When More Words Say Less: Decoupling Length and Specificity 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 accessible via text-based descriptions. In current systems, however, description specificity is often conflated with their length. We argue that these two concepts must be disentangled: descriptions can be concise yet dense with information, or lengthy yet vacuous. We define specificity relative to a contrast set, where a description is more specific to the extent that it picks out the target image better than other possible images. We construct a dataset that controls for length while varying information content, and validate that people reliably prefer more specific descriptions regardless of length. We find that controlling for length alone cannot account for differences in specificity; it matters *how* the length budget is applied. These results support evaluation approaches that directly prioritize specificity over verbosity.

1 Introduction

Vision–language models (VLMs) are increasingly deployed to produce textual descriptions of visual content (Zhang et al., 2024; Wang et al., 2024; Deitke et al., 2025; Ghandi et al., 2023), with consequences for blind, low-vision, and sighted users alike (Morris et al., 2016; Gleason et al., 2019; Stangl et al., 2020). When generating descriptions, a central challenge is deciding how *specific* to be: which pieces of information should be included, and at what level of detail? Whatever determines the appropriate level of specificity for a given context (cf. Grice, 1975), missing that target causes problems: underspecific descriptions fail to support necessary distinctions, while overspecific ones reduce communicative efficiency (Goodman and Frank, 2016) and can trigger unintended inferences (e.g., Sedivy, 2003; Tourtouri et al., 2019).

Specificity is central to communicative effectiveness yet notoriously difficult to measure. A natural intuition, grounded in information theory, is that (all else being equal), longer descriptions pack more detailed information (Shannon, 1948). This intuition motivates a common practice of treating description length as a proxy for specificity, conditional on the description being accurate and well-formed. This practice appears across studies of human description preferences (Williams et al., 2022; Kreiss et al., 2022), dataset construction (Urbanek et al., 2024; Wang et al., 2025), accessibility guidelines (McCall and Chagnon, 2022), and evaluation methods (Kapur and Kreiss, 2024).

Yet the relationship between length and specificity is far from straightforward (Chen et al., 2022): descriptions can be lengthy yet vacuous, or concise yet dense with details. In fact, we often want to improve a system's specificity and ability to produce distinct outputs while controlling for excessive verbosity (Singhal et al., 2024; Dubois et al., 2024; Nayab et al., 2024; Hu et al., 2024). For principled evaluation, we must operationalize specificity as independent of length. Following classic possible world semantics (Carnap, 1947; Kripke, 1959; Montague et al., 1970), we suggest treating an utterance as more specific when it is compatible with fewer possible worlds. In visually grounded settings, a description is more specific when it truthfully describes fewer possible images (Young et al., 2014; Nie et al., 2020). This is an entailment relationship that holds across contexts: "small red chair" is strictly more specific than "red chair".

In this paper, we construct a dataset that manipulates length independently of information content, pairing images with descriptions that are lengthy yet vacuous (verbose) or concise yet information-dense (composite). We operationalize specificity via contrastive image compatibility: a description is more specific to the extent that it picks out the target image from a large set of alternatives. Using this framework, we show that human preferences track specificity, not length, and characterize how

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 410–422
July 2-7, 2026

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 App. C for additional examples.

different specificity constraints in the prompt (e.g., requesting conciseness versus imposing a character limit) affect the specificity of VLM-generated descriptions. Our central finding is that controlling for length alone cannot account for differences in specificity; it matters how the length budget is applied. These results support evaluation approaches that directly measure specificity rather than relying on length as a proxy.

2 Related Work

Defining specificity via contrast sets Recent referring expression generation (REG) models formalize specificity with respect to a contrast set of alternatives (Krahmer and Van Deemter, 2012). In the Rational Speech Act (RSA) framework (Goodman and Frank, 2016; Degen et al., 2020; Degen, 2023), speakers select utterances that maximize the likelihood of a listener identifying the intended referent from these alternatives while minimizing production costs. A key insight from this work is that not all words contribute equally to specificity: it is the inclusion of distinguishing features that differentiate the target from alternatives, not sheer quantity. 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 is absent from common image description datasets (Ilinykh et al., 2018, 2019; Pezzelle, 2023; Takmaz et al., 2022). The idea also has precedent in the denotation graph of Young et al. (2014), where caption entailment and similarity relationships are defined by the sets of images a pair of descriptions truthfully applies to. We adopt the contrast-set approach to operationalize specificity using image-text compatibility scores: a description's specificity is determined by how well it distinguishes the target image from an *implicit* set of alternatives.

Limitations of evaluation metrics Existing evaluation metrics for image captioning fail to disentangle length from specificity. Reference-based metrics like BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005) primarily assess similarity to human-written references but fail to capture human judgments of distinction (Kapur and Kreiss, 2024). Reference-less 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). None of these metrics capture specificity independent of length in communication-theoretic terms (Newman et al., 2020; Tang et al., 2024; Coppock et al., 2020). By constructing a dataset that manipulates length and information content independently, we provide a framework for evaluating specificity directly.

3 Approach

3.1 Dataset construction

To systematically investigate the relationship between length and specificity, we sampled 5,000 images uniformly across MS COCO's 80 categories (Lin et al., 2014). Our core theoretical contrast draws on possible-world semantics: descriptions expressing the same propositions should have equivalent specificity regardless of length, as they rule out the same possible images. Descriptions that incorporate distinct informational content should rule out more alternatives, yielding higher specificity. A metric that successfully disentangles length from specificity should detect this difference. For each image, we generated multiple description variants that deliberately vary in length and content (see Table 1):

***Original*:** A single human-written description for the image from MS COCO.

***Verbose*:** A longer rephrasing of the original that preserves the same semantic content, increasing length without adding new information.

***Composite*:** A longer description that combines content from all five COCO reference descriptions, incorporating additional distinct details.

***Image-to-Text*:** A VLM-generated description based on the image and minimal instructions.

The latter three description types were generated using OpenAI's `GPT-4o-mini` (OpenAI et al., 2024, prompts in App. A). While the verbose and composite conditions provide a theoretical frame for analysis, the image-to-text condition provides a practical baseline for how VLMs balance length and specificity in practice. We make our complete dataset, experiments, and analyses available.[1]

3.2 Measuring specificity

The central challenge of measuring specificity is that it is not defined on an absolute scale (Nie et al., 2020; Degen et al., 2020). Following the possible-worlds framework above, specificity must be defined relative to a contrast set: a description is more specific to the extent that it picks out the target from a set of implicit alternatives. For each image-description pair, we define the contrast set as the remaining 4,999 images. We then quantify specificity as a description's ability to discriminate the target image from these competitor images. The intuition, grounded in entailment relationships (see, e.g., Montague et al., 1970; Urquhart, 1973), is that more specific descriptions apply more selectively to a single image: e.g., all images showing "an albacore" show "a fish" but not vice versa.

To investigate this, we operationalize this idea using CLIPScore (Hessel et al., 2021). Its contrastive training objective makes it well-suited for measuring image-text compatibility in a discriminative setting (Ou et al., 2023; Takmaz et al., 2022). For each description, we compute its CLIPScore against the target image and all 4,999 alternatives. The rank of the target image is our specificity measure, where lower ranks indicate higher specificity (i.e., the description is less compatible with competitor images and more uniquely picks out the target). See App. B for technical details.

[1] `https://github.com/rkapur102/vision-language-specificity`

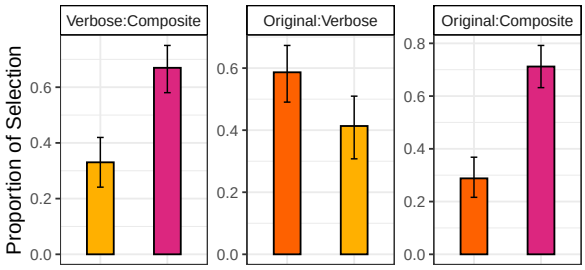

Figure 1: Pairwise human preferences by description type (95% bootstrapped CIs). Full results in App. D.

4 Results

4.1 Validating human specificity preferences

With conditions and the metric established, we validated that human preferences track specificity over length. We recruited 30 participants on Prolific. Participants saw an image paired with two descriptions and selected which they preferred. To isolate specificity, we sampled stimuli where verbose and composite descriptions were matched for length (see App. D). Using logistic regression with length as a predictor, we found participants preferred composite descriptions over the original ($\beta = 1.85$, $z = 2.54$, $p = .01$) and verbose descriptions ($\beta = -1.40$, $z = -4.6$, $p < .001$; see Fig. 1).

To rule out potential confounds, we included average word frequency (excluding stopwords) as an additional fixed effect in our statistical models. Because formality and intended purpose are difficult to operationalize directly, we also included Flesch-Kincaid readability scores as a proxy, following prior work linking readability scores to text formality (Graesser et al., 2014). Neither feature significantly predicts preferences ($p \in [0.25, 0.94]$ across models), whether incorporated individually or jointly, and all across-condition effects remain significant. These results confirm that the observed preferences are not explained by surface-level differences in word frequency or readability.

4.2 Validating the specificity metric

Having established that humans prefer more specific descriptions (not simply longer ones), we next test whether our specificity measure captures the same distinctions. Fig. 2 shows the cumulative distribution of target image ranks by description types. A steeper initial slope indicates that the descriptions more often receive low ranks (i.e., the target ranks highly), suggesting greater specificity. The metric recovers a clear specificity hierarchy consistent with human preferences: composite de-

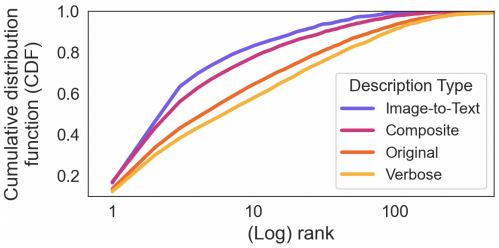

Figure 2: Cumulative distribution of ranks across description types relative to all other images.

scriptions yield significantly lower ranks than original ($\beta = -14.61$, $z = -10.33$, $p < 0.001$) and verbose ($\beta = -21.96$, $z = -13.72$, $p < 0.001$) descriptions. The difference between verbose and original ranks is also significant ($\beta = -7.35$, $z = -4.08$, $p < 0.001$), i.e., the verbose paraphrase appears to slightly decrease specificity rather than leaving it unchanged. Crucially, however, this does not affect the key comparison between length-matched composite and verbose conditions. We discuss a potential explanation related to CLIP's training data in the Limitations.

4.2.1 Testing hallucination robustness

A natural concern is that hallucinated details could inflate specificity scores if they happen to be discriminative. To test this, we selected 37 composite captions manually verified to contain no hallucinations and used GPT-4o-mini to introduce targeted hallucinations, prompting: "Change one detail in this description so that it becomes incompatible with the original. Only output the changed description and nothing else." We generated three hallucinated variants per caption and computed the CLIPScore of each against its matched image.

Introducing hallucinations significantly worsened the specificity ranks ($t(36) = 2.50$, $p = 0.017$) with mean rank increasing from 17.05 (original composite) to 30.91 (average across three hallucinated variants per image). The drop in image-text alignment (as assigned by CLIPScore) outweighs any discriminative value gained from the hallucinated detail. That said, this measure is not a substitute for hallucination detection: it will be most reliable when hallucinations are minimal. We discuss this interaction further in the Limitations.

4.2.2 Analyzing contrast set sensitivity

While we conducted our analysis on a large contrast set of 4,999 images to obtain high-resolution results, this may be too computationally expensive in practice. To investigate the sensitivity of our results

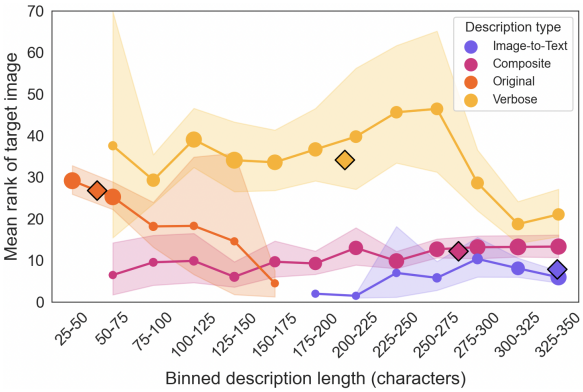

Figure 3: Mean rank vs. description length by type. Point size: bin sample size; ribbons: 95% CIs; diamonds: overall means. Range excludes outliers.

to the size of the contrast set, we recomputed specificity ranks for all captions across the four main conditions (Image-to-Text, Composite, Original, Verbose) using subsampled contrast sets of $N = 3, 10, 25, 50, 100, 250, 500, 1000, 2500, 4999[full]$. Each subsampling was repeated 100 times per caption, and Table 2 reports the mean rank across all captions per condition.

We find that all pairwise between-condition differences are statistically significant (all $p < 0.001$) at every contrast set size tested and the ordering of conditions is preserved throughout (Image-to-Text, then Composite, then Original, then Verbose). This suggests that when aggregated across a dataset, the relative specificity differences are robust to the size of the contrast set. Larger contrast sets simply provide greater resolution in effect size magnitude. For example, Verbose vs. Image-to-Text yields $\beta = 0.016$ at $N = 3$ and $\beta = 26.28$ at $N = 4,999$, but both are highly significant. For application in open-ended settings, this suggests that a contrast set could be sampled or constructed from a large corpus like LAION-5B (Schuhmann et al., 2022) in a manner that trades off contrast set size against computational resources, making these experiments more sustainable for subsequent research without sacrificing the robustness of relative specificity differences.

4.3 Distinguishing specificity from length

The preceding analyses are agnostic to the length of the descriptions. We now ask directly: does controlling for length eliminate the specificity differences between conditions? Fig. 3 shows the mean rank as a function of description length for each condition, indicating that the specificity hierarchy persists across length bins. In a regression

Contrast set sizes	3	10	25	50	100	250	500	1000	2500	4999
Image-to-Text	1.004	1.014	1.034	1.069	1.139	1.345	1.693	2.391	4.456	7.911
Composite	1.007	1.022	1.056	1.112	1.224	1.563	2.123	3.247	6.620	12.233
Original	1.016	1.051	1.129	1.259	1.516	2.293	3.583	6.169	13.917	26.842
Verbose	1.020	1.066	1.165	1.332	1.665	2.659	4.324	7.640	17.595	34.192

Table 2: Mean rank by description type across contrast set sizes N.

model controlling for length as a covariate, composite descriptions remain more specific than original ($\beta = -16.97$, $z = -4.60$, $p < 0.001$, $\Delta R^2 = 0.002$) and verbose ($\beta = -21.35$, $z = -12.24$, $p < 0.001$, $\Delta R^2 = 0.018$) variants.

Beyond these differences, the within-condition relationship between length and specificity is itself revealing. Only original (human-written) descriptions show the expected relationship where longer descriptions are more specific ($\beta = -0.22$, $z = -2.42$, $p < 0.05$), suggesting humans genuinely add information with length. The other conditions lack this trend (verbose, image-to-text) or even show a reversal (composite: $\beta = 0.02$, $z = 2.86$, $p < 0.01$). These patterns underscore that the "longer means more specific" heuristic cannot be assumed across data sources, particularly for synthetic or VLM-generated descriptions. Finally, without specificity or length constraints, GPT-4o-mini produces descriptions significantly longer ($\beta = 60.97$, $z = 26.56$, $p < 0.001$) and more specific ($\beta = -4.32$, $z = -4.15$, $p < 0.001$) than even composite descriptions. Composite descriptions are bound by what *annotators* chose to write; VLM-generated captions reflect model choice in description, which shapes this difference. Evaluating how length constraints shape that choice is a natural next step toward understanding this interesting facet of model behavior.

4.4 Evaluating VLM length constraints

Having established that our approach distinguishes specificity from length, we can now investigate questions that the length-as-a-proxy assumption precluded. In particular, when we prompt VLMs to constrain their output length, does specificity decrease proportionally, or does it depend on how the constraint is imposed? As a case study, we tested GPT-4o-mini under three length-constraint instructions:

***Concise*:** Be as concise as possible.

***Hard Constraint*:** Do not exceed 200 characters.

***k-Limited*:** Do not exceed k (i.e., the mean COCO caption length for that image).

The k-limited condition accounts for per-image variation in content, rather than imposing a uniform length across images. Full prompts are in App. A.

All conditions significantly reduced description length, but their impact on specificity varied (Fig. 4). Counterintuitively, the concise condition is not associated with decreased specificity; it actually increased it ($\beta = -1.97$, $z = -3.12$, $p < 0.01$), suggesting that prompting for conciseness encourages the model to prioritize discriminative information. Constraint strategy dictates specificity even at *matched* lengths. The hard 200-character-limited descriptions are significantly less specific than concise ones ($\beta = 10.19$, $z = 9.49$, $p < 0.001$, $\Delta R^2 = 0.013$). This indicates allocation matters: explicit length caps may lead to arbitrary truncation, while conciseness prompts allow the model to select what information to prioritize.

Notably, even the k-limited condition, despite being calibrated to image-specific COCO description lengths, does not replicate this human pattern; if anything, the VLM conditions show flat or reversed relationships between length and specificity. This echoes our earlier finding about the composite condition and underscores that matching length targets alone is insufficient to align VLM behavior with human patterns. Together, these results demonstrate the practical value of measuring specificity independent of length: they reveal that prompt design choices have downstream consequences for specificity that length metrics alone would miss.

5 Conclusion

As VLMs become increasingly critical for making visual content accessible through image descriptions, we show that description length is not a reliable proxy for specificity, even though the two are frequently conflated. Using a contrast-set approach, we demonstrate descriptions can be lengthy yet vacuous, or concise yet dense. These differences matter for both human preferences and automated evaluation. Our findings call for evaluation metrics that measure specificity directly rather than relying on length as a surrogate, and for prompt design strategies that directly optimize for appropriate levels of specificity and relevance to context.

Limitations

Specificity is only one dimension among many that may matter for description quality. A maximally discriminative description could simply list every visible object in exhaustive detail, which would be accurate and specific, but potentially unreadable and irrelevant. Our focus on specificity complements rather than replaces attention to fluency, coherence, and user needs.

Our approach has two key dependencies, each of which brings its respective limitations. First, we operationalize specificity using CLIPScore, whose contrastive training objective made it a promising candidate. However, CLIPScore has known biases and practical constraints. Prior work has shown CLIP exhibits concept association biases (Tang et al., 2023; Ahmadi and Agrawal, 2024) and struggles with spatial relationships (Kamath et al., 2023), which may affect which descriptive details register as discriminative under our metric. Additionally, the CLIP training data likely contained MSCOCO images and captions, thereby slightly inflating rank scores of the original caption-image pairs. The rewritten verbose descriptions may therefore have taken a minimal hit in specificity scores since they were never seen during training. While this can explain why descriptions from the verbose condition may receive slightly higher ranks than originals, the main effect of composite descriptions outranking verbose and originals cannot (see Subsec. 4.2). CLIPScore also has a 77-token input limit that required us to exclude longer descriptions. Our framework is not committed to CLIPScore specifically; any model providing image-text compatibility scores could be substituted, potentially offering different sensitivity profiles. Importantly, because we compare conditions (verbose vs. composite vs. original) using the same encoder, any encoder blindness is a constant across conditions (as per the standard "within-subjects" logic) and doesn't explain the real differences in specificity we observe between conditions (averaged over lots of data). We confirmed this by running a linear mixed-effects model predicting rank from description type with image as a random intercept, finding that all pairwise condition differences remain significant ($p < 0.001$, across statistical models). While the blindspots are a clear limitation (especially when trying to draw item-level instead of dataset-level inferences) these results suggest that they don't significantly alter our main findings.

Though research shows CLIPScore and its extensions have general utility in detecting hallucinations (Oh and Hwang, 2026; Petryk et al., 2024), specific hallucinations that happen to be discriminative are most likely to survive the filter. Our specificity analysis is therefore complementary to hallucination detection, not a substitute for it. In our experimental data, hallucinated content is minimal because conditions are derived from human-written ground-truth captions, and we confirm in Section 4.2.1 that artificially introduced hallucinations significantly worsen specificity scores. In practical applications, however, hallucinated content can meaningfully interfere with specificity estimates. It is precisely because specificity and faithfulness can interact in this way that independent tools for each are valuable: characterizing when hallucinations are specific versus generic requires an independent measure of specificity of the kind we propose.

Second, our operationalization of specificity is fundamentally relative to an implicit contrast set. There is no "view from nowhere." The COCO-based contrast set we constructed was well-suited for our purposes: with 5,000 images sampled uniformly across 80 categories, we observed differences between conditions at the aggregate level. More generally, the sensitivity of this approach will depend on the contrast set size (see subsubsection 4.2.2) and composition. If the set is too small and/or lacks coverage, ceiling effects emerge; if there are too many images of a particular type (e.g., a specific kind of bird), it becomes disproportionately sensitive to variation within that dimension relative to other dimensions.

We did observe ceiling effects in some cases, suggesting that a larger set may obtain more sensitive estimates. And our uniform sampling approach was intended to mitigate biases in feature overrepresentation, though we cannot rule it out entirely. For example, if green walls happen to be rarer in COCO than dual-toilet bathrooms, "a bathroom with a green wall" would counterintuitively rank as more specific than "a bathroom with two toilet seats," even though the latter would statistically rule out higher proportion of real-world bathrooms. Importantly, however, such biases should not be correlated with description type, and should wash out in the aggregate analyses. This concern underscores that our approach was tailored for dataset-level analysis rather than individual caption scoring. Generalizing to absolute judgments about individ-

ual descriptions would require more sophisticated approaches, such as synthesizing contrastive images along specific semantic dimensions.

Finally, our LLM-generated conditions (verbose, composite, and the VLM ablations) may introduce stylistic artifacts beyond the intended manipulation, for example, formal hedging language (e.g. "in the current situation, it appears...") and potential differences in word frequency or register. Our key comparison between verbose and composite at matched lengths partially controls for this, as both are subject to similar LLM stylistic tendencies, as well as our inclusion of some of these confounds as fixed effects in our statistical modeling (see Subsec. 4.1). However, individual items likely vary in how cleanly they instantiate the intended manipulation.

The verbose condition is a controlled manipulation rather than a naturalistic sample of VLM verbosity. This is a deliberate feature of our design: isolating the length-specificity confound requires a condition that adds length without adding information, which we can then compare against the composite condition and, subsequently, against naturalistic VLM outputs (Subsec. 4.4). More broadly, our contrast-based ranking targets specificity and is agnostic about relevance. Consider two descriptions: one stating "in the background, there is a house with a red roof" and another stating "in the foreground, there is a person with a hat." Under our framework, either could be more specific depending on how many images in the contrast set match each description. A background detail, while potentially less relevant to a user's needs, may be highly discriminative — and would score higher than non-specific padding that merely rephrases existing content. This is by design: we aim to capture whether added content narrows the set of compatible images, regardless of whether that content is relevant to a particular user's goals. Modeling relevance is beyond the scope of this paper.

Taken together, there are many challenging design choices when designing an approach to disentangle length and specificity. Despite that, we show that when making informed decisions about these constraints, such analyses provide insights that *length as a proxy* can never deliver.

Acknowledgments

We thank Google's GiG program for supporting this research. We are also grateful to Stanford's Social Interaction Lab, and the Models and Linguistic Theory Reading Group for feedback on earlier versions of this work.

References

Saba Ahmadi and Aishwarya Agrawal. 2024. An examination of the robustness of reference-free image captioning evaluation metrics. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 196–208.

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.

Rudolf Carnap. 1947. *Meaning and necessity: A study in semantics and modal logic*. University of Chicago Press.

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, Nathanial 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.

Judith Degen. 2023. The Rational Speech Act Framework. *Annual Review of Linguistics*, 9:519–540.

Judith Degen, Robert D. Hawkins, Caroline Graf, Elisa Kreiss, and Noah D. Goodman. 2020. When Redundancy Is Useful: A Bayesian Approach to "Over-informative" 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, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, and 31 others. 2025. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 91–104.

Yann Dubois, Percy Liang, and Tatsunori Hashimoto. 2024. Length-Controlled AlpacaEval: A Simple Debiasing of Automatic Evaluators. In *First Conference on Language Modeling*.

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.

Noah D Goodman and Michael C Frank. 2016. Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences*, 20(11):818–829.

Arthur C Graesser, Danielle S McNamara, Zhiqiang Cai, Mark Conley, Haiying Li, and James Pennebaker. 2014. Coh-metrix measures text characteristics at multiple levels of language and discourse. *The Elementary School Journal*, 115(2):210–229.

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.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. 2023. "What's "up" with vision-language models? Investigating their struggle with spatial reasoning". In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9161–9175, Singapore. 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.

Saul A Kripke. 1959. A completeness theorem in modal logic1. *The journal of symbolic logic*, 24(1):1–14.

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, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Chaitanya Malaviya, Sudeep Bhatia, and Mark Yatskar. 2022. Cascading biases: Investigating the effect of heuristic annotation strategies on data and models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6525–6540.

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.

Richard Montague and 1 others. 1970. Universal grammar. *1974*, pages 222–46.

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.

H. Oh and W. Hwang. 2026. Do vision encoders truly explain object hallucination?: Mitigating object hallucination via simple fine-grained clipscore. *arXiv preprint arXiv:2502.20034*.

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 Krojer, 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.

Suzanne Petryk, David M. Chan, Anish Kachinthaya, Haodi Zou, John Canny, Joseph E. Gonzalez, and Trevor Darrell. 2024. ALOHa: A new measure for hallucination in captioning models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 342–357. Association for Computational Linguistics.

Sandro Pezzelle. 2023. Dealing with semantic underspecification 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.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Preprint*, arXiv:2210.08402.

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.

Wei Shen, Rui Zheng, Wenyu Zhan, Jun Zhao, Shihan Dou, Tao Gui, Qi Zhang, and Xuan-Jing Huang. 2023. Loose lips sink ships: Mitigating length bias in reinforcement learning from human feedback. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2859–2873.

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.

Yingtian Tang, Yutaro Yamada, Yoyo Zhang, and Ilker Yildirim. 2023. When are lemons purple? The concept association bias of vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14333–14348.

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.

Elli N Tourtouri, Francesca Delogu, Les Sikos, and Matthew W Crocker. 2019. Rational over-specification in visually-situated comprehension and

production. *Journal of Cultural Cognitive Science*, 3(2):175–202.

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.

Alasdair Ian Fenton Urquhart. 1973. The semantics of entailment.

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*.

Xinran Wang, Muxi Diao, Yuanzhi Liu, Chunyu Wang, Kongming Liang, Zhanyu Ma, and Jun Guo. 2025. Harnessing caption detailness for data-efficient text-to-image generation. *arXiv preprint arXiv:2505.15172*.

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.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Jingyi Zhang, Jiaxing 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 specificity 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.

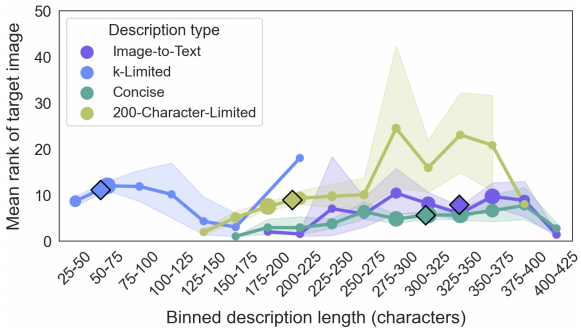

Figure 4: Mean rank vs. description length by each type in the length-constrained case study (Subsec. 4.4). Point size: bin sample size; ribbons: 95% CIs; diamonds: overall means. Range excludes outliers.

A.2 Composite

To increase the expected level of specificity 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-mini 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 specificity changes when a VLM is instructed to constrain its length *at its own discretion*, we simply passed the target image to GPT-4o-mini 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.

IMAGE		
ORIGINAL	A den with a table, chairs and a tripod camera.	A man wearing glasses while eating a hot dog.
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.	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	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.	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 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.	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 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.	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	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.	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: Example set of descriptions for images 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.

A.5 Character-limited

Finally, to understand how specificity changes when a VLM is instructed to constrain its length *to a hard cutoff*, we simply passed the target image to GPT-4o-mini with the Image-to-Text prompt from above and the additional instruction "don't exceed 200 characters" (example shown) or "don't exceed k characters", where k was passed in as the average COCO caption length for an image:

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

B Computational Details

Computing specificity requires calculating CLIPScores between each description and all images in the contrast set. For our primary dataset (4 description types × 5,000 images × 5,000 contrast images), this amounts to 100 million pairwise scores. Using a single NVIDIA RTX 6000 Ada Generation GPU, this computation completed in approximately 5 hours. The additional length-constrained conditions (Subsec. 4.4) required an additional 2.5 hours. While this particular size is feasible for research-scale evaluation, we have shown that our specificity metric is robust to varying contrast set sizes (Section 4.2.2), which greatly improves scalability for

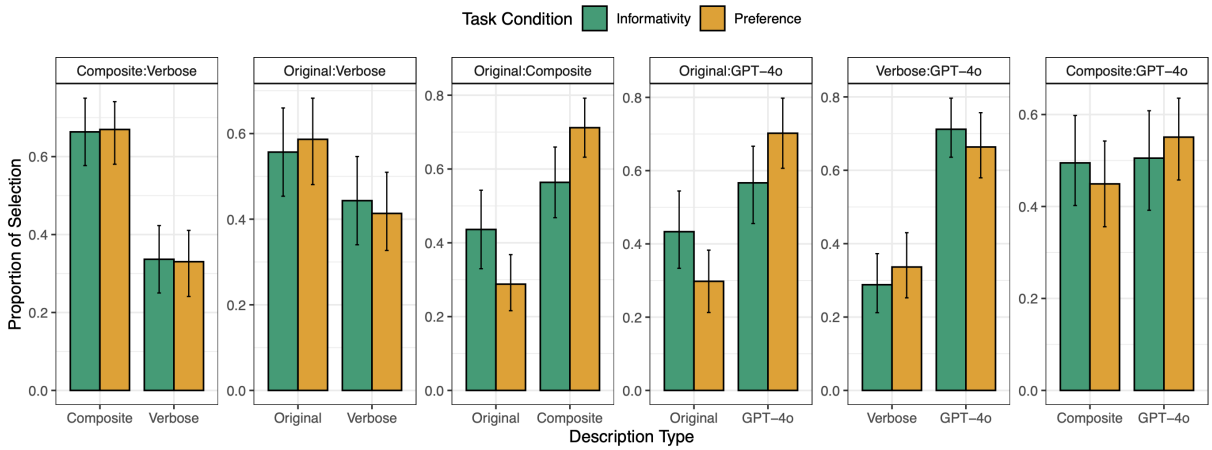

Figure 5: Experimental results of the preference study (in Subsec. 4.1 and Subsec. D.1) and the specificity study (in Subsec. D.2).

larger deployments. There are several further optimizations to consider: parallelization across multiple GPUs, caching of image embeddings (which remain constant across descriptions), or sampling representative subsets from the contrast set rather than exhaustive comparison. We also note that our approach is not committed to CLIPScore specifically; any model providing image-text compatibility scores could be substituted, potentially offering different tradeoffs between computational cost and sensitivity.

C 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 Subsec. 4.4. These can be found in Table 3.

D Human experiment details

We conducted two human subject experiments to investigate people's understanding and preferences of the original and edited (i.e., composite and verbose) image descriptions from our dataset. Subsec. D.1 introduces the main study (reported in Subsec. 4.1) where we elicited people's preferences, and Subsec. D.2 supplements these findings with data on people's specificity judgments.

D.1 Eliciting people's description preferences

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.

We included the paragraph that highlighted the distinction between length and specificity due to the well-attested phenomenon that human raters in annotation studies often themselves use length as a heuristic for other measures in order to minimize cognitive load (Shen et al., 2023; Malaviya et al., 2022).

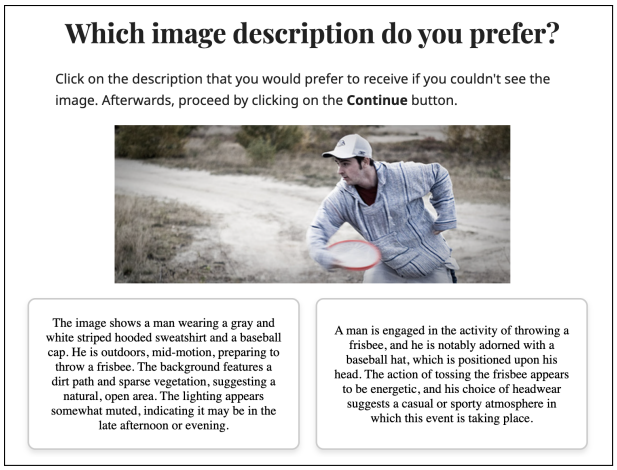

Figure 6: Screenshot of preference study description choice interface as seen by participants.

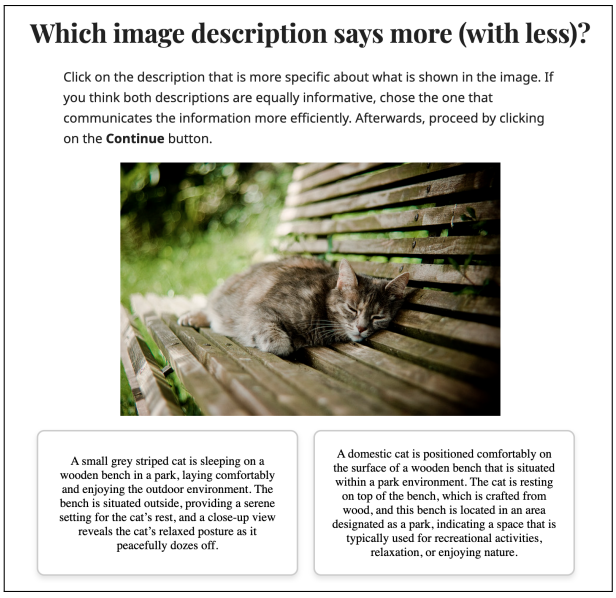

Figure 7: Screenshot of specificity study description choice interface as seen by participants.

Participants saw 30 images each with a pair of descriptions and chose the description they preferred (see a screenshot of this interface in Fig. 6). Further results for the 3 VLM-generated description types can be found in Fig. 4. With 30 participants each completing 30 trials, our design was powered for aggregate validation rather than detailed individual-differences analysis; characterizing sources of individual variation in description preferences remains a direction for future work.

D.2 Eliciting people's specificity understanding

We conducted a second study that more directly tested whether the descriptions differed in their perceived specificity. As shown in Fig. 7, the design is identical to the preference study, only the objective for participants changed. While participants in the first study were asked to select the description they *preferred*, participants in this second study were asked to select the description that *contains more information*. The only change to the main instructions was done to the second paragraph, which then read:

> [...] Your task is to **choose the description that is more specific about the image content**. [...]

D.3 Results

Fig. 5 presents the proportion of description selections from each study. All analyses were conducted using the `glm` function in `R` (`chosen` $\sim$ `condition + length`). Similar to the preference study, we find that participants consider the composite descriptions more informative than the original descriptions ($\beta = 1.63, z(185) = 2.4, p = 0.02$) and their verbose counterparts ($\beta = -1.30, z(205) = -4.1, p < 0.001$). These results clearly show that our data successfully manipulates perceived specificity.

Participant choices across studies are significantly correlated ($r = 0.21$, $p < 0.001$), suggesting that participants across conditions preferred descriptions that are more specific. This is confirmed when we use the empirically elicited average specificity rate as a predictor for each item. The average specificity rate is a significant predictor for preference data in the verbose-composite ($\beta_{spec} = 1.73, z(282) = 3.95, p < 0.001$) and original-composite settings ($\beta_{spec} = 1.10, z(232) = 2.34, p = 0.02$), and marginally significant in the original-verbose setting ($\beta_{spec} = 0.86, z(204) = 1.90, p = 0.057$). These results show a strong association between the descriptions people prefer and how specific they perceive them to be, and that this goes beyond what can be explained through length.