.76^{\dagger}$	0.20	0.61	-0.045**
DeepSeek-VL2-1B	0.03	0.91	0.88	0.46	0.33	$0.72^{\dagger}$	0.16	0.61	-0.045**
Gemini 2.0 Flash	0.10	0.84	0.82	0.48	0.48	$0.78^{\dagger}$	0.11	0.57	0.352**
GPT4o	0.10	0.85	0.83	0.54	0.60	$0.76^{\dagger}$	0.21	0.65	-0.009
InternVL2.5-78B	0.13	0.80	0.78	0.42	0.58	$0.74^{\dagger}$	0.18	0.64	0.012
Qwen2-VL-72B	0.12	0.82	0.80	0.45	0.63	$0.79^{\dagger}$	0.29	0.64	0.113**
Qwen2-VL-7B	0.15	0.77	0.74	0.37	0.61	$0.78^{\dagger}$	0.22	0.57	0.131**
Human	0.23	0.67	0.63	0.27	0.54	0.84	0.29	0.65	0.005

Table 3: Aggregated qualitative metric values and referring expression change (REC) across models. MCC is shown with mean (μ) and minimum (↓). Pearson correlation (ρ) shows correlation between REC and text length. Values marked with ** denote statistical significance ($p < 0.05$). Mean values marked with † differ significantly from human according to a two-sample t-test.

Referring expression change (REC) The REC metric captures how consistently a character chain is realised across mentions (e.g., as a proper name or pronoun). For a character chain c, let the mention sequence be $\mathrm{MS}(c) = [m_1, m_2, \ldots, m_k]$, where each m_i is a proper name (N), pronoun (P), or both. We set $\mathtt{REC}(c) = 0$ if all mentions are realised the same way ($|\{\mathrm{MS}(c)\}| = 1$), and $\mathtt{REC}(c) = 1$ if the form changes at least once. Higher REC values indicate more variation in referring expressions, while lower values suggest consistent usage. We also compute Pearson correlation between REC and text length (word count).

3 Results and analysis

We begin by examining the results in Table 2, with example outputs provided in Appendices E – G. Humans refer to fewer entities on average (3.85), but each is mentioned multiple times (6.87) suggesting a focused narrative. The chain crossing index (CCI) measures how often two chains intersect, excluding overlaps or disjoint chains. A low human CCI (0.89) reflects consistent reference to key characters, with less frequent switching between entities.

Gemini 2.0 Flash is closest to humans across all metrics, though it sometimes produces 0 mentions, indicating inconsistency in referencing characters. While its structure appears human-like, it may omit key entities entirely. In contrast, DeepSeek-VL2-4.5B generates more entities, longer chains, and more frequent cross-chain intersections (high CCI), reflecting overgeneration. GPT4o produces fewer chains overall but tracks characters more consistently than DeepSeek-VL2-4.5B.

The results in Table 3 show that *human*-generated stories show stronger character continuity across sentences. Humans have higher transition scores (e.g., 0.23), drop and add characters less frequently (e.g., 0.67 and 0.63), and rarely switch to entirely new sets of characters between sentences. Characters are also reintroduced after shorter gaps (mean 0.54) unlike in many model outputs. These patterns suggest that human narratives maintain a more coherent and trackable set of characters, while models tend to drop, add, or switch characters more often.

This trend is reinforced by MCC scores, where humans outperform all models, indicating stronger alignment between text and image character mentions. Two-sample t-tests confirm the difference is significant in every case ($p < 0.05$), with effect sizes (Cohen's d) ranging from 0.45 to 0.83, reflecting moderate to large differences. Finally, while average `REC` across models is similar, their relationship with text length differs. We observe that only models show increasing variation in referring expressions with longer text, e.g., `Gemini 2.0 Flash`. Human stories remain consistent in how they refer to characters, regardless of length.

An interesting trend is that larger models often change characters more frequently than smaller ones, indicating weaker character consistency. However, as Table 1 shows, they also generate longer outputs with more content, likely introducing more characters and switches. In contrast, smaller models produce shorter, simpler outputs – sometimes just a sentence or a word – leading to lower CharCh scores that do not necessarily imply better coherence. For instance, `DeepSeek-VL2-1B` often produces no mentions, while `DeepSeek-VL2-4.5B` generates many (Table 2). Larger models tend to over-generate, resulting in dynamic but less grounded stories.

To support this, we report Pearson correlation between CharCh and MCC in Table 4. These consistently negative correlations (except in human stories) suggest that higher character turnover is linked to lower coherence. Human-authored stories do not show this trend, suggesting that humans can manage frequent character changes without sacrificing clarity – something models still struggle with.

Model	ρ
`DeepSeek-VL2-4.5B`	-0.236**
`DeepSeek-VL2-1B`	-0.263**
`Gemini 2.0 Flash`	-0.189**
`GPT4o`	-0.248**
`InternVL2.5-78B`	-0.251**
`Qwen2-VL-72B`	-0.238**
`Qwen2-VL-7B`	-0.188**
Human	-0.040

Table 4: Pearson correlation (ρ) between CharCh scores and MCC scores. Values marked with ** are statistically significant ($p < 0.001$).

4 Conclusions and future work

We found substantial differences in coreferential patterns between LLM and human outputs. The fact that humans subjectively perceive model outputs as human-like does not imply that the models actually behave in a human-like way. Coreference helps humans structure narratives, and its divergence in LLMs has implications for AI-human interaction that require further exploration. Our metrics allow us to measure the effects of the fact that, unlike humans, models are not explicitly required to caption each image.

The proposed metrics operate on textual references and can be applied to a range of formats, including dialogues, text-only stories (Fan et al., 2018), and collaboratively written narratives (Akoury et al., 2020). We plan to incorporate VIST (Huang et al., 2016) and VIST-Character (Liu and Keller, 2023), which include detailed visual and textual coreference chains and importance ratings for characters, providing a strong basis for further evaluation of coreferential coherence.

Future work will explore LLM attention patterns to better understand biases in reference and coreference generation. We are also actively considering interpretability experiments to probe how models internally represent characters during generation, for example, whether attention aligns with character mentions in images. In addition, we aim to conduct human evaluations to better understand what makes a "good" story.

Limitations

This work deals with the quality of generation in English. In addition, presented metrics rely on the output of an automatic coreference system. If a reliable model of coreference does not exist the metrics cannot be computed reliably. While our main focus is on the analysis of coreferential patterns produced by recent multimodal models, we use data from only one visual storytelling dataset, VWP (Hong et al., 2023b). Our metrics are also affected by the quality of automatically generated texts which we do not explicitly evaluate with automatic metrics or regenerate with different decoding methods. We also note that prompt design can impact Instruction-following ability of the models which in turn can affect coherence of the generated stories.

Acknowledgments

The work reported in this paper has been supported by a grant from the Swedish Research Council (VR project 2014-39) for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg. The computations and data storage were enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council through grant agreement no. 2022-06725.

References

Nader Akoury, Shufan Wang, Josh Whiting, Stephen Hood, Nanyun Peng, and Mohit Iyyer. 2020. STORIUM: A Dataset and Evaluation Platform for Machine-in-the-Loop Story Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6470–6484, Online. Association for Computational Linguistics.

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

Bernd Bohnet, Chris Alberti, and Michael Collins. 2023. Coreference Resolution through a seq2seq Transition-Based System. *Transactions of the Association for Computational Linguistics*, 11:212–226.

Haixia Chai and Michael Strube. 2022. Incorporating centering theory into neural coreference resolution. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 2996–3002. Association for Computational Linguistics.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. 2025. Expanding Performance Boundaries of Open-Source Multimodal Models with Model, Data, and Test-Time Scaling. *arXiv preprint*. ArXiv:2412.05271 [cs].

Luca Dini, Dominique Brunato, Felice Dell'Orletta, and Tommaso Caselli. 2025. TEXT-CAKE: Challenging language models on local text coherence. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4384–4398, Abu Dhabi, UAE. Association for Computational Linguistics.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Seraphina Goldfarb-Tarrant, Tuhin Chakrabarty, Ralph Weischedel, and Nanyun Peng. 2020. Content planning for neural story generation with aristotelian rescoring. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4319–4338, Online. Association for Computational Linguistics.

Barbara J. Grosz, Aravind K. Joshi, and Scott Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225.

M. A. K. Halliday and Ruqaiya Hasan. 1976. *Cohesion in English*. Longman, London.

Xudong Hong, Vera Demberg, Asad Sayeed, Qiankun Zheng, and Bernt Schiele. 2023a. Visual coherence loss for coherent and visually grounded story generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, Toronto, Canada. Association for Computational Linguistics.

Xudong Hong, Asad Sayeed, Khushboo Mehra, Vera Demberg, and Bernt Schiele. 2023b. Visual Writing Prompts: Character-Grounded Story Generation with Curated Image Sequences. *Transactions of the Association for Computational Linguistics*, 11:565–581. Place: Cambridge, MA Publisher: MIT Press.

Chi-Yang Hsu, Yun-Wei Chu, Vincent Chen, Kuan-Chieh Lo, Chacha Chen, Ting-Hao Huang, and Lun-Wei Ku. 2022. Learning to rank visual stories from human ranking data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6365–6378, Dublin, Ireland. Association for Computational Linguistics.

Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. 2020. Movienet: A holistic dataset for movie understanding. *Preprint*, arXiv:2007.10937.

Ting-Hao Kenneth Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. 2016. Visual storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1233–1239, San Diego, California. Association for Computational Linguistics.

Shalom Lappin and Herbert J. Leass. 1994. An algorithm for pronominal anaphora resolution. *Comput. Linguistics*, 20(4):535–561.

Haeju Lee, Oh Joon Kwon, Yunseon Choi, Minho Park, Ran Han, Yoonhyung Kim, Jinhyeon Kim, Youngjune Lee, Haebin Shin, Kangwook Lee, and Kee-Eung Kim. 2022. Learning to embed multi-modal contexts for situated conversational agents. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 813–830, Seattle, United States. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Danyang Liu and Frank Keller. 2023. Detecting and grounding important characters in visual stories. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press.

Danyang Liu, Mirella Lapata, and Frank Keller. 2024. Generating Visual Stories with Grounded and Coreferent Characters. *arXiv preprint*. ArXiv:2409.13555 [cs].

Ruicheng Liu, Rui Mao, Anh Tuan Luu, and Erik Cambria. 2023. A brief survey on recent advances in coreference resolution. *Artificial Intelligence Review*, page 14439–14481.

Giuliano Martinelli, Edoardo Barba, and Roberto Navigli. 2024. Maverick: Efficient and accurate coreference resolution defying recent trends. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13380–13394, Bangkok, Thailand. Association for Computational Linguistics.

Gail McKoon and Roger Ratcliff. 1992. Inference during reading. *Psychological review*, 99 3:440–66.

Vincent Ng. 2016. Advanced machine learning models for coreference resolution. In *Anaphora Resolution: Algorithms, Resources, and Applications*, pages 283–313. Springer Nature Link.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaïd Harchaoui. 2021. MAUVE: measuring the gap between neural text and human text using divergence frontiers. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 4816–4828.

Massimo Poesio, Rosemary Stevenson, Barbara Di Eugenio, and Janet Hitzeman. 2004. Centering: A parametric theory and its instantiations. *Comput. Linguistics*, 30(3):309–363.

Ellen F. Prince. 1992. The zpg letter: Subjects, definiteness, and information status. In W. Mann and S. Thompson, editors, *Discourse description: Diverse linguistic analysis of a fund-raising text*, pages 223–255. John Benjamins, Amsterdam.

Jenna Russell, Marzena Karpinska, and Mohit Iyyer. 2025. People who frequently use chatgpt for writing tasks are accurate and robust detectors of ai-generated text. *Preprint*, arXiv:2501.15654.

Aili Shen, Meladel Mistica, Bahar Salehi, Hang Li, Timothy Baldwin, and Jianzhong Qi. 2021. Evaluating document coherence modeling. *Transactions of the Association for Computational Linguistics*, 9:621–640.

Zineng Tang, Lingjun Mao, and Alane Suhr. 2024. Grounding language in multi-perspective referential communication. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19727–19741, Miami, Florida, USA. Association for Computational Linguistics.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. *arXiv preprint*. ArXiv:2409.12191 [cs].

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. 2024. DeepSeek-VL2: Mixture-of-Experts Vision-Language Models for Advanced Multimodal Understanding. *arXiv preprint*. ArXiv:2412.10302 [cs].

Jingjing Xu, Xuancheng Ren, Yi Zhang, Qi Zeng, Xiaoyan Cai, and Xu Sun. 2018. A skeleton-based model for promoting coherence among sentences in narrative story generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4306–4315, Brussels, Belgium. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

A Multimodal character continuity: technical details

We propose a metric for evaluating character continuity across visual and textual modalities in image-grounded stories. Our approach leverages existing character annotations from MovieNet (Huang et al., 2020), the dataset that also provides the images used in Visual Writing Prompts (VWP) (Hong et al., 2023b). In MovieNet, each image is associated with a movie and includes character detections labeled either with artificial names (e.g., `nm000004`) or as `Unknown`. For each story X with an image sequence $\mathcal{I} = \{i_1, i_2, \ldots, i_K\}$, we process character annotations from MovieNet that include: (i) character bounding boxes $B = \{b_1, b_2, \ldots, b_n\}$ for each image, (ii) character identifiers (PIDs) mapped to actor names using MovieNet's cast metadata, and (iii) the relative size of bounding boxes (computed as a ratio to the image area). We then record the character names, effectively capturing which characters appear in which images. Mentions of characters in the texts are identified by matching these names to those in the annotations. e.g. "russell" in text is mapped with "Russell" in annotations. Finally, `LinkAppend` allows us to determine when a character is mentioned in texts by analysing the coreference chains that include those character names.

B Model size and budget

We used A100 40GB GPUs and A100 80GB GPUs to run models for our tasks (visual story generation and coreference resolution). For story generation, the time required for the models took up to 14 GPU hours. For coreference resolution, the time required was up to 12 GPU hours. Closed models were prompted through their API.

Prompt text A

View a sequence of N images and figure out the content. Then write a story with it. View a sequence of images as many times as you wish. Figure out who were involved and what happened. Then write a story that fits the image sequence. You should write the story using at least 5 images. You need to write at least 50 but no more than 300 words. You do not need to write text without a corresponding image unless it is necessary. The story should be related to the image sequence. Describe only the most important character(s) and event(s). You can use either the first name, a pronoun, or a noun phrase according to the context. If the character you want to mention is not there, name the characters as you want, but please be consistent. Use punctuation and letter case correctly. Do not mention that you are describing an image. Avoid using phrases like "In this image, ...". Do not write a monologue of a character or a dialogue between characters.

We adjusted the prompts to the input format of each model according to their official documentation. We used scripts from `https://github.com/boberle/corefconversion` to convert the output of `Link-Append` to appropriate format for our analysis (from .conll format to .jsonlines).

C Prompts

Below are two texts that were given to the models for visual story generation. The two texts differ slightly in phrasing depending on whether the data had images of characters from the story and their names. The differences are highlighted in red.

D Dataset distribution

A general distribution of visual stories in our dataset in terms of movies and number of images is shown in Figure 1.

Prompt text B

View a sequence of N images followed by K character images and figure out the content. Then write a story with it. View a sequence of images as many times as you wish. Figure out who were involved and what happened. Then write a story that fits the image sequence. You should write the story using at least 5 images. You need to write at least 50 but no more than 300 words. You do not need to write text without a corresponding image unless it is necessary. The story should be related to the image sequence. Describe only the most important character(s) and event(s). When mentioning the characters, please follow their names which are provided in the order that the character images were given: [character names]. You can use either the first name, a pronoun, or a noun phrase according to the context. If the character you want to mention is not there, name the characters as you want, but please be consistent. Use punctuation and letter case correctly. Do not mention that you are describing an image. Avoid using phrases like "In this image, ...". Do not write a monologue of a character or a dialogue between characters.

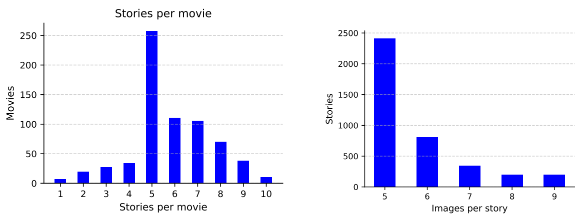


Figure 1: Distributions in our evaluation dataset.

E Example I

Below you see a sequence of images depicting a visual story alongside bounding boxes of characters that were given to models and humans. Next, we show texts produced by humans and models, including coreference chains produced by `Link-Append`. In this example human text has one of the highest CharTr values across all examples in the dataset (2.33). We provide corresponding results for CharTr metric for each model alongside text that they generated.

Image 1

Image 2

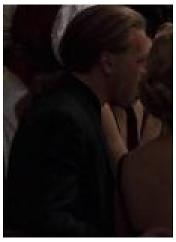
Image 3

Image 4

Image 5

Character image 1: Ashley

Character image 2: Val



Character image 3: Tom

Character image 4: Robert

Character image 5: Amy

Human, CharTr 2.33**:**

[[Ashley][1] and [Val][2]][3] are sitting together at [a table][4] at [the Gala][5] .
[Tom][6] was sitting across from [them][3] , and [Tom][6] leans in to ask what [their][3] secret to a loving and long marriage is
.
Flattered , [[Ashley][1] and [Val][2]][3] smile .
[Ashley][1] starts to tell [Tom][6] that [Val][2] does everything [she][1] tells [him][2] to .
[Robert][7] got stuck sitting at the kids ' table at [the Gala][5] , but [he][7] is starting at [[[Val][2] and [Ashley][1] 's][3] table][4] while wishing [he][7] could be sitting there with [them][3] .
After [the Gala][5] , [[Val][2] , [Ashley][1] , and everyone at [[their][8] table][4]][8] are standing outside the hotel discussing when and where [their][8] next party will be .
[Robert][7] is walking up behind [them][8] to join [their][8] conversation .

`DeepSeekVL2-1B`, CharTr 0.46**:**

[The sequence of images][1] depicts [a group of people gathered at a dinner table in what appears to be a restaurant setting][2] .
The atmosphere is warm and intimate , with soft lighting and floral centerpieces adding to the ambiance .
[The individuals][2] are engaged in conversation , some smiling and others looking at each other , suggesting a lively and friendly gathering .
[Ashley , Val , Tom , Robert , and [Amy][3]][4] are the main characters in [this scene][5] .
[They][4] are seated around a table , with plates of food and drinks in front of [them][4] .
The setting is indoors , likely in a restaurant , as indicated by the decor and lighting .
The background shows other diners and tables , creating a bustling environment .
The story that fits [the sequence of images][1] is as follows : [[Ashley][6] , [Val][7] , [Tom][8] , [Robert][9] , and [Amy][3]][4] are friends who have gathered for a dinner party .
[They][4] are enjoying each other 's company , sharing stories and laughter .
The atmosphere is relaxed and convivial , with everyone feeling comfortable and at ease .
The presence of wine and the elegant table setting suggest that this is a special occasion , perhaps a birthday or anniversary dinner .
[The characters][4] are engaged in conversation , with some looking at each other and [others][10] focused on [their][10] food .
[The scene][5] captures a moment of connection and camaraderie among [the group][4] .

`DeepSeek-VL2-4.5B`, CharTr 0.66**:**

The evening began with a sense of anticipation as [the group of six][1] gathered around [the table in [the dimly lit restaurant][3]][2] .
[Ashley , the elegant woman with blonde hair][4] , sat at the head of [the table][2] , [her][4] eyes scanning [the room][3] with a mixture of curiosity and caution .
[Val , the man in the dark suit][5] , sat beside [her][4] , [his][5] hands clasped together as [he][5] listened intently to [the conversation][6] .
Across from them , [Tom , the man with the neatly combed hair][7] , leaned forward , [his][7] eyes locked onto [Robert , the man with the serious expression , who was deep in conversation with [Amy , the woman with the dark hair][9]][8] .
As [the night][10] wore on , [[the group 's][1] conversation][6] turned more serious .
[Robert 's][8] words were laced with urgency , and [Amy 's][9] eyes reflected a mix of concern and determination .
[Tom , ever the observer][7] , noticed the tension in the air and tried to steer [the conversation][6] towards lighter topics .
[Val][5] , however , remained focused on the matter at hand , [his][5] mind racing with possibilities .
Outside , [the night][10] was quiet , the only sound the distant hum of the city .
[[The group 's][1] discussion][6] had reached a critical point , and the stakes were high .
As [they][1] prepared to leave , [the group 's][1] expressions were a mix of relief and resolve .
[They][1] knew that the decisions [they][1] made that night would have far-reaching consequences .
As [they][1] stepped into [the night][10] , [the group 's][1] unity was palpable .
[They][1] moved as one , [their][1] purpose clear .
The city lights flickered around [them][1] , casting long shadows that seemed to stretch into the future .
[The night][10] had brought [them][1] together , and now , [they][1] were ready to face whatever challenges lay ahead .

`Gemini 2.0 Flash`, CharTr None:

Val and Tom attend a dinner party at a local restaurant .

`GPT4o`, CharTr 0.75**:**

[[Ashley][1] , [Val][2] , [Tom][3] , [Robert][4] , and [Amy][5]][6] gathered at an elegant restaurant to celebrate a special occasion .
The atmosphere was lively , with [everyone][6] enjoying the company and the food .

[Ashley][1] and [Val][2] , sitting closely , shared a laugh over a shared memory , creating a warm and joyful mood at [the table][7] .
[Tom][3] , engaged in conversation with [Robert][4] , gestured animatedly , drawing laughter from both .
Across [the table][7] , [Amy][5] was sharing a story that held [everyone 's][6] attention , [her][5] expressions enchanting [the group][6] .
After dinner , [they [all][6]][6] decided to step outside for a breath of fresh air .
The cool night air filled with [their][6] lively chatter created a serene end to [a wonderful evening][8] .
As [they][6] stood together , [they][6] realized the depth of [their][6] camaraderie .
Meanwhile , [Amy][5] received an unexpected phone call that seemed urgent .
[Her][5] expression turned serious , reminding [her][5] of the responsibilities awaiting [her][5] return .
[She][5] promised to meet [the others][6] again soon , cherishing [the evening 's][8] joyful moments .
[Their][6] gathering ended with warm goodbyes , [everyone][6] cherishing the memories [they][6] had made together .
The shared laughter and stories would linger , a testament to [their][6] enduring friendship .

`InternVL2.5-78B`, CharTr 0.64**:**

It was a chilly evening in the small town of Millfield , where [a formal banquet][1] had gathered a group of old friends for [a reunion at [a luxurious restaurant][3]][2] .
As [guests][4] mingled near [their][4] tables , [[Ashley][5] and Robert][6] were seated at the center of [the banquet][1] , engaged in a lively conversation .
The table was adorned with a vase of vibrant blossoms , and the air was filled with the hum of laughter and clinking glasses .
In the background , [[Val][7] and [Tom][8]][9] were deeply engrossed in [their][9] conversation , sharing stories and memories of [their][9] youth .
[Their][9] affectionate exchanges brought warmth to [their][9] surroundings , eliciting gentle smiles from fellow guests .
As [the evening][10] unfolded , the atmosphere was filled with camaraderie and joyous reunion .
[[Ashley][5] and [Val][7]][6] shared light-hearted banter with the other guests around [their][6] table .
The conversation was animated , as [they][6] reminisced on shared experiences and adventures from the past .
In a quiet corner , an unexpected moment caught everyone ' s attention .
[[Val][7] and [Tom][8]][9] were tenderly holding hands , [their][9] eyes meeting with a love so pure and unmarred by the passage of time .
This tender display of affection brought a wave of warmth even to those who had merely observed from afar .
As night deepened , [the gathering][2] began to wind down , but the heartfelt connections and cherished moments lingered in the air long after [the guests][4] had bid [their][4] farewells and left [the restaurant 's][3] warm embrace .
[The evening][10] had been a poignant reminder of the power of friendship and the enduring bonds that time can not extinguish .

`Qwen2-VL-72B`, CharTr 0.17**:**

[The evening][1] began with a lively gathering at [a restaurant , where [friends and family][2] sat around a table adorned with flowers and enjoyed each other 's company][3] .
[Ashley][4] , dressed elegantly , engaged in animated conversation with [her][4] companions , while Val and Tom listened intently .
As the night progressed , [the group][2] decided to move on to another location for further celebrations .
Outside [the restaurant][3] , [they][2] formed a small circle , continuing [their][2] discussions .
Amy , who had been busy earlier , joined [them][2] later , bringing along a cheerful energy .
[The group][2] laughed and shared stories , creating a warm atmosphere despite the cool evening air .
Meanwhile , [Robert , who was not part of the initial gathering][5] , found [himself][5] in a different setting , engrossed in a phone call .
[His][5] expression suggested [he][5] was deep in thought , possibly dealing with some pressing matter .
Unbeknownst to [him][5] , [his][5] path would soon cross with [the lively group][2] , adding a new dynamic to [the evening 's][1] events .

`Qwen2-VL-7B`, CharTr 1.29**:**

[The story][1] revolves around [a group of people who gather for a dinner party at a restaurant][2] .
Among [them][2] are [Ashley][3] , [Val][4] , [Tom][5] , [Robert][6] , and [Amy][7] .
The scene begins with [a large group of people seated at [a long table][8] , engaged in lively conversation and enjoying [their][2] meal][2] .
As the evening progresses , the atmosphere becomes more intimate , with smaller groups forming around [the table][8] .
In one of the images , [[Tom][5] and [Robert][6]][9] are seen in a quiet conversation , suggesting a deeper connection between [them][9] .
Meanwhile , [Ashley][3] and [Val][4] are seated together , sharing a moment of laughter and camaraderie .
[Amy][7] , on the other hand , is seen in a more solitary moment , talking on the phone , perhaps discussing something important or personal .
As the night draws to a close , [the group][2] begins to disperse .
[[Robert][6] and [Tom][5]][9] are seen in a final conversation , perhaps reflecting on [the evening][10] or discussing something

of mutual interest .
[Amy][7] is also seen in a moment of solitude , deep in thought , possibly contemplating the events of [the evening][10] .
[The story][1] is one of camaraderie , reflection , and the passage of time , as [the group of friends and acquaintances][2] gather , share stories , and then part ways , each with [their][2] own thoughts and memories of [the evening][10] .

F Example II

In this story example we have one of the highest multimodal character continuity (MCC) scores in the dataset for human-generated texts. The scores for both humans and models are provided alongside generated texts.

Image 1

Image 2

Image 3

Image 4

Image 5

Character image 1: Keanu

Character image 2: Charlize

Human, MCC 1**:**

[Keanu][1] tells [Charlize][2] how much [she][1] loves [her][2] and asks for forgiveness .
[Charlize][2] says it is very difficult to forgive [him][1] because [he][1] hurt [her][2] so much .
[Charlize][2] cries a lot and [Keanu][1] Loses forgiveness .
[Charlize][2] stops crying and says [she][2] loves [him][1] so much .
who will trust [him][1] one more time .

DeepSeek-VL2-1B, MCC 0.33**:**

[The sequence of images][1] depicts [a dramatic scene with [two individuals][2] engaged in an intense conversation][3] .
The first image shows [a man in a dark suit][4] looking at [a woman who appears distressed , [[her][5] face][6] contorted with emotion][5] .
The second image captures [the same woman][5] , now wrapped in [a pink blanket][7] , [[her][5] eyes][8] closed and [mouth][9] open as if [she][5] is crying .
The third image shows [the woman][5] still wrapped in [the blanket][7] , [her][5] expression more subdued , with [her][5] eyes partially closed .
The fourth image focuses on [the woman 's][5] face , showing [her][5] in a state of distress , with [her][5] eyes wide open and mouth slightly agape .
The fifth image is a close-up of [the woman 's][5] face , emphasizing [her][5] emotional state .
The story that fits [the sequence of images][1] is as follows : In [the dimly lit room][10] , [Keanu , a man in a dark suit][4] , stands beside [a woman wrapped in [a pink blanket][7]][5] .
[She][5] appears to be in distress , [[her][5] face][6] contorted with emotion .
[The room][10] is dimly lit , with shadows cast on the walls , creating an atmosphere of tension .
[Keanu][4] looks at [her][5] with concern , [his][4] expression serious .
[[The woman 's][5] eyes][8] are closed , and [[her][5] mouth][9] is open as if [she][5] is crying .
[She][5] seems to be in a state of emotional turmoil , possibly due to a recent event or revelation .
[The scene][3] suggests a moment of high tension , with [the characters][2] deeply engaged in a conversation that could have significant consequences .

DeepSeek-VL2-4.5B, MCC 0.68**:**

The story begins with [Keanu , a man in a black suit][1] , standing close to [Charlize , a woman wrapped in [a pink blanket][3]][2] , in [a dimly lit room][4] .
[[Keanu 's][1] expression][5] is serious as [he][1] leans in , [his][1] eyes locked onto [Charlize 's][2] .
[She][2] looks up at [him][1] , [[her][2] face][6] a mix of concern and confusion .
The tension between [them][7] is palpable , as if [they][7] are discussing something of great importance .
[[Charlize 's][2] eyes][8] well up with tears as [she][2] speaks , [her][2] voice trembling .
“ [Keanu][1] , [I][2] ca n't keep doing [this][9] , ” [she][2] says , [her][2] words barely audible .
“ [I][2] need to find a way out of [this][9] . ”
[[Keanu 's][1] expression][5] softens slightly , but [he][1] remains firm .
“ [You][2] ca n't just leave , ” [he][1] replies , [his][1] tone gentle yet resolute .
As the conversation continues , [Charlize 's][2] distress becomes more evident .
[She][2] clutches [the blanket][3] around [her][2] , [her][2] body shaking with emotion .
[Keanu][1] watches [her][2] , [[his][1] heart][10] aching for [her][2] pain .
“ [I][1] know it 's not easy , ” [he][1] says , [his][1] voice filled with empathy .
“ But [we][7] have to find a way . ”
In the next image , [[Charlize 's][2] eyes][8] are closed , [[her][2] face][6] contorted in pain .
[She][2] takes a deep breath , trying to compose [herself][2] .
[Keanu][1] stands beside [her][2] , [his][1] hand gently resting on [her][2] shoulder .
“ [You][2] 're not alone , ” [he][1] whispers , [his][1] voice a soothing balm to [her][2] turmoil .
Finally , [Charlize][2] opens [[her][2] eyes][8] , [her][2] gaze filled with determination .
[She][2] reaches out with [her][2] hand , [her][2] fingers brushing against [Keanu 's][1] .
“ [We][7] 'll figure this out together , ” [she][2] says , [her][2] voice steady .
[Keanu][1] smiles , a sense of hope filling [[his][1] heart][10] .
“ Yes , ” [he][1] replies , [his][1] eyes shining with resolve .
As [they][7] stand together , the dim light of [the room][4] casting a warm glow around [them][7] , [they][7] know that [whatever challenges lie ahead][11] , [they][7] will face [them][11] as a team .

Gemini 2.0 Flash, MCC 0.66**:**

[Keanu][1] looks intently at [Charlize , who seems distressed with [[her][2] hand][3] to [her face][4]][2] .
[Charlize][2] appears more upset with tears streaming down [[her][2] face][4] and a blanket wrapped around [her][2] shoulders .
[She][2] looks up , perhaps pleading for answers .
In the third image , [she][2] closes [her][2] eyes , clearly in pain .
Then , in the next image [she][2] raises [[her][2] hand][3] as if to speak .
It seems a tragic event has left [Charlize][2] distraught , while [Keanu][1] looks like [he][1] is trying to help .

GPT4o, MCC 0.67**:**

[Keanu][1] walked briskly through the echoing halls , urgency marking [his][1] steps .
[He][1] found [Charlize][2] wrapped in [a pink blanket][3] , tears streaming down [her][2] face .
[[Her][2] eyes][4] met [his][1] , seeking comfort and understanding .
[Keanu][1] kneeled beside [her][2] , concern etched on [his][1] features .

[Charlize][2] looked up , [her][2] voice trembling as [she][2] tried to speak about what had shaken [her][2] so deeply .
[Her][2] vulnerability reached out to [him][1] , and [he][1] offered a steady hand , encouraging [her][2] to share [her][2] burden .
[Charlize][2] struggled , [her][2] words breaking under the weight of emotion .
But [Keanu][1] remained patient , listening intently , [his][1] presence a silent promise of support .
As [she][2] continued , [her][2] words began to flow more freely .
[The fear that had gripped [her][2]][5] started to loosen [its][5] hold .
[She][2] clenched [[her][2] blanket][3] tighter , as if drawing strength from the warmth .
[Keanu ’ s][1] soft gaze and reassuring nods communicated what words couldn’t—she was not alone .
With each passing moment , the trust between [them][6] grew .
[[Charlize ’s][2] eyes][4] , though red from tears , began to shine with a glimmer of relief as [Keanu][1] reached out , bridging the space between [them][6] with understanding and care .

`InternVL2.5-78B`, MCC 0.72**:**

[[Keanu][1] and [Charlize][2]][3] find [themselves][3] alone in a dimly lit room , [[their][3] conversation][4] marked by tension and emotion .
[Keanu][1] , dressed in a formal black suit , seems desperate to explain [himself][1] .
[He][1] leans forward , looking serious but hopeful as [he][1] speaks to [Charlize][2] .
The flickering light behind [him][1] casts a somber glow , adding to the gravity of [the conversation][4] .
[Charlize][2] , wrapped in [a pink blanket][5] , sits tensely .
[[Her][2] eyes][6] are filled with a mix of sadness and confusion as [she][2] listens to [Keanu][1] .
[She][2] touches [her][2] head , [her][2] hand trembling slightly .
Tears well up in [[her][2] eyes][6] as [she][2] processes [his][1] words , [her][2] gaze occasionally flickering away from [him][1] .
The atmosphere thickens with emotion .
[Keanu ’ s][1] expression shifts from hopeful to anxious as the significance of what [he][1] is saying becomes more intense .
[He][1] reaches out , as if to bridge the emotional gap between [them][3] , but pulls back , realizing the depth of [Charlize ’ s][2] turmoil .
[Charlize][2] , clearly overwhelmed , looks down , [her][2] face reflecting the inner battle [she][2] is fighting .
The soft pink of [[her][2] blanket][5] contrasts sharply with the seriousness of [her][2] emotions .
[Keanu][1] , sensing [her][2] distress , attempts to comfort [her][2] , [his][1] voice taking on a more soothing tone .
Finally , [Charlize][2] raises [[her][2] tear-filled eyes][6] , a mix of sadness and acceptance visible in [her][2] gaze .
[Keanu][1] looks back at [her][2] , [his][1] expression softening into relief and love .
The image sequence ends with a tender moment , as [Charlize][2] , despite [her][2] sorrow , gives a faint nod , signaling that [she][2] understands and forgives [him][1] .
The tension between [them][3] eases , replaced by a poignant understanding .

`Qwen2-VL-72B`, MCC 0.76

[[Keanu][1] and [Charlize][2]][3] found [themselves][3] in a mysterious , dimly lit room with high arches and intricate designs on the walls .
[They][3] had stumbled upon [an ancient artifact][4] during [their][3] travels , and [they][3] were both captivated by [[its][4] power][5] .
As [they][3] examined [the artifact][4] , [Charlize][2] suddenly felt a sharp pain in [her][2] chest .
[She][2] clutched [her][2] heart and staggered backward , [[her][2] face][6] contorted in agony .
[Keanu][1] rushed to [her][2] side , [[his][1] eyes][7] filled with concern .
[He][1] tried to help [her][2] up , but [she][2] was too weak to stand .
[[Charlize ’s][2] eyes][8] rolled back , and [she][2] collapsed onto the floor .
[Keanu][1] knelt beside [her][2] , gently cradling [her][2] head in [his][1] hands .
[He][1] could see the fear and confusion in [[her][2] eyes][8] as [she][2] struggled to breathe .
With a deep breath , [Charlize][2] managed to speak .
“ What ... what ’s happening to [me][2] ? ”
[she][2] gasped .
[Keanu][1] looked around frantically , searching for any clue that might explain [her][2] sudden collapse .
Suddenly , [he][1] noticed [a faint glow][9] emanating from [the artifact][4] .
As [he][1] reached out to touch [it][4] , [he][1] felt a surge of energy coursing through [his][1] body .
[He][1] realized that [the artifact][4] was somehow linked to [Charlize ’s][2] condition .
[Keanu][1] focused all [his][1] energy on [the artifact][4] , willing [it][4] to reveal [its][4] secrets .
Slowly , [the glow][9] intensified , and [a shimmering image][10] appeared before [him][1] .
[It][10] showed [Charlize][2] , healthy and vibrant , standing in front of [the same artifact][4] .
[Keanu][1] knew what [he][1] had to do .
[He][1] closed [[his][1] eyes][7] and concentrated , channeling [[the artifact ’s][4] power][5] into [Charlize ’s][2] body .
Slowly , [her][2] breathing steadied , and [[her][2] eyes][8] fluttered open .
[She][2] looked up at [Keanu][1] , a smile spreading across [[her][2] face][6] .
“ Thank [you][1] , ” [she][2] whispered .
[Keanu][1] smiled back , relieved that [she][2] was safe .
Together , [they][3] stood up , ready to continue [their][3] journey , knowing that [they][3] had uncovered something truly extraordinary .

`Qwen2-VL-7B`, MCC 0.83:

In the quiet of [the room][1] , [[Keanu][2] and [Charlize][3]][4] stood facing each other , [their][4] eyes locked in a tense conversation .
[Charlize 's][3] face was marked with [tears][5] , and [[her][3] body language][6] suggested [she][3] was in distress .
[Keanu][2] , with a concerned expression , seemed to be trying to comfort [her][3] .
As the scene progressed , [[Charlize 's][3] tears][5] flowed freely , and [[her][3] body language][6] became more vulnerable .
[Keanu][2] , still looking at [her][3] with a mixture of concern and empathy , seemed to be trying to understand [her][3] pain .
[The room][1] was filled with an intense atmosphere , and the tension between [the two characters][4] was palpable .

G Example III

Image 1

Image 2

Image 3

Image 4

Image 5

Character image 1: Bruce

Human:

[The police officer][1] closed the zipper of the body bag .
[Journalists][2] were already flooding the place .
[Bruce][3] asked [the officer][1] how the hell [all these journalists][2] know already about the dead body .
[The police officer][1] looked desperate .
[He][1] said that every time had been like that in the last months .
They mast have a snitch in the station .
[He][1] left [Bruce][3] and headed to [the crowd of journalists][2] .
It was about time for some order there .

DeepSeek-VL2-1B

[The sequence of images][1] depicts a chaotic scene with multiple individuals involved in what appears to be a public disturbance or protest .
[The central figure , Bruce][2] , is a police officer who seems to be addressing [the crowd][3] .
[He][2] is surrounded by other officers and a group of people , some of whom are holding cameras and microphones , indicating media presence .
The atmosphere is tense , with [the crowd][3] appearing agitated and [the officer 's][2] expression serious .
The story that fits [the sequence of images][1] is as follows : [Bruce , a dedicated police officer][2] , is assigned to a high-profile case involving [a viral social media video that has captured the attention of [the public][5]][4] .
[The video , which depicts a young woman being attacked][4] , has gone viral , sparking widespread outrage and [protests][6] .
As [the officer][2] , [Bruce][2] is tasked with maintaining order and ensuring the safety of the protesters while also handling [the media 's][7] demands for information .
As the situation escalates , [Bruce][2] finds [himself][2] at the center of the conflict , facing pressure from both [the public][5] and [the media][7] .
[He][2] must navigate the complexities of balancing [his][2] duty to protect [the community][8] with the demands of [the press][7] .
Despite the challenges , [Bruce][2] remains committed to [doing][9] what is right , even if [it][9] means facing criticism and scrutiny .
In the end , [Bruce 's][2] actions lead to a resolution that addresses the underlying issues of [the video][4] and [the protests][6] , ultimately bringing closure to [the community][8] while also highlighting the importance of responsible journalism and public service .

DeepSeek-VL2-4.5B

[Bruce][1] stood in [the dimly lit room][2] , [his][1] hands trembling as [he][1] clutched [the black plastic bag][3] .
[The weight of the unknown inside][4] pressed down on [him][1] , a heavy burden [he][1] had to carry .
[The faint blue lines on [the floor][5]][6] seemed to mock [him][1] , a reminder of the mundane world [he][1] was leaving behind .
[He][1] glanced at [the metal cart][7] , [its][7] wheels cold and unyielding , a symbol of the journey ahead .
The shadows danced around [him][1] , whispering secrets [he][1] could n't yet decipher .
As [he][1] stepped into [the bustling crowd][8] , [the weight of [the bag][3]][4] felt even heavier .
[The cameras][9] flashed , [their][9] lights blinding [him][1] momentarily .
[He][1] saw [the man in the blue jacket , the word “ POLICE ” emblazoned on [his][10] back][10] , and [[the crowd 's][8] murmurs][11] grew louder .
[Bruce 's][1] heart raced , [his][1] mind a whirlwind of questions .
Who was [he][10] ?
What had [he][10] done ?
[The police officer 's][10] stern gaze met [his][1] , and for a moment , time stood still .
[Bruce 's][1] voice trembled as [he][1] spoke , “ [I][1] ...
[I][1] do n't know what 's inside . ”
[The officer 's][10] expression softened , but [[the crowd 's][8] murmurs][11] grew louder .
[The man in the gray sweater][12] stepped forward , [his][12] eyes filled with concern .
“ We 'll help [you][1] , ” [he][12] said , [his][12] voice steady .
[Bruce][1] nodded , a small smile tugging at [his][1] lips .
[The officer][10] led [Bruce][1] through [the crowd][8] , [the cameras][9] capturing every moment .
[The man in the gray sweater][12] followed , [his][12] presence a comforting anchor .
As they moved through [the crowd][8] , [Bruce][1] felt a sense of relief wash over [him][1] .
[The bag , once a symbol of [his][1] burden][3] , now felt lighter , as if [the weight][4] had been lifted .
In the end , [Bruce][1] stood in [the dimly lit room][2] , [the bag][3] now empty .
[The blue lines on [the floor][5]][6] seemed to glow , a reminder of the journey [he][1] had taken .
[The cameras][9] were gone , [the crowd][8] dispersed , but [Bruce][1] felt a sense of peace .
[He][1] had faced [his][1] fears , and in doing so , had found a new sense of purpose .

Gemini 2.0 Flash

[A black body bag][1] was lying on a stretcher .
Someone 's hands were touching [the body][2] , perhaps to see if [that person][2] was still alive .
Later that day , [a coroner][3] was surrounded by a press conference .
[The coroner 's][3] actions were probably the focus of the media .
The same scene showed [Bruce][4] arguing with [a police officer][5] .
[He][4] seemed worried , or maybe just angry , but certainly not happy with [the officer][5] .
[The police officer][5] was not listening , or had [his][5] own agenda .
[He][5] seemed unwilling to help [Bruce][4] , or maybe just was doing [his][5] job , blocking access to the crime scene .
Many reporters and cops were present , creating [a hectic situation][6] .
[The situation][6] seemed to be about [the body in [the black bag][1]][2] and perhaps the circumstances that lead to this tragedy .
[Bruce 's][4] body language seemed to indicate that [he][4] was not pleased with the current state of things .

GPT4o

[A mysterious situation][1] unfolded late one night in the bustling corridors of the city hospital .
A black body bag lay ominously on a stretcher , drawing worried glances .
The coroner , surrounded by flashing cameras and an eager crowd of reporters , quietly spoke to the press , trying to maintain some level of order amidst [the chaos][2] .
In the middle of [this confusion][2] , [Bruce , a concerned onlooker who found himself amidst [the growing throng][4]][3] , tried to gather information .
[His][3] brow furrowed with concern as [he][3] spoke directly with [a nearby police officer][5] .
Tension buzzed through the air as [Bruce][3] pressed for answers .
[The officer][5] , surrounded by [his][5] colleagues , remained firm but sympathetic .
[He][5] communicated with [Bruce][3] , explaining the procedural details and what little information could be shared .
[He][5] ensured [Bruce][3] understood that [the situation][1] was under control , even if answers were elusive .
As [the crowd][4] continued to swell , [reporters][6] pushed closer , eager for any snippets of information [they][6] could relay to the world outside .
[Bruce][3] stood steadfast amidst the bustle , determined to uncover more .
[The situation][1] was intense , but [he][3] remained resolute in seeking the truth .
Through [the restless chaos][2] , [the police officer][5] gestured [Bruce][3] to follow [him][5] to a quieter corner , hoping to provide clearer insights into [the situation][1] .
The night was far from over , and the truth lay hidden within the depths of the shadows .

InternVL2.5-78B

The chaotic scene unfolded with intense fervor , as a diverse group of reporters clamored for information .
The news was breaking , and it was clear that this moment would go down in history .
Amidst the bustling media presence , [a black police officer , distinguished by his stern expression and authoritative demeanor][1] , was the focal point of the frenzy—a veritable island of order in the midst of the storm .
Known to many by [his][1] name , Bruce , [he][1] was well-versed in handling crowd control .
[A man in [a gray sweater][3]][2] approached [Bruce][1] with evident urgency , [his][2] eyes betraying a mixture of concern and determination .
The exchange between [them][4] was brief but tense , encapsulating the gravity of [the situation][5] .
With a stern look , [Bruce][1] listened intently , a silent sentinel assessing [the man 's][2] words with a professional detachment .
As [Bruce][1] gave [his][1] serious reply , [the man in [the gray sweater][3]][2] seemed to plead earnestly for cooperation .
The urgency was palpable , underscoring the high stakes of [their][4] interaction .
The camera lights flickered in the background , casting a stark , almost surreal light on the unfolding drama .
Just then , [a man in a leather jacket and cap][6] maneuvered through the crowd , seemingly oblivious to the surrounding commotion .
With a serious expression , [he][6] seemed to have just entered from a nearby door , [his][6] presence yet another layer of complexity to [the situation][5] .
[The man in [the gray sweater][3]][2] looked at [Bruce][1] with unflinching resolve as [they][4] continued [their][4] conversation , the tension between [them][4] thickening the electric atmosphere .
[The camera crews , including one with a large ' W ' sign on the back of [their][7] jacket][7] , were recording every word—a testament to the importance of the discussion .
Amidst all this , [another black police officer][8] , [his][8] badge gleaming under the harsh lights , managed the chaotic throngs of reporters .
[His][8] calm yet vigilant presence served as a counterbalance to the frenzied energy swirling around [him][8] .
[The scene][9] captured a moment fraught with tension , as [[the man in [the gray sweater][3]][2] and [Officer Bruce][1]][4] navigated a world where the stakes were high and every word counted .
[It][9] was a snapshot of history in the making , a testament to the intersection of media , law enforcement , and the relentless pursuit of the truth .

Qwen2-VL-72B

[Bruce][1] was lying on the floor , unconscious , after being shot .
[He][1] had been involved in a dangerous situation , and things had taken a turn for the worse .
As [he][1] lay there , a group of people gathered around [him][1] , including [a police officer][2] .
[The officer][2] was trying to maintain order amidst the chaos .
The scene was chaotic , with cameras flashing and reporters shouting questions .
[Bruce 's][1] condition was critical , and [the police officer][2] was doing [his][2] best to keep everyone at bay .
Despite the commotion , [the officer][2] remained calm and collected , ensuring that medical help arrived quickly .
As the situation unfolded , it became clear that [Bruce][1] had been targeted .
[The police officer][2] questioned witnesses and gathered evidence , determined to find out who was responsible for the attack .
[[Bruce 's][1] friends and colleagues][3] were shocked and devastated by the news , but [they][3] knew that [the police officer][2] would do everything in [his][2] power to bring [the perpetrator][4] to justice .
Despite the danger and uncertainty , [the police officer][2] remained steadfast in [his][2] duty .
[He][2] worked tirelessly to ensure that [Bruce][1] received the care [he][1] needed and that justice was served .
In the end , thanks to [his][2] efforts , [the culprit][4] was caught , and [Bruce][1] made a full recovery .
[The police officer][2] had proven [himself][2] to be a true hero , risking [his][2] own safety to protect others and uphold the law .

Qwen2-VL-7B

The sequence of images depicts [a tense and chaotic scene involving [a police officer][1] and [a crowd of people][2]][3] .
[The officer][1] , wearing [a police uniform][4] , appears to be in a confrontational situation with [a group of individuals , including [a man in a gray sweater][5]][2] .
[The crowd][2] seems to be gathered around [the officer][1] , with some people holding cameras and microphones , suggesting a media presence .
[The story][6] begins with [a police officer][1] standing in the middle of [a crowd][2] , surrounded by reporters and photographers .
[The officer][1] looks tired and stressed , with dark circles under [his][1] eyes .
[He][1] is wearing [a police uniform][4] and a badge on [his][1] chest .
[The crowd][2] is , with people asking questions and taking pictures .
As [the officer][1] tries to maintain control , [a man in [a gray sweater][7]][5] approaches [him][1] .
[The man][5] appears to be angry and confrontational , yelling at [the officer][1] and [the crowd][2] .
[The officer][1] remains calm and composed , but [the situation][8] escalates quickly .
Suddenly , a camera flash goes off , and [the officer][1] turns to look at the source .
[He][1] seems to be on edge , as if anticipating trouble .
[The crowd][2] continues to press in , and [the officer 's][1] body language suggests [he][1] is preparing for a physical altercation .
In the midst of [the chaos][9] , [a police officer in [a hat and uniform][11]][10] steps forward , trying to calm [the situation][8] .
[He][10] speaks firmly to [the crowd][2] , but the tension remains high .
[The officer in [the gray sweater][7]][5] continues to argue , and [the crowd][2] seems to be divided , with some people supporting [the officer][1] and others siding with [the man in [the gray sweater][7]][5] .
As [the scene][3] reaches [its][3] climax , [the officer in [the hat and uniform][11]][10] tries to intervene , but [the situation][8] quickly degrades into a physical altercation .
[The officer in [the gray sweater][7]][5] appears to be physically attacking [the police officer][1] , and [the crowd][2] becomes increasingly hostile .
[The story][6] ends with [the police officer][1] being overwhelmed by [the crowd][2] , and [the situation][8] spiraling out of control .
[The officer in [the hat and uniform][11]][10] tries to help , but [the chaos][9] is too much to handle .
[The scene][3] is a vivid portrayal of a tense and dangerous situation , with [the police officer][1] struggling to maintain order in the face of [a hostile crowd][2] .