Too Long or Too Fake? Disentangling the Causes of Hallucination in Vision–Language Models

Anonymous EMNLP submission

Abstract

Training vision–language models (VLMs) on long, synthetic captions has been shown to increase hallucination compared to using short, human-written ones. Prior work attributes this to errors in synthetic data, but confounds caption origin (human vs. synthetic) and caption length. We disentangle these factors through controlled experiments on three matched set of captions: short human-written, long human-written, and long synthetic ones. VLMs trained on these datasets are evaluated using recent advanced metrics, with a breakdown by objects, attributes, and relations. We find that caption length is the main driver of hallucination, though synthetic origin also contributes, particularly through object and attribute errors.

1 Introduction

Synthetic data generation has emerged as a powerful paradigm for distilling knowledge from large language models (LLMs) into smaller ones (Peng et al., 2023; Hsieh et al., 2023). In the space of vision-language modeling, where short image captioning is giving way to the more complex task of long image captioning (Onoe et al., 2024; Garg et al., 2024), generative caption enrichment (GCE) is proving very effective for creating relevant training data (Chen et al., 2023; Singla et al., 2024).

Yet, recent work has shown that while also improving descriptiveness and object recall, GCE can lead also to increased hallucination. In particular, Hirota et al. (2024) found that vision–language models (VLMs) trained on long, synthetically enriched captions, such as those from the ShareGPT-4V dataset (Chen et al., 2023), exhibit higher hallucination rates in terms of the CHAIR metric (Rohrbach et al., 2019), compared to models trained on concise human-authored captions from the COCO datatset (Lin et al., 2015). They attributed this phenomenon to likely error propagation from the teacher model used to generate the data to the student model, trained on it.

The experimental design in prior work overlooks a key confound: caption length. As the length of generated text increases, there is naturally more room for errors. And as the level of detail described becomes more fine-grained, the vision-language alignment abilities of the model are strained. It is thus unclear to what extent hallucination in models trained with GCE is driven by the length of captions or by their synthetic origin.

To resolve this confound, we conduct experiments on three matched datasets, all using the same 118k images annotated with different captions: short human-written (COCO), long human-written (Pont-Tuset et al., 2020, Localized Narratives), and long synthetic captions (ShareGPT-4V.) We train three vision–language models and study their hallucination behavior using two advanced recent metrics: CAPTURE (Dong et al., 2024) and HalFscore (Chen et al., 2025).

By analyzing object, attribute, and relation-level correctness and coverage, we show how caption properties shape model behavior. We find that the shift from human-authored to synthetic data has a smaller impact on hallucination rates than the shift from short to long human-written captions, although both contribute considerably. While the correctness of relations is not affected by the origin of the data, hallucination in objects and attributes rise.

The field of image captioning has shifted from short, generic descriptions to long-form captions with richer visual grounding. Early datasets like COCO (118K) (Lin et al., 2015) offered concise, crowd-sourced captions but lacked detail. Localized Narratives (849K) (Pont-Tuset et al., 2020) addressed this by aligning spoken descriptions with mouse traces over images from COCO, Flickr30k, ADE20K, and Open Images (Lin et al., 2015; Plummer et al., 2016; Zhou et al., 2017; Kuznetsova et al., 2020). While effective, this annotation method is costly to scale. To enable larger datasets, synthetic captioning has emerged. ShareGPT-

4V (Chen et al., 2023) was created via a two-stage process: GPT-4V generated captions for 100K images, which were then used to train a ShareCaptioner for scalable generation (1.2M). Though scalable, synthetic captions risk hallucinations due to LLM errors.

Hallucination in VLMs. Hallucination refers to models mentioning visual details not present in the image (Rohrbach et al., 2019; Bai et al., 2025; Liu et al., 2024). Benchmarks like CHAIR (Rohrbach et al., 2019) and POPE (Li et al., 2023a) categorize hallucinations at the object, attribute, and relation levels, linking them to visual-text misalignment or language priors.

Hirota et al. (2024) fine-tune BLIP-2 on COCO and ShareGPT-4V and evaluate hallucination using CHAIR on the COCO validation set. They find that models trained on synthetic captions hallucinate more. However, CHAIR only measures object-level hallucination, and comparing COCO (which omits many present objects) to ShareGPT-4V (which is longer and more detailed) introduces two variables, caption origin and length, making it unclear whether hallucination increases due to synthetic data or verbosity. Since COCO captions average 10 words and ShareGPT-4V captions 143, this mismatch increases the likelihood of false positives from correctly grounded but unannotated objects.

In contrast, we isolate the effects of length and origin, and evaluate objects, attributes, and relations using LLM- and graph-based metrics.

2 Methodology and Experimental Design

Our methodology was designed to systematically evaluate the impact of caption length and origin on hallucination in vision-language models. We established a controlled experimental framework with three key components: (1) a dataset curation with multiple caption types for the same images, (2) a vision-language model architecture designed to minimize pre-existing biases, and (3) a comprehensive evaluation pipeline that measures both coverage and hallucination.

2.1 Datasets

The training data is based on the 118,000 images shared between COCO, Localized Narratives and ShareGPT-4V. For every image we thus have a short, human-written caption, a long, human-written caption, and long, model-generated one. We fine-tune VLMs separately on each set of captions, with images being matched across experiments, to isolate the effects of caption types.

The captions between the three datasets vary substantially in length: COCO captions averaged 10 words, Localized Narratives around 42 words, and ShareGPT-4V captions, approximately 143 words per image. While the latter two were intentionally chosen to be longer than COCO, between themselves, they differ both in origin and also in length. We thus truncated the ShareGPT-4V captions to include only the first three sentences. This reduced their average length to 52 words, bringing them closer to Localized Narratives, while preserving their grammaticality and coherence. The final length distribution of the training data is visualized in Figure 7 in the Appendix and an example of a training sample can be seen in Figure 8 (Appendix).

2.2 Model Architecture and Training

A SigLIP vision encoder (Zhai et al., 2023) is paired with a Qwen-2.5 (3B) language decoder (Qwen et al., 2025), using a 2-layer MLP projector. The model is trained in two stages to support the gradual learning of fine-grained cross-modal alignment. In the first stage, we train only the visual projection layer on COCO, to establish initial visual-textual alignment.[1] In the second stage, we independently train three models on each dataset. Here, all model parameters are updated. Each training run uses the same 118,000 images, differing only in the caption source, to ensure controlled comparisons. The details on optimisation and computational resources are in in Appendix F.

2.3 Evaluation Pipeline

Evaluation is based on the Visual Genome (VG) dataset (Krishna et al., 2016), specifically the subset of 2,186 images from it, which do not overlap with the COCO training set. VG provides structured annotations of objects, attributes, and relations, as well as region-level descriptions. These annotations serve as ground truth for computing correctness and coverage in the evaluations below.

Following prior work (Hirota et al., 2024; Dong et al., 2024), we adopt the terms correctness and coverage, which extend the hard surface-level matching precision and recall scores to semantic matching, more appropriate in image captioning evaluation: **Correctness** ($\mathcal{C}$) measures the proportion of concepts in the generated caption that are

[1] We found that without this step, the long-caption training was not converging.

Model	HalFscore		CAPTURE								Length	
	Average		Objects		Attributes		Relations		Average		Train	Gen.
	$\mathcal{C}$	$\mathcal{V}$	$\mathcal{C}$	$\mathcal{V}$	$\mathcal{C}$	$\mathcal{V}$	$\mathcal{C}$	$\mathcal{V}$	$\mathcal{C}$	$\mathcal{V}$		
COCO	**61.04**	38.60	85.6	26.0	76.0	19.0	64.5	33.4	**75.4**	26.1	10	11
Loc. Narratives	$58.81_{2.23}$	$42.39_{3.79}$	$81.3_{4.3}$	$39.7_{13.7}$	$75.9_{0.1}$	$19.9_{0.9}$	$59.2_{5.3}$	$38.6_{5.2}$	$72.1_{3.3}$	$32.7_{6.6}$	42	33
ShareGPT-4V	$56.83_{1.98}$	**43.19**$_{0.80}$	$78.7_{2.6}$	$34.3_{5.4}$	$73.8_{2.1}$	$39.1_{19.2}$	$59.4_{0.2}$	$36.1_{2.5}$	$70.6_{1.5}$	**36.5**$_{3.8}$	56	34

Table 1: HalFscore on full test set and CAPTURE on subset of 307 images. $\mathcal{C}$: correctness, $\mathcal{V}$: coverage. Right: CAPTURE scores broken down by objects, attributes, relations, and their average. The subscripts indicate the change from the number above, with red indicating a drop and green indicating an increase. "Train" and "Gen." list the mean lengths of the training and generated captions, respectively.

Model	Constraint	Objects		Attributes		Relations		Average		Length	
		$\mathcal{C}$	$\mathcal{V}$	$\mathcal{C}$	$\mathcal{V}$	$\mathcal{C}$	$\mathcal{V}$	$\mathcal{C}$	$\mathcal{V}$	Train	Gen.
Localized Narratives	Constrained	81.0	38.3	76.2	20.4	59.8	38.1	**72.3**	32.3	42	33
	Unconstrained	79.0	41.2	75.2	22.2	57.6	39.4	70.6	**34.3**	42	37
ShareGPT-4V	Constrained	78.3	34.8	73.5	39.2	58.8	36.4	**70.2**	36.8	56	34
	Unconstrained	76.2	40.2	71.3	44.3	57.6	40.4	68.4	**41.6**	56	49

Table 2: CAPTURE evaluation on the 487 intersected captions that contain attributes. Bolded values mark the higher average within each model pair.

present in the image; higher values indicate fewer hallucinations. **Coverage** ($\mathcal{V}$) measures the proportion of ground truth concepts from the image that are successfully captured in the caption.

HalFscore Chen et al. (2025) relies on GPT-4o to extract triplets of concepts (objects, attributes, relations) from both generated and ground-truth (GT) captions, and uses binary LLM decisions to match these when computing correctness and coverage. In our evaluations, GT captions are built by concatenating all region-level descriptions for an image from our VG test set. The resulting score is highly reliable, as we manually observe that the extracted objects, attributes and relations as well as their matching, is largely correct. The final score coarsely indicates the overall correctness and coverage, with no per-type breakdown.

CAPTURE Dong et al. (2024) evaluates grounding by aligning generated captions with Visual Genome scene graphs to compute correctness and coverage for objects, attributes, and relations. In its original setup, both generated and reference captions are parsed into scene graphs using the FACTUAL parser (Li et al., 2023b). In our version, we use ground-truth triplets from Visual Genome for the reference and parse only the generated captions, enabling per-type evaluation via soft matching (exact match, synonym expansion, and SBERT similarity). However, reliance on FACTUAL to extract triplets from generated captions can miss elements, limiting reliability. Implementation details for both metrics are in Appendix C.

3 Results and Discussion

Effect of Caption Length To investigate our first research question, we compare captions generated by models fine-tuned on short human captions from the COCO dataset versus longer ones from Localized Narratives (LN).

We begin with a coarse-grained analysis using HalFscore on overall concept correctness and coverage. As shown in Table 1, models trained on shorter COCO captions achieve the highest correctness scores, meaning they include fewer hallucinated concepts. However, their coverage is lower, as many relevant concepts are omitted. This reflects a correctness–coverage trade-off introduced by verbosity: longer captions improve coverage but are more prone to hallucination (Appendix A).

For fine-grained analysis, we use CAPTURE. In our initial evaluation, most captions from COCO- and LN-finetuned models lacked attributes and relations, focusing almost exclusively on objects. As a result, average correctness scores were skewed by zero-valued entries (Appendix B). To better assess model behavior when these concept types are present, we restrict CAPTURE evaluation to a filtered subset of captions mentioning at least one attribute and one relation. In this subset (Table 1),

COCO: A man in a shirt holding a Wii remote..
Localized Narratives: I can see a man is standing and holding a remote in his hand. I can see he is wearing a shirt and a tie. I can see colour of his shirt is blue .
ShareGPT-4V: A man is standing in a room with a rustic charm. He's wearing a blue shirt that has a green star on it, adding a pop of color to his attire.

COCO: A truck on a snowy street next to a house..
Localized Narratives: In the image we can see a car which is of red color and a person wearing white color dress is standing near the car and at the background of the image there are some houses, trees, snow and a clear sky.
ShareGPT-4V: a red pickup truck is parked on the side of a street. The truck, facing away from us, is equipped with a white utility sled attached to its back.

Figure 1: Captions generated by different models for the same image; hallucinations are marked in red.

the same trend holds: COCO-trained models yield higher correctness, while LN-trained ones achieve higher coverage. This supports the hypothesis that increased caption length introduces hallucination across objects, attributes, and relations. See Appendix G for a comparison of how CAPTURE and HalFscore handle hallucinated concepts.

Effect of Caption Origin To investigate our second research question, we compare captions generated by models fine-tuned on human-written versus synthetic data. Both models, one trained on Localized Narratives (LN) and the other on ShareGPT-4V (SG-4V), were constrained to produce captions of similar average lengths (33 vs. 34 tokens).

Coarse-grained results from HalFscore show that, despite similar lengths, LN-trained models achieve higher correctness, while SG-4V-trained models yield higher coverage, suggesting synthetic captions are more detailed but also more prone to hallucination.

Fine-grained CAPTURE results (Table 1) reinforce these findings. The LN-finetuned model attains higher correctness for objects and attributes (fewer hallucinations), whereas relation correctness is comparable with the SG-4V trained model. These results suggest that while the model trained on synthetic captions tends to have slightly lower correctness on average, its performance in relation correctness remains competitive with human-trained models.

We also find that SG-4V trained model achieves higher average coverage across all concept types due to a significant increase in attribute coverage, but LN trained model performs better on object and relation coverage. This reflects the nature of each dataset (Figure 8): LN tends to list most scene objects and their spatial relations, while SG-4V emphasizes stylistic detail for prominent entities. These trends are reflected in the qualitative example in Figure 1. The LN caption captures more of the scene than COCO and maintains correctness overall, but hallucinates a *box*. SG-4V adds further details, including stylistic and attribute-rich descriptions, but hallucinates objects like *fork* and *knife* (marked in red). Appendix H provides additional examples, and Appendix I offers further analysis of how scores vary with object count.

Overall, we find that while Hirota et al. (2024) attribute increased hallucination to synthetic data, reporting a rise when switching from human written COCO to LLM-generated SG-4V captions, our controlled setup shows caption length also plays a significant role. In both metrics, moving from short COCO captions to longer ones (LN) increased hallucination by 3.3 points, while switching from LN to SG-4V (at similar length) caused a smaller 1.5-point increase. This suggests that length, not just origin, is a key driver of hallucination.

Analysis of Constrained and Unconstrained Generation A complementary analysis using CAPTURE in Table 2 compares the effect of length control within the same dataset, i.e., constrained and unconstrained generations from models fine-tuned on both LN and SG-4V. In both cases, we observe a consistent trend across all concept types, objects, attributes, and relations. Restricting the caption length leads to a drop in coverage but a noticeable improvement in correctness. This further reinforces the role of verbosity as a key factor in hallucination.

4 Conclusion and Future Work

Our results show that both caption length and origin influence hallucination and coverage in vision–language models. Longer captions consistently improve coverage but also increase hallucination, while synthetic captions tend to describe more attributes and relations but with slightly lower correctness than human-written ones. These trends hold across both fine- and coarse-grained evaluations. Future work should explore training and decoding strategies that balance correctness with coverage, and develop evaluation methods that better account for semantic nuance across concept types.

5 Limitations

This study focuses on caption-level grounding and relies on automated tools for evaluation, which introduces two key limitations. First, our use of FACTUAL and GPT-4o for triplet extraction depends on the accuracy of these parsers. While they support large-scale evaluation, they may occasionally overlook or misclassify fine-grained concepts, particularly in visually complex scenes. Second, our analysis is limited to image captioning and does not extend to downstream tasks such as visual question answering or retrieval. Exploring how hallucination affects these applications remains an important avenue for future research.

References

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2025. Hallucination of multimodal large language models: A survey. *Preprint*, arXiv:2404.18930.

Cong Chen, Mingyu Liu, Chenchen Jing, Yizhou Zhou, Fengyun Rao, Hao Chen, Bo Zhang, and Chunhua Shen. 2025. Perturbollava: Reducing multimodal hallucinations with perturbative visual training. *arXiv preprint arXiv:2503.06486*.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023. Sharegpt4v: Improving large multimodal models with better captions. *Preprint*, arXiv:2311.12793.

Hongyuan Dong, Jiawen Li, Bohong Wu, Jiacong Wang, Yuan Zhang, and Haoyuan Guo. 2024. Benchmarking and improving detail image caption. *Preprint*, arXiv:2405.19092.

Roopal Garg, Andrea Burns, Burcu Karagol Ayan, Yonatan Bitton, Ceslee Montgomery, Yasumasa Onoe, Andrew Bunner, Ranjay Krishna, Jason Baldridge, and Radu Soricut. 2024. Imagein-words: Unlocking hyper-detailed image descriptions. *Preprint*, arXiv:2405.02793.

Yusuke Hirota, Ryo Hachiuma, Chao-Han Huck Yang, and Yuta Nakashima. 2024. From descriptive richness to bias: Unveiling the dark side of generative image caption enrichment. *Preprint*, arXiv:2406.13912.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8003–8017, Toronto, Canada. Association for Computational Linguistics.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Preprint*, arXiv:1602.07332.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. 2020. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023a. Evaluating object hallucination in large vision-language models. *Preprint*, arXiv:2305.10355.

Zhuang Li, Yuyang Chai, Terry Yue Zhuo, Lizhen Qu, Gholamreza Haffari, Fei Li, Donghong Ji, and Quan Hung Tran. 2023b. Factual: A benchmark for faithful and consistent textual scene graph parsing. *Preprint*, arXiv:2305.17497.

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.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024. A survey on hallucination in large vision-language models. *Preprint*, arXiv:2402.00253.

Yasumasa Onoe, Sunayana Rane, Zachary Berger, Yonatan Bitton, Jaemin Cho, Roopal Garg, Alexander Ku, Zarana Parekh, Jordi Pont-Tuset, Garrett Tanzer, Su Wang, and Jason Baldridge. 2024. Docci: Descriptions of connected and contrasting images. *Preprint*, arXiv:2404.19753.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *Preprint*, arXiv:2304.03277.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2016. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *Preprint*, arXiv:1505.04870.

Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. 2020. Connecting vision and language with localized narratives. *Preprint*, arXiv:1912.03098.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang,

Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2019. Object hallucination in image captioning. *Preprint*, arXiv:1809.02156.

Vasu Singla, Kaiyu Yue, Sukriti Paul, Reza Shirkavand, Mayuka Jayawardhana, Alireza Ganjdanesh, Heng Huang, Abhinav Bhatele, Gowthami Somepalli, and Tom Goldstein. 2024. From pixels to prose: A large dataset of dense image captions. *Preprint*, arXiv:2406.10328.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. *Preprint*, arXiv:2303.15343.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2017. Scene parsing through ade20k dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5122–5130.

A Visualization of Correctness-Coverage Tradeoff

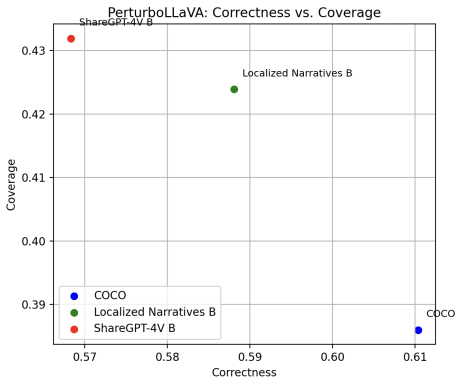

Figure 2: Correctness vs. Coverage plot for overall Concepts using HalFscore

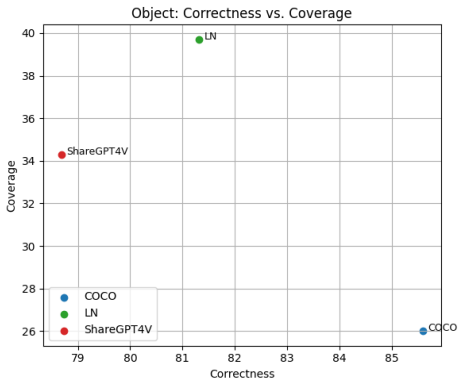

Figure 3: Correctness vs. Coverage for object grounding using Capture

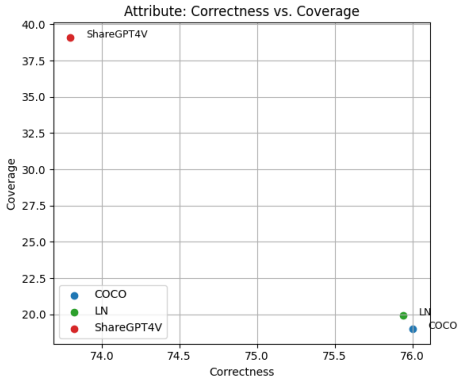

Figure 4: Correctness vs. Coverage for attribute grounding using Capture

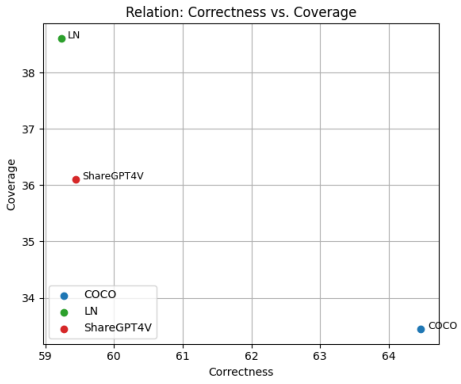

Figure 5: Correctness vs. Coverage for relation grounding using Capture

B Attribute and Relation Metrics (Non-Zero Subset)

We report detailed grounding performance for attributes and relations on the full set of 1,872 Visual Genome images, which we get from the pipeline in Appendix C, used during evaluation. However,

many captions, particularly those from models fine-tuned on COCO and Localized Narratives, omit attributes or relations entirely, focusing primarily on object mentions. As a result, many entries receive zero scores for these concepts, inflating variance and reducing the reliability of comparisons.

This issue is evident in the wider confidence intervals for attribute correctness (e.g., ± 2.02 for LN-C), especially compared to more stable object-level scores. To mitigate this, we report results on a filtered subset of 307 captions containing at least one attribute and one relation (Section 3). This controlled evaluation enables a fairer and more meaningful comparison across models.

Notably, the SG-4V-finetuned model shows substantially higher attribute coverage, with attribute mentions in nearly all captions (1857/1872), while the LN-finetuned model includes them in only 922/1872. However, when attributes are present, the LN model achieves higher correctness, as shown in Table 1, suggesting a trade-off between correctness and coverage. This aligns with broader stylistic differences across datasets: SG-4V captions tend to be more verbose and descriptive, while LN captions are more selective and concise.

Table 3 presents full correctness and coverage scores with 95% confidence intervals across all concept types, computed on the entire 1,872-image set.

C Evaluation Pipelines

The Capture Evaluation Pipeline The Capture evaluation pipeline consists of the following steps:

1. **Ground-truth extraction:** We extract objects, attributes, and relations from Visual Genome annotations.

2. **Ground-truth cleaning:** To ensure high-quality reference data, we apply a filtering step that removes samples with fewer than 10 distinct objects or that contain no attributes or relations. During this process, all concept types are normalized using the same preprocessing applied to the generated captions: lowercasing, lemmatization, stop-word removal, and deduplication. This reduces noise and prevents false positives when evaluating hallucination. After filtering, the number of usable images from Visual Genome decreased from 2,186 to 1,872. This filtering step was necessary for Capture evaluation, which relies on objects, attributes and relations. In contrast, the HalFscore evaluation used the region descriptions which did not require such filtering. As a result, all 2,186 images were retained in HalFscore. Full data statistics after filtration are provided in Appendix E.

3. **Caption generation:** Each fine-tuned model generates one caption per test image.

4. **Scene graph parsing:** Captions are converted into scene graphs that capture mentioned entities and their relations.

5. **Preprocessing:** Concept types in the generated scene graphs undergo the same normalization as the ground truth: lowercasing, lemmatization, deduplication, and stop-word removal.

6. **Semantic matching:** Caption and ground-truth elements are aligned using a combination of exact matching, synonym expansion, and SBERT-based similarity.

7. **Metric computation:** We compute correctness and coverage separately for objects, attributes, and relations.

The HalFcore Evaluation Pipeline The HalFscore evaluation pipeline is as follows:

1. **Caption generation:** Each image is captioned once by the model under evaluation.

2. **Entity-relation extraction:** We use GPT-4o with a structured prompt to convert each caption (GT and generated) into a set of structured triplets: `(object, attribute)` and $(\texttt{object}_1, \texttt{object}_2, \texttt{relation})$. These represent visual concepts and their relationships.

3. **Hallucination detection:** GPT-4o compares the generated caption triplets against the GT triplets to identify hallucinated concepts (i.e., objects, attributes, or relations not supported by GT). To avoid double counting, any object counted as hallucination is not counted again if its repeated as there can be repeats it the region descriptions used to make the GT caption. Moreover, any attribute or relation linked to a hallucinated object is excluded from further consideration.

Table 3: 95% confidence intervals for correctness ($\mathcal{C}$) and coverage ($\mathcal{V}$) across all concept types, computed over the full 1,872 Visual Genome images.

Model	Objects		Attributes		Relations		Train	Gen.
	$\mathcal{C}$	$\mathcal{V}$	$\mathcal{C}$	$\mathcal{V}$	$\mathcal{C}$	$\mathcal{V}$	len	len
COCO	84.15±0.73	26.48±0.58	28.09±1.76	07.20±0.58	60.64±1.06	31.95±0.66	10	11
LN-C	81.44±0.61	38.78±0.51	38.03±2.02	09.85±0.62	56.91±0.84	37.86±0.61	42	33
SG-4V-C	77.95±0.67	35.31±0.48	71.69±0.83	38.28±0.63	57.24±0.79	35.90±0.53	56	34

4. **Omission detection:** GPT-4o identifies concepts present in the GT caption but missing in the generated caption. As with hallucination, attributes or relations linked to omitted objects are not counted again.

5. **Metric computation:** Precision, recall, and HalFscore are computed from the counts of hallucinated and omitted concepts (see below).

This setup provides a granular understanding of model behavior, allowing us to assess both the factual accuracy and completeness of generated captions.

D HalFscore Metric

Following triplet extraction and comparison using an LLM, we compute the (HalFscore) from (Chen et al., 2025) as follows:

1. **Precision:** Measures the proportion of correctly grounded concepts in the generated caption, this is equivalent to *correctness score*:

$$\text{Precision} = 1 - \frac{|\text{Hallucinated Concepts}|}{|\text{Generated Concepts}|}$$

2. **Recall:** Measures the proportion of ground truth concepts recovered by the model, this is equivalent to *coverage score*:

$$\text{Recall} = 1 - \frac{|\text{Omitted Concepts}|}{|\text{GT Concepts}|}$$

3. **HalFscore:** The harmonic mean of precision and recall:

$$\text{HalFscore} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

This metric penalizes both hallucination and incompleteness, encouraging models to generate captions that are both accurate and exhaustive. The scoring is based on LLM reasoning, enabling semantic matching rather than brittle string comparisons, and avoids double-counting by ignoring attributes and relations connected to hallucinated or omitted objects.

E Dataset Statistics

After applying our filtering and normalization steps to the Visual Genome data (as described in Section 2.3), we retained 1,872 high-quality examples used for evaluation using CAPTURE metric. Figure 6 shows the distribution of object, attribute, and relation counts per image using a shared density plot. On average, images contain 18.35 objects (median 18), 19.74 attributes (median 18.5), and 16.88 relations (median 16), indicating rich semantic annotation across all concept types.

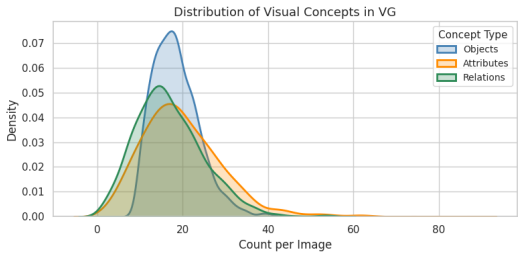

Figure 6: Overlaid KDE plot showing the distribution of object, attribute, and relation counts per image in the cleaned Visual Genome subset.

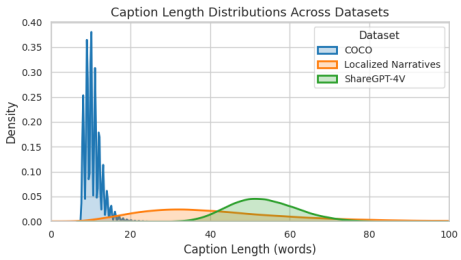

Figure 7: Caption length across the training datasets.

F Optimiser Hyper-Parameters

Both stages are trained with AdamW, using a learning rate of 1×10^{-5} and a batch size of 32. The learning rate is linearly warmed up during the first 10% of updates and then follows a cosine decay. The first stage runs for three epochs, and the second stage for one epoch. Experiments were run on 2×A100-SXM4-40GB GPUs.

G Comparison of CAPTURE and HalFscore Behavior

It is worth noting that while HalFscore and CAPTURE both avoid penalizing repeated mentions of the same hallucinated object, HalFscore also does not double-count hallucinated attributes or relations if their head object was already marked incorrect. This contrasts with CAPTURE, which treats each concept type independently and evaluates them regardless of their connection to hallucinated entities. As such, attribute coverage is more stable in HalFscore, as the evaluation avoids over-penalizing missing attributes when the associated object was also incorrect hence why the complete scores using HalFscore in Table 1 are not skewed like those of CAPTURE in Table 3 in Appendix B.

H Dataset Examples

Figure 8 shows example images with captions from the three different datasets: COCO, Localized Narratives, and ShareGPT-4V.

COCO captions are short and concise, typically mentioning only the main objects and actions in the image. Localized Narratives aim for more exhaustive coverage of the scene by referring to as many objects as possible and anchoring them spatially within the image. In contrast, ShareGPT-4V captions are generated by large language models and tend to be stylistically elaborate, often emphasizing visual attributes, setting details, and scene composition. and Figure **??** shows more qualitative examples.

I Object Mention Analysis by Caption Length

Further insight comes from analyzing the relationship between caption length, the number of object mentions, and grounding performance. As shown in Figure 9, Localized Narratives exhibit a clear trade-off: longer captions mention more objects, which boosts coverage but leads to reduced correctness. This pattern reflects an increase in hallucinations as verbosity rises.

In contrast, Figure 10 shows that ShareGPT-4V captions maintain a nearly constant number of object mentions across caption lengths. This leads to stable correctness and coverage, suggesting that ShareGPT-4V avoids the precision–recall trade-off by elaborating on existing entities rather than introducing new ones.

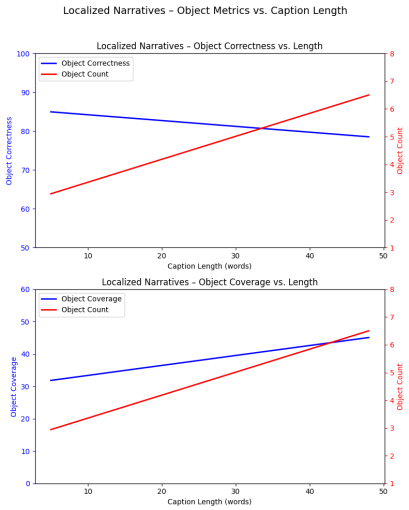

Figure 9: Analysis of object grounding behavior as caption length increases. Localized Narratives show a trade-off: longer captions increase object coverage but reduce correctness by introducing more hallucinated objects.

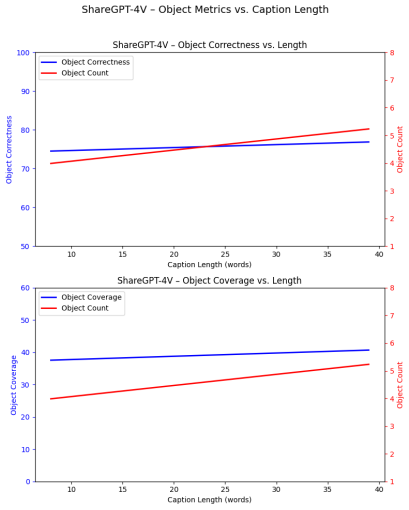

Figure 10: Analysis of object grounding behavior as caption length increases. ShareGPT-4V (bottom row) maintains stable correctness and coverage, as it tends to elaborate on existing objects rather than introduce new ones.

COCO: an image of a cars driving on the highway.

Localized Narratives: This is a picture taken of the outdoors. It is sunny. There are some vehicles moving on the road. On the left side of the vehicles there is a fence, electric poles, cables and some vehicles are parked. On the right side of the vehicles there is a fence, traffic cone and two cars are parked on the path. In front of the vehicles there are a bridge, trees and buildings.

ShareGPT-4V: The image captures a bustling intersection in Japan, teeming with life and activity. From a high vantage point, we can see a multitude of cars, predominantly in shades of white, orange, and blue, navigating their way through the intersection. The traffic lights, glowing in hues of red and yellow, guide the flow of vehicles, adding a sense of rhythm to the scene. The backdrop of the image is a blend of urban and natural elements. Buildings of varying heights stretch into the distance, interspersed with trees that bring a touch of green to the concrete jungle. The sky above is a clear blue, suggesting a bright and sunny day. The image is a snapshot of everyday life in Japan, encapsulating the organized chaos that characterizes its roads. It's a testament to the country's efficient traffic management system, where every element has its place and purpose.

Figure 8: Example captions from the original datasets used for fine-tuning.