From Descriptive Richness to Bias: Unveiling the Dark Side of Generative Image Caption Enrichment

Anonymous ACL submission

Abstract

Large language models (LLMs) have enhanced the capacity of vision-language models to caption visual text. This generative approach to image caption enrichment further makes textual captions more descriptive, improving alignment with the visual context. However, while many studies focus on benefits of generative caption enrichment (GCE), are there any negative side effects? We compare standard-format captions and recent GCE processes from the perspectives of "gender bias" and "hallucination", showing that enriched captions suffer from increased gender bias and hallucination. Furthermore, models trained on these enriched captions amplify gender bias by an average of 30.9% and increase hallucination by 59.5%. This study serves as a caution against the trend of making captions more descriptive.

1 Introduction

Large vision-language models (VLMs), such as BLIP (Li et al., 2023a), with superior performance in multi-modal understanding (Lüddecke and Ecker, 2022; Tewel et al., 2022), benefiting from millions of image-caption pairs. Improving training paradigms (Wang et al., 2023; Liu et al., 2024) and data augmentation strategies (Rotstein et al., 2024; Li et al., 2024) are crucial topics to enhance VLM performance in image captioning.

Among these techniques, Generative language models based Caption Enrichment (GCE) methods (Chen et al., 2023; Chan et al., 2023) have achieved some of the latest state-of-the-art performances. Unlike standard caption benchmarks, which concisely describe the salient parts (Misra et al., 2016) of an image (*e.g.*, COCO captions (Chen et al., 2015)), GCE methods create more descriptive and semantically enhanced captions. These enhanced textual captions are aligned to boost downstream performance with large language models (LLMs).

While many studies emphasize improving caption quality, issues such as *societal bias* and *hallucination* are significant yet often overlooked (Zhou et al., 2023) in image captioning. For example, Zhao et al. (2021) found that the COCO dataset is *skewed* towards men, and Hirota et al. (2022) showed that models trained on this biased data generate gender-stereotypical captions (*e.g.*, describing a *pink dress* for women not wearing one). These studies have highlighted potential biases in datasets like COCO and the models trained on them. Addressing this bias is crucial as it can exacerbate unfairness and risks towards underrepresented groups.

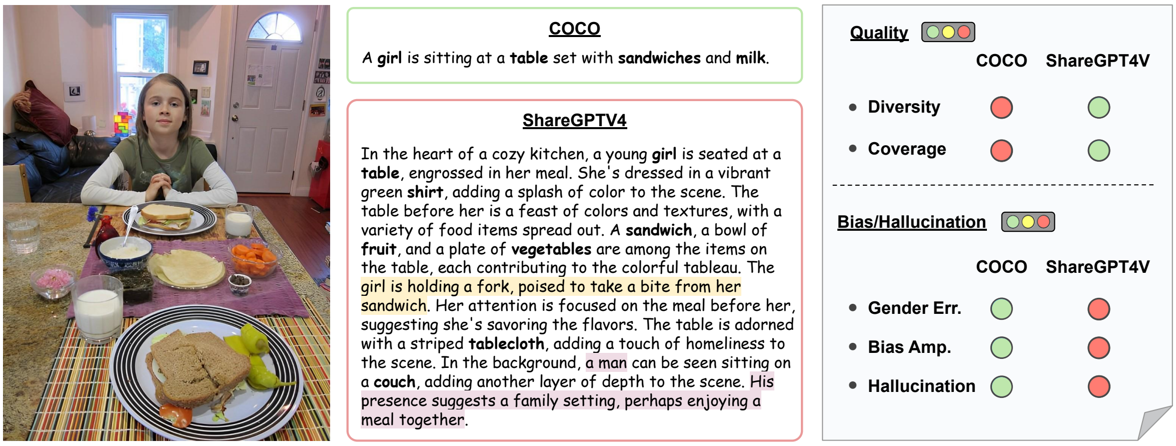
We aim to examine one critical question that has been overlooked in GCE works: *"Although LLM-enriched captions boost VLM performance, do they have negative effects, regarding societal bias and object hallucination?"* To answer this, we investigate gender bias and hallucination using comprehensive metrics, examining both datasets and models trained on these datasets for standard captions (COCO captions) and enriched captions (ShareGPT4V (Chen et al., 2023), FuseCap (Rotstein et al., 2024), CapsFusion (Yu et al., 2024)).

Our analysis reveals that LLM-enriched captions indeed have negative side effects, worsening issues of gender bias and hallucination by making captions more descriptive. Meanwhile, models trained on these enriched captions tend to amplify these problems further. Finally, we discuss possible causes of these negative effects and warn against the trend of making captions more descriptive.

2 Evaluation Framework

2.1 GCE approaches

We introduce recent representative approaches to generate enriched captions: ShareGPT4V (Chen et al., 2023), FuseCap (Rotstein et al., 2024), and Capsfusion (Yu et al., 2024). All these GCE methods utilize LLMs (Brown et al., 2020) or Large



COCO

A **girl** is sitting at a **table** set with **sandwiches** and **milk**.

ShareGPTV4

In the heart of a cozy kitchen, a young **girl** is seated at a **table**, engrossed in her meal. She's dressed in a vibrant green **shirt**, adding a splash of color to the scene. The table before her is a feast of colors and textures, with a variety of food items spread out. A **sandwich**, a bowl of **fruit**, and a plate of **vegetables** are among the items on the table, each contributing to the colorful tableau. The girl is holding a fork, poised to take a bite from her sandwich. Her attention is focused on the meal before her, suggesting she's savoring the flavors. The table is adorned with a striped **tablecloth**, adding a touch of homeliness to the scene. In the background, a man can be seen sitting on a **couch**, adding another layer of depth to the scene. His presence suggests a family setting, perhaps enjoying a meal together.

Quality

	COCO	ShareGPT4V
Diversity	🟥	🟩
Coverage	🟥	🟩

Bias/Hallucination

	COCO	ShareGPT4V
Gender Err.	🟩	🟥
Bias Amp.	🟩	🟥
Hallucination	🟩	🟥

Figure 1: Left: an overview of our analysis. Although the "LLM-enriched" caption (ShareGPT4V) covers more content than standard COCO (objects described in captions are **bolded**), it exhibits hallucination (in yellow) and gender bias, including describing gender not exist in the image and possible gender-stereotypical sentence (in purple). Right: a comparison between standard and enriched captions on caption quality, bias, and hallucination.

Multi-modal Models (LMMs) (OpenAI, 2023) to describe images in detail or summarize different sources of the information.

ShareGPT4V utilizes GPT4-Vision (OpenAI, 2023) to generate 1.2M large scale high-quality captions for incremental training on a strong 7B VLM with strong generalization and SOTA results. **FuseCap** uses several pre-trained off-the-shelf vision models (*e.g.*, object detector) to extract diverse visual information. The outputs from these models and original captions are fused using ChatGPT (Ouyang et al., 2022) to generate enriched captions. **Capsfusion** generates captions using a pre-trained captioner, BLIP (Li et al., 2023a), then fuses them with original captions using ChatGPT.

2.2 Evaluation metrics

Our analysis focuses on caption quality, societal bias, and hallucination. Let $\mathcal{D}$ be a dataset of n samples, $\mathcal{D} = \{(I_i, c_i, a_i) \mid 1 \leq i \leq n\}$, where each sample includes an image I_i, a caption c_i, and an optional binary gender label a_i (woman or man). We introduce the metrics to evaluate each aspect.

Caption quality. We evaluate caption quality in three aspects: ***Vocabulary diversity*** is the total number of unique words across all captions in $\mathcal{D}$. ***Caption length*** is the average number of tokens per caption in $\mathcal{D}$. ***Recall*** measures the proportion of objects mentioned in captions to the total objects in the images. For each caption $c_i \in \mathcal{D}$, recall is calculated as:

$$\text{Recall} = \frac{1}{n}\sum_{i=1}^{n}\frac{r_i}{o_i}, \qquad (1)$$

where o_i is the total number of objects in I_i, and r_i is the number of relevant objects mentioned in c_i. Note that conventional reference-based metrics like CIDEr (Vedantam et al., 2015) cannot be applied to descriptive captions (Chan et al., 2023).

Societal bias. We focus on gender bias as gender terms are more frequently described in captions than other attributes. We adopt three metrics to measure gender bias: ***Gender error*** (Burns et al., 2018) measures the rate of incorrect gender predictions in captions. If a caption $c_i \in \mathcal{D}$ with gender label a_i refers to a woman as a *man* or vice versa, it counts as an error. The gender error is the proportion of such errors in $\mathcal{D}$. ***Recall disparity*** (Hall et al., 2023) evaluates the recall disparity between genders. Consider two subsets based on a_i: $\mathcal{D}_{\text{woman}}$ and $\mathcal{D}_{\text{man}}$. Recall disparity is the average absolute difference in recall for each object j:

$$\text{Disparity} = \tfrac{1}{m}\textstyle\sum_{i=1}^{m} |\text{Recall}_{\text{man},j} - \text{Recall}_{\text{woman},j}| \qquad (2)$$

where m is the total number of COCO objects (Lin et al., 2014), $\text{Recall}_{\text{man},j}$ is the recall of COCO objects j in $\mathcal{D}_{\text{man}}$, and vice versa. ***LIC*** (Hirota et al., 2022) quantifies how gender-stereotypical captions in $\mathcal{D}$ are compared to human-written captions. It compares the accuracies of two gender classifiers: one trained on $c_i \in \mathcal{D}$ and the other on ground-truth captions. Higher accuracy for the classifier trained on c_i indicates more gender-stereotypical information in these captions.

Hallucination. We use the CHAIR metric (Rohrbach et al., 2018) to evaluate hallucination in captions. CHAIR has two components: ***CHAIR***$_i$ is the fraction of mentioned objects in the captions

Table 1: Caption quality, gender bias, and hallucination for upstream and downstream analysis. Red/green indicates the worst/best score for each metric. Recall, gender bias, and hallucination metrics are scaled by 100.

	Caption Quality ↑			Gender bias ↓			Hallucination ↓	
Captions	Diversity	Length	Recall	Gender Err.	LIC	Recall Disp.	CHAIR_s	CHAIR_i
Upstream								
COCO captions	**12,834**	**11.3**	**42.6**	**0**	**0**	**7.0**	**0**	**0**
ShareGPT4V	25,349	**166.1**	**61.7**	2.5	**17.4**	**24.9**	**20.7**	**5.7**
FuseCap	**25,892**	39.8	59.4	**3.2**	14.3	9.9	9.2	4.0
CapsFusion	13,158	16.9	44.6	1.4	1.2	7.6	3.5	2.2
Downstream								
COCO captions	**3,312**	**10.9**	**45.7**	**3.1**	**5.5**	**7.8**	**4.7**	**3.1**
ShareGPT4V	**9,573**	**153.8**	56.3	3.4	14.3	**30.5**	**21.5**	**6.9**
FuseCap	6,341	42.0	**56.9**	**4.8**	**17.3**	16.3	13.2	6.3
CapsFusion	3,385	15.3	48.0	4.2	6.3	8.8	7.2	4.4

Table 2: Difference in gender bias and hallucination between upstream and downstream (downstream - upstream). Red/green is bias amplification/mitigation.

	ΔGender bias			ΔHallucination	
Captions	Err.	LIC	Disp.	C_s	C_i
COCO cap.	3.1	5.5	0.8	4.7	3.1
ShareGPT4V	0.9	-3.1	5.6	0.8	1.2
FuseCap	1.6	3.0	6.4	4.0	2.3
CapsFusion	2.8	5.1	1.2	3.7	2.2

c_i that do not appear in images I_i:

$$\text{CHAIR}_\text{i} = \frac{H}{M}, \tag{3}$$

where H is the number of hallucinated objects, and M is the total number of objects mentioned in the captions. ***CHAIR***$_s$ is the fraction of captions c_i with at least one hallucinated object:

$$\text{CHAIR}_\text{s} = \frac{S_h}{n}, \tag{4}$$

where S_h is the number of captions with hallucinated objects. We focus on 80 objects in COCO.

3 Evaluation

Setup. We analyze concise (COCO captions) and enriched captions (ShareGPT4V, FuseCap, CapsFusion) based on the metrics in Section 2.2. Enriched captions are generated for the COCO training set using these approaches. We first conduct an *upstream* analysis of the four datasets and then a *downstream* analysis of captions generated by a captioner trained on each dataset. For downstream analysis, we fine-tune a pre-trained BLIP for 5 epochs with the AdamW optimizer, generating captions for the COCO validation set. Detailed experimental settings are in Appendix A.

3.1 Upstream & downstream analysis

We present qualitative results in Figure 1 and Appendix B, with key observations below.

Observation 1. **More descriptive, more gender bias.** Table 1 (upstream) shows a clear tendency for gender bias to increase as captions become more descriptive. For instance, COCO captions have the lowest object coverage (*i.e.*, recall: 42.6) but exhibit the least bias. In contrast, ShareGPT4V and FuseCap have higher object coverage but higher gender bias than COCO captions (*e.g.*, LIC is 0 for COCO and 17.4 for ShareGPT4V). This observation is further confirmed by Figure 2 (left), showing a strong correlation between LIC and recall ($R^2 = 0.99$). In other words, making captions more descriptive increases the risk of gender bias.

Observation 2. **Enriched captions exhibit greater recall disparity.** In Figure 4, we visualize the difference in recall ($\text{Recall}_\text{man} - \text{Recall}_\text{woman}$) for the top-10 objects that co-occur with images in $\mathcal{D}_\text{woman}$ and $\mathcal{D}_\text{man}$. The results show that ShareGPT4V exhibits a more significant recall disparity for all objects. For example, for the *handbag* object, COCO captions show almost no gender difference, while ShareGPT4V exhibits a strong bias towards men. This further validates the risk of gender bias in enriched captions.

Observation 3. **More descriptive, more hallucination.** A similar trend between descriptiveness and hallucination is also evident in Table 1 (upstream). COCO captions, which has the lowest object coverage, exhibits the lowest hallucination rates. Conversely, ShareGPT4V, with the highest object coverage, shows significantly increased hallucination rates compared to COCO captions (*e.g.*, CHAIR$_s$ is 0 for COCO and 20.7 for ShareGPT4V). This trend is corroborated by Figure 3 (left), high-

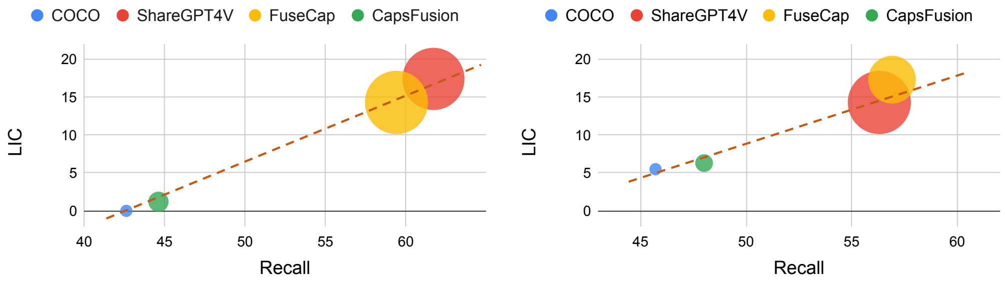


Figure 2: LIC vs. Recall (left: upstream, right: downstream). The bubble size indicates vocabulary size. LIC tends to increase with higher recall, shown by strong trends (dotted lines) with $R^2 = 0.99$ (left) and $R^2 = 0.97$ (right).

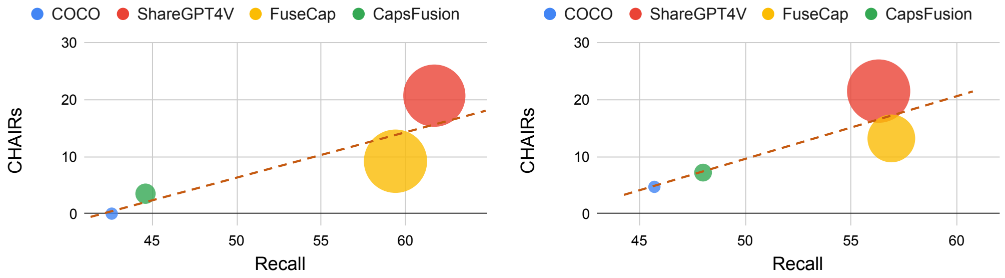


Figure 3: CHAIR$_s$ vs. Recall (left: upstream, right: downstream). The bubble size indicates vocabulary size. CHAIR$_s$ tends to increase with higher recall, shown by strong trends with $R^2 = 0.80$ (left) and $R^2 = 0.76$ (right).

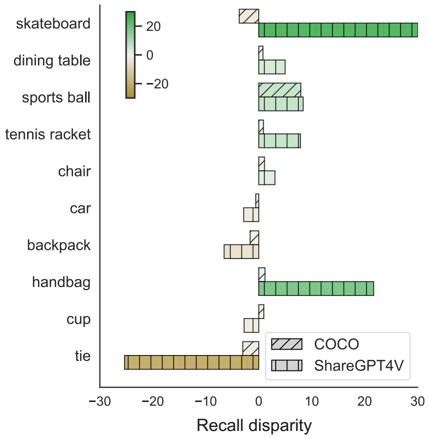


Figure 4: Recall disparity by visual object.

lighting a strong correlation between hallucination rates and recall ($R^2 = 0.80$). Thus, making captions more descriptive increases hallucination risks.

Observation 4. **Models trained on the datasets inherit/amplify bias and hallucination.** Table 1 (downstream) shows that models inherit the dataset's bias tendencies. Specifically, the model trained on the least descriptive captions (*i.e.*, COCO captions) exhibits the smallest bias and hallucination, while models trained on the most descriptive captions, ShareGPT4V and FuseCap, show significant bias and hallucination. Figures 2 and 3 (right) further demonstrate that the models inherit the datasets' bias and hallucination. Furthermore, Table 2 shows that in most cases, the models amplify the dataset's biases. For example, ShareGPT4V's recall disparity worsens from 24.9 to 30.5 ($\Delta = 5.6$), and CHAIR$_s$ increases from 20.7 to 21.5 ($\Delta = 0.8$). These results highlight the severe issue of dataset bias, as it directly affects the outcomes of the trained models.

4 Discussion on Possible Sources of Bias

To enhance descriptiveness, GEC methods heavily rely on LLMs to improve textual alignment. However, issues with gender bias and hallucination (Gunjal et al., 2024) have been explored in these LLMs. The enrichment process, which depends on text representations, risks incorporating these inherent biases into the final captions. Furthermore, the lack of human oversight in the caption generation process can exacerbate these issues. Without iterative human intervention to correct biases, the inaccuracies of LLMs remain unaddressed, leading to increased bias and hallucination. Introducing human-in-the-loop (Yang et al., 2019) could mitigate these problems by ensuring that captions are free from gender-stereotypical descriptions.

5 Conclusion

We examined standard and LLM-enriched captions for gender bias and hallucination, deriving key insights: GCE-based image captioning exacerbates these bias, which are further amplified in downstream models. We argue that further efforts must be invested to the problems to strike a balance between *descriptive richness* and *incremental bias*.

Limitations

Attributes other than gender. We focused our analysis on gender bias for societal bias. This is because gender-related terms are more frequently described in captions compared to other attributes, making gender bias particularly prominent in captions (Hirota et al., 2022). However, previous works have shown that racial bias, though not as pronounced as gender bias, is also present in captioning models (Zhao et al., 2021). Analyzing racial bias and bias of other attributes requires future studies and efforts.

Evaluation metrics. While our analysis demonstrated various critical problems in enriched captions (*e.g.*, they exacerbate bias and hallucination), there may be aspects that we can further investigate. For example, we can consider other attributes for societal bias analysis and utilize hallucination metrics that account for elements *beyond objects*. However, our analysis is robust and highlights critical issues in enriched captions, serving as a counterpoint to the trend of making captions more descriptive and benefiting the community.

Source datasets other than COCO. In our analysis, we used COCO as the source for images for two reasons: (1) COCO images come with high-quality, human-annotated concise captions, providing a solid basis for evaluating concise captions; and (2) COCO has been extensively analyzed in existing research for societal bias and hallucination (Li et al., 2023b). We did not use other image-caption datasets (*e.g.*, Google Conceptual Captions (Sharma et al., 2018), LAION (Schuhmann et al., 2022)) because the quality of the accompanying captions is lower, making the analysis results less reliable.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *NeurIPS*.

Kaylee Burns, Lisa Anne Hendricks, Kate Saenko, Trevor Darrell, and Anna Rohrbach. 2018. Women also snowboard: Overcoming bias in captioning models. In *ECCV*.

David Chan, Austin Myers, Sudheendra Vijayanarasimhan, David A Ross, and John Canny. 2023. Ic3: Image captioning by committee consensus. In *EMNLP*.

Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Jacob Devlin, Hao Cheng, Hao Fang, Saurabh Gupta, Li Deng, Xiaodong He, Geoffrey Zweig, and Margaret Mitchell. 2015. Language models for image captioning: The quirks and what works. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Association for Computational Linguistics.

Anisha Gunjal, Jihan Yin, and Erhan Bas. 2024. Detecting and preventing hallucinations in large vision language models. In *AAAI*.

Melissa Hall, Laura Gustafson, Aaron Adcock, Ishan Misra, and Candace Ross. 2023. Vision-language models performing zero-shot tasks exhibit gender-based disparities. In *ICCV Workshops*.

Yusuke Hirota, Yuta Nakashima, and Noa Garcia. 2022. Quantifying societal bias amplification in image captioning. In *CVPR*.

Jiaxuan Li, Duc Minh Vo, Akihiro Sugimoto, and Hideki Nakayama. 2024. Evcap: Retrieval-augmented image captioning with external visual-name memory for open-world comprehension. In *CVPR*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023b. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *EMNLP*.

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 *ECCV*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *CVPR*.

Timo Lüddecke and Alexander Ecker. 2022. Image segmentation using text and image prompts. In *CVPR*.

Ishan Misra, C Lawrence Zitnick, Margaret Mitchell, and Ross Girshick. 2016. Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels. In *CVPR*.

OpenAI. 2023. Gpt-4v(ision) system card. In *OpenAI Blog*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *EMNLP*.

Noam Rotstein, David Bensaïd, Shaked Brody, Roy Ganz, and Ron Kimmel. 2024. Fusecap: Leveraging large language models for enriched fused image captions. In *WACV*.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. In *NeurIPS*.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*.

Yoad Tewel, Yoav Shalev, Idan Schwartz, and Lior Wolf. 2022. Zerocap: Zero-shot image-to-text generation for visual-semantic arithmetic. In *CVPR*.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *CVPR*.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *CVPR*.

Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. 2023. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. In *NeurIPS*.

Yiwei Yang, Eser Kandogan, Yunyao Li, Prithviraj Sen, and Walter S Lasecki. 2019. A study on interaction in human-in-the-loop machine learning for text analytics. In *IUI Workshops*.

Qiying Yu, Quan Sun, Xiaosong Zhang, Yufeng Cui, Fan Zhang, Xinlong Wang, and Jingjing Liu. 2024. Capsfusion: Rethinking image-text data at scale. In *CVPR*.

Dora Zhao, Angelina Wang, and Olga Russakovsky. 2021. Understanding and evaluating racial biases in image captioning. In *ICCV*.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2023. Analyzing and mitigating object hallucination in large vision-language models. In *ICLR*.

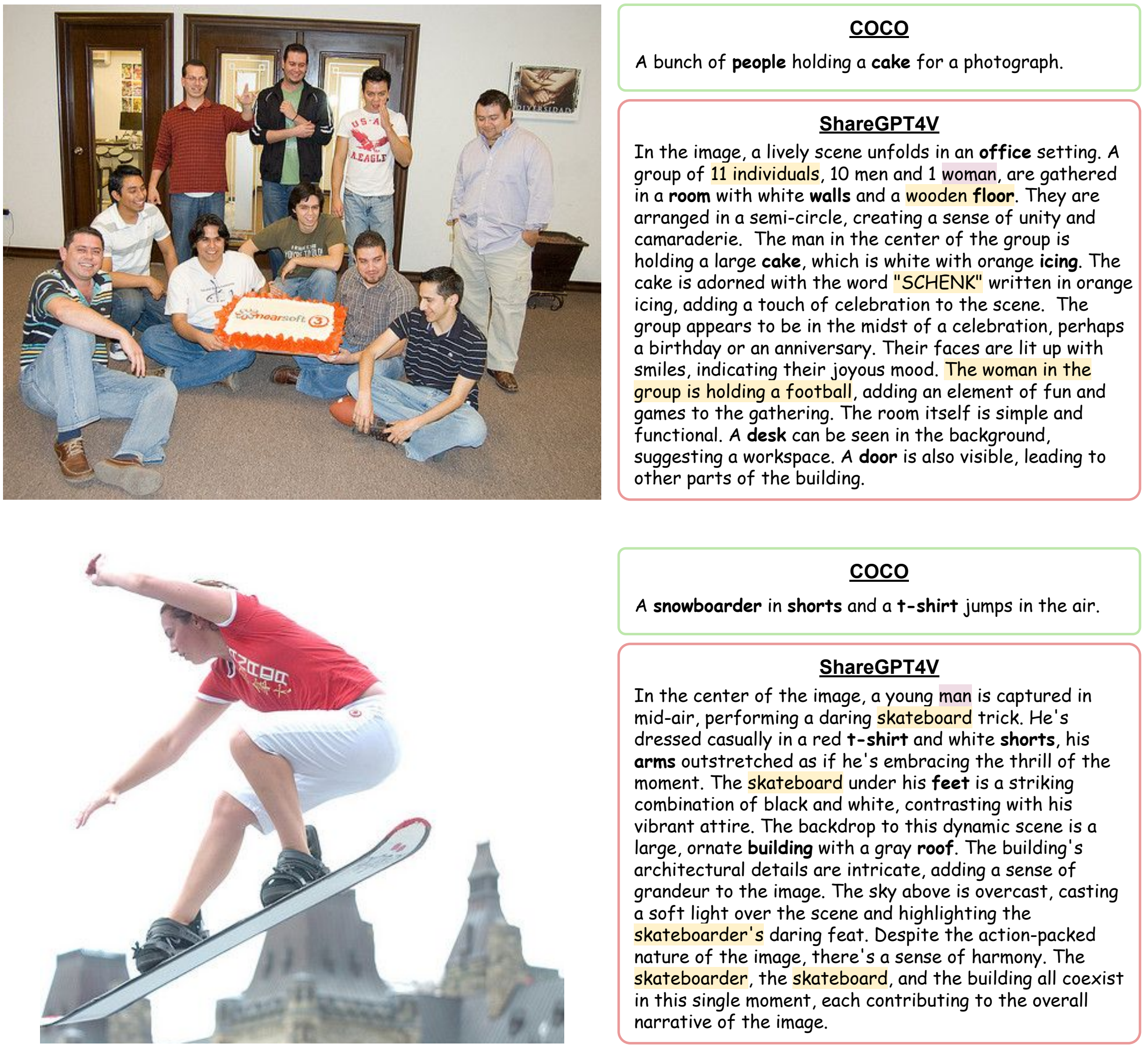
A Appendix Settings of the Experiments

A.1 Upstream analysis

For the upstream analysis, we use images of the COCO training set as an image source. For every image, we apply GCE approaches introduced in Section 2.1, obtaining enriched captions for each approach. We compare these enriched captions with concise captions (COCO captions) in terms of the metrics introduced in Section 2.2.

A.2 Downstream analysis

For the downstream analysis, we fine-tune pre-trained BLIP on one of the four datasets (*i.e.*, COCO captions, ShareGPT4V, FuseCap, or CapsFusion). We fine-tune the models for 5 epochs with the AdamW optimizer. Then, the fine-tuned models are used to generate captions for the COCO validation set. The generated captions for the validation set are used for the downstream analysis.



COCO

A bunch of **people** holding a **cake** for a photograph.

ShareGPT4V

In the image, a lively scene unfolds in an **office** setting. A group of 11 individuals, 10 men and 1 woman, are gathered in a **room** with white **walls** and a wooden **floor**. They are arranged in a semi-circle, creating a sense of unity and camaraderie. The man in the center of the group is holding a large **cake**, which is white with orange **icing**. The cake is adorned with the word "SCHENK" written in orange icing, adding a touch of celebration to the scene. The group appears to be in the midst of a celebration, perhaps a birthday or an anniversary. Their faces are lit up with smiles, indicating their joyous mood. The woman in the group is holding a football, adding an element of fun and games to the gathering. The room itself is simple and functional. A **desk** can be seen in the background, suggesting a workspace. A **door** is also visible, leading to other parts of the building.

COCO

A **snowboarder** in **shorts** and a **t-shirt** jumps in the air.

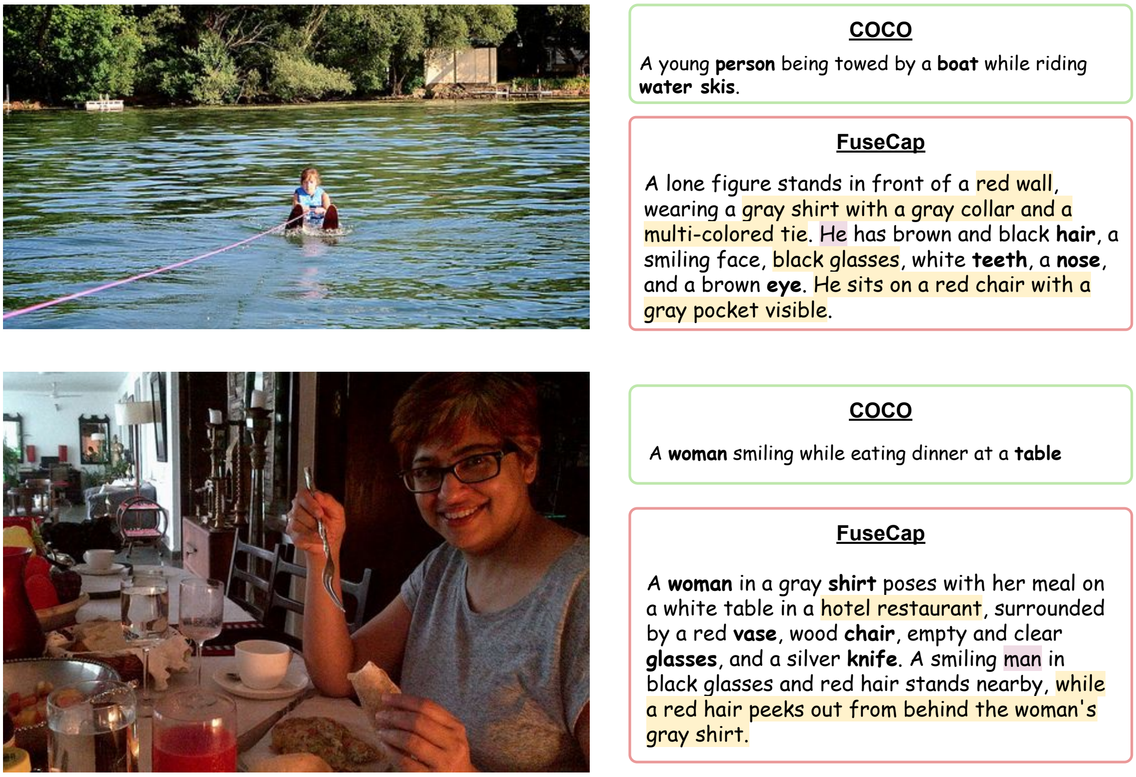
ShareGPT4V

In the center of the image, a young man is captured in mid-air, performing a daring skateboard trick. He's dressed casually in a red **t-shirt** and white **shorts**, his **arms** outstretched as if he's embracing the thrill of the moment. The skateboard under his **feet** is a striking combination of black and white, contrasting with his vibrant attire. The backdrop to this dynamic scene is a large, ornate **building** with a gray **roof**. The building's architectural details are intricate, adding a sense of grandeur to the image. The sky above is overcast, casting a soft light over the scene and highlighting the skateboarder's daring feat. Despite the action-packed nature of the image, there's a sense of harmony. The skateboarder, the skateboard, and the building all coexist in this single moment, each contributing to the overall narrative of the image.

Figure 5: Qualitative examples of the comparison between COCO captions and ShareGPT4V. Objects described in captions are **bolded**. Gender bias and hallucination are highlighted in purple and yellow, respectively.

B Additional Qualitative Examples

We show some qualitative examples where enriched captions generated by GCE methods exhibit gender bias and hallucination. The examples are shown in Figures 5 to 7 (ShareGPT4V, FuseCap, and CapsFusion, respectively).



COCO

A young **person** being towed by a **boat** while riding **water skis**.

FuseCap

A lone figure stands in front of a red wall, wearing a gray shirt with a gray collar and a multi-colored tie. He has brown and black **hair**, a smiling face, black glasses, white **teeth**, a **nose**, and a brown **eye**. He sits on a red chair with a gray pocket visible.

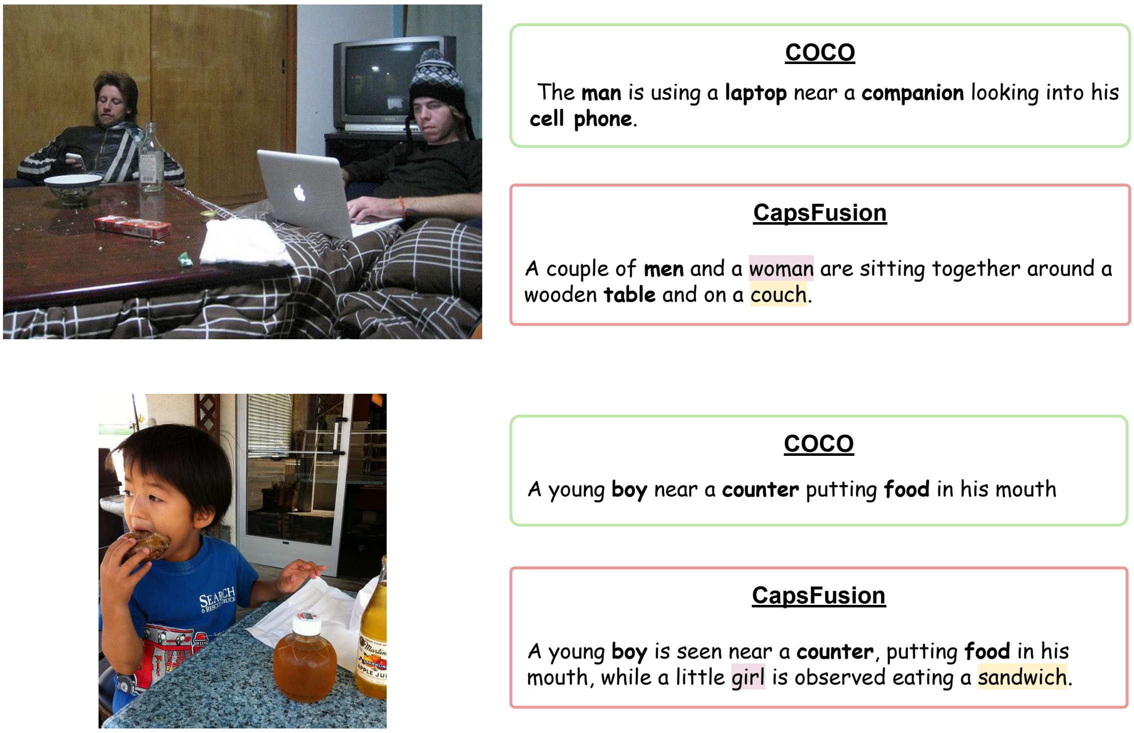
COCO

A **woman** smiling while eating dinner at a **table**

FuseCap

A **woman** in a gray **shirt** poses with her meal on a white table in a hotel restaurant, surrounded by a red **vase**, wood **chair**, empty and clear **glasses**, and a silver **knife**. A smiling man in black glasses and red hair stands nearby, while a red hair peeks out from behind the woman's gray shirt.

Figure 6: Qualitative examples of the comparison between COCO captions and FuseCap. Objects described in captions are **bolded**. Gender bias and hallucination are highlighted in purple and yellow, respectively.



COCO

The **man** is using a **laptop** near a **companion** looking into his **cell phone**.

CapsFusion

A couple of **men** and a woman are sitting together around a wooden **table** and on a couch.

COCO

A young **boy** near a **counter** putting **food** in his mouth

CapsFusion

A young **boy** is seen near a **counter**, putting **food** in his mouth, while a little girl is observed eating a sandwich.

Figure 7: Qualitative examples of the comparison between COCO captions and CapsFusion. Objects described in captions are **bolded**. Gender bias and hallucination are highlighted in purple and yellow, respectively.

The enriched captions suffer from gender misclassification (*e.g.*, bottom of Figure 5), incorrectly describing two people in the image as a couple (*e.g.*, top of Figure 7), describing nonexistent individuals with different genders (*e.g.*, bottom of Figure 6), and object hallucination (in all the figures). These results further confirm the negative impacts of GCE.

B.1 Non-LLM based Caption Enrichment

We also would like to credit previous works (Devlin et al., 2015; Vinyals et al., 2015) of non pre-training based language modeling to enhance image captioning by providing structured linguistic patterns and vocabulary. However, without the depth of large language models, such systems may exhibit bias and limited expressiveness, struggling to generate diverse and contextually nuanced captions. These models often rely on statistical techniques, which can constrain their descriptive capabilities compared to their more advanced counterparts. In other words, how to incorporate structure knowledge refinement or graphical structure would also be important for LLM-based caption enrichment in future studies.