No Detail Left Behind: Revisiting Self-Retrieval for Fine-Grained Image Captioning

Anonymous authors
Paper under double-blind review

Abstract

Image captioning systems are unable to generate fine-grained captions as they are trained on data that is either noisy (alt-text) or generic (human annotations). This is further exacerbated by maximum likelihood training that encourages generation of frequently occurring phrases. Previous works have tried to address this limitation by fine-tuning captioners with a self-retrieval (SR) reward. However, we find that SR fine-tuning has a tendency to reduce caption faithfulness and even hallucinate. In this work, we circumvent this bottleneck by improving the MLE initialization of the captioning system and designing a curriculum for the SR fine-tuning process. To this extent, we present (1) Visual Caption Boosting, a novel framework to instill fine-grainedness in generic image captioning datasets while remaining anchored in human annotations; and (2) BagCurri, a carefully designed training curriculum that more optimally leverages the contrastive nature of the self-retrieval reward. Jointly, they enable the captioner to describe fine-grained aspects in the image while preserving faithfulness to ground-truth captions. Our approach outperforms previous work by +8.9% on SR against 99 random distractors (RD100) (Dessì et al., 2023); and +7.6% on ImageCoDe.

Additionally, existing metrics to evaluate captioning systems fail to reward diversity or evaluate a model's fine-grained understanding ability. Our third contribution addresses this by proposing self-retrieval from the lens of evaluation. We introduce TrueMatch, a benchmark comprising bags of highly similar images that uses SR to assess the captioner's ability to capture subtle visual distinctions. We evaluate and compare several state-of-the-art open-source MLLMs on TrueMatch, and find that our SR approach outperforms them all by a significant margin (*e.g.* +4.8% - 7.1% over Cambrian) while having 1-2 orders of magnitude fewer parameters. We also outperform vanilla SR by +14.4% to +19.5%.

1 Introduction

Image captioning, or generating natural language image descriptions, has witnessed remarkable progress over the last decade. Today's captioning systems are composed of sophisticated deep learning architectures (Mokady et al., 2021; Stefanini et al., 2022; Dai et al., 2024; Liu et al., 2023) trained on vast datasets (Lin et al., 2014; Sharma et al., 2018; Thomee et al., 2016; Desai et al., 2021; Schuhmann et al., 2022). However, even with these advances, approaches often generate generic captions that are unable to differentiate similar images (see Figure 1), violating the fundamental purpose of a caption: *to facilitate accurate and efficient communication of visual content* (Fisch et al., 2020; Kreiss et al., 2022; Dessì et al., 2023). We attribute the shortcomings of image captioning systems to three key factors: (i) the nature of their training data, (ii) captioning evaluation metrics, and (iii) the maximum likelihood estimation (MLE) training approach.

I. Challenges with the training data. Training datasets may be divided into two: curated datasets (COCO (Lin et al., 2014), Flickr30k (Plummer et al., 2015)) or large-scale alt-text data (*e.g.* CC3M (Sharma et al., 2018)). While alt-text is noisy and may be unaligned with the image, human-annotated captions (COCO) may be generic (Kornblith et al., 2023) and lack world knowledge (Bavishi et al., 2023) as annotators describe visual concepts in a simplistic manner (*e.g.* labeling a *Golden Retriever* as a dog). Recently, foundation models are being used to enhance large-scale alt-text data by making them denser (Doveh et al.,

COCO MLE	A man flying through the air while riding a bike.	A man flying through the air while riding a bike.	A man flying through the air while riding a bike.
COCO SR (CVPR'23)	A man flying through the air while riding a bike.	Two people jumping in the air on their dirt bikes.	A man flying through the air while riding a bike.
OURS SR	A man wearing a helmet and gloves performs a stunt on a blue dirt bike, while riding on a cloudy day.	A motorcyclist wearing a helmet and black attire performs tricks on his bike, jumping high in the air.	A person wearing a helmet performs an aerial stunt on a red and white dirt bike, flying through cloudy sky.

Figure 1: For a set of similar images, captioning systems struggle to generate meaningful captions that uniquely describe each image. In this example, **COCO MLE**: A model trained on COCO with MLE generates the same generic description for all images. **COCO SR** (Dessì et al., 2023): While the SR objective may help, the COCO captions are not rich enough and lead to hallucinations of incorrect attributes such as "two people" (middle). **OURS SR**: Our improved data and training recipe result in captions that are able to correctly distinguish 2 of 3 images (middle and right). Even though correct fine-grained details are captured for the left image, they may be insufficient, indicating the challenging nature of SR on our benchmark.

2024; Urbanek et al., 2024). However, such methods are prone to inherit biases (*e.g.* gender, geography) present in foundation models (Hall et al., 2023; Sirotkin et al., 2022; Basu et al., 2023; Salman et al.) and exhibit verbose language modeling priors (Liu et al., 2024).

To address these data inadequacies, we propose *Visual Caption Boosting* (VCB), a model agnostic framework designed to generate dense captions that holistically capture different aspects of the image (objects, attributes, relations, scene, *etc.*) while remaining anchored in human annotations (see Section 2). In brief, multiple human annotated captions are blended together using a Large Language Model (LLM) and expanded with an image description generated from a Multimodal Large Language Model (MLLM). To prevent conflicting visual details, we prompt the LLM to prefer the blended caption over the description from the MLLM. Thus, VCB creates fine-grained captions that are grounded in human annotations, enabling rich and informative datasets to train image captioning systems.

II. Image caption evaluation metrics can be broadly classified into reference-based and reference-free. Reference-based metrics such as BLEU (Papineni et al., 2002), CIDEr (Vedantam et al., 2015), and SPICE (Anderson et al., 2016) rely on comparisons to ground-truth (GT) captions that may not include image details beyond salient objects. These metrics tend to penalize captions that are more specific than the ground-truth, often favoring generic descriptions (Wang et al., 2020). Reference-free metrics like CLIPScore (Hessel et al., 2021) alleviate this issue by directly measuring image-text similarity, but may fail to encourage diverse and discriminant captions. These limitations necessitate the development of new evaluation strategies that incentivize models to produce fine-grained and distinct captions.

In this work, we ask what makes a "good" description when evaluating captioning systems? We posit that a caption should enable a listener to pick out a target image from a bag of images with similar visual elements (Liu et al., 2018). To this extent, we propose to evaluate captioning systems through the lens of self-retrieval (SR) *i.e.*, their ability to retrieve an image using its generated caption against of bag of highly similar distractor images. Traditionally, improvements in text-to-image (T2I) retrieval have focused on either

enhancing the embedding spaces of scoring functions (Faghri et al., 2017) or recently, making denser, more detailed text queries (Lai et al., 2023; Li et al., 2024; Singla et al., 2024; Shvetsova et al., 2023). In this work, we introduce a novel perspective: by curating image sets of highly similar images, we can use T2I retrieval to assess whether the query (caption) captures fine-grained visual concepts necessary to retrieve the correct image. Specifically, we construct *bags* of highly similar images, where uniquely describing each image within a bag requires describing various facets of visual understanding. For example, retrieving the target image from Figure 1 requires the caption to incorporate information about attributes (*red* bike) or orientation (*inverted* body). Hence, we evaluate whether the captioning system captures subtle visual distinctions by retrieving the target image using its generated caption *without looking at the other candidate images in the bag*.

While previous works have incorporated SR as a reward signal during training (Dessì et al., 2023), to our knowledge, we are the first to curate bags of similar images and leverage SR to evaluate fine-grained understanding exhibited by captioning systems. To this end, we introduce *TrueMatch*, a benchmark of image sets with varying size. TrueMatch offers a comprehensive evaluation framework to assess the ability of captioning systems to capture various aspects of visual discrimination such as positioning, action, orientation.

III. Guiding captioners away from their language modeling priors. MLE training incentivizes captioning systems to overuse common concepts and statistically probable phrases when describing visually similar images. While previous works have optimized for self-retrieval (SR) reward to learn discriminant captioners (Luo et al., 2018; Liu et al., 2018; Rennie et al., 2017; Dessì et al., 2023), the models are first trained via MLE on generic caption datasets (*e.g.* COCO). This is suboptimal and we show that SR fine-tuning is sensitive to initialization: it is necessary to start with a captioning system that captures fine-grained details in order to better preserve faithfulness to GT captions. In fact, we discover a propensity of captioning systems to hallucinate (Section 4.4) when trained via SR on generic captions (Dessì et al., 2023). Thus, current SR approaches face two distinct challenges: a trade-off between retrieval performance and caption faithfulness, and sub-optimal fine-grained retrieval performance.

To enhance SR's ability to instill fine-grained visual information, we (i) fine-tune both components (language and visual) of the captioning system (Section 4.5), and (ii) mine multiple hard negatives (visually similar sets of images) to create training bags. We also introduce a curriculum learning approach that progressively increases the bag size during training to leverage the contrastive nature of our retrieval-based reward (Section 4.6). Our carefully designed curriculum enables SR to leverage the rich initialization provided by *VCB*, improving retrieval performance as well as the faithfulness of the generated captions. Through our training procedure, we are able to circumvent the aforementioned trade-off and achieve substantial performance gains, surpassing (Dessì et al., 2023) by 20% on TrueMatch without making any modifications to the model architecture or reward. Our results demonstrate that a well-crafted training paradigm achieves performance comparable to MLLMs that are orders of magnitude larger. Figure 1 shows images of a bag from TrueMatch and captions generated by the same model architecture trained with different datasets (COCO *vs.* Ours) and paradigms (MLE *vs.* vanilla SR *vs.* Ours SR).

Key contributions. We address three challenges plaguing current image captioning systems: Data, Evaluation, and the MLE training. (i) We begin by identifying the inadequacies of image captioning datasets in Section 2, and propose *Visual Caption Boosting*, a novel caption enhancement strategy that leverages LLMs and MLLMs to generate dense, informative, and unbiased captions that are anchored in human annotations. (ii) We then introduce TrueMatch (Section 3), a benchmark consisting of curated *bags* of highly similar images that uses self-retrieval (SR) to assess the ability of captioning systems to capture fine-grained visual distinctions. (iii) Section 4 leverages the rich data and effective evaluation methods introduced in the previous two sections to train fine-grained captioning systems with SR. We offer new insights into SR fine-tuning through extensive ablations revealing its sensitivity to MLE initialization, failure to preserve caption faithfulness, and tendency to hallucinate. (iv) Finally, we design a simple "plug and play" training recipe that enables SR to improve caption faithfulness while significantly outperforming vanilla SR (Dessì et al., 2023) on TrueMatch and achieving state-of-the-art results on ImageCoDe (Krojer et al., 2022), another benchmark for fine-grained image retrieval.

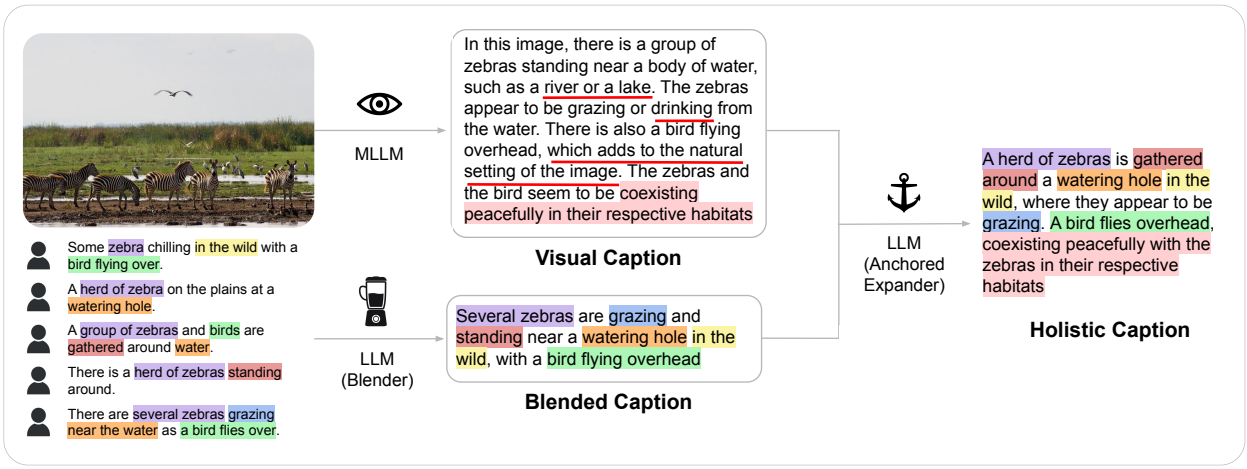


Figure 2: Example of Visual Caption Boosting transforming the original human annotated captions to a Holistic Caption. First, an LLM blends the human annotations to create a Blended Caption. Next, an MLLM generates a dense visual caption that may be noisy. Finally, we create a Holistic Caption by instructing the LLM to incorporate fine-grained details from the Visual Caption with the Blended Caption, while staying anchored in human annotations in case of conflicts. Specific prompts are shared in Appendix A.1. The colors indicate various concepts extracted from the human annotations or the visual caption. The red underlined text (illustrated by us for ease of understanding), indicating hallucinations or verbose text, is ignored in the holistic caption as we anchor the visual caption to human annotations.

2 Improving Datasets through Visual Caption Boosting

Training with dense ground-truth descriptions benefits vision-language models (Doveh et al., 2024; Urbanek et al., 2024; Lai et al., 2023). Although existing image captioning datasets like COCO (Lin et al., 2014) provide multiple annotations per image, they do not elicit detailed captions. This leads to generic annotations such as *"There is a herd of zebras standing around"* (see Figure 2) that ignore finer visual concepts and fail to describe the image holistically. This bottleneck reflects in trained captioning systems as well.

To address this, recent works leverage foundation models to synthetically expand visual information within captions (Doveh et al., 2024; Li et al., 2024; Singla et al., 2024; Lai et al., 2023), instilling a wider range of visual details. However these methods are prone to inherit biases present in foundation models (Sirotkin et al., 2022; Basu et al., 2023; Salman et al.). Moreover, the descriptions generated by some of these methods are excessively long making them susceptible to hallucinations (Favero et al., 2024). They also pose challenges to training captioning systems as they exceed the token capacity of current VLMs (*e.g.* CLIP with 77 tokens) (Urbanek et al., 2024). Nevertheless, while the individual COCO captions are sparse, we find they describe complementary facets of the image, *e.g.* *"watering hole", "bird flying over", "herd of zebras"* (Figure 2) that may provide additional details. This also corroborates findings by Ye et al. (2023) that annotators from different cultural backgrounds describe different visual concepts when viewing the same image.

Building upon these findings, we introduce *Visual Caption Boosting* (VCB), a novel two-stage approach to enrich the training data with dense, more informative captions that encourage captioning systems to learn and generate rich descriptions. VCB leverages foundation models and the diverse perspectives offered by human annotators to generate rich descriptions while being anchored in human data.

BlendCap leverages an off-the-shelf LLM to create a blended caption that combines multiple facets of visual information that the human annotators describe. Figure 2 shows an example of how a few captions can be blended together into a comprehensive description of the image using our method. Notably, we prompt the LLM to minimize redundant information resulting in short descriptions.

HolisticCap builds upon BlendCap by incorporating a fine-grained visual description produced by an MLLM[1]. Specifically, we prompt the LLM to instill the visual caption into the blended caption while preferring human-grounded BlendCap in case of conflicting visual information. As seen in Figure 2, this enables the LLM to ignore hallucinations present in Visual Caption such as *"river or lake"* and *"drinking"*, as they conflict with BlendCap's description like *"watering hole"* and *"grazing and standing"*. Additionally, the anchoring of semantic visual information in human annotations encourages the LLM to eliminate verbose tendencies of MLLMs, producing rich and succinct captions that capture fine-grained details. Although recent works (Urbanek et al., 2024; Lai et al., 2023) address similar challenges, they have some drawbacks as they anchor synthetic visual information in alt-text (Lai et al., 2023) or depend on manual annotation for incorporating dense visual information (Urbanek et al., 2024).

Anchored captions hallucinate less. We conduct a human study on 60 randomly sampled images to measure the degree of hallucinations in BlendCap and HolisticCap (see Appendix A.2 for more details). We find that 7/60 Blended Captions contain hallucinations. Of these, 5 stem from incorrect COCO human annotations and only 2 are induced by the LLM. Furthermore, anchoring the Visual Captions in human annotations reduces the hallucinations from 12 (in VisualCap) to 7 (in HolisticCap), a reduction of 41.7%.

In summary, Visual Caption Boosting is able to efficiently instill fine-grained visual information into image caption datasets while being ~~anchored in human annotations~~ less prone to hallucinations. Appendix A.1 presents the LLM and MLLM prompts and Appendix A.3 analyzes the caption lengths distributions.

3 TrueMatch: Fine-grained evaluation through Self-Retrieval

Existing self-retrieval (SR) approaches require models to select the target image from a set of N *random* distractor images in the dataset (Dessì et al., 2023). However, randomly chosen distractors often have simple differences (*e.g.* the primary object or scene), making it easy for captioning systems to distinguish between them. This evaluation is suboptimal as they neither encourage the model to generate detailed captions nor do they evaluate fine-grained abilities of captioning systems.

In this section, we present the SR setup used in our work. We propose *TrueMatch*, a benchmark of carefully curated *bags* of highly similar images that enables SR to evaluate whether captioning systems capture different facets of fine-grained visual discrimination. The results in Section 3.2, show that most captioning systems (including MLLMs) struggle to generate captions that would allow distinguishing fine-grained visual details.

Self-retrieval setup. Within a bag of images $\mathcal{B}$ (or a minibatch during training), we require the generated caption c for an image i to retrieve itself (image i) from $\mathcal{B}$. Note, $\mathcal{B}$ contains i and a set of visually similar distractor images $\mathcal{D}$. The caption c and all images in $\mathcal{B}$ are encoded using CLIP text and vision encoders and ranking is performed by computing the cosine similarity $\text{sim}(c, i)$ in the CLIP embedding space. Consequently, the caption is deemed to be good $\iff \text{sim}(c, i) > \text{sim}(c, i'), \forall i' \in D$.

3.1 Benchmark Creation

Given a dataset, the benchmark creation process involves curating bags of highly similar images. We use 10,000 images from COCO's validation and test sets (Karpathy & Fei-Fei, 2015) for our benchmark.

Creating candidate bags of images. We use the fine-grained descriptions generated through HolisticCap and encode visual and textual features in the CLIP embedding space. We treat each multimodal embedding (concatenation of the two modalities) as a query and use a simple nearest neighbour search to create bags of highly similar images. Algorithm 1 provides details on the bag creation process, especially for creating bags of variable size. Having bags of varying size facilitates SR evaluation at multiple levels of difficulty as increasing the bag size limits the number visual concepts that can be used to uniquely describe each image. This is different from previous works that typically use image pairs (Jhamtani & Berg-Kirkpatrick, 2018; Park et al., 2019; Tong et al., 2024b) or random distractors (Dessì et al., 2023) to evaluate captioners.

[1] https://huggingface.co/Salesforce/instructblip-vicuna-7b

Automated bag curation. We compute *intra-bag similarity* α, an average pairwise cosine-similarity between multimodal embeddings of all images in the bag to quantify the difficulty of uniquely describing each image in the bag (details in Algorithm 2). To create TrueMatch, we sort the list of bags created above in the descending order based on α and use a set $\mathcal{V}$ to keep track of images that are included in TrueMatch. Going down the sorted list, a bag is added to the benchmark only if none of its images exist in $\mathcal{V}$. This ensures that only the hardest bags with highest visual similarity are added to the benchmark for each visual concept in the embedding space. Bags within TrueMatch unlike previous works involving image sets, are fine-grained (Dunlap et al., 2024) and capture different aspects of visual discrimination beyond object positioning and negation (Krojer et al., 2022; Jhamtani & Berg-Kirkpatrick, 2018; Park et al., 2019).

A manual filtering process is performed to retain only those bags that would require captioning systems to capture some aspect of fine-grained visual discrimination (see example in Figure 10). For example, for bag size 3, this reduces the number of bags in TrueMatch from 680 to 254.

3.2 Experiment 1: Results on TrueMatch

We evaluate several open-source captioning approaches, MLLMs, and SR trained models on TrueMatch. Recall@1 is reported in Table 1 for bag sizes 3, 5, and 7.

Table 1: Recall@1 for several open-source captioning models, MLLMs, and SR-based methods on TrueMatch. The number of bags in #3 is 254, #5 is 104, and #7 is 93.

Method	Params	TrueMatch		
		#3	#5	#7
Random chance	-	33.3	20.0	14.3
OFA (Wang et al., 2022)	180M	50.4	36.3	37.2
ClipCap (Mokady et al., 2021)	240M	50.8	36.5	33.8
CoCa (Yu et al., 2022)	640M	52.2	40.4	38.2
PaliGemma-224 (Beyer et al., 2022)	3B	48.3	35.6	32.7
PaliGemma-448 (Beyer et al., 2022)	3B	49.3	38.5	33.3
InstructBLIP (Dai et al., 2024)	13B	53.7	42.7	42.1
LLaVA 1.6 (Liu et al., 2023)	34B	57.9	46.9	47.9
Cambrian-1 (Tong et al., 2024a)	3B	58.3	48.8	52.8
Cambrian-1 (Tong et al., 2024a)	8B	**60.6**	**53.3**	**53.1**
Discritune (Dessì et al., 2023)	240M	53.3	42.3	38.4
Ours (best)	240M	**67.7**	**58.7**	**57.9**

Approaches lack fine-grained details. Irrespective of their size, captioning approaches struggle to capture fine-grained visual details leading to poor performance on TrueMatch.

Doing more with less. Although billion-parameter MLLMs have achieved impressive results (Liu et al., 2023; Tong et al., 2024a), Table 1 demonstrates that they still struggle with fine-grained visual discrimination. In fact DiscriTune (Dessì et al., 2023), that trains ClipCap with the vanilla SR setup matches InstructBLIP despite being two orders of magnitude smaller. Cambrian-1 is the best-performing open-source model and although it surpasses DiscriTune, our proposed approach outperforms it by a significant margin. This demonstrates the effectiveness of both: TrueMatch for evaluating captioning systems, and our approach of using SR to improve captioning (Section 4).

3.3 Experiment 2: Benchmarking Visual Caption Boosting with Self-Retrieval

We evaluate the quality of captions adopted in VCB in Table 2. Along with TrueMatch, we also adopt the SR evaluation strategy of Dessì et al. (2023) with 100 random distractors (RD100). On RD100, BlendCap outperforms original COCO captions by a large margin of 8% on R@1 and 6.2% on ClipScore, confirming that human annotations capture complementary visual aspects of the same image. However, this sizeable gap shrinks to 2.5% on TrueMatch (bags of size #3). This confirms that even though annotated captions blended together work better, they inherently lack fine-grained details (see Figure 2) necessary to perform well on TrueMatch. HolisticCap, on the other hand, yields remarkable performance gains over COCO and BlendCap across both setups: RD100 and all bag sizes of TrueMatch. This demonstrates the effectiveness of VCB in instilling fine-grained information into standard image captioning datasets.

Table 2: R@1 scores for COCO and VCB captions evaluated on RD100 (100 random distractors) and TrueMatch.

Caption	RD100		TrueMatch		
	R@1	ClipSc	#3	#5	#7
COCO	80.9	26.4	51.1	41.3	39.6
VisualCap	86.9	32.2	53.7	42.7	42.1
BlendCap	88.9	32.6	53.6	44.8	43.4
HolisticCap	**91.3**	**33.4**	**57.5**	**48.1**	**49.1**

Effectiveness of anchoring visual captions. Table 2 shows that the standalone visual caption (generated using InstructBLIP) is slightly worse than BlendCap. However, creating HolisticCap by anchoring the visual caption in human annotations results in about +4% improvement over BlendCap on TrueMatch. By drawing comparisons to Table 1, we observe that HolisticCap, which uses InstructBLIP, achieves performance on par to larger (LLaVa 1.6) and newer (Cambrian-1-3B) MLLMs.

> ***Finding* 1.** Self-retrieval with *TrueMatch* evaluates models' ability to produce fine-grained captions and *Visual Caption Boosting* is a promising way to enrich captioning datasets with granular details.

In summary, SR within TrueMatch enables fine-grained evaluation of captioning systems, probing different facets of visual discrimination. It encourages captions be to unique while also capturing semantic similarity with images, addressing the limitations of traditional captioning metrics.

4 Improving Captioning Systems with Self-Retrieval as a Training Objective

In this section, we present key insights on improving training with self-retrieval (SR), resulting in significant performance gains over previous works (Dessì et al., 2023). We begin by outlining the method (Section 4.1) and experimental setup (Section 4.2). Our experiments reveal that captioners trained with SR are highly sensitive to their MLE initialization (Section 4.3). In fact, we discover a trade-off between caption faithfulness and retrieval performance plaguing captioning systems fine-tuned with previous SR approaches (Dessì et al., 2023; Liu et al., 2018). To address this we: (1) initialize our model with more detailed captions from HolisticCap, (2) fine-tune the visual encoder with SR (Section 4.5), and (3) mine hard training bags and design a curriculum over bag sizes (Section 4.6). Finally we show that our training strategy is complementary to CIDEr optimization (Rennie et al., 2017) resulting in further improvements (Section 4.8).

4.1 Methodology

Model architecture. Similar to Dessì et al. (2023), we adopt ClipCap (Mokady et al., 2021), a lightweight simplification of the modern MLLMs (*e.g.* LLaVA, InstructBLIP, Cambrian). ClipCap connects a pretrained visual encoder (CLIP (Radford et al., 2021)) to a pretrained language model (GPT-2 (Radford et al., 2019)) through a simple MLP adapter. The adapter is tasked with mapping the rich visual embeddings from CLIP into a fixed number of *prefix tokens*. These tokens capture essential visual information and guide the language model towards generating an image-conditioned caption.

Training this captioning system has two steps: (1) model pretraining with maximum likelihood estimation (MLE), and (2) model fine-tuning by maximizing the SR reward with Reinforce (Williams, 1992). We outline both training objectives, followed by an experimental investigation of their properties and limitations.

Maximum Likelihood pretraining models the training data distribution. Specifically, captioning models are often teacher-forced to learn the word (token) distribution that maximizes the log-likelihood of the ground-truth captions given an input image. However, this results in strong language modeling priors resulting in generic captions that use a small vocabulary (see Table 6).

Maximizing self-retrieval with Reinforce. A fine-tuning step that optimizes a metric (*e.g.* CIDEr) is popular in training captioners (Rennie et al., 2017). We adopt Dessì et al. (2023)'s contrastive formulation of SR that compares the generated caption against distractors, providing an optimal learning signal:

$$\text{Reward } \; R(c, i, \mathcal{D}) = \log \frac{e^{\text{sim}(c,i)}}{\sum_{i' \in \mathcal{D} \cup \{i\}} e^{\text{sim}(c,i')}}. \tag{1}$$

We use Reinforce to optimize the captioning system by estimating the gradient of the expected reward $\mathbb{E}_{c \sim P}\left[\mathcal{R}(c, i, D) \nabla_\theta \log P(c|i;\theta)\right]$, since backpropagation is broken due to auto-regressive sampling. Further discussion on using Reinforce and variance reduction of the gradient estimates is in Appendices C.1 and C.2.

4.2 Experiment Setup and Metrics

We present results for all our experiments on ClipCap. During MLE pretraining, the ClipCap MLP adapter is trained from scratch, CLIP is frozen, and GPT-2 is fine-tuned. When fine-tuning with REINFORCE, ClipCap's MLP adapter is trained normally while CLIP and GPT-2 are fine-tuned with LoRA adapters (Hu et al., 2021) with rank 32. The model is trained with the AdamW optimizer (Loshchilov & Hutter, 2018) on a single A6000 GPU. Further details about the hyperparameters are provided in Appendix C.3.

We evaluate the quality of captions with NLG metrics: BLEU-4 and CIDEr. On RD100, a setup with 100 random distractors (Dessì et al., 2023), we use R@1 and ClipScore. On TrueMatch, we report R@1.

4.3 Experiment 3: Improving SR with Richer MLE Initialization

Following the findings of Section 3.3, we use TrueMatch to investigate the impact of different MLE initializations on SR fine-tuning in Table 3. We study SR in the same setting as previous work (Dessì et al., 2023), *i.e.* we only fine-tune the language model (SR-L), but using LoRA adapters.

Table 3: Results for different training methods and captions show the effectiveness of improved initialization with HolisticCap and SR-L.

	MLE Dataset	Training	CIDEr	RD100		TrueMatch		
				R@1	ClipSc	#3	#5	#7
1	COCO	MLE	107.8	76.5	29.9	50.8	36.5	33.8
2	BlendCap	MLE	80.2	84.9	31.6	51.8	42.5	42.2
3	HolisticCap	MLE	26.6	84.7	31.7	55.9	43.7	42.2
4	COCO	SR-L	108.2	83.7	30.2	53.3	42.3	38.4
5	BlendCap	SR-L	80.1	87.5	31.7	54.7	47.5	44.2
6	HolisticCap	SR-L	27.8	**87.6**	**31.9**	**59.1**	**47.7**	**46.5**

MLE pretraining yields generic captions. Table 3 rows 1-3 report the performance of MLE training on captioning datasets with increasing level of detail. We observe that improving the training captions results in a significant performance boost. Training with HolisticCap beats COCO by +8.2% on RD100 R@1 and +5.1% on TrueMatch #3. However, comparing performance against the ground-truth HolisticCap directly (from Table 2), we see that MLE training on HolisticCap yields -6.6% on RD100 R@1 and -1.6% on TrueMatch #3. This indicates that even though HolisticCap improves performance, MLE training tends to generate more generic and non-discriminative descriptions.

VCB breaks CIDEr evaluation. We see a significant drop in CIDEr scores for BlendCap and HolisticCap compared to COCO in Table 3. Upon further analysis, we believe that CIDEr rewards brevity and struggles to effectively evaluate with reference sets containing longer descriptive captions. For example with HolisticCap as ground-truth, Appendix C.6 shows some valid captions that unfortunately obtain scores close to 0.

Self-retrieval fine-tuning benefits from VCB. SR-L reduces MLE's preference towards generic concepts resulting in more discriminant captions and consistent improvements in RD100 and TrueMatch across datasets (compare rows 1-4, 2-5, 3-6). Providing a richer initialization to SR-L through VCB enhances fine-grained visual discrimination. Row 6 outperforms DiscriTune (row 4) with +3.9% RD100 and +5.8% TrueMatch #3. This illustrates the importance of a good MLE initialization for SR and the effectiveness of *VCB*.

Self-retrieval unlocks latent semantic information when fine-tuning the LLM. Although BlendCap does not capture any new information that is not present in one of the COCO annotations, being descriptive is sufficient to achieve superior retrieval performance against random distractors (BlendCap improves over COCO by +8% on RD100 R@1, Table 2). Further, MLE pretraining is able to model these data distributions as the performance gap is maintained at +8.4% on RD100 R@1 between rows 1 and 2 in Table 3. The model trained on COCO, due to the MLE objective, generates sparse captions that resemble the independent annotations of COCO. Thus, even though the semantic information in the embedding spaces of both models (trained on COCO or BlendCap) are similar, their ability to access this information is bottlenecked by the MLE objective, leading to a large gap between COCO and BlendCap on RD100 in Table 3.

Interestingly, SR-L fine-tuning narrows this gap dramatically from +8.4% to a mere +3.8% (rows 4, 5). This indicates that SR fine-tuning teaches the captioner to uncover semantic information within its embedding

space, even when it was initially obscured by MLE. Furthermore, the mostly unchanged BLEU-4 and CIDEr scores hint that the captioner remains faithful to the GT captions for both COCO (rows 1, 4) and BlendCap (rows 2, 5). This is remarkable, as it demonstrates that SR fine-tuning unlocks latent semantic information and steers the language model while preserving the initial data distribution.

***Finding* 2.** Self-retrieval guides the captioner away from its language modeling priors by unlocking latent semantic information and steering the language model to generate discriminative captions.

4.4 Trade-off plaguing SR fine-tuning: Retrieval Performance vs Faithfulness

While SR fine-tuning removes language modeling priors from captioners by making them more discriminant, we observe that they have a tendency to become less faithful to the GT captions upon extended training. We contrast these two tendencies of SR fine-tuning in Figure 3, comparing CIDEr and RD100 R@1 scores.

While SR performance continually improves (+7.9%) with extended training, a corresponding degradation in CIDEr (-15.4%) is observed. This trade-off reveals that SR has the capacity to elicit better retrieval from the captioning system, even at the cost of generating *lower-quality captions*. We further investigate with qualitative examples in Appendix C.8. Dessì et al. (2023) attribute this deterioration to the model diverging from the GT distribution during SR-L fine-tuning and argue that it is desirable. While this is partially true (see Figure 7), our investigation also reveals that captioners trained with vanilla SR-L, in a dash to enhance retrieval performance, build a propensity to hallucinate (see Figure 8) resulting in poor CIDEr scores. This tendency is only exacerbated with extended training due to further degradation in NLG metrics. These findings reveal an important bottleneck in SR fine-tuning:

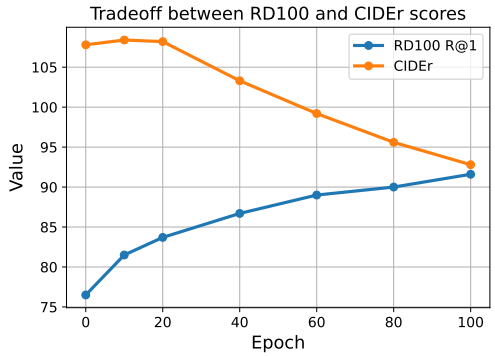


Figure 3: RD100 R@1 continually increases while CIDEr degrades when fine-tuning ClipCap with SR-L on COCO for 100 epochs.

***Finding* 3.** Captioners suffer from a trade-off between retrieval performance and caption faithfulness, showing a tendency to hallucinate details and deviate from ground-truth captions upon extended fine-tuning with SR-L.

Finally, the continued improvement in RD100 R@1 over 100 epochs indicates that vanilla SR-L fine-tuning fails to saturate SR's capacity to instill fine-grained visual discrimination in captioning systems. This motivates us to explore efficient SR fine-tuning strategies while mitigating their propensity to hallucinate. Since a higher learning rate destabilizes the policy gradients, we turn to: (1) fine-tuning the visual encoder (CLIP) with SR; and (2) adopting a curriculum over bags of hard negatives.

4.5 Experiment 4: Fine-tuning CLIP with Self-Retrieval

We present a comprehensive analysis of fine-tuning language and vision modules of the captioning system with LoRA in Table 4. Fine-tuning CLIP (SR-V) makes the captions discriminative. For example, with BlendCap MLE pretraining, SR-V fine-tuning results in +2.6% on RD100 R@1 and +6.6% on TrueMatch #3 (row 4, 5). However this comes at the cost of deteriorating NLG metrics (-7.7% CIDEr), suggesting that SR-V is unable to preserve the faithfulness of generated captions. We further verify that even after extended SR-L fine-tuning to match SR-V's retrieval scores, SR-L has a higher CIDEr (see Appendix C.4). This suggests that CIDEr deteriorates due to CLIP fine-tuning and not due to NLG metrics penalizing discriminative captions (Wang et al., 2020). This is a notable bottleneck in fine-tuning CLIP with SR: while it enables superior retrieval performance, it makes the captioner less faithful to the GT captions.

Further qualitative analysis of the captions generated by SR-L and SR-V in Figure 9, reveal that captioners fine-tuned with SR-V struggle to bind attributes to the correct objects. For instance, while SR-V captures

the *red* color of the *crane*, it binds it with the incorrect object (*train*). This is likely due to two reasons: (1) the retrieval objective treats the caption as a bag of words and does not promote compositionality; and (2) CLIP has been shown to struggle with attribute binding (Lewis et al., 2022; Hsieh et al., 2024).

RD100 is suboptimal for training fine-grained captioners. On TrueMatch, while SR-V fine-tuning provides substantial improvements over SR-L for COCO and BlendCap (rows 1-2, 4-5), we observe relatively smaller gains for HolisticCap (rows 7-8). This indicates that the vanilla SR-L objective of retrieving against 99 random distractors is not challenging enough to provide a good learning signal for models pre-trained on fine-grained HolisticCap. While SR-LV does better than SR-V on HolisticCap (rows 8-9), the gap between BlendCap (row 6) and HolisticCap (row 9) still remains small. To address this, we devise a curriculum with hard negatives, discussed in the next section, to increase the difficulty of SR fine-tuning objective.

Table 4: Ablation for fine-tuning different modules with SR: GPT-2 (SR-L), CLIP (SR-V), both (SR-LV). Row 1 is Dessì et al. (2023). **Bold** indicates best, *italics* is second. (x) presents improvement against SR-L.

	MLE Dataset	Training	NLG C	RD100 R@1	TrueMatch #3	TrueMatch #5	TrueMatch #7
1	COCO	SR-L	108.2	83.7	53.3	42.3	38.4
2	COCO	SR-V	97.2	86.1 (2.4)	59.3 (6.0)	49.2 (6.9)	45.0 (6.6)
3	COCO	SR-LV	96.6	87.4 (3.7)	59.6 (6.3)	51.2 (8.9)	47.2 (8.8)
4	BlendCap	SR-L	80.1	87.5	54.7	47.5	44.2
5	BlendCap	SR-V	72.4	90.1 (2.6)	61.3 (6.6)	**52.9** (5.4)	51.0 (6.8)
6	BlendCap	SR-LV	72.7	**90.7** (3.2)	61.3 (6.6)	52.3 (4.8)	50.1 (5.9)
7	HolisticCap	SR-L	27.8	87.6	59.1	47.7	46.5
8	HolisticCap	SR-V	25.6	*90.2* (2.6)	*61.9* (2.8)	52.5 (4.8)	**51.5** (5.0)
9	HolisticCap	SR-LV	26.5	**90.7** (3.1)	**62.6** (3.5)	*52.7* (5.0)	*51.2* (4.7)

***Finding* 4.** Fine-tuning the visual module with self-retrieval instills fine-grained visual details in the captioning system but struggles with attribute binding ~~and increased hallucinations~~.

Among methods in Table 4, SR-LV demonstrates the best retrieval performance while also preserving CIDEr slightly better than SR-V. Thus, we adopt SR-LV for subsequent experiments with a training curriculum.

4.6 Experiment 5: Self-Retrieval Fine-tuning with Bags and Curriculum

The SR experiments above were trained using 99 random distractors (Dessì et al., 2023). Instead, in this experiment, we fine-tune with bags of highly similar images within a training minibatch (see Appendix B.2). Retrieving against multiple hard negatives flattens the softmax distribution resulting in a stronger learning signal. We also propose a curriculum over bag sizes (BagCurri) to more optimally leverage the contrastive reward (see Appendix B.3). SR fine-tuning with BagCurri instills fine-grained visual details in the model, essential for discriminating highly similar images.

Table 5: Ablation of fine-tuning with the bag curriculum (BagCurri) as compared against vanilla SR (SR-L). Row 1 is Dessì et al. (2023). (Green) numbers indicate absolute improvement and (red), a degradation.

	MLE Dataset	Training	Bag Curri	NLG CIDEr	RD100 R@1	TrueMatch #3	TrueMatch #5	TrueMatch #7
1	COCO	SR-L	-	108.2	83.7	53.3	42.3	38.4
2	COCO	SR-L	✓	103.4	84.5	53.7	46.2	41.5
3	COCO	SR-V	✓	85.8	87.4	63.1	51.9	49.3
4	COCO	SR-LV	✓	84.2	85.8	67.2	56.3	53.6
5	Holistic-Cap	SR-L	-	27.8	87.6	59.1	47.7	46.5
6	Holistic-Cap	SR-L	✓	31.3	88.5	58.4	50.0	49.5
7	Holistic-Cap	SR-V	✓	25.8	90.6	64.0	54.6	54.1
8	Holistic-Cap	SR-LV	✓	29.7	91.3	65.2	57.1	57.1

SR-V versus SR-L with our bag curriculum. We evaluate the impact of fine-tuning different components of the captioning system (CLIP and GPT-2) with BagCurri compared to vanilla SR-L in Table 5. We find that SR-V fine-tuning with our curriculum forces CLIP to learn fine-grained visual features resulting in a substantial +5% to +11% improvement over vanilla SR-L (rows 3 *vs.* 1, 7 *vs.* 5) on TrueMatch. Interestingly, SR-V with BagCurri renders COCO (row 3) comparable to HolisticCap (row 7) on TrueMatch #3. However, these disproportionate gains for COCO come at the expense of CIDEr scores, which plummet -22.4% compared to SR-L. Further qualitative analysis of the generated captions shows that SR-V fine-tuning with BagCurri amplifies the trends

observed in Section 4.5 – while it boosts retrieval performance, it exacerbates the inability of the captioner to bind attributes to the correct objects and increases hallucinations (see Figure 9). In contrast, BagCurri may be too challenging for SR-L fine-tuning, as evident by the small improvements of +0% to +4% over vanilla SR-L (rows 2 *vs.* 1, 6 *vs.* 5) on RD100 or TrueMatch. However, similar to the findings of Section 4.3, we find that fine-tuning SR-L with BagCurri benefits heavily from the rich initialization provided by VCB. While the CIDEr score drops for COCO when using SR-L with BagCurri (-4.8%, row 2), for HolisticCap we observe an improvement (+3.5%, row 6).

Best of both worlds: SR-LV with BagCurri. Above findings indicates that BagCurri has the capacity to enhance both: (1) visual discrimination ability of the captioner with SR-V; and (2) adherence to GT captions with SR-L when provided with a rich initialization. Motivated by this, we initialize the captioner with HolisticCap and fine-tune both the language and visual components (SR-LV) with BagCurri (row 8), retaining the best of both worlds. SR-LV+BagCurri further enhances SR-V's already impressive performance on TrueMatch (achieving +6-10% over vanilla SR-L) while also improving the CIDEr score over vanilla SR-L by +1.9% (rows 5, 8). This is momentous, as without modifying the model's architecture or reward function, we outperform vanilla SR-L (Dessì et al., 2023) by an average +15% across all bag sizes on TrueMatch (row 1, 8), while circumventing the caption faithfulness trade-off plaguing current SR approaches.

Importance of the curriculum. In Appendix B.3, we investigate the role our curriculum plays in overcoming the retrieval and caption quality trade-off compared to directly using bags as hard negatives. Table 10 reveals that BagCurri is responsible for preserving caption faithfulness, as indicated by a decrease in CIDEr when bags are used without a curriculum.

***Finding* 5.** Providing a rich initialization with VCB and fine-tuning SR-LV with BagCurri instills fine-grained visual details in the captioner while preserving its faithfulness to ground-truth captions.

4.7 Impact on Generating Diverse Caption

We evaluate how VCB and SR increase caption diversity by measuring the number of words that appear ≥ 5 times on the COCO test set. In Table 6, we observe a clear increase in the number of unique words as we progress from COCO to BlendCap to HolisticCap, across all optimization and training strategies. Furthermore, within each individual dataset, there is a substantial increase in the usage of diverse vocabulary with transitions from MLE to SR-L, SR-V, and SR-LV. Notably, SR-LV with BagCurri yields significant improvements over MLE: relative +80% on COCO, +17.6% on BlendCap, and +21% on HolisticCap. This demonstrates the effectiveness of VCB and our curriculum in guiding the captioning system away from the language modelling priors and improving caption diversity.

Table 6: Number of words with frequency >= 5 on the COCO test set for captioners trained with different optimization strategies, modules, and datasets.

Dataset	MLE	Self-Retrieval			+BagCurri
		SR-L	SR-V	SR-LV	SR-LV
COCO	428	689	719	762	770
BlendCap	1220	1306	1345	1366	1435
HolisticCap	1431	1535	1606	1628	1732

4.8 Experiment 6: Combining Self-Retrieval and CIDEr Optimization

CIDEr optimization with Reinforce is commonly used to improve captioning system's adherence to the GT captions (Rennie et al., 2017). As the final experiment, we present results of combining it with SR in Table 7.

CIDEr optimization (rows 3, 8) improves faithfulness to GT captions compared to MLE pretraining (rows 1, 6): CIDEr improves +10.7% on COCO and +29.1% on HolisticCap. Importantly, CIDEr optimization on top of HolisticCap MLE pretraining provides +4.8% on TrueMatch #7 (rows 6, 8), while we see only a smaller +2.5% with COCO captions. Even without SR fine-tuning, the larger improvement with HolisticCap illustrates the effectiveness of VCB and importance of MLE initialization with captions containing fine-grained visual details.

Furthermore, compared to vanilla SR-L (rows 2, 7), CIDEr opt. (rows 3, 8) fairs worse on both RD100 and TrueMatch. This presents an opportunity to optimize both rewards simultaneously (Luo et al., 2018).

CIDEr meets SR. Building upon the findings of previous sections, we push the boundaries of SR by explicitly adding CIDEr to our discriminative reward $R = \text{SR} + \lambda \cdot \text{CIDEr}$. We conduct an ablation over various λ values (see Appendix C.5), and select $\lambda = 0.5$ for our experiments.

Table 7: Impact of combining SR with CIDEr optimization. Results are presented for various combinations of RL rewards. C: CIDEr, SR: Self-Retrieval, and BC: SR with BagCurri. Row 1 is Mokady et al. (2021), row 2 is Dessì et al. (2023), and row 3 is Rennie et al. (2017) with ClipCap.

	MLE Dataset	Reward C	Reward SR	Reward BC	NLG CIDEr	RD100 R@1	TrueMatch #3	TrueMatch #5	TrueMatch #7
1	COCO	-	-	-	107.8	76.5	50.8	36.5	33.8
2		-	SR-L	-	108.2	83.7	53.3	42.3	38.4
3		✓	-	-	118.5	79.1	50.7	37.3	36.3
4		-	SR-LV	✓	84.2	85.8	*67.2*	56.3	53.6
5		✓	SR-LV	✓	92.8	89.0	59.6	51.2	47.2
6	Holistic-Cap	-	-	-	26.6	84.7	55.9	43.7	42.2
7		-	SR-L	-	27.8	87.6	59.1	47.7	46.5
8		✓	-	-	**55.7**	86.9	57.6	47.1	47.0
9		-	SR-LV	✓	29.7	*91.3*	65.2	*57.1*	*57.1*
10		✓	SR-LV	✓	*38.0*	**92.6**	**67.7**	**58.7**	**57.9**

With HolisticCap, we see that joint optimization (row 10) outperforms the most discriminant SR model (row 9) achieving improvements on all metrics: +8.3% on CIDEr, +1.3% on RD100 R@1, and +0.8% to +2.5% on TrueMatch. Conversely, on COCO captions (rows 4, 5), we observe that adding CIDEr optimization adversely affects SR metrics on TrueMatch with -5.1% to -7.6% performance drops. This aligns with previous findings and underscores the importance of initialization during SR fine-tuning.

Recall, our best model of row 10 also achieves highest performance on TrueMatch as compared to several MLLMs that are orders of magnitude larger and trained on more data (see Table 1).

5 Generalization Experiments

In this section, we evaluate the generalization of our approach on ImageCoDe (Krojer et al., 2022), a benchmark for text-based image retrieval where image sets are created with 10 sequential video frames. Different from TrueMatch, ImageCoDe focuses on fine-grainedness found across video frames such as negation, occlusion, *etc.* We follow the same experimental setup of Dessì et al. (2023). Table 8 shows that SR fine-tuning of both the vision and text modules with our bag curriculum (SR-LV BagCurri) achieves state-of-the-art results. In fact, it outperforms Dessì et al. (2023)'s work on using SR-L after MLE pretraining on COCO or Conceptual Captions (a 30× larger dataset).

Table 8: Zero-shot evaluation on the ImageCoDe validation set (R@1). Rows 2 and 3 are from Dessì et al. (2023).

	Dataset	Method	R@1
1	COCO	MLE	28.8
2	COCO	SR-L	30.3
3	ConCap	SR-L	36.2
4	HolisticCap	MLE	31.9
5		SR-L	33.0
6		SR-LV BagCurri	**37.9**

6 Conclusion

We investigated self-retrieval (SR) as a way to comprehensively improve current captioning systems on all three fronts: (1) data, (2) evaluation, and (3) training. (1) We proposed Visual Caption Boosting (VCB), a model agnostic framework to instill fine-grained visual information into generic image captioning datasets while remaining anchored in human annotations. (2) Furthermore, we also created TrueMatch, a benchmark comprised of highly similar image bags. SR evaluation with TrueMatch enabled fine-grained evaluation of captioning systems and unlike traditional captioning metrics encouraged diversity in generated captions. (3) We also adopted SR fine-tuning and showed that the rich MLE initialization provided by VCB is important. We uncovered a trade-off between caption faithfulness and retrieval performance plaguing current SR approaches and designed a training recipe to address it. Specifically, we employed a curriculum over the number of hard negatives to more optimally leverage the SR reward, and fine-tune both the visual and language components of the captioning system. We demonstrated that this training recipe and VCB circumvents the aforementioned trade-off. Our approach achieved significant performance improvements over vanilla SR (Dessì et al., 2023) across a dataset with 99 random distractors, and set the state-of-the-art on ImageCoDe (Krojer et al., 2022). On TrueMatch, our approach outperforms powerful MLLMs that are 1-2 orders of magnitude larger and trained on large datasets, while improving over vanilla SR by up to +15%.

7 Limitations and Future Scope

Self-retrieval (SR) is heavily reliant on a scorer to identify and retrieve the image that best matches the generated text. We employ CLIP as a scorer, which has been shown to struggle with capturing extremely fine-grained and compositional visual details. However, the reliability of SR evaluation is set to improve steadily over time. Additionally, as Visual Caption Boosting uses foundation models for generating fine-grained descriptions, it is prone to hallucinations. However, our anchoring strategy mitigates this to a large extent.

As future work, exploring further applications of the SR fine-tuning presents a promising research avenue. As we show, MLE initialization plays an important role and this too deserves more investigation.

References

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. SPICE: Semantic Propositional Image Caption Evaluation. In *European Conference on Computer Vision (ECCV)*, pp. 382–398. Springer, 2016. 2

Abhipsa Basu, R Venkatesh Babu, and Danish Pruthi. Inspecting the Geographical Representativeness of Images from Text-to-Image Models. In *International Conference on Computer Vision (ICCV)*, pp. 5136–5147, 2023. 2, 4

Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sağnak Taşırlar. Introducing our multimodal models, 2023. URL https://www.adept.ai/blog/fuyu-8b. 1

Lucas Beyer, Xiaohua Zhai, and Alexander Kolesnikov. Big vision. https://github.com/google-research/big_vision, 2022. 6

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. *Advances in Neural Information Processing Systems (NeurIPS)*, 36, 2024. 1, 6, 17

Karan Desai, Gaurav Kaul, Zubin Trivadi Aysola, and Justin Johnson. RedCaps: Web-curated image-text data created by the people, for the people. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 1

Roberto Dessì, Michele Bevilacqua, Eleonora Gualdoni, Nathanaël Carraz Rakotonirina, Francesca Franzon, and Marco Baroni. Cross-Domain Image Captioning With Discriminative Finetuning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6935–6944, 2023. 1, 2, 3, 5, 6, 7, 8, 9, 10, 11, 12, 21, 22, 24

Sivan Doveh, Assaf Arbelle, Sivan Harary, Roei Herzig, Donghyun Kim, Paola Cascante-Bonilla, Amit Alfassy, Rameswar Panda, Raja Giryes, Rogerio Feris, et al. Dense and Aligned Captions (DAC) Promote Compositional Reasoning in VL Models. *Advances in Neural Information Processing Systems (NeurIPS)*, 36, 2024. 1, 4

Lisa Dunlap, Yuhui Zhang, Xiaohan Wang, Ruiqi Zhong, Trevor Darrell, Jacob Steinhardt, Joseph E Gonzalez, and Serena Yeung-Levy. Describing Differences in Image Sets with Natural Language. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 6

Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. VSE++: Improving Visual-Semantic Embeddings with Hard Negatives. In *British Machine Vision Conference (BMVC)*, 2017. 3

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. Multi-modal hallucination control by visual information grounding. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14303–14312, 2024. 4

Adam Fisch, Kenton Lee, Ming-Wei Chang, Jonathan H Clark, and Regina Barzilay. CapWAP: Image Captioning with a Purpose. In *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 8755–8768, 2020. 1

Melissa Hall, Laura Gustafson, Aaron Adcock, Ishan Misra, and Candace Ross. Vision-Language Models Performing Zero-Shot Tasks Exhibit Gender-based Disparities. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7514–7528, 2021. 2

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality. *Advances in Neural Information Processing Systems (NeurIPS)*, 36, 2024. 10

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations (ICLR)*, 2021. 8

Harsh Jhamtani and Taylor Berg-Kirkpatrick. Learning to Describe Differences Between Pairs of Similar Images. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2018. 5, 6

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023. 17

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling Laws for Neural Language Models. *arXiv preprint arXiv:2001.08361*, 2020. 21

Andrej Karpathy and Li Fei-Fei. Deep Visual-Semantic Alignments for Generating Image Descriptions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3128–3137, 2015. 5

Simon Kornblith, Lala Li, Zirui Wang, and Thao Nguyen. Guiding Image Captioning Models Toward More Specific Captions. In *International Conference on Computer Vision (ICCV)*, pp. 15259–15269, 2023. 1

Elisa Kreiss, Fei Fang, Noah Goodman, and Christopher Potts. Concadia: Towards Image-Based Text Generation with a Purpose. In *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4667–4684, 2022. 1

Benno Krojer, Vaibhav Adlakha, Vibhav Vineet, Yash Goyal, Edoardo Ponti, and Siva Reddy. Image Retrieval from Contextual Descriptions. In *Association of Computational Linguistics (ACL)*, pp. 3426–3440, 2022. 3, 6, 12

Zhengfeng Lai, Haotian Zhang, Wentao Wu, Haoping Bai, Aleksei Timofeev, Xianzhi Du, Zhe Gan, Jiulong Shan, Chen-Nee Chuah, Yinfei Yang, et al. From scarcity to efficiency: Improving clip training via visual-enriched captions. *arXiv preprint arXiv:2310.07699*, 2023. 3, 4, 5

Martha Lewis, Nihal V Nayak, Peilin Yu, Qinan Yu, Jack Merullo, Stephen H Bach, and Ellie Pavlick. Does clip bind concepts? probing compositionality in large image models. In *European Chapter of the Association for Computational Linguistics (EACL)*, 2022. 10

Xianhang Li, Haoqin Tu, Mude Hui, Zeyu Wang, Bingchen Zhao, Junfei Xiao, Sucheng Ren, Jieru Mei, Qing Liu, Huangjie Zheng, et al. What If We Recaption Billions of Web Images with LLaMA-3? *arXiv preprint arXiv:2406.08478*, 2024. 3, 4

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision (ECCV)*, pp. 740–755. Springer, 2014. 1, 4

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*, 2024. 2

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 6

Xihui Liu, Hongsheng Li, Jing Shao, Dapeng Chen, and Xiaogang Wang. Show, Tell and Discriminate: Image Captioning by Self-retrieval with Partially Labeled Data. In *European Conference on Computer Vision (ECCV)*, pp. 338–354, 2018. 2, 3, 7

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations (ICLR)*, 2018. 8

Ruotian Luo, Brian Price, Scott Cohen, and Gregory Shakhnarovich. Discriminability Objective for Training Descriptive Captions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6964–6974, 2018. 3, 11, 23

Ron Mokady, Amir Hertz, and Amit H Bermano. ClipCap: CLIP Prefix for Image Captioning. *arXiv preprint arXiv:2111.09734*, 2021. 1, 6, 7, 12

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Association of Computational Linguistics (ACL)*, pp. 311–318, 2002. 2

Dong Huk Park, Trevor Darrell, and Anna Rohrbach. Robust Change Captioning. In *International Conference on Computer Vision (ICCV)*, pp. 4624–4633, 2019. 5, 6

André Susano Pinto, Alexander Kolesnikov, Yuge Shi, Lucas Beyer, and Xiaohua Zhai. Tuning Computer Vision Models with Task Rewards. In *International Conference on Machine Learning (ICML)*, pp. 33229–33239, 2023. 21

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. In *International Conference on Computer Vision (ICCV)*, pp. 2641–2649, 2015. 1

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners. 2019. 7

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning (ICML)*, pp. 8748–8763. PMLR, 2021. 7

Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-Critical Sequence Training for Image Captioning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7008–7024, 2017. 3, 7, 11, 12, 22

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2018. 23

Hadi Salman, Saachi Jain, Andrew Ilyas, Logan Engstrom, Eric Wong, and Aleksander Madry. When does bias transfer in transfer learning? *arXiv preprint arXiv:2207.02842*. 2, 4

Gabriel Oliveira dos Santos, Esther Luna Colombini, and Sandra Avila. Cider-r: Robust consensus-based image description evaluation. *arXiv preprint arXiv:2109.13701*, 2021. 23

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. LAION-5B: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:25278–25294, 2022. 1

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Association of Computational Linguistics (ACL)*, pp. 2556–2565, 2018. 1

Nina Shvetsova, Anna Kukleva, Xudong Hong, Christian Rupprecht, Bernt Schiele, and Hilde Kuehne. HowToCaption: Prompting LLMs to Transform Video Annotations at Scale. *arXiv preprint arXiv:2310.04900*, 2023. 3

Vasu Singla, Kaiyu Yue, Sukriti Paul, Reza Shirkavand, Mayuka Jayawardhana, Alireza Ganjdanesh, Heng Huang, Abhinav Bhatele, Gowthami Somepalli, and Tom Goldstein. From Pixels to Prose: A Large Dataset of Dense Image Captions. *arXiv preprint arXiv:2406.10328*, 2024. 3, 4

Kirill Sirotkin, Pablo Carballeira, and Marcos Escudero-Viñolo. A study on the distribution of social biases in self-supervised learning visual models. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10442–10451, 2022. 2, 4

Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. From Show to Tell: A Survey on Deep Learning-Based Image Captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 45(1):539–559, 2022. 1

Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59 (2):64–73, 2016. 1

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*, 2024a. 6

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024b. 5

Jack Urbanek, Florian Bordes, Pietro Astolfi, Mary Williamson, Vasu Sharma, and Adriana Romero-Soriano. A Picture is Worth More Than 77 Text Tokens: Evaluating CLIP-Style Models on Dense Captions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 4, 5

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-Based Image Description Evaluation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4566–4575, 2015. 2

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning (ICML)*, pp. 23318–23340. PMLR, 2022. 6

Zeyu Wang, Berthy Feng, Karthik Narasimhan, and Olga Russakovsky. Towards Unique and Informative Captioning of Images. In *European Conference on Computer Vision (ECCV)*, pp. 629–644. Springer, 2020. 2, 9, 21, 22

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8:229–256, 1992. 7

Andre Ye, Sebastin Santy, Jena D Hwang, Amy X Zhang, and Ranjay Krishna. Computer Vision Datasets and Models Exhibit Cultural and Linguistic Diversity in Perception. *arXiv preprint arXiv:2310.14356*, 2023. 4

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 6

A Visual Caption Boosting

A.1 Prompting Strategy

In this section, we discuss prompting strategies used to produce BlendCap, the visual caption, and HolisticCap. We provide the system prompt along with the in-context examples we utilize. In order to generate captions that holistically describe the image, we use the following two step process:

(a) **BlendCap.** First, we prompt Mistral-7B (Jiang et al., 2023) to reduce redundancy, preserve all factual information obtained from human annotated captions, and not introduce any new information other than what is mentioned in the reference set. This ensures that the blended caption remains factual and anchored to the human annotations:

```
System prompt : You are given a reference set of 5 image captions. Create a holistic
caption to describe the same image by incorporating relevant information from all the
reference captions. Ensure that the generated caption minimizes redundancy by excluding
overlapping information found in the reference set. Do not include additional information
and adjectives other than what is given in reference set. Stop after the first period and do
not ask any additional questions. Just complete the text. Three examples are shown below :

In context examples :
User : Reference caption 1: A plate full of food with an assortment of food on it.
Reference caption 2: There is meat and vegetables on a white and brown plate.
Reference caption 3: A plate on a table that hasfood on it.
Reference caption 4: A plate with some steak, carrots, and sliced fried potatoes.
Reference caption 5: A plate of food on a table.
Assistant : A white and brown plate on a table with an assortment of steak, carrots,
sliced fried potatoes, and vegetables.

User : Reference caption 1: A park bench on the side of a lake.
Reference caption 2: A lone bench sits atop a hill looking over the river.
Reference caption 3: A wooden bench sitting on top of a sandy beach.
Reference caption 4: A bench on a river bank in the countryside.
Reference caption 5: A wood bench is sitting in front of a river.
Assistant : A wooden bench sits atop a hill in front of a river overlooking the water and
the surrounding countryside.

User : Reference caption 1: A bathroom with a toilet sitting next to a sink.
Reference caption 2: A white sink and toilet in a room.
Reference caption 3: A bathroom with a mirror, sinks, toilet and toilet roll.
Reference caption 4: A bathroom that has a toilet sink and mirror in it.
Reference caption 5: A bathroom with a toilet next to a sink.
Assistant : A bathroom with a toilet next to a white sink, mirror and toilet roll.
```

(b) **Visual Caption.** We concurrently use InstructBLIP (Dai et al., 2024) to produce dense visual descriptions for the given image. We find the following prompt performs the best.

```
System prompt : Briefly describe the content of the image.
```

(c) **HolisticCap.** We then prompt Mistral-7B to expand the blended caption with the visual description generated by the MLLM. Notably, the LLM is prompted to preserve human annotated information when

there is a conflict of information between *BlendCap* and *Visual Caption*. Only 3 few shot examples are shown for brevity.

System prompt : In this task, your objective is to improve "Correct caption" by incorporating any additional information present in "New caption". If there are conflicting details between the two captions regarding the objects or their attributes, strictly use information from "Correct caption". "New caption" can have incorrect information sometimes but "Correct caption" will always be correct. The focus is on transferring missing information, including adjectives and nouns, from "New caption" to "Correct caption". Refrain from seeking further clarification and explanations; simply generate the caption. Three examples are shown below :

In context examples :
User : Correct caption: A man reaches up with a tennis racquet to hit an approaching ball in a tennis court.
New noisy caption: In this image, a woman is playing tennis on a purple court. She is wearing a white shirt and a blue short, and she is holding a tennis racket. The tennis ball is in the air as she prepares to hit it with her racket.
Assistant : A man wearing a white shirt and blue shorts reaches up with a tennis racquet to hit an approaching ball in a purple tennis court.

User : Correct caption: A white fire hydrant sits in front of an old couch on a sidewalk in front of a house.
New noisy caption: In this image, there is a black fire hydrant sitting on the sidewalk in front of a brick building. The fire hydrant is positioned next to a couch, which is placed on the sidewalk in front of the building.
Assistant : A white fire hydrant sits in front of an old couch on a sidewalk in front of a brick house.

User : Correct caption: A young boy poses next to a wall with writing on it, smiling and holding bags.
New noisy caption: In this image, a young girl is sitting in front of a graffiti-covered wall, wearing a red shirt. She is holding a box of crayons and smiling at the camera.
Assistant : A young boy wearing a red shirt poses next to a graffiti-covered wall with writing on it. He is holding bags of crayons and smiling at the camera.

A.2 Human Study: Quantifying Hallucinations in VCB

We conduct a human study to investigate the extent of hallucinations in BlendCap and HolisticCap. We randomly sample 60 images from MSCOCO and for each image-caption pair, a human is asked to count the number of hallucinations present in the caption. Three types of hallucinations are considered: (i) object, (ii) attribute, (iii) egregious statements, *i.e.*, a glaringly incorrect statement that is unambiguously false.

BlendCap contains only 4 objects and 3 attribute hallucinations. Out of the 7 hallucinations, 5 are due to incorrect COCO human annotations and only 2 (3.3%) are induced by the LLM. This demonstrates the reliability of BlendCap in combining human annotations into a denser caption.

HolisticCap. contains only 9 instances of hallucinations (4 objects and 5 attributes) across 7 captions, compared to VisualCap's 16 instances (12 captions). Although VisualCap contained 3 egregious statements, HolisticCap removed them all. Furthermore, only 7/60 captions in HolisticCap contain hallucinations, reflecting a 41.6% reduction compared to VisualCap's 12/60 captions. This demonstrates VCB's effectiveness in anchoring the MLLM descriptions (Visual captions) into human annotations (BlendCap) to create HolisticCap.

A.3 Token Statistics for Captions

We present the caption statistics for Visual Caption Boosting compared to original COCO captions in Table 9. While Visual Captions are longer than HolisticCap, they may contain erroneous details and verbose language modelling priors as shown in Figure 2. Both BlendCap and HolisticCap are generally more descriptive than COCO, containing more words. Figure 4 shows the histogram of the number of tokens, tokenized using the CLIP tokenizer, for COCO, BlendCap and HolisticCap.

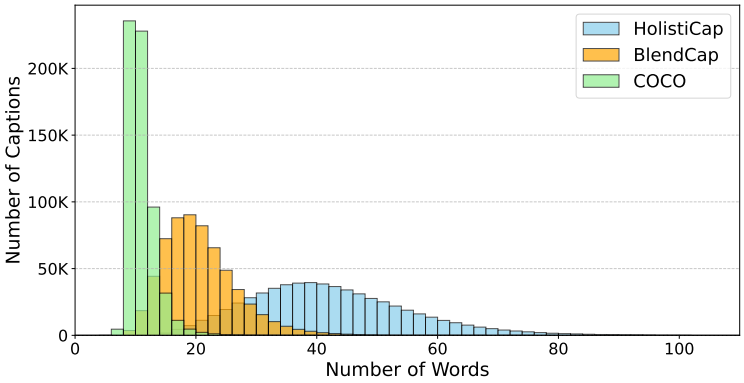


Figure 4: Histogram of number of words for COCO, BlendCap, and HolisticCap. HolisticCap is more descriptive with 41.5 words per caption on average, while COCO only has 10.5 words on average.

Table 9: Statistics of different captions. The mean and standard deviation of the number of words and tokens (post CLIP tokenization) are reported. Note the 4× difference between original COCO and HolisticCap.

	Dataset	Words/Cap	Toks/Cap
1	COCO	10.5 ±2.4	13.5 ±2.7
2	BlendCap	20.1 ±6.0	25.4 ±7.0
3	HolisticCap	41.5 ±13.1	50.2 ±15.4
4	Visual Caption	51.1 ±14.4	59.2 ±12.8

B Bag Creation, Curation, and Curriculum Details

B.1 Bag Creation and Curation

Algorithm 1 presents the process for creating bags of highly similar images that is used for both TrueMatch (Section 3.1) and SR fine-tuning (Section 4.6).

We take the bags created by Algorithm 1 on the validation and test sets of COCO (10,000 images) and use Algorithm 2 to curate bags of highly similar images for TrueMatch. The curated bags only consider the bag with the highest intra-bag similarity for a given visual concept.

B.2 Training Bags used for Self-Retrieval

We utilize our bag creation algorithm (Algorithm 1) to create training bags. However, for each row in the sorted cosine similarity matrix D we only consider the top 200 retrievals, *i.e.* `D[:, :200]`. Each training bag of size s is comprised of a given query image and the top $s-1$ images it retrieves from D such that none of the retrieved images have been added to any other bag before. This ensures that all the bags are unique and each image is seen exactly once in an epoch.

Unlike TrueMatch where we select the hardest bag corresponding to specific visual concepts in the CLIP embedding space, we relax the constraint of only sampling the sub-cluster (bag of images) with the highest intra-cluster similarity.

Algorithm 1 Candidate Bag Creation

Input:
Output bag size s.
Set of N images each with M captions: $\mathcal{I} = \{(i_1, c_1^1, \ldots, c_1^M), \ldots, (i_N, c_N^1, \ldots, c_N^M)\}$.

Output:
List of T bags $\mathcal{B} = [b_1, b_2, \ldots, b_T]$ where each bag $b \subset \mathcal{I}$.
$\mathbf{A} = [\alpha_1, \alpha_2, \ldots, \alpha_T]$ are the intra-bag similarities.

Multimodal feature extraction:
for each image $(i, c^1, \ldots, c^M)$ **do**
  Compute image embedding $\mathbf{z} = \texttt{CLIP.encode_image}(i)$.
  Compute text embedding $\mathbf{t} = \frac{1}{M}\sum_{j=1}^{M} \texttt{CLIP.encode_text}(c^j)$.
  Compute multimodal embedding $\mathbf{m} = \texttt{concat}(\mathbf{z}, \mathbf{t})$, $\mathbf{m} \in \mathbb{R}^d$.
end for

Compute the cosine similarity matrix $D \in \mathbb{R}^{N \times N}$ based on multimodal embeddings $[\mathbf{m}_1, \ldots, \mathbf{m}_N]$.

Bag Creation:
$\mathcal{B} \leftarrow [], \mathbf{A} \leftarrow []$
for each image i_r, $r = [1, \ldots, N]$ **do**
  Create a bag b_r with i_r and corresponding top-scoring $s{-}1$ images from row D_r.
  Compute intra-bag similarity between i_r and $s{-}1$ images in the bag: $\alpha_r = \text{mean}_k(\cos(\mathbf{m}_r, \mathbf{m}_k))$.
  Update $\mathcal{B} \leftarrow [\mathcal{B}, b_r]$ and $\mathbf{A} \leftarrow [\mathbf{A}, \alpha_r]$.
end for

Algorithm 2 TrueMatch: Automated Bag Curation

Input: List of T bags $\mathcal{B} = [b_1, \ldots, b_T]$ and corresponding intra-bag similarities $\mathbf{A} = [\alpha_1, \ldots, \alpha_T]$
Output: A curated list of bags $\mathcal{Q}$ with highly similar images used in TrueMatch.

Benchmark Creation:
Sort all bags $\mathcal{B}$ in descending intra-bag similarity $\mathbf{A}$.
Initialize the set of visited images $\mathcal{V} = \{\}$ and curated list of bags $\mathcal{Q} = []$
for each bag $b \in \mathcal{B}$ **do**
  if $b \cap \mathcal{V} = \varnothing$ **then**
    Update $\mathcal{V} \leftarrow \mathcal{V} \cup b$
    Update $\mathcal{Q} \leftarrow [\mathcal{Q}, b]$
  end if
end for

Table 10: Self-retrieval fine-tuning of GPT-2 and CLIP (SR-LV) for ClipCap that is MLE initialized with HolisticCap. Values indicate R@1. (Green) numbers indicate absolute improvement and (red), a degradation.

			TrueMatch			
	Method	CIDEr	RD100	#3	#5	#7
1	Rand	26.5	90.2	62.6	52.7	51.2
2	Bag2	26.0 (0.5)	90.5	64.8	54.0	55.1
3	Bag3	26.0 (0.5)	90.7	64.8	55.4	55.3
4	Bag5	25.2 (1.3)	90.9	65.0	55.8	52.8
5	Bag7	25.7 (0.8)	90.8	65.1	56.7	55.3
6	Bag10	25.3 (1.2)	90.9	65.1	55.8	54.1
7	Bag15	24.9 (1.6)	91.0	66.5	59.2	56.5
8	Bag20	24.8 (1.7)	91.1	66.5	58.8	57.3
9	BagCurri	29.7 (3.2)	91.3	65.2	57.1	57.1

B.3 BagCurri: Designing a Self-Retrieval Learning Curriculum with Bags

We design a curriculum that varies the bag sizes during training, gradually increasing the bag size with each epoch as shown in Figure 5. Since the reward is the cross entropy of matching the generated caption with the target image Dessì et al. (2023), mining multiple hard negatives yields a flatter softmax distribution leading to stronger learning signal. Hence, increasing the bag size introduces makes the task of SR more challenging. However, it is important to increase bag sizes gradually to ensure that the task is not too hard, and prevent a collapse of caption faithfulness (Section 4.6).

To assess the role our curriculum plays in preventing this collapse, we conduct an ablation study. Specifically, we compare our curriculum against training with bags of a fixed size for the entire SR fine-tuning process. Table 10 demonstrates that our carefully designed curriculum is responsible for preserving caption faithfulness. While training with hard negatives without a curriculum improves retrieval performance on TrueMatch (rows 2-8), it fails to maintain the CIDEr score. In contrast, BagCurri not only yields substantial gains over training with random distractors (row 1) but also enhances CIDEr by +3.2%. This highlights the effectiveness of our carefully designed curriculum in preventing the collapse of NLG metrics while improving retrieval performance. Note, this is different to previous SR approaches (Dessì et al., 2023) as presented in Section 4.4.

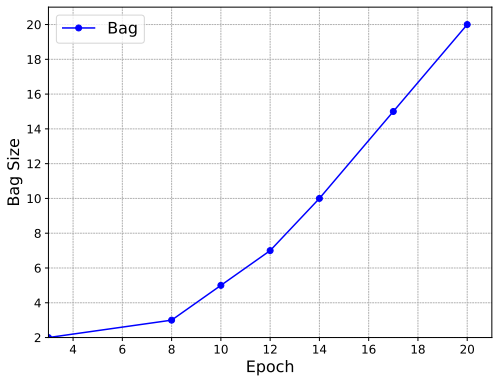


Figure 5: Our curriculum over bag sizes during training.

C Self-Retrieval Fine-tuning

C.1 Discussion on Optimization Strategies

Similar to (Pinto et al., 2023), we find both optimization strategies discussed in Section 4.1 have complementary strengths and weaknesses for the task of image captioning.

MLE. Training the captioner on MLE results in powerful models capable of capturing complex data distributions. The task of next-token prediction with current language models is both efficient (due to teacher forcing) and scalable (Kaplan et al., 2020). However, it creates captioning systems that prefer generic captions (Wang et al., 2020), reusing a small vocabulary to describe different images with fine-grained visual differences (see Figure 1 and Section 4.3).

Reinforce. While fine-tuning with SR alleviates these issues (Section 4), it comes with its own set of challenges. Not only are the tokens sampled sequentially but optimizing Reinforce is unstable, necessitating

smaller learning rates. This results in an extremely slow training process. Notably, using REINFORCE from scratch is infeasbile due to the vast action space of sampling a token and reward sparsity. However, using a pretrained MLE model provides a good initial sampling strategy and requires relatively few optimization steps to perform well on SR. This is consistent with previous works such as Rennie et al. (2017).

In Reinforcement Learning jargon, the gradients computed with REINFORCE are used directly to optimize the policy network (captioning system) in prioritizing actions (next-token prediction) that minimize the negative expected reward $\mathbb{E}_{c\sim P_\theta(\cdot|i)}[-R(c, i, \mathcal{D})]$.

C.2 REINFORCE baseline

The variance of the gradient estimate in REINFORCE is commonly reduced by subtracting a baseline b from the reward R. This stabilizes training by reducing the variance of the estimated gradient. We adopt a running mean of past rewards as our baseline similar to Dessì et al. (2023) as they find that it outperforms computing the reward with greedy decoding as a baseline. The latter requires sampling two outputs for one training input, one to estimate the gradient and other to compute the baseline. However, when the reward includes CIDEr, we adopt a greedy baseline.

C.3 Overview of Hyperparameters

We provide a brief overview of hyperparameters for different optimization strategies: MLE (Table 11) and REINFORCE (Table 12). Notably, CIDEr optimization for HolisticCap is done with learning rate 10^{-7}.

Table 11: MLE pretraining settings.

MLE Pretraining	
Batch size:	40
Schedule:	Linear Decay
Learning-rate:	$2 \cdot 10^{-5}$
Total steps:	30 000
Warmup steps:	1 000

Table 12: REINFORCE optimization settings.

Base params	
Batch size:	100
Schedule:	constant
Total steps:	23 000
Warmup steps:	0
SR reward	
Learning-rate:	$9 \cdot 10^{-8}$
CIDEr reward	
Learning-rate:	$1 \cdot 10^{-6}$
CIDEr + SR reward	
Learning-rate:	$1 \cdot 10^{-7}$

C.4 Fine-tuning CLIP hurts Caption Faithfulness

Table 4 shows that SR-V is substantially more discriminative compared to SR-L. However unlike SR-L, SR-V fails to preserve caption faithfulness through early stopping. To verify that the observed deterioration isn't due to the tendency of NLG metrics to penalize discriminability (Wang et al., 2020), we train SR-L until it achieves the same RD100 score as SR-V for each dataset. This ensures a fair comparison of both methods' abilities to preserve CIDEr, as presented in Table 13. We clearly see that SR-L (col 2) attains a significantly higher CIDEr score compared to SR-V (col 3) for all caption types, demonstrating that the deterioration in NLG metrics observed in Table 4 is directly caused by CLIP fine-tuning.

Table 13: Extended SR-L fine-tuning until its R@1 on RD100 becomes equal to SR-V fine-tuning R@1 when trained for 20 epochs.

Data	MLE	SR-L	SR-V
COCO	107.8	104	97.2
BlendCap	80.2	77.3	72.4
HolisticCap	26.6	26.4	25.6

383443

Caption : A spacious and well-appointed bathroom features two sinks, a bathtub, and a mirror.

CIDEr : 0

470005

Caption : A baby wearing pajamas sits on a bed while using a laptop computer.

CIDEr : 4

410428

Caption : Three sheep, two of which are adults and one a baby, stand together in a grassy field.

CIDEr : 0

Figure 6: **Failure of CIDEr**. COCOIDs in the test set where CIDEr fails to correctly evaluate captions against HolisticCap reference set.

C.5 Scaling CIDEr during SR optimization

The weighting of the rewards when jointly optimizing CIDEr and SR (Luo et al., 2018) determines the trade-off between faithfulness and discriminative ability of the captioning system. We MLE pretrain the captioner with HolisticCap and fine-tune both the LLM and CLIP components (SR-LV) with BagCurri. Table 14 presents an ablation over different values of λ while scaling CIDEr in the joint reward R = SR $+\lambda\cdot$ CIDEr. As expected, increasing λ *i.e.* heavily weighing CIDEr leads to improved caption faithfulness. Interestingly, using a smaller λ however doesn't always improve TrueMatch scores, as $\lambda = 0.7$ (row 5) outperforms $\lambda = 0.1$ (row 2). We adopt $\lambda = 0.5$ for our experiments (Section 4.8), as it results in the best performing model achieving an optimal balance between retrieval performance and caption faithfulness.

Table 14: Ablation over different lambda values for HolisticCap pretraining and SR-LV with BagCurri fine-tuning. We choose λ=0.5 as our best model.

	λ	NLG	RD100	TrueMatch		
		CIDEr	R@1	#3	#5	#7
1	0	29.7	91.3	65.2	57.1	57.1
2	0.1	29.8	92.4	64.8	54.0	55.1
3	0.3	33.3	92.0	68.2	61.2	57.1
4	**0.5**	38.0	92.6	67.7	58.7	57.9
5	0.7	41.4	92.7	68.1	58.3	56.8
6	1	45.4	92.5	66.5	56.2	56.2

C.6 CIDEr Struggles with Longer Descriptions

Notably, we observe that the scale of CIDEr scores for MLE pretrained models vary significantly across COCO, BlendCap, and HolisticCap (rows 1-3, Table 3). Consistent with Santos et al. (2021), we find that CIDEr rewards brevity, while reference sets containing longer captions have substantially lower scores. While the inefficacy of MLE in modeling longer descriptions does contribute to lower CIDEr scores, we also find that CIDEr fails to effectively evaluate MLE pretraining with HolisticCap, even assigning scores of 0 to correct captions, as illustrated in Figure 6.

C.7 CHAIR: Quantifying Object Hallucinations during Extended SR-L Fine-tuning

We generate captions for all the 10,000 images of MSCOCO test and validation set and compute object level hallucinations using CHAIR (Rohrbach et al., 2018). CHAIR compares model-generated captions with human annotations and provides 2 scores CHAIR_I and CHAIR_S to quantify object and sentence level hallucinations respectively:

$$\text{CHAIR}_S = \frac{|\{\text{captions with hallucinated objects}\}|}{|\{\text{all captions}\}|}, \quad \text{CHAIR}_I = \frac{|\{\text{hallucinated objects}\}|}{|\{\text{all objects mentioned}\}|}. \tag{2}$$

Table 15: **Evaluating object hallucinations**. We use CHAIR to study the extent of object hallucinations induced by MLE training and SR-L fine-tuning.

Dataset	Method	$CHAIR_I$	$CHAIR_S$	Hallucinated Objects	CIDEr
COCO	MLE	11.66	8.38	1255	107.8
	SR-L	9.79	7.03	1033	108.2
	SR-L 100 epochs	9.63	7.33	1102	92.4

MLE training encourages the captioner to generate generic phrases for different images that may not be aligned with the image content, resulting in increased object hallucinations.

SR-L fine-tuning may struggle with attribute binding, but it makes the generated captions more discriminant, resulting in a significant reduction in object hallucinations.

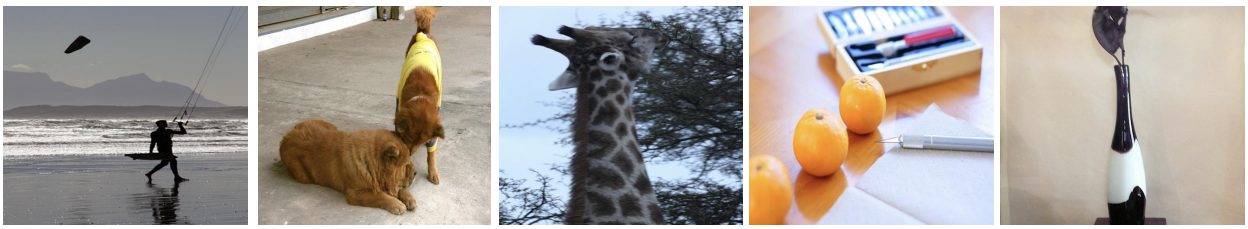
Extended fine-tuning with SR-L for 100 epochs in addition to reducing caption faithfulness, also increases both the absolute count and percentage of hallucinated objects.

C.8 Investigating Caption Unfaithfulness in SR Fine-tuning

To improve our understanding of the trade-off presented in Section 4.4 during SR fine-tuning, we further investigate the deterioration of the NLG metrics. Dessì et al. (2023) attribute this deterioration to the model diverging from the GT distribution during SR-L fine-tuning. They argue it to be desirable property as it allows the captioner to generalize to a wider range of captioning distributions. We investigate this claim by qualitatively analyzing image-caption pairs (COCO test set) where SR-L exhibits poor CIDEr scores compared to MLE. We find several image-caption pairs (some are shown in Figure 7) corroborating this claim, highlighting the inadequacies of reference-based metrics. For instance, a more discriminative caption "A woman serving a hot dog to a man" scores lower on CIDEr than a more generic one, "A couple of people standing in front of a food truck". However, our analysis also reveals that models fine-tuned on SR, in an attempt to enhance retrieval, have a propensity to hallucinate which results in poor CIDEr scores (see Figure 8). This differs from the findings of Section 4.3, since the model is incorporating non-factual information to the generated captions instead of semantic details. We observe this tendency under the same experimental setting as (Dessì et al., 2023), noting a modest dip in NLG metrics due to early stopping. However, we find this effect exacerbates with longer training durations, as illustrated by the dramatic drop in CIDEr scores in Figure 3.

cocoid: 482436	cocoid: 222757	cocoid: 558137	cocoid: 472216	cocoid: 10428
MLE: A man standing in front of a counter in a restaurant	**MLE**: A view of an empty room with a window	**MLE**: A couple of people that are standing in front of a food truck	**MLE**: A group of people riding ski down a snow covered slope	**MLE**: A bathroom with a sink and mirror
CIDEr: 56	**CIDEr:** 71	**CIDEr:** 109	**CIDEr:** 139	**CIDEr:** 170
SR: A black and white photo of people in a restaurant	**SR**: A room with a window and a black and white floor	**SR**: A woman serving a hot dog to a man	**SR**: A group of women cross country skiing in a competition	**SR**: A gallery filled with portraits of men and women
CIDEr: 0	**CIDEr:** 12	**CIDEr:** 11	**CIDEr:** 8	**CIDEr:** 0

Figure 7: **Low CIDEr due to generalization**. COCOIDs in test set, where COCO SR fine-tuned caption scores poorly on CIDEr compared to COCO MLE trained model due to generated captions being different from COCO's ground truth distribution.



cocoid: 503311	cocoid: 310757	cocoid: 186747	cocoid: 326344	cocoid: 375078
MLE: A man on a beach flying a kite	**MLE**: A brown dog standing next to a yellow dog	**MLE**: A tall giraffe standing next to a tree	**MLE**: A close up of an orange on a table	**MLE**: A pair of scissors sitting on top of a table
CIDEr: 209	**CIDEr:** 92	**CIDEr:** 107	**CIDEr:** 4	**CIDEr:** 13
SR: A man and a woman windsurfing in the ocean	**SR**: Two brown dogs standing next to yellow frisbee	**SR**: Two giraffes standing next to each other in a tree	**SR**: A pair of scissors and an orange on a table	**SR**: A black and white stylized black and white stylized stylized
CIDEr: 24	**CIDEr:** 14	**CIDEr:** 24	**CIDEr:** 1	**CIDEr:** 0

Figure 8: **Unfaithfulness through hallucination.** Some COCOIDs from the test set where SR fine-tuned captioner scores poorly on CIDEr compared to COCO MLE trained model due to nonfactual information.

cocoid: 536589

SR-L: A bicycle is parked next to a boat on a dock.

CIDEr: 77

SR-V: Two bicycles that are next to each other.

CIDEr: 0

SR-V curri: A bicycle and a car are displayed in a display case.

CIDEr: 12

cocoid: 72397

SR-L: A blue fire hydrant sitting on the side of a street.

CIDEr: 275

SR-V: A blue fire hydrant sitting on the side of a street.

CIDEr: 275

SR-V curri: A blue and blue blue blue and yellow blue and yellow blue and yellow blue and yellow blue…

CIDEr: 0

cocoid: 560756

SR-L: A living room filled with furniture and a couch.

CIDEr: 83

SR-V: A living room filled with red chairs and orange walls.

CIDEr: 13

SR-V curri: Three orange chairs sitting next to each other in a room.

CIDEr: 0

cocoid: 137727

SR-L: A blue and grey passenger train is parked on the tracks at a train station, with a crane nearby.
CIDEr: 22

SR-V: A blue and orange train sits on the tracks at a train station, ready for passengers to board
CIDEr: 7

SR-V curri: Two blue and orange train cars are parked next to each other on the train tracks.

CIDEr: 0

Figure 9: SR-V fine-tuned captioner struggles with attribute binding and hallucination. Training with BagCurri (SR-V curri) worsens this tendency.

Figure 10: **Manual filtering of bags in TrueMatch**. We manually remove bags that do not capture some aspect of fine-grained visual discrimination, even if they represent the same visual concept.