# Image Embedding Sampling Method for Diverse Captioning

**Anonymous ACL submission**

## Abstract

Image Captioning for state-of-the-art VLMs has significantly improved over time; however, this comes at the cost of increased computational complexity, making them less accessible for resource-constrained applications such as mobile devices and assistive technologies. Alternatively, smaller VLMs prioritize high-level scene descriptions, overlooking finer details that contribute to a richer understanding of an image. In this paper, we introduce a training-free framework that enhances caption diversity and informativeness by explicitly attending to distinct image regions using a comparably small VLM, BLIP, as the backbone. Our approach leverages structured segmentation to produce hierarchical representations that capture both global and localized semantics. Without requiring additional model training, we demonstrate that our method allows smaller VLMs to achieve performance comparable to larger models in terms of image-caption alignment, semantic integrity, and diversity. We evaluate our framework on MSCOCO, Flickr30k, and Nocaps test datasets, achieving a Div-2 score of 0.735, 0.750, and 0.748 for each dataset respectively, while maintaining strong image-caption relevancy and semantic integrity with the human-annotated captions.

## 1 Introduction

Visual-Language Models (VLMs) have seen rapid advancements in image captioning, benefiting from increasingly sophisticated architectures and larger training datasets (Alayrac et al., 2022; Li et al., 2022b; Radford et al., 2021a; Wang et al., 2022a). State-of-the-art large-scale models generate highly detailed and diverse captions, yet their extensive computational requirements can be prohibitive in resource-constrained settings. Conversely, smaller VLMs, while more efficient, often prioritize dominant visual elements and overlook fine-grained details, resulting in captions that lack the depth and specificity seen in human-generated captions (Aneja et al., 2019a; Bianco et al., 2023; Chen et al., 2023; Yuksekgonul et al., 2022).

Inspired by previous work (Ji et al., 2021; Shao et al., 2023; Shukor et al., 2022) that demonstrates the advantages of hierarchical approaches in image understanding, our method leverages structured segmentation to capture both global and regional aspects of an image. We sample segmentation-driven embeddings to explicitly attend to distinct image regions while preserving contextual relationships, generating captions at multiple levels of granularity. This approach enables smaller VLMs to achieve performance comparable to larger models in terms of caption diversity and image-caption alignment.

We validate our approach, namely, **HBoP** - **H**ierarchical **B**ags **o**f **P**hrases, by evaluating generated captions for MSCOCO, Flickr30k, and Nocaps datasets on conventional diversity metrics such as mBLEU-4, n-gram diversity (Aneja et al., 2019b), and newly presented pairwise cosine distance (PCD). Our findings show that structured caption generation effectively improves diversity while maintaining relevancy with images and human-generated captions (*compare* BLIP (Li et al., 2022a), HBoP, and gold captions in Figure 1).

## 2 Related Works

Vision-language models have demonstrated remarkable capabilities in various multi-modal tasks, with caption generation being a primary benchmark for evaluating their performance. Models such as CLIP (Radford et al., 2021a), Flamingo (Alayrac et al., 2022), and BLIP-2 (Li et al., 2023) leverage contrastive learning and large-scale pre-training to improve vision-language alignment. Although these approaches improve caption fluency and coherence, they often generate high-level scene descriptions without capturing fine-grained details, limiting their utility for applications requiring de-

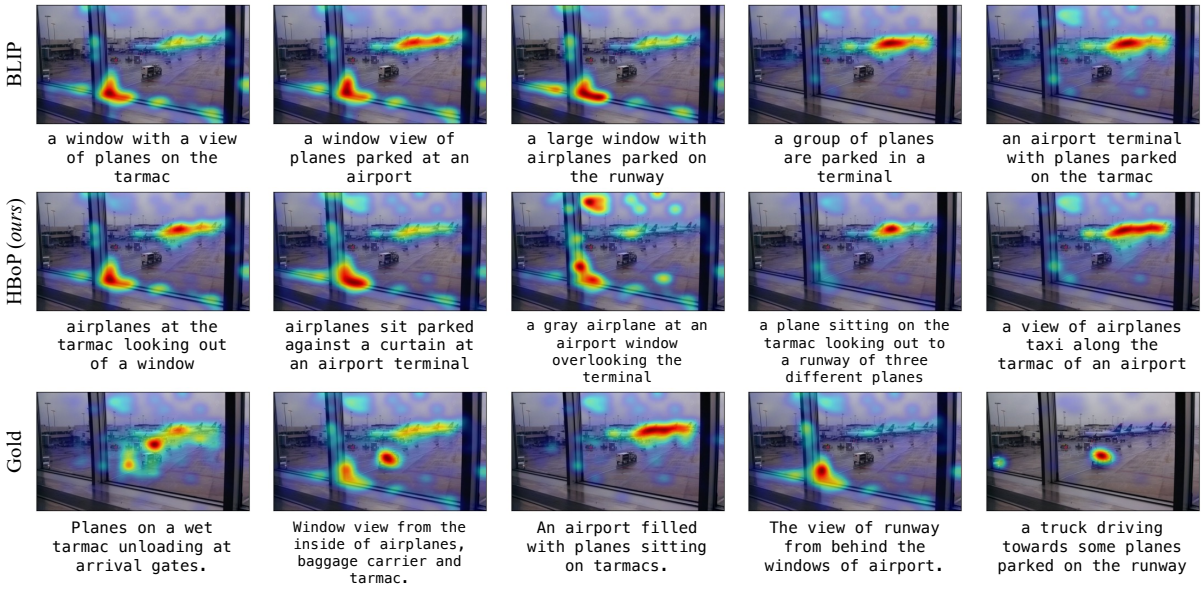


Figure 1: Comparison of captions generated by BLIP, HBoP, and human annotations. The images are overlaid with GradCAM heatmaps to highlight the regions focused on by the pretrained image-text matching model (Li et al., 2022a). HBoP captions exhibit greater diversity compared to BLIP captions and are closer to human-annotated gold captions.

tailed image understanding. Traditional captioning models treat images holistically, often overlooking fine-grained hierarchical details (Xu et al., 2021), unless trained with an explicit training objective of improving diversity, such as ModeCap (Chen et al., 2022) which explores various modes in the training corpus and Seq-CVAE (Aneja et al., 2019b) that uses latent variables to generate every word within a sentence.

Inspired by hierarchical representation techniques (Ji et al., 2021; Shao et al., 2023; Shukor et al., 2022), our approach samples the latent image embeddings corresponding to structured segmentation to generate captions that reflect multiple levels of contextual details. PnP-VQA (Tiong et al., 2022) proposes a similar methodology but focuses primarily on visual question answering. While PnP-VQA leverages image segmentation to generate captions, its sampling approach results in less meaningful captions due to its reliance on high-activation regions extracted using Grad-CAMs (Selvaraju et al., 2017).

## 3 Methodology

In this section, we introduce our proposed framework, HBoP (depicted in Fig 2), a modular architecture that uses pre-trained segmentation and captioning models. We show that HBoP ensures multiple levels of captions (*i.e.*, global, regional, fine-grained) by inducing a hierarchical structure for image understanding.

### 3.1 Image Segmentation Module (ISM)

The first component of HBoP, ISM, selects patch embeddings ($E_X$) corresponding to image regions ($X = (X_1, X_2, ..., X_n)$) from the original image embeddings extracted using a Vision Transformer (ViT) (Dosovitskiy et al., 2020) encoder. These regions are selected based on the segmentation masks produced by the state-of-the-art Segment Anything model (SAM) (Kirillov et al., 2023). If we select a set of $p$ segmentation masks for the image, the resulting masks for the selected image regions would be: $M_X = \{M_{X_1}, M_{X_2}, ..., M_{X_p}\} = \text{SAM}(X), X \in \mathbb{R}^{H \times W \times C}$, where $H$, $W$, and $C$ represent the height, width, and channels of $X$.

### 3.2 Hierarchical Composition Module (HCM)

The second component, HCM, is a key component that can control the level of captions. Specifically, we present three types of captions that can be derived using HCM.

***Global/Fine-grained level* captions** The global segmentation masks ($M_G$) are selected by choosing the top-$k$ (5 in our case) largest segmentation masks from $M_X$ after applying non-maximum sup-

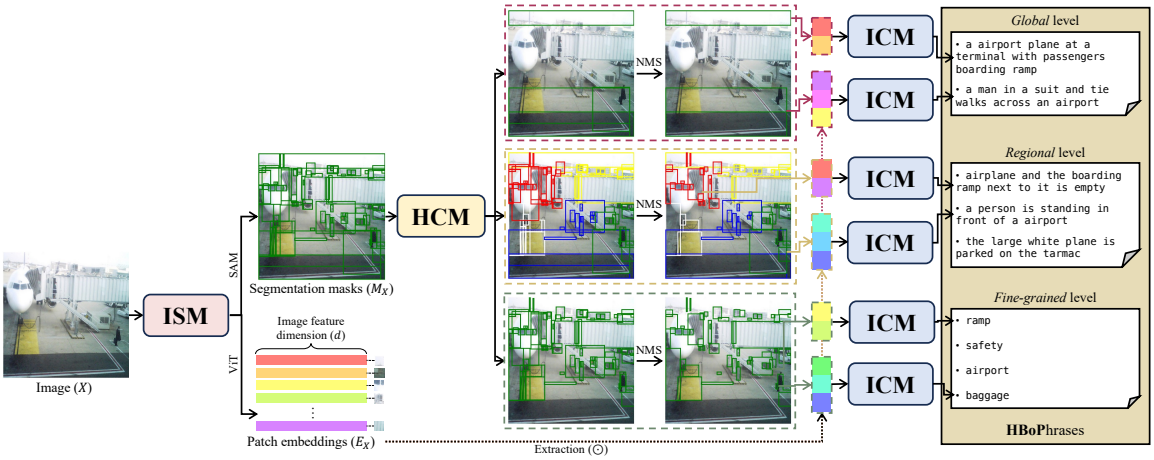


Figure 2: The proposed HBoP framework consists of three components: (1) Image Segmentation Module (ISM), (2) Hierarchical Composition Module (HCM), and (3) Image Captioning Module (ICM). HBoP controls caption granularity by selecting meaningful patch embeddings of varying sizes from the segmentation model.

pression (NMS) (Hosang et al., 2017):

$$M_G = \{M_{g_1}, M_{g_2}, ..., M_{n_g}\},$$
$$M_{g_i} = \text{NMS}(\text{Top-}k(M_X)), \quad i = 1, ..., n_g$$

NMS removes multiple segmentation masks with overlapping, similar contexts using the Intersection over Union (IoU) and predicted confidence from SAM. The remaining masks, after applying NMS, can also be used to generate fine-grained captions (discussed in Appendix D.3):

$$M_F = \{M_{f_1}, M_{f_2}, ..., M_{n_f}\},$$
$$M_{f_i} = \text{NMS}(M_X) \setminus M_G, \quad i = 1, ..., n_f$$

***Regional level* captions** To create regional-level segmentation masks, $M_R$, we use $K$-means clustering to partition all the segmentation masks ($M_X$) and apply NMS to each cluster individually:

$$M_R = \{M_{r_1}, M_{r_2}, ..., M_K\},$$
$$M_{r_i} = \text{NMS}(\text{K-means}(M_X)), i = 1, ..., K$$

The hierarchical segmentation masks ($M_g$, $M_r$ and $M_f$) are used to extract relevant patch embeddings, $E_g$, $E_r$ and $E_f$ using $E_X$ from the first stage. We extract ($\odot$) the corresponding embeddings by concatenating the extracted patch embeddings of different levels. Thus, the final selected image embeddings can be categorized as:

$$E_G = \{E_{g_1}, E_{g_2}, ..., E_{g_{n_g}}\}, E_{g_i} = E_X \odot M_{g_i}$$
$$E_R = \{E_{r_1}, E_{r_2}, ..., E_K\}, E_{r_i} = E_X \odot M_{r_i}$$
$$E_F = \{E_{f_1}, E_{f_2}, ..., E_{n_f}\}, E_{f_i} = E_X \odot M_{f_i}$$

### 3.3 Image Captioning Module (ICM)

To generate captions for different levels of image embeddings, we use BLIP fine-tuned on image captioning (Li et al., 2022a) with the stochastic sampling method, following the same procedure as (Tiong et al., 2022). The caption generation process is repeated for $n_g$, $n_r$, and $n_f$ patch embeddings corresponding to the number of selected hierarchical masks. Since the patch embedding size may vary due to the different mask sizes, we use zero padding before using the captioning module. Our final HBoP captions would be:

$$\text{HBoP}_G = \{s_{g_1}, ..., s_{n_g}\}, s_{g_i} = \text{BLIP}(E_{g_i})$$
$$\text{HBoP}_R = \{s_{r_1}, ..., s_K\}, s_{r_i} = \text{BLIP}(E_{r_i})$$
$$\text{HBoP}_F = \{s_{f_1}, ..., s_{n_f}\}, s_{f_i} = \text{BLIP}(E_{f_i})$$

## 4 Results

**HBoP achieves the best diversity scores while maintaining relevance among smaller VLMs.** We compare the diversity and relevance of captions generated by different models, as shown in Table 1. While larger VLMs with LLM-based text encoders generally achieve stronger overall performance, HBoP remains competitive, achieving the closest diversity scores to the gold captions among smaller VLMs in terms of PCD (see Appendix C for details), mBLEU-4 and Div-2 (Aneja et al., 2019b).

Specifically, compared to captions generated by BLIP (Li et al., 2022a) and BLIP-2 (Li et al., 2023), HBoP achieves a PCD score closest to that of the

| | LLM Encoder | # of Param | MSCOCO (5k test set) | | | | | Flickr30K (1k test set) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Relevancy | | Diversity | | | Relevancy | | Diversity | | |
| | | | SBERT ↑ | CLIP-S ↑ | PCD | mBLEU-4 ↓ | Div-2 ↑ | SBERT ↑ | CLIP-S ↑ | PCD | mBLEU-4 ↓ | Div-2 ↑ |
| *Random* | - | - | - | 17.77 | 0.963 | 0.001 | 0.868 | - | 17.54 | 0.962 | 0.003 | 0.860 |
| BLIP (−NS) | ✗ | 446M | 56.00 | 29.98 | 0.600 | 1.000 | 0.179 | 55.78 | 28.58 | 0.600 | 1.000 | 0.179 |
| BLIP (+NS) | ✗ | 446M | 57.23 | 30.33 | 0.668 | 0.658 | 0.387 | 46.99 | 29.56 | 0.690 | 0.664 | 0.384 |
| Seq-CVAE | ✗ | - | - | - | - | 0.640 | 0.480 | - | - | - | - | - |
| ModeCap | ✗ | - | - | 29.35 | 0.714 | 0.281 | 0.594 | - | - | - | - | - |
| BLIP-2 | ✓ | 3.9B | 65.47 | 30.66 | 0.651 | 0.712 | 0.345 | 57.81 | 30.37 | 0.667 | 0.732 | 0.336 |
| Honeybee | ✓ | 7B | 53.55 | 28.21 | 0.792 | 0.062 | 0.716 | 47.41 | 27.65 | 0.827 | 0.057 | 0.732 |
| Honeybee | ✓ | 13B | 55.11 | 27.41 | - | 0.014 | 0.872 | 50.41 | 27.27 | - | 0.013 | 0.875 |
| LLaVA-1.5 | ✓ | 13B | 59.61 | 30.08 | - | 0.180 | 0.658 | 54.74 | 29.54 | - | 0.176 | 0.680 |
| LLaVA-1.6 | ✓ | 7B | 55.99 | 29.36 | - | 0.046 | 0.787 | 51.00 | 27.46 | - | 0.028 | 0.809 |
| *Gold* | - | - | - | 30.33 | 0.753 | 0.043 | 0.748 | - | 30.87 | 0.776 | 0.049 | 0.760 |
| HBoP (*ours*) | ✗ | 1B | 56.30 | 29.12 | 0.772 | 0.049 | 0.735 | 54.00 | 28.46 | 0.815 | 0.042 | 0.750 |
| HBoP *Ranking* | | | 4/8 | 8/11 | 1/7 | 5/12 | 5/12 | 4/8 | 6/10 | 1/6 | 4/10 | 5/10 |

Table 1: Relevancy and diversity scores across different model captions. HBoP achieves higher diversity scores with higher Div-2 and PCD values, and a lower mBLEU-4 score, on both the MSCOCO and Flickr30K datasets compared to smaller VLMs or models trained to enhance diversity. It also maintains comparable relevance scores (SBERT and CLIP-S). Simultaneously, HBoP scores competitive results compared to much larger VLMs with LLM encoders. The cells are colored based on comparison to HBoP, with blue indicating lower values and red indicating higher values.

| | PnP-VQA | BLIP | BLIP-2 | Gold | HBoP |
|---|---|---|---|---|---|
| LLama-2-13B | 7.70 (±0.09) | 9.36 (±0.05) | 9.69 (±0.05) | 9.17 (±0.06) | 8.56 (±0.07) |
| GPT-4 | 2.18 (±0.84) | 2.97 (±0.10) | 2.96 (±0.19) | 2.94 (±0.49) | 2.48 (±0.73) |

Table 2: Semantic Integrity scores exhibit a similar trend across two LLM evaluations for the Flickr30K dataset (1k test set).

gold captions, along with the lowest mBLEU-4 and highest Div-2 scores among BLIP (both with and without nucleus sampling (NS) (Li et al., 2022a)), Seq-CVAE (Aneja et al., 2019b), and ModeCAP[1] (Chen et al., 2022). Notably, in some cases, HBoP also surpasses larger VLMs with LLM-based encoders, such as BLIP-2 (Li et al., 2023), Honeybee-7B (Cha et al., 2023), and LLaVA-1.5 (Liu et al., 2023a), in diversity metrics, as indicated by lower mBLEU-4 and higher Div-2 scores (*notice* red mBLEU-4 and blue Div-2, marking stronger diversity performance for HBoP).

At the same time, HBoP maintains strong similarity between generated captions and reference texts, as measured by SBERT (Reimers and Gurevych, 2019), and preserves high image-text alignment, as indicated by CLIP-Score (Hessel et al., 2021). Notably, HBoP achieves SBERT and CLIP-Score values that are comparable to BLIP, BLIP-NS, and LLaVA, while outperforming HoneyBee in both metrics. Although BLIP-2 scores the highest, HBoP remains competitive, demonstrating a strong balance between relevance and diversity.

**HBoP generates semantically meaningful captions.** We evaluate the semantic integrity of HBoP captions against those generated by other models, using LLama-2-13b (Touvron et al., 2023) and GPT-4 (Fu et al., 2023). As shown in Table 2, HBoP achieves semantic integrity scores close to the gold captions. Notably, HBoP outperforms models like PnP-VQA in this metric. We attribute this surpass to our method, which samples more meaningful image embeddings using the proposed HCM component.

## 5 Conclusion

We propose HBoP, a hierarchical caption generation framework that leverages a modular architecture combining lightweight pre-trained VLMs and segmentation models to generate semantically meaningful yet diverse captions. Our experimental results demonstrate HBoP's ability to produce meaningful image embeddings for captioning, achieving performance comparable to larger VLMs and human-generated captions. HBoP sets a solid baseline for future work aiming to extract more relevant knowledge by controlling the intermediate image embeddings.

[1]The dataset annotations and features necessary to train ModeCap are exclusively available for the MSCOCO dataset, making it difficult to replicate the experiments for fair comparison on the NoCaps and Flickr30k datasets.

## 6 Limitations

The HBoP architecture uses bounding box information to extract image embeddings, where the bounding boxes may contain additional information beyond the segmented object. The exploration of caption generation solely based on irregular-shaped segmentation masks is deferred as future work.

## 7 Ethical Statement

Captions generated with HBoP might inadvertently contain harmful content. However, the final caption outputs mainly depend on the image content and pretrained image captioning model. Therefore, unless the images themselves are harmful or the pretrained model produces unsafe captions, HBoP captions are expected to pose minimal risk.

## References


Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2019. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8948–8957.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.

Jyoti Aneja, Harsh Agrawal, Dhruv Batra, and Alexander Schwing. 2019a. Sequential latent spaces for modeling the intention during diverse image captioning. *Preprint*, arXiv:1908.08529.

Jyoti Aneja, Harsh Agrawal, Dhruv Batra, and Alexander Schwing. 2019b. Sequential latent spaces for modeling the intention during diverse image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4261–4270.

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.

Simone Bianco, Luigi Celona, Marco Donzella, and Paolo Napoletano. 2023. Improving image captioning descriptiveness by ranking and llm-based fusion. *Preprint*, arXiv:2306.11593.

Junbum Cha, Wooyoung Kang, Jonghwan Mun, and Byungseok Roh. 2023. Honeybee: Locality-enhanced projector for multimodal llm. *arXiv preprint arXiv:2312.06742*.

Qi Chen, Chaorui Deng, and Qi Wu. 2022. Learning distinct and representative modes for image captioning. *Advances in Neural Information Processing Systems*, 35:9472–9485.

Qi Chen, Chaorui Deng, and Qi Wu. 2023. Learning distinct and representative styles for image captioning. *Preprint*, arXiv:2209.08231.

Cheng-Han Chiang and Hung yi Lee. 2023. Can large language models be an alternative to human evaluations? *Preprint*, arXiv:2305.01937.

Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10578–10587.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. *Preprint*, arXiv:2010.11929.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Zhiyuan Fang, Jianfeng Wang, Xiaowei Hu, Lin Liang, Zhe Gan, Lijuan Wang, Yezhou Yang, and Zicheng Liu. 2022. Injecting semantic concepts into end-to-end image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18009–18019.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *Preprint*, arXiv:2302.04166.

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.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Jan Hosang, Rodrigo Benenson, and Bernt Schiele. 2017. Learning non-maximum suppression. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4507–4515.

Zhong Ji, Kexin Chen, and Haoran Wang. 2021. Stepwise hierarchical alignment network for image-text matching. In *IJCAI*.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. *Preprint*, arXiv:1412.2306.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023. Segment anything. *arXiv:2304.02643*.

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

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022a. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022b. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *Preprint*, arXiv:2201.12086.

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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Shilong Liu, Hao Cheng, Haotian Liu, Hao Zhang, Feng Li, Tianhe Ren, Xueyan Zou, Jianwei Yang, Hang Su, Jun Zhu, et al. 2023a. Llava-plus: Learning to use tools for creating multimodal agents. *arXiv preprint arXiv:2311.05437*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: Nlg evaluation using gpt-4 with better human alignment. *Preprint*, arXiv:2303.16634.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002a. 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.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021a. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021b. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626.

Bin Shao, Jianzhuang Liu, Renjing Pei, Songcen Xu, Peng Dai, Juwei Lu, Weimian Li, and Youliang Yan. 2023. Hivlp: Hierarchical interactive video-language pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13756–13766.

Mustafa Shukor, Guillaume Couairon, and Matthieu Cord. 2022. Efficient vision-language pretraining with visual concepts and hierarchical alignment. In *33rd British Machine Vision Conference (BMVC)*.

Anthony Meng Huat Tiong, Junnan Li, Boyang Li, Silvio Savarese, and Steven CH Hoi. 2022. Plug-and-play vqa: Zero-shot vqa by conjoining large pre-trained models with zero training. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 951–967.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. 2022a. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *Preprint*, arXiv:2208.10442.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. 2022b. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*.

Guanghui Xu, Shuaicheng Niu, Mingkui Tan, Yucheng Luo, Qing Du, and Qi Wu. 2021. Towards accurate text-based image captioning with content diversity exploration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12637–12646.

Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. 2019. Hierarchy parsing for image captioning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2621–2629.

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.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 2022. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*.

Xuying Zhang, Xiaoshuai Sun, Yunpeng Luo, Jiayi Ji, Yiyi Zhou, Yongjian Wu, Feiyue Huang, and Rongrong Ji. 2021. Rstnet: Captioning with adaptive attention on visual and non-visual words. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15465–15474.

# A Appendix

# B Additional Related Works

## B.1 Vision-Language Models (VLMs)

A growth of interest in VLMs has continued due to the wide availability of multi-modal data on the web. Foundation VLMs can be applied to a range of tasks in a zero-shot manner. Notably, CLIP (Radford et al., 2021b) jointly pre-trains an image encoder and a text encoder by maximizing and minimizing the cosine similarity of correct and incorrect image-text pair embeddings respectively with image-text contrastive (ITC) loss. In contrast, BLIP (Li et al., 2022a) uses both ITC and image-text matching (ITM) loss for enhanced image-text data representation. Additionally, the BLIP (Li et al., 2022a) *captioner* uses language modeling (LM) loss for autoregressive image caption generation along with a filter, *capfilt* to improve the quality of image-text pairs for training.

Flamingo (Alayrac et al., 2022) shows remarkable zero-shot ability in image captioning, visual question-answering (VQA), and image-text retrieval (ITR) tasks by leveraging the few-shot learning ability of pre-trained vision-only and language-only models. It simply interleaves input visual data with task-specific text examples, producing free-form texts for unseen visual data. Another general-purpose model, BEIT3 (Wang et al., 2022b) with Multiway Transformer structure, uses different types of modality experts to perform fusion and modality-specific training. A masked modeling objective on images only and image-text pairs is performed for computer vision tasks (*e.g.*, image classification, semantic segmentation, object detection) and vision-language tasks (*e.g.*, VQA), respectively. Whereas the VQA task uses a fused encoder for image-text pairs, the ITR task encodes images and texts independently with ITC loss. Lastly, sequence-to-sequence learning is applied to generate texts from images for the image captioning task. Inspired by these previous works, we propose a meta-VLM model that utilizes a pre-trained BLIP (Li et al., 2022a) image captioning module to generate enhanced textual representations, which can later serve as useful data for various downstream tasks.

## B.2 Hierarchical Representation

Identifying and extracting regions of interest within images is crucial for a hierarchical representation. The most intuitive way to achieve this would typically involve the use of object detectors (Yao et al., 2019; Cornia et al., 2020; Zhang et al., 2021). However, the heavy computational demands of the object detectors inevitably lead to inefficiency during the inference stage (Yao et al., 2019; Cornia et al., 2020; Zhang et al., 2021). In response, recent works sought to replace these cumbersome detectors by adopting visual concepts in the form of object tags (Fang et al., 2022; Shukor et al., 2022) as an alternative. However, this detector-free approach is contingent upon the availability of object-specific data within the dataset. Employing pre-trained models is a more efficient way to identify areas of interest within images. GradCAM (Selvaraju et al., 2017) highlights essential regions that the pre-trained models used to predict any target concept using its gradients with respect to feature map activations of the final convolutional layer. DINOv2 (Oquab et al., 2023) capitalizes on existing self-supervised pre-trained models to generate robust, all-purpose visual features, supporting a wide array of tasks ranging from image-level classification to pixel-level segmentation. However, the image regions/features delineated by GradCAM/DINOv2 tend to show saliency for specific tasks and are unable to capture the full spectrum of visual representations. Conversely, SAM (Kirillov et al., 2023) intricately segments every semantically significant component of an image into high-quality segmentation masks generated by prompting with various inputs such as point, box, mask, or free-form text, unrestricted with types of tasks. In our framework, we integrate SAM (Kirillov et al., 2023) to create semantically meaningful segmentation masks for an entire image automatically.

Several prior studies have incorporated the principles of hierarchy or multi-scale representation into their model architectures, aiming to enhance the alignment between images and texts (Ji et al., 2021; Shao et al., 2023; Shukor et al., 2022). SHAN (Ji et al., 2021) deconstructs the image-text matching process into two distinct facets: fragment-level and context-level alignments enabling matches across three different scopes: local-to-local, global-to-local, and global-to-global. HiVLP (Shao et al., 2023) leverages both low-

and high-dimensional features to represent coarse and fine details. ViCHA (Shukor et al., 2022) aligns images and texts across various layers of neural network encoders with the underlying assumption that each layer reflects varying semantic levels. Unlike these approaches, we divide the segmentation masks hierarchically and use the embeddings of the extracted individual image patches for caption generation.

### B.3 Caption Evaluation

Common image captioning evaluation metrics, including BLEU (Papineni et al., 2002a), METEOR (Banerjee and Lavie, 2005), ROUGE (Lin, 2004), and CIDEr (Vedantam et al., 2015) scores are primarily n-gram approaches that assess the quality of generated captions by considering their overlap with human-generated captions. Most SOTA VLMs frequently exhibit promising scores across these conventional evaluation metrics. However, these metrics are limited in their capabilities to measure the diversity of the generated captions. This limitation leads to a bias in these models towards generating an "average" and "safe" caption reflecting the most basic information in the image, rendering them less informative than human-generated captions. To address this gap, we incorporate several diversity metrics, including mBLEU-4, Div-2 (Aneja et al., 2019b), and the proposed pairwise cosine distance (PCD), along with semantic integrity and relevance scores to ensure that the captions generated by our framework are not only diverse but also meaningful and directly relevant to the given image and human-annotated captions.

## C Experiments

### C.1 Implementation Details

The ISM (Section 3.1) employs the fully automated SAM with no prompting (Kirillov et al., 2023), along with the image encoder initialized from ViT (`ViT-L/16`) pre-trained on ImageNet (Dosovitskiy et al., 2021), following the same settings as BLIP (Li et al., 2022a). Note that we use BLIP (Li et al., 2022a) for captioning instead of BLIP-2 (Li et al., 2023) since BLIP-2 uses intermediate representations trained on pairs of entire images and texts for caption generation using an LLM, which is not directly applicable to HBoP that uses pairs of image patches and texts. The HCM (Section 3.2) creates the global level by selecting the top ($k$ =) 5 masks with the largest areas and designating the remaining masks as fine-grained. To create the regional level, K-means clustering, with ($K$ =) 5 clusters per image, is applied to the bounding boxes of the segmentation masks. NMS with a threshold of 0.1 is applied at all three levels. Lastly, the ICM (Section 3.3) follows the methodology outlined in Tiong et al., 2022.

Although HBoP presents a three-tier hierarchical structure, it is crucial to note that we adjust the different hierarchy levels depending on a given dataset. A dataset with information-rich complex images would require using all three hierarchy levels. However, a dataset with relatively simpler images, such as the MSCOCO dataset (Lin et al., 2014), would benefit from a two-tier hierarchy with just the global and regional captions. We use the first two levels during evaluations unless specified otherwise.

All the model captions in Tables 1 and 3 are regenerated, except for Seq-CVAE (Aneja et al., 2019b), where the results are taken directly from the original paper. While HBoP benefits from bounding box information, it is important to note that other baseline methods (*e.g.*, ModeCap) have the additional advantage of explicit learning objectives to improve diversity. The exact prompts we use for Honeybee (Cha et al., 2023) (top) and LLaVA-1.5/1.6 (Liu et al., 2023a) are in Table 4.

### C.2 Evaluation

We evaluate the model captions using three distinct metrics: 1) diversity across captions per image, 2) relevancy with images, and 3) semantic coherence and meaningfulness. The datasets we use for evaluation are: the Karpathy test split (Karpathy and Fei-Fei, 2015) of MSCOCO (5k images) (Lin et al., 2014), Flickr30K zero-shot (1k test images) (Young et al., 2014), and NoCaps validation (4.5k images) (Agrawal et al., 2019).

#### C.2.1 Diversity

We measure the diversity in the generated captions using the cosine similarity between the sentence embeddings of all the corresponding captions per image. The comparison baselines are random captions,

where each caption corresponds to different images, BLIP (Li et al., 2022a) with and without nucleus sampling (NS[2]) (Holtzman et al., 2019), BLIP-2 (Li et al., 2023), ModeCap (Chen et al., 2022), Honeybee (Cha et al., 2023), and gold captions[3]. The diversity of the generated captions $(s_1, s_2, ...s_n)$ per dataset instance[4] is measured using pairwise cosine distance (PCD):

$$\text{PCD}(s_1, s_2, ...s_n) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{j<i}(1 - cos(M(s_i), M(s_j))) \quad (1)$$

In the above equation, $cos$ represents the cosine similarity of the input embeddings. We use sentence embeddings from a pre-trained sentence transformer model (`all-MiniLM-L6-v2`) (Reimers and Gurevych, 2019), denoted as $M$ in the Eq. 1 that can capture the semantic relationships between captions. This measure evaluates the extent to which the generated captions differ from each other per image. We report the final diversity score for each dataset as the averaged PCD scores of all images in the dataset. Ideally, the PCD score should be lower than that of random captions that serve as the upper bound of the diversity score, but it should be higher than that for captions generated by existing baselines.

Additionally, we use mBLEU-4 and n-gram diversity (*e.g.*, Div-1, Div-2) (Aneja et al., 2019b), to compare with more challenging baseline models, such as ModeCap (Chen et al., 2022) and Seq-CVAE (Aneja et al., 2019b) that are built to achieve diversity within captions per image. For ModeCap (Chen et al., 2022), we follow the default settings from the original paper to reproduce the results based on training the Transformer-DML model. We also prompt a recently introduced multimodal LLM called Honeybee (Cha et al., 2023) as follows: "Describe this image with 5 diverse captions, using less than 20 words for each caption."

#### C.2.2 Relevancy

While confirming that each dataset contains captions with high semantic integrity is crucial, the captions must also be relevant to the corresponding images. We employ CLIP-Score (Hessel et al., 2021) that calculates the correlation between visual and textual CLIP embeddings (Radford et al., 2021b) using pre-trained ViT (`openai/ clip-vit-base-patch32`) without relying on human-generated references. Similar to the comparison baseline datasets for semantic integrity evaluation, we compare HBoP with PnP-VQA (Tiong et al., 2022), BLIP (Li et al., 2022a), BLIP-2 (Li et al., 2023), gold captions, along with random captions. We generate random captions by selecting five random captions for each image from a pool of HBoP captions corresponding to different images. In other words, although the random caption itself should make sense, they depict mismatched images. We randomly select one out of a total of five captions per image for each dataset and compute the correlation between CLIPScores of generated captions and gold captions.

Additionally, we measure the semantic similarity between ground-truth (or *gold*) captions and captions generated with models using transformer-based SBERT (Reimers and Gurevych, 2019). Note that this metric is robust to synonyms or paraphrasing, unlike n-gram metrics (Papineni et al., 2002b; Lin, 2004).

#### C.2.3 Semantic Integrity

We measure the semantic integrity of the generated captions (Table 2) to ensure meaningfulness and cohesiveness using LLM evaluation as it has empirically shown high correlation with human judgment (Chiang and yi Lee, 2023; Liu et al., 2023b; Fu et al., 2023). We prompt Llama-2-13B (`Llama-2-13b-chat-hf`) (Touvron et al., 2023) to access the semantic integrity of HBoP captions along with gold and other baselines (PnP-VQA (Tiong et al., 2022), BLIP (Li et al., 2022a), BLIP-2 (Li et al., 2023)) captions. Specifically, we randomly select two captions out of a total of five captions per image for each dataset and evaluate the semantic integrity by averaging the coherency and meaningfulness scores for each caption using the prompt shown in Table 5. We use the prompt "This is a picture of" to generate captions for all models in our experiments. This deliberate choice ensures a fair comparison of the general caption

[2]Unless otherwise specified, all the BLIP models in this paper refer to BLIP with NS.

[3]We exclude PnP-VQA since the captions are generated per question instead of per image, unlike other baselines.

[4]Note that $n = 5$ for all dataset instances, and we use one global caption and five regional captions for HBoP.

| | # Param | PCD | mBLEU-4 ↓ | Div-2 ↑ |
|---|---|---|---|---|
| *Random* | - | 0.962 (+0.223) | 0.001 | 0.867 |
| BLIP (−NS) | 446M | 0.600 (-0.129) | 1.000 | 0.178 |
| BLIP-2 | 3.9B | 0.654 (-0.075) | 0.715 | 0.340 |
| BLIP (+NS) | 446M | 0.679 (-0.050) | 0.629 | 0.400 |
| Honeybee | 7B | 0.791 (+0.062) | 0.080 | 0.705 |
| *Gold* | - | 0.729 | 0.078 | 0.666 |
| HBoP (*ours*) | 1B | 0.783 (+0.054) | 0.041 | 0.748 |
| HBoP *Ranking* | | 2/6 | 2/7 | 2/7 |

Table 3: Diversity scores for Nocaps test set. We observe a similar diversity trend across model captions as Table 1.

generation ability across models, as altering the prompt can yield significantly different results, making fair evaluation challenging.

Similarly, we use GPT-4 (Fu et al., 2023) for additional Semantic Integrity evaluation using only a single caption per image with the prompt shown in Table 6. Note that we sample the first 1k image instances in each dataset for this evaluation due to the cost limitations.

# D Additional Results

## D.1 Relevancy

In Figure 3, HBoP captions (y-axis values in the last column) show comparable relevance scores with gold captions (x-axis values in the last column) with the slope of a linear regression line[5] being close to 0.5. Although the slopes of these regression lines (MSCOCO (Lin et al., 2014): 0.42, Flickr30k (Young et al., 2014): 0.39, Nocaps (Agrawal et al., 2019): 0.34) are less than those of BLIP (Li et al., 2022a) (0.49, 0.44, and 0.45) and BLIP-2 (Li et al., 2023) (0.51, 0.45, 0.43), we observe a trend of having relevance scores in the range of 20 to 40 for both x and y axes values. On the other hand, relevance scores for random and PnP-VQA (Tiong et al., 2022) captions have a spurious and less-correlated relation with those of gold captions.

## D.2 GradCAM Results

In addition to the evaluation results of the generated captions (samples in Figure 4), we illustrate how the generated captions correlate with specific image regions through GradCAMs (Selvaraju et al., 2017). The visual representation identifies the image regions on which the generated captions are based. Specifically, we aggregate the gradients from all cross-attention layers of the pre-trained ITM model in PnP-VQA (Tiong et al., 2022). Whereas PnP-VQA (Tiong et al., 2022) feeds the question for the textual input, we input BLIP (Li et al., 2022a) and gold captions, along with HBoP captions. As shown in Figures 1 and 5, the highlighted regions in the image for HBoP captions closely resemble the same pattern as those observed using human-generated captions. On the contrary, BLIP exhibits a more constrained range, predominantly concentrating on specific image regions.

## D.3 Fine-grained Captions

Although not evaluated in the perspectives of three main evaluation metrics, we can also create what we refer to as fine-grained captions that can serve as image tags using our proposed methodology. These serve as supplementary information, enhancing the depth of understanding of the image. They are more vital when dealing with complex images containing various small or intricate objects, which conventional caption generation processes may often overlook. By introducing the additional layer of granularity, our approach ensures a more detailed and inclusive interpretation of the image.

[5]The p-values for all the regression lines are less than 0.001, except for the those of lines in the first columns, which are not statistically significant

The following is a conversation between a curious human and AI assistant. The assistant gives helpful, detailed, and polite answers to the user's questions.
Human: <image>
Human: Describe this image with 5 captions with numberings.
AI:

A chat between a curious human and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the human's questions.
Human: <im_start><image><im_end>
Human: Describe this image with 5 captions.###Assistant:

---

[INST] <image> What is shown in this image? Describe this image with 5 captions. [/INST]

Table 4: Image caption generation prompts for Honeybee (top) and LLaVA-1.5/1.6 (bottom).

[INST] <<SYS>>
You will be given a caption generated from an image. Given criteria and rating options, rate the response. Respond with a number only.
Evaluation Criteria: **[CRITERION]**: **[DEFINITION]**
Scale: from 1 to 10
Answer: <</SYS>>
INPUT [/INST]

---

**[CRITERION]**: Coherence/Meaningfulness
**[DEFINITION]**: the logical and clear connection between ideas or elements within a context. It is characterized by the consistency, integrity, and clarity of information or arguments presented./the relevance and significance of the content in the caption. A meaningful caption goes beyond a literal description, providing insight, context, or emotion that enhances the viewer's understanding or appreciation of the image.

Table 5: The prompt for evaluating semantic integrity (coherence + meaningfulness) of generated model captions using Llama-2-13B.

You will be given one caption written for describing an image.

Your task is to rate the caption on one metric.

Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.

Evaluation Criteria:
Fluency (1-3): the quality of the caption in terms of grammar, spelling, punctuation, word choice, and sentence structure.
- 1: Poor. The caption has many errors that make it hard to understand or sound unnatural.
- 2: Fair. The caption has some errors that affect the clarity or smoothness of the text, but the main points are still comprehensible.
- 3: Good. The caption has few or no errors and is easy to read and follow.

Example:
Caption:
Caption

Evaluation Form (scores ONLY):
- Fluency (1-3):

Table 6: The prompt for evaluating semantic integrity (*i.e.*, fluency) of generated model captions using GPT-4.

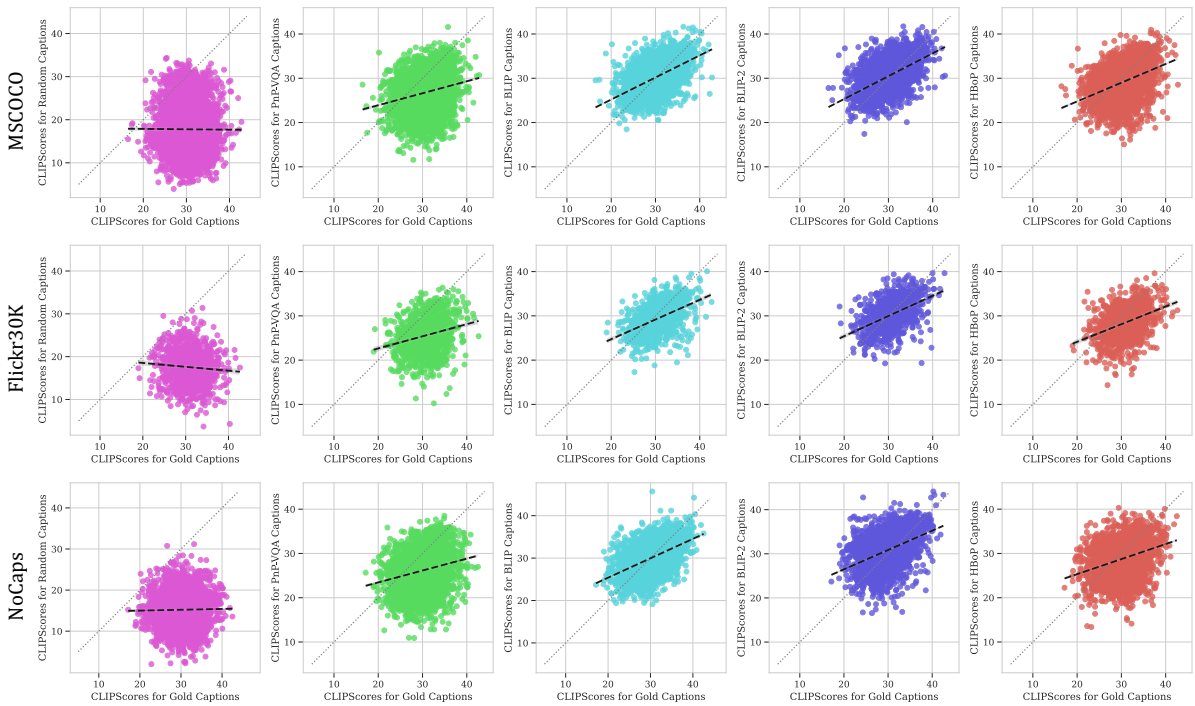


Figure 3: Correlation of relevance scores between gold captions and model captions. We observe higher correlations for HBoP, BLIP, and BLIP-2 captions as comapred to random and PnP-VQA captions.

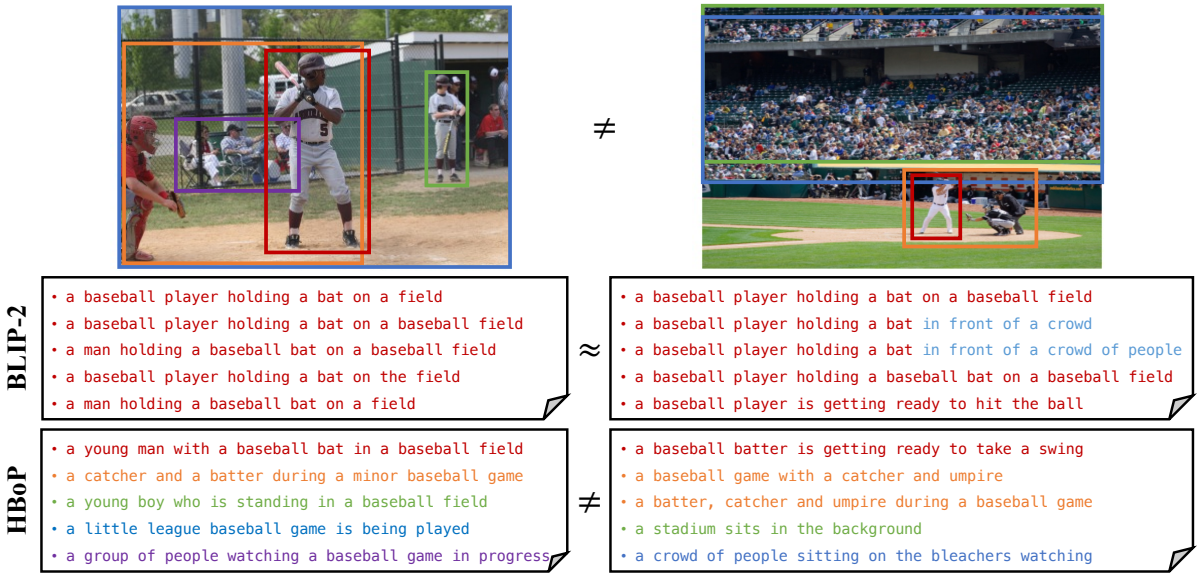


Figure 4: Comparison between captions generated using BLIP-2 (Li et al., 2023) and HBoP. Our captions contain more diverse interpretations of the images while maintaining high relevancy.

Figure 5: Additional visualizations of GradCAMs across different model captions.