

When More Words Say Less: Decoupling Length and Informativity in Image Description Evaluation

Rhea Kapur
Stanford University
rheak@stanford.edu

Robert Hawkins
Stanford University
rdhawkins@stanford.edu

Elisa Kreiss
University of California, Los Angeles
ekreiss@ucla.edu

Abstract

Vision-language models (VLMs) are increasingly used to make visual content more accessible via text-based descriptions. In current systems, however, informativity is often conflated with *utterance length*. We argue that these two concepts ought to be disentangled: descriptions can be concise yet information-dense or, conversely, lengthy yet vacuous. Drawing from literature on pragmatic language use, we define informativity in terms of a contrast set: a description is informative to the extent that it describes the target image better than other images. To operationalize this, we introduce a dataset that controls for length while varying information content, as well as a diagnostic tool for measuring informativity directly. Our human preference study shows that people reliably prefer more informative descriptions regardless of length, and our tool enables fine-grained analysis of informativity as the core variable of interest. These results support the design and evaluation of VLMs that prioritize informativeness over verbosity, with implications for accessibility applications.

1 Introduction

Vision-language models (VLMs) strive to establish bidirectional correspondences between images and text, raising deep questions about how to represent the same information across modalities (Zhang et al., 2024; Wang et al., 2024; Deitke et al., 2024; Ghandi et al., 2023). With today’s models, these questions are not only of theoretical interest but have real consequences for downstream application, such as image accessibility for blind, low-vision, and sighted individuals (Morris et al., 2016; Gleason et al., 2019; Stangl et al., 2020). When generating descriptions, VLMs must navigate challenging issues about the appropriate level of *informativity*: what information should be included, in what level of detail, and how should that information vary across different contexts and user needs?

In the growing body of research that discusses what makes a “good” image description, we often use *length* as a proxy measure: we assume that longer outputs are more informative (e.g., Ding et al., 2024; Nitta et al., 2024; Singhal et al., 2024). This holds for studying people’s description preferences (Williams et al., 2022; Kreiss et al., 2022), for establishing novel image description/captioning datasets (e.g., Urbanek et al., 2024), for developing evaluation metrics (Kreiss et al., 2022; Kapur and Kreiss, 2024), or for the design of accessibility guidelines (see McCall and Chagnon, 2022). While there is a natural correlation between length and information content — longer descriptions can, all else being equal, pack in more information (Shannon, 1948) — this relationship is far from straightforward (Chen et al., 2022). In fact, we often want to improve a system’s informativeness while controlling for excessive verbosity (Singhal et al., 2024; Dubois et al., 2024; Nayab et al., 2024; Hu et al., 2024). In this paper, we aim to delineate these concepts and propose a design tool that starts getting at the true variable of interest.

In pragmatics, informativity is commonly defined using the Gricean maxim of quantity. Speakers should provide just as much information as needed — no more, no less — to achieve their communicative goals (Grice, 1975). This maxim has been formalized in computational pragmatics in relative terms. We say a description is informative to the extent that it is able to distinguish a referent from contextually salient alternatives (e.g., Dale and Reiter, 1995; Krahmer and Van Deemter, 2012; Rubio-Fernandez, 2019; Sedivy, 2003; Chen and van Deemter, 2023).

Based on these fundamental linguistic insights, we developed a dataset containing a diverse set of image descriptions that were designed to manipulate length independent of information content, and a tool that can quantify informativity itself. Our findings, firstly, replicate our main assumption



Source	Description
ORIGINAL	There are three girls playing a video game together.
COMPOSITE	Three young girls are sitting next to each other, playing video games together, specifically using Nintendo Wii with wheels.
VERBOSE	In the current situation, there are a total of three girls who are engaged in the activity of playing a video game together.
IMAGE-TO-TEXT	The image shows three girls sitting together on a white stool. The girl on the left is wearing a red onesie, the middle girl is dressed in a pink top, and the girl on the right is wearing a blue top. Each girl is holding a game controller. The background features a blue wall and appears to be a living space, likely a playroom or family room. The lighting in the image is warm and soft.

Table 1: Example set of descriptions for an image in our dataset. Composite and verbose descriptions are longer variants of the original description, but vary in the amount of additional information provided. Image-to-Text is an example output of a VLM (here, GPT-4o-mini) with minimal instructions. See Appendix B for additional examples.

that people’s description preferences are driven by informativity rather than length. Secondly, our proposed tool that quantifies informativity based on a contrast set manages to delineate length and informativity, and we show an example of how this can provide novel insights for developing better aligned VLM-generated descriptions.

2 Related Work

2.1 Defining informativity via contrast sets

Recent models of referring expression generation (REG) have formalized informativity with respect to a listener and context (Krahmer and Van Deemter, 2012). For example, in the Rational Speech Act (RSA) framework (Degen et al., 2020; Degen, 2023), speakers select utterances that maximize the likelihood of a listener identifying the intended referent while minimizing production costs. Under this framework, not all words contribute equally to increasing informativity. It is not the sheer quantity of information, but the inclusion of distinguishing features that differentiate the target from contextually relevant alternatives in a contrast set. This contrast set is made explicit in discriminative or issue-sensitive captioning tasks (Ou et al., 2023; Cohn-Gordon et al., 2018; Nie et al., 2020; Andreas and Klein, 2016), but it is not part of common image description datasets (Illykh et al., 2018, 2019; Pezzelle, 2023; Takmaz et al., 2022). We argue that to quantify informativity in isolation, we must consider an *implicit* contrast set.

2.2 Limitations of evaluation metrics

Existing evaluation metrics for image captioning fail to disentangle length from informativity. Reference-based metrics like BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005) primarily assess simi-

larity to human-written references, often implicitly favoring captions of lengths similar to those of the references and failing to capture human judgments of informativity (Kapur and Kreiss, 2024). Referenceless metrics such as CLIPScore (Hessel et al., 2021) measure image-text alignment but do not explicitly account for the contrastive value of the information provided (Kreiss et al., 2022). We argue that none of these metrics adequately capture the informativity of a description independent of its length in communication-theoretic terms (Newman et al., 2020; Tang et al., 2024; Coppock et al., 2020). By systematically examining this relationship, we aim to caution against the assumption that longer descriptions are inherently more informative and to encourage more nuanced approaches to both generating and evaluating image descriptions.

3 Dataset

To systematically investigate the relationship between length and informativity, we created a controlled dataset based on MS COCO data (Lin et al., 2015). We randomly selected 5,000 images balanced across COCO’s 80 object categories and generated multiple description variants for each image that deliberately varied in length and information content. Each image is paired with four descriptions (Table 2):

Original: A randomly selected human-written description for the image from MS COCO.

Verbose: A longer and more elaborate rephrasing of the original description without adding content.

Composite: An extended version of the original that combines content from all five COCO references, adding more distinct details.

Image-to-Text: A VLM-generated description for the image, using minimal instructions.

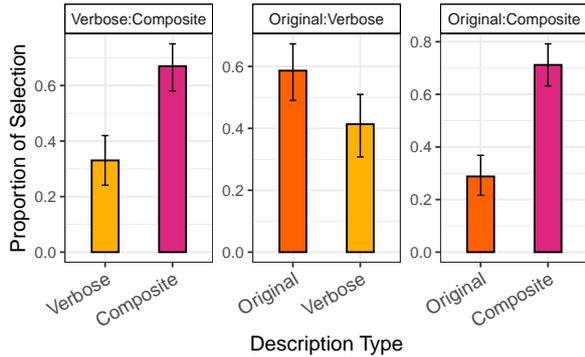


Figure 1: Average proportion of times each description type was selected in human pairwise preference comparisons. Error bars indicate 95% bootstrapped confidence intervals. Full results, including the VLM conditions are in Appendix C.

The latter three description types were generated using OpenAI’s GPT-4o-mini (OpenAI et al., 2024). Verbose and composite descriptions are both longer variants of the original description, but they differ in their informativity. An informativity tool should be able to pick up this difference. The image-to-text description condition enables us to compare the default VLM outputs in this framework. The specific prompts used for each condition are provided in Appendix A. Together, this dataset creates a rich testbed for analyzing the interaction between length and informativity.

4 People Prefer Informativity

We conducted a human subject experiment to test whether informativity—not length—drives preferences and to validate our dataset conditions.

We recruited 30 participants on Prolific. Participants saw an image and were asked to select which out of two potential descriptions they preferred. We randomly selected the stimuli from all images in our synthesized dataset where the verbose and composite descriptions were of similar length so that we could clearly separate out effects of length and informativity. The complete participant prompt, a UX screenshot, and further details about this study are included in Appendix C.

In Figure 1, we compare the proportion of times each description type was selected as preferred in the six possible pairwise matchups. Crucially, participants show a strong and statistically significant preference for the composite descriptions over both the verbose ($\beta = -0.71$, $SE = 0.20$, $z = -3.52$, $p < .001$) and original variants ($\beta = 0.91$, $SE = 0.20$, $z = 4.6$, $p < .001$).

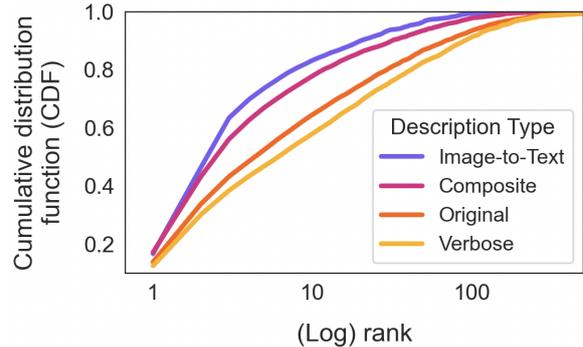


Figure 2: Cumulative distribution of ranks across description types relative to all other images.

The generalized linear model results support a ranking of composite over original and verbose, but we don’t find statistical support for a preference order between original and verbose ($\beta = -0.35$, $SE = 0.20$, $z = -1.76$, $p = .08$). These preference results are in line with our expectations: Descriptions from the higher informativity condition (composite) are preferred over lower informativity (verbose) and shorter length descriptions (original).

5 Method: The Informativity Tool

The main challenge of developing an informativity metric is that the concept of informativity isn’t defined on an absolute scale (Nie et al., 2020; Degen et al., 2020). For this reason, we have to define a contrast set: We use a set of contrasting images to the target image and define the informativity of a description as the extent to which the description allows us to single out the target (e.g., Ou et al., 2023; Takmaz et al., 2022).

For each image-description pair, we define the contrast set as the remaining 4,999 images from our dataset. We quantify the informativity of a description as its ability to discriminate the target image from its 4,999 competitors. The intuition is that the more informative a description, the more uniquely does it apply to only a single image (consider that all images that show “an albacore” show “a fish” but not vice versa).

Due to its discriminative loss training objective, CLIPScore (Hessel et al., 2021) provides one opportunity to quantify this link (Ou et al., 2023; Takmaz et al., 2022). We use CLIPScore to obtain compatibility ratings of the description with the true and all contrasting images in the dataset, yielding a rank distribution over all images. We calculated all CLIPScores on a NVIDIA RTX 6000 Ada Genera-

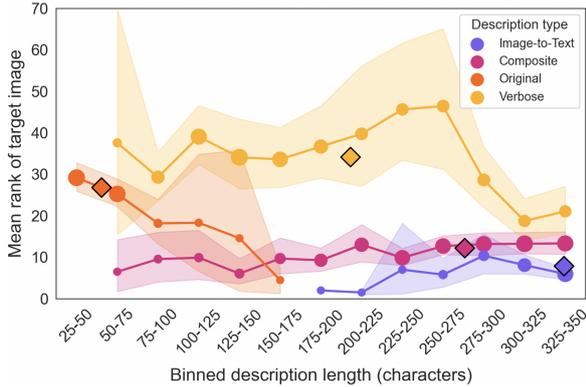


Figure 3: Mean rank as a function of description length (binned by character count) for each description type. Diamonds represent means across length.

tion GPU in ~ 5 hours. We take the rank of the true image as the approximation of the description’s informativity level, such that lower average ranks indicate higher informativity and vice versa.

6 Evaluating the Informativity Tool

In order to evaluate this approach for quantifying informativity, we first validate that it separates out the description conditions. To that end, we show the cumulative distribution function (CDF) of the target image’s rank across description conditions in Figure 2. This figure visualizes for each description condition the proportion of cases where the target image ranked at or above a given rank. There is a clear informativity hierarchy which, as predicted in our human subject study (Section 4), finds that target images for composite descriptions receive lower ranks than original and verbose descriptions. This provides first evidence that this method picks up on a notion of informativity.

To specifically delineate the length and informativity level of descriptions across the different conditions, we show the mean rank across description lengths in Figure 3. Overall, we observe again the informativity hierarchy illustrated in Figure 2, which largely persists across description lengths. However, there are a number of additional insightful patterns that emerge. First, only the original descriptions become more informative with increasing length ($r = -0.034, p < 0.05$), i.e., longer descriptions correlate with higher levels of informativity. All other description conditions either do not show this trend (verbose, image-to-text) or in fact show a reverse trend where longer descriptions become comparatively less informative (composite: $r = 0.047, p < 0.01$).

In sum, our results suggest that this discriminative approach for quantifying informativity is a promising step towards moving away from length-as-proxy and towards metrics that can quantify informativity more directly. The following case study starts to highlight how insights from this methodology can directly inform generated image description evaluation and prompt design.

6.1 Case study: Evaluating informativity of VLM-generated descriptions

As shown in Figure 3, the VLM-generated descriptions are difficult to compare with the original data due to the stark length difference. But what happens to informativity levels when we instruct the model to constrain its length? To answer this, we used VLM-generated descriptions produced under two length constraints: one prompting the model to be “as **concise** as possible,” and another limiting the response “to not exceed **200 characters**” (see Appendix A for full prompts). We calculated CLIP-Scores with these and all images on a NVIDIA RTX 6000 Ada Generation GPU in ~ 2.5 hours.

Using our informativity tool, we find the descriptions from the 200 character condition are less informative than the unrestricted condition ($t = 1.014, p = 3.106e - 01$), which is on par with the constraint condition — a pattern that holds across description lengths (see Appendix 4). Intriguingly, these results suggest that different strategies for constraining lengths might result in generated descriptions of varying information density. We can only obtain these types of insights with the availability of tools that go beyond length, opening an exciting opportunity to align model behavior.

7 Conclusion

In the context of VLMs becoming increasingly critical for making online visual content accessible through image descriptions, we have shown that description length is not a consistent proxy for informativity even though the two are frequently conflated. We provide a proposal for a tool that can delineate the two and show how it can successfully be applied to VLM-generated descriptions. Our findings call for models and corresponding evaluation metrics that move beyond using length as a surrogate for informativity, instead optimizing for informativeness relative to context (via a contrast set or another means) and accessibility-minded user preferences.

Limitations

While being able to reliably quantify informativity brings much analytical power for model alignment, it’s important to clarify that this doesn’t mean that higher informativity is always better. A description can be maximally discriminative by simply listing all objects in the most detailed way but one which is unreadable to a user. Informativity is only one dimension out of many that appears to play a significant role in our description preferences. This tool aims to complement these other efforts.

But also the setup of the tool itself has clear limitations. Our evaluation of informativity relies on 1) CLIPScore and 2) a defined contrast set of images that defines the sensitivity of CLIPScore as a metric. Firstly, previous work has shown that description length is not a consistent predictor of CLIPScores (Kreiss et al., 2022) and together with its discriminative loss training objective, CLIPScore was a promising candidate for this analysis. However, CLIPScore also has constraints such as a 77-token input limit (which required us to exclude longer descriptions), which significantly limits its practical use. However, we would like to highlight that our approach doesn’t commit to the CLIPScore metric — on the contrary. Alternative models can be flexibly used as long as they provide an image-text compatibility score.

Secondly, the contrast set fundamentally constrains the resolution of informativity that the tool is going to provide. There are two dimensions of interest here: the number of overall contrast images, and the similarity of those images in relation to a particular feature. If the overall number of images in the contrast set is small, the tool will quickly reach ceiling effects as the target is likely to be easily picked out from the limited number of contrasts. If the data is artificially dense, e.g., with regards to types of birds, it will be more sensitive to information about birds than other features. In this work, we do find that ranks approach ceiling in certain conditions but we’re achieving sufficient variation to find patterns between conditions, suggesting that our contrast set is sufficiently complex for our needs. To counteract the overrepresentation of certain features, we sampled contrast images uniformly across categories. However, we can’t exclude the possibility that certain features are over-represented and therefore weigh more heavily in the analysis. Nevertheless, these variations should average out on sufficiently large data samples.

Taken together, there are many challenging design choices when implementing an informativity tool. Despite that, we show that when making informed decisions about these constraints, an informativity tool can provide insights that *length as a proxy* can never deliver.

Acknowledgments

We thank Google’s GiG program for supporting this research.

References

- Jacob Andreas and Dan Klein. 2016. Reasoning about pragmatics with neural listeners and speakers. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1173–1182.
- Satanjeev Banerjee and Alon Lavie. 2005. **METEOR: An automatic metric for MT evaluation with improved correlation with human judgments**. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.
- Guanyi Chen and Kees van Deemter. 2023. Varieties of specification: Redefining over-and under-specification. *Journal of Pragmatics*, 216:21–42.
- Yuyan Chen, Yanghua Xiao, and Bang Liu. 2022. Grow-and-clip: Informative-yet-concise evidence distillation for answer explanation. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)*, pages 741–754. IEEE.
- Reuben Cohn-Gordon, Noah Goodman, and Christopher Potts. 2018. Pragmatically informative image captioning with character-level inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 439–443.
- Elizabeth Coppock, Danielle Dionne, Nathaniel Graham, Elias Ganem, Shijie Zhao, Shawn Lin, Wenxing Liu, and Derry Wijaya. 2020. **Informativity in image captions vs. referring expressions**. In *Proceedings of the Probability and Meaning Conference (PaM 2020)*, pages 104–108, Gothenburg. Association for Computational Linguistics.
- Robert Dale and Ehud Reiter. 1995. Computational interpretations of the gricean maxims in the generation of referring expressions. *Cognitive Science*, 19(2):233–263.
- Judith Degen. 2023. **The rational speech act framework**. *Annual Review of Linguistics*, 9(Volume 9, 2023):519–540.

- Judith Degen, Robert D. Hawkins, Caroline Graf, Elisa Kreiss, and Noah D. Goodman. 2020. [When redundancy is useful: A bayesian approach to “overinformative” referring expressions](#). *Psychological Review*, 127(4):591–621.
- Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, and 1 others. 2024. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. *arXiv preprint arXiv:2409.17146*.
- Ning Ding, Chaorui Deng, Mingkui Tan, Qing Du, Zhiwei Ge, and Qi Wu. 2024. [Image captioning with controllable and adaptive length levels](#). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(2):764–779.
- Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. 2024. Length-controlled alpaca-eval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*.
- Taraneh Ghandi, Hamidreza Pourreza, and Hamidreza Mahyar. 2023. [Deep learning approaches on image captioning: A review](#). *ACM Comput. Surv.*, 56(3).
- Cole Gleason, Patrick Carrington, Cameron Cassidy, Meredith Ringel Morris, Kris M. Kitani, and Jeffrey P. Bigham. 2019. [“it’s almost like they’re trying to hide it”: How user-provided image descriptions have failed to make twitter accessible](#). In *The World Wide Web Conference, WWW ’19*, page 549–559, New York, NY, USA. Association for Computing Machinery.
- H. Paul Grice. 1975. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Syntax and Semantics, Volume 3: Speech Acts*, pages 41–58. Academic Press.
- Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. [CLIPScore: A reference-free evaluation metric for image captioning](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Zhengyu Hu, Linxin Song, Jieyu Zhang, Zheyuan Xiao, Zhengyu Chen, and Hui Xiong. 2024. Explaining length bias in llm-based preference evaluations. In *ICLR 2025 Workshop on Navigating and Addressing Data Problems for Foundation Models*.
- Nikolai Ilinykh, Sina Zarrieß, and David Schlangen. 2018. [The task matters: Comparing image captioning and task-based dialogical image description](#). In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 397–402, Tilburg University, The Netherlands. Association for Computational Linguistics.
- Nikolai Ilinykh, Sina Zarrieß, and David Schlangen. 2019. [Tell me more: A dataset of visual scene description sequences](#). In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 152–157, Tokyo, Japan. Association for Computational Linguistics.
- Rhea Kapur and Elisa Kreiss. 2024. [Reference-based metrics are biased against blind and low-vision users’ image description preferences](#). In *Proceedings of the Third Workshop on NLP for Positive Impact*, pages 308–314, Miami, Florida, USA. Association for Computational Linguistics.
- Emiel Krahmer and Kees Van Deemter. 2012. Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218.
- Elisa Kreiss, Cynthia Bennett, Shayan Hooshmand, Eric Zelikman, Meredith Ringel Morris, and Christopher Potts. 2022. [Context matters for image descriptions for accessibility: Challenges for referenceless evaluation metrics](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4685–4697, Abu Dhabi, United Arab Emirates. 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.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. [Microsoft coco: Common objects in context](#). *Preprint*, arXiv:1405.0312.
- Karen McCall and Beverly Chagnon. 2022. Rethinking Alt Text to Improve Its Effectiveness. In *International Conference on Computers Helping People with Special Needs*, pages 26–33. Springer.
- Meredith Ringel Morris, Annuska Zolyomi, Catherine Yao, Sina Bahram, Jeffrey P. Bigham, and Shaun K. Kane. 2016. [“with most of it being pictures now, i rarely use it”: Understanding twitter’s evolving accessibility to blind users](#). In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, CHI ’16*, page 5506–5516, New York, NY, USA. Association for Computing Machinery.
- Sania Nayab, Giulio Rossolini, Giorgio C. Buttazzo, Nicolamaria Manes, and Fabrizio Giacomelli. 2024. [Concise thoughts: Impact of output length on llm reasoning and cost](#). *CoRR*, abs/2407.19825.
- Benjamin Newman, Reuben Cohn-Gordon, and Christopher Potts. 2020. [Communication-based evaluation for natural language generation](#). In *Proceedings of the Society for Computation in Linguistics 2020*, pages 116–126, New York, New York. Association for Computational Linguistics.
- Allen Nie, Reuben Cohn-Gordon, and Christopher Potts. 2020. [Pragmatic issue-sensitive image captioning](#).

- In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1924–1938, Online. Association for Computational Linguistics.
- Tomoya Nitta, Takumi Fukuzawa, and Toru Tamaki. 2024. [Fine-grained length controllable video captioning with ordinal embeddings](#). *CoRR*, abs/2408.15447.
- OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. [Gpt-4o system card](#). *Preprint*, arXiv:2410.21276.
- Jiefu Ou, Benno Kroger, and Daniel Fried. 2023. Pragmatic inference with a clip listener for contrastive captioning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1904–1917.
- 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 on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.
- Sandro Pezzelle. 2023. [Dealing with semantic under-specification in multimodal NLP](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12098–12112, Toronto, Canada. Association for Computational Linguistics.
- Paula Rubio-Fernandez. 2019. [Overinformative speakers are cooperative: Revisiting the gricean maxim of quantity](#). *Cognitive Science*, 43(11):e12797.
- Julie C Sedivy. 2003. Pragmatic versus form-based accounts of referential contrast: Evidence for effects of informativity expectations. *Journal of Psycholinguistic Research*, 32:3–23.
- Claude E Shannon. 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423.
- Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. 2024. A long way to go: Investigating length correlations in rlhf. In *First Conference on Language Modeling*.
- Abigale Stangl, Meredith Ringel Morris, and Danna Gurari. 2020. ["person, shoes, tree. is the person naked?" what people with vision impairments want in image descriptions](#). In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–13, New York, NY, USA. Association for Computing Machinery.
- Ece Takmaz, Sandro Pezzelle, and Raquel Fernández. 2022. [Less descriptive yet discriminative: Quantifying the properties of multimodal referring utterances via CLIP](#). In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 36–42, Dublin, Ireland. Association for Computational Linguistics.
- 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.
- Jack Urbanek, Florian Bordes, Pietro Astolfi, Mary Williamson, Vasu Sharma, and Adriana Romero-Soriano. 2024. A picture is worth more than 77 text tokens: Evaluating clip-style models on dense captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26700–26709.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.
- Candace Williams, Lilian de Greef, Ed Harris III, Leah Findlater, Amy Pavel, and Cynthia Bennett. 2022. Toward supporting quality alt text in computing publications. In *Proceedings of the 19th International Web for All Conference*, pages 1–12.
- Jingyi Zhang, Jiaying Huang, Sheng Jin, and Shijian Lu. 2024. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

A Prompts

In this appendix, we describe the prompts used to generate each description type.

A.1 Verbose

To synthesize a description with the same level of informativity as the *original* description while increasing the length, we passed the *coco* description to GPT-4o-mini with the following prompt:

Given this description, generate one longer description that expresses the same information as in the original description but in a more verbose way. In other words, use more words but say the same thing as given. Do not augment the description with any emotional or made-up information. Only output the longer description and nothing else.

A.2 Composite

To increase the expected level of informativity above and beyond the *verbose* and *original* conditions, we synthesized a *composite* description by passing in *all five* of the original descriptions from COCO with the following prompt:

Given these 5 descriptions, generate one longer, final description that combines all information in the individual descriptions. Do not augment the description with any emotional or made-up information. Only output the longer description and nothing else.

A.3 Image-to-Text

To obtain a baseline for how a VLM describes the image (without reference to a human description), we simply passed the target image to GPT-4o with the following prompt:

Describe this image and don't introduce any emotional information. Just describe what's there.

A.4 Concise

To understand how informativity changes when a VLM is instructed to constrain its length *at its own discretion*, we simply passed the target image to GPT-4o with the Image-to-Text prompt from above with the addition to "be as concise as possible":

Describe this image and don't introduce any emotional information. Just describe what's there. Be as concise as possible.

A.5 Character-limited

Finally, to understand how informativity changes when a VLM is instructed to constrain its length *to a hard cutoff*, we simply passed the target image to GPT-4o with the Image-to-Text prompt from above and the additional instruction "don't exceed 200 characters":

Describe this image and don't introduce any emotional information. Just describe what's there. Don't exceed 200 characters.

B Example Descriptions

We include 2 additional example sets of descriptions for images in our dataset, this time including the length-constrained ablations as well from Section 6.1. These can be found in Tables 2 and 3.

C Human Subject Experiment: Details and Results

Participants were recruited from the crowdsourcing platform Prolific, and recruitment was restricted to within the US, UK and Canada. Participants spent on average 10 minutes on the task and were paid \$14/hr. All data was anonymized before analysis. The anonymized data will be shared upon publication. The study was conducted under the lead author's institution's IRB protocol. The participant prompt for the preference study read as follows:

Thank you for participating in our study!

In this study, you will see 30 images, each paired with two potential descriptions of the image. Your task is to determine which of the two descriptions you prefer. The whole study should take no longer than **10 minutes**.

Please do **not** participate on a mobile device, as the page may not display properly.

If you have any questions or concerns, please contact me at *lead.author@email.address*

Please, enter your **Prolific ID**:

After that, we displayed legal and IRB information for the participants to read. Then, once they clicked "Begin Experiment," we displayed the following set of instructions:

In this study, you will see one image at a time, each paired with two potential descriptions.

Your task is to **choose the description that you would prefer to receive if you couldn't see the image**.

The descriptions you'll see vary in length and how much information they contain. Please note that **some descriptions might be long but still contain less information than shorter ones** and take that into account in your decision.

Are you ready?

Finally, participants saw 30 images each with a pair of descriptions and chose the description they preferred (see a screenshot of this interface in Figure

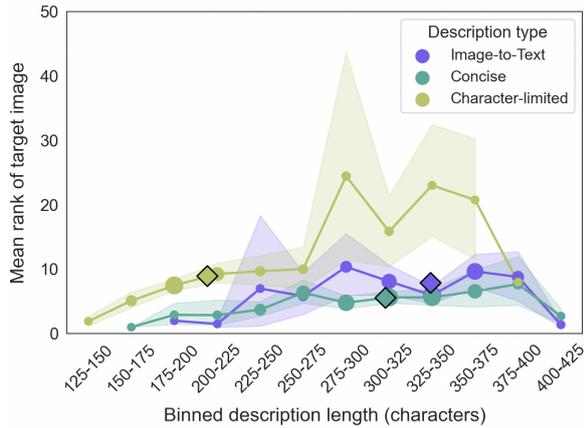


Figure 4: Mean rank as a function of description length (binned by character count) for each description type in the length-constrained case study described in Section 6.1. Diamonds represent means across length.

5). Results of this study (mean rank vs. binned description length for the 3 VLM-generated description types) can be found in Figure 4.



Source	Description
COCO	A den with a table, chairs and a tripod camera.
COMPOSITE	The den features a table and chairs, alongside a tripod camera. In the living room, there is a glass table, complemented by a black and white picture of a room that showcases three chairs and a coffee table. The living room itself includes a bay window, additional chairs, and a coffee table, creating a spacious area. The entire living area is very neatly organized, characterized by a black and white color scheme.
VERBOSE	The space referred to as a den is furnished with a table that serves as a central piece for various activities or gatherings, accompanied by several chairs that are arranged around it to provide seating. Additionally, this den is equipped with a tripod camera, which is a device typically used for stabilization while taking photographs or recording video, ensuring that it remains steady for better quality results.
IMAGE-TO-TEXT	The image shows a living room featuring natural light filtered through window blinds. There are three chairs arranged around a low, round coffee table at the center of the room. The table has several items on it, including a few books and candles. A small vase with plant stems is also present on the table. On one side, there is a bookshelf filled with books, and a lamp is positioned next to it. The floor is wooden, and there is a patterned rug beneath the coffee table. A tripod camera stands against the wall near the window. The walls are light-colored, and framed pictures are hung on one of the walls. The overall setting appears simple and organized.
CONCISE	The image depicts a living room with natural light filtering through window blinds. Three chairs are positioned around a low, round coffee table in the center of the room. The table holds several books, candles, and a small vase with plant stems. On one side, there is a bookshelf filled with books, accompanied by a lamp. The floor is wooden, and a patterned rug is placed beneath the coffee table. A tripod camera stands against the wall near the window. The walls are light-colored, with framed pictures hung on one of them. The setting appears simple and organized.
CHARACTER-LIMITED	The living room features natural light through blinds, three chairs around a round coffee table with books, candles, and a vase. A bookshelf with a lamp, wooden floor, patterned rug, tripod camera, and framed pictures are present.

Table 2: Second example set of descriptions for an image in our dataset. Composite and verbose descriptions are longer variants of the original description, but vary in the amount of additional information provided. IMAGE-TO-TEXT, CONCISE, and CHARACTER-LIMITED are generated by a VLM (here, GPT-4o-mini) under different instruction conditions.



Source	Description
COCO	A man wearing glasses while eating a hot dog.
COMPOSITE	A man wearing glasses and a hat is taking a bite from a hot dog while dressed in a gingham shirt. The scene features a close-up of him eating the hot dog, highlighting his actions.
VERBOSE	A gentleman who is donning a pair of glasses is in the process of consuming a hot dog, which is a popular food item typically consisting of a cooked sausage served in a sliced bun.
IMAGE-TO-TEXT	The image shows a man holding a hot dog with various toppings, including green and red items, appearing to prepare to take a bite. He is wearing glasses and a dark cap and is dressed in a checkered shirt with a collar. In the background, there are some outdoor elements, including greenery and a container or object that looks like a storage bin. The lighting suggests it is a casual outdoor setting.
CONCISE	The image depicts a man holding a hot dog adorned with various toppings, including green and red ingredients. He is wearing glasses, a dark cap, and a checkered shirt with a collar. In the background, there is greenery and a storage bin or container, indicating an outdoor environment. The lighting suggests a casual setting.
CHARACTER-LIMITED	A man wearing glasses and a dark cap holds a hot dog with green and red toppings. He is dressed in a checkered shirt. The background features greenery and a storage bin.

Table 3: Third example set of descriptions for an image in our dataset. Composite and verbose descriptions are longer variants of the original description, but vary in the amount of additional information provided. IMAGE-TO-TEXT, CONCISE, and CHARACTER-LIMITED are generated by a VLM (here, GPT-4o-mini) under different instruction conditions.

Which image description do you prefer?

Click on the description that you would prefer to receive if you couldn't see the image. Afterwards, proceed by clicking on the **Continue** button.



In a scene depicting the natural world, there are two zebras positioned closely together, standing side by side, on a field characterized by its dirt surface. The two zebras are neighboring each other, creating a visual representation of their cohabitation in this particular environment, which consists primarily of earth rather than grass or other forms of vegetation.

Two zebras stand next to each other on a dirt field, while a couple of other zebras explore in the sand next to some chopped wood. In a zoo setting, zebras wander around their pen, with two of them walking around as well. Among the group, baby zebras are also present, moving around to explore their environment.

CONTINUE

Figure 5: Screenshot of preference study description choice interface as seen by participants.