TROPE: TRaining-Free Object-Part Enhancement for Seamlessly Improving Fine-Grained Zero-Shot Image Captioning

Joshua Feinglass and **Yezhou Yang**
Arizona State University
{joshua.feinglass,yz.yang}@asu.edu

Abstract

Zero-shot inference, where pre-trained models perform tasks without specific training data, is an exciting emergent ability of large models like CLIP. Although there has been considerable exploration into enhancing zero-shot abilities in image captioning (IC) for popular datasets such as MSCOCO and Flickr8k, these approaches fall short with fine-grained datasets like CUB, FLO, UCM-Captions, and Sydney-Captions. These datasets require captions to discern between visually and semantically similar classes, focusing on detailed object parts and their attributes. To overcome this challenge, we introduce TRaining-Free Object-Part Enhancement (TROPE). TROPE enriches a base caption with additional object-part details using object detector proposals and Natural Language Processing techniques. It complements rather than alters the base caption, allowing seamless integration with other captioning methods and offering users enhanced flexibility. Our evaluations show that TROPE consistently boosts performance across all tested zero-shot IC approaches and achieves state-of-the-art results on fine-grained IC datasets[1].

1 Introduction

Object parts and their attributes have been shown to play a critical role in distinguishing between classes in tasks like fine-grained classification (Liu et al., 2024; Zhang and Feng, 2023; Feinglass et al., 2024). Despite their importance, previous works in image captioning (IC) have instead focused primarily on objects, their attributes, and their interactions, as seen in by common utilized semantic structures like scene graphs (Zhao et al., 2020; Zhang et al., 2022; Chen et al., 2020). This focus is partly because IC is often applied to general domain datasets like MSCOCO (Lin et al., 2014) and Flickr8k (Hodosh et al., 2013), where images typically contain various object classes, and captions provide a high-level scene description.

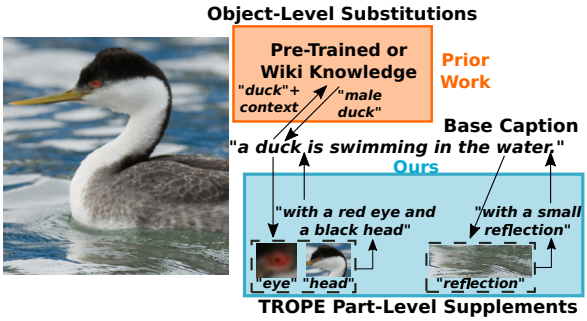

Figure 1: An example differentiating TROPE from prior work in image caption enhancement, which substitute existing words in the sentence with more contextually appropriate alternatives. TROPE instead inserts supplemental information after key objects by mapping nouns to a region of the image and constructing semantic part proposals based on object parts and attributes found within this region.

Existing training-free Zero-Shot methods, such as ZeroCap (Tewel et al., 2022) and ConZIC (Zeng et al., 2023), enhance captions by substituting words in a base caption with more a contextually appropriate ones, utilizing scores from CLIP (Radford et al., 2021) and a Large Language Model (LLM). However, since these models are pre-trained on general domain data, the resulting captions often lack fine-grained detail. This is especially problematic for fine-grained datasets, such as bird species in CUB (Welinder et al., 2010), flower species in FLO (Welinder et al., 2010), or aerial scenes in UCM-Captions (Yang and Newsam, 2010) and Sydney-Captions (Yang and Newsam, 2010), which require distinguishing between visually and semantically similar classes. Ren et al. (2023) showed that that visual and semantic feature representations from these fine-grained datasets differ significantly from those of general domain

[1]TROPE source codes and data: https://github.com/JoshuaFeinglass/TROPE.

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 3644–3655
November 12-16, 2024

datasets, leading to poor performance in domain generalization benchmarks due to task misalignment (Feinglass and Yang, 2024).

We conjecture that effective zero-shot IC in fine-grained contexts necessitates robust primitives that are consistent across both the training and test domains. Following this line of reasoning, we proposed TRaining-free Object-Part Enhancement (TROPE), which adapts pre-trained models to fine-grained datasets by supplementing captions with object part information. TROPE effectively augments the base captions of existing zero-shot IC methods with fine-grained details as shown in Figure 1. Our evaluations demonstrate that adding information from object part semantic proposals consistently enhances IC performance across all tested methods, datasets, and metrics. Precision-recall curves indicate that TROPE significantly improves recall with a minimal impact on precision, particularly in datasets where there is substantial overlap between the object detector's vocabulary and the terms commonly used by human annotators. We also present examples of TROPE's application to an enterprise captioner, GPT4, and discuss two failure cases: one involving a lack of recognizable objects and another featuring redundant or incorrect part information.

To further explore the bias of general domain datasets, we conducted an analytical study on the frequency of terms in human-annotated and machine-generated texts across both general domain and fine-grained datasets. We found that semantic indicator words, such as "with", "has", and "have", which introduce object part descriptions, are much more common in fine-grained datasets. This finding underscores the strong relationship between the semantic structure of images and the captions used to describe them, reinforcing the need to adapt models trained on general domain datasets to fine-grained settings using techniques like TROPE.

Contributions: Our work introduces the setting of fine-grained zero-shot captioning, extending zero-shot capabilities to four fine-grained captioning datasets. Our analyses reveal that existing zero-shot benchmarks cater predominantly to general domains and fail to meet the specific needs of fine-grained settings. We propose TROPE as a solution to enhance zero-shot captioning performance by incorporating detailed information from a pre-trained object detector, consistently enriching caption detail and improving performance across all methods, evaluation metrics, and datasets.

2 Related Work

The detection of objects and attributes, facilitated by large datasets of human-labeled regions (Krishna et al., 2017; Ramanathan et al., 2023), has historically been a cornerstone for various vision-language tasks (Zhang et al., 2021). Previous works have integrated this object and attribute information into training (Zhang et al., 2021), labels (Anderson et al., 2018), and text generation (Li et al., 2020). TROPE builds on this foundation by extracting hierarchical object relationships to improve the detail of image captions.

Enhancing the level of detail presented in image captions is a popular and multi-faceted topic. Entity-aware captioning seeks to replace generic nouns with context-specific entities from Wiki text based either on a base (template) caption (Lu et al., 2018; Biten et al., 2019; Jing et al., 2020) or optimized generation (Hu et al., 2020; Tran et al., 2020). Similarly, stylized captioning is also performed by either modifying a base caption (Zhao et al., 2020) or optimized generation (Yang and Jin, 2023). Lastly, scene graphs have been used to either enhance a base caption (Zhao et al., 2020) or as an additional feature for optimized generation (Zhang et al., 2022; Chen et al., 2020).

Zero-shot captioning presents distinct challenges, as it operates without direct access to image-text pairs for training, relying instead on the intrinsic capabilities of pre-trained models like CLIP (Radford et al., 2021) and SimCTG (Su et al., 2022b). Several works have tried to enhance zero-shot performance using training-free (Li et al., 2023a; Zeng et al., 2023) methods or text-only training (Su et al., 2022a; Li et al., 2023b; Tu et al., 2023; Nukrai et al., 2022; Fei et al., 2023) strategies that assume access to target dataset captions. Ren et al. (2023) introduced a domain generalization IC benchmark spanning general and fine-grained datasets, where a large gap in performance on general datasets and fine-grained datasets could be observed. TROPE aims to bridge the gap between general and fine-grained datasets by adding detailed object descriptions without requiring additional training data, improving the applicability and effectiveness of zero-shot captioning methods in more challenging environments.

3 Preliminaries

3.1 Image Captioning Task

The image captioning (IC) task involves an image captioner that takes an image as input and outputs a caption, typically a single sentence, that describes the image. The process begins with a vision module **V** that extracts features w from the image as a pre-processing step. This is followed by a cross-modal understanding module **VL**, which integrates the pre-processed image information to generate the caption y as shown

$$w = \mathbf{V}(image), \quad y = \mathbf{VL}(w). \tag{1}$$

3.2 Object Detectors in Image Captioning

n Vision-Language (VL) tasks such as IC, object detectors play a crucial role. These detectors are specialized to not only identify objects within an image but also provide detailed labels and attributes for these objects (Anderson et al., 2018; Zhang et al., 2021). The information about specific regions provided by these detectors is essential for many IC methods. For instance, Oscar (Li et al., 2020) model utilizes this detailed, region-specific data to facilitate cross-modal understanding when generating captions. Our work utilizes the object detector VinVL (Zhang et al., 2021), which provides bounding boxes b_r, regional features θ_r, object labels l_r^o, and attribute labels l_r^a (the most confident attribute label for an object) for all proposed regions of interest $r \in \mathcal{R}$ of an image

$$\{b_r, \theta_r, l_r^o, l_r^a\}_{r \in \mathcal{R}} = \text{VinVL}(image). \tag{2}$$

The integration of VinVL with the Oscar model is one of the approaches used in our work to generate base captions. We also select VinVL to serve as the source of object part information used by TROPE to enhance base captions because of its extremely large vocabulary of 1848, which encompasses objects present in all of the fine-grained datasets included in our benchmark.

3.3 Measures of Object Proposal Similarity

In assessing object proposals from VINVL, we focus on the spatial relationships and characteristics of the regions outlined by the bounding boxes. Operations such as intersection and union are used to evaluate the overlap between different bounding boxes, while the area of individual bounding boxes is calculated to assess their size. These measures help in determining the similarity and relevance of object proposals.

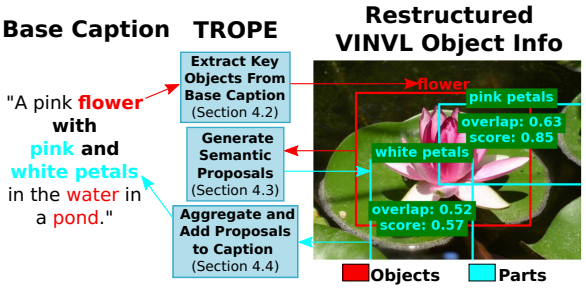

Figure 2: A high-level visualization of the TROPE methodology expanded upon in Algorithm 1. Detailed descriptions of each TROPE function block can be found in their corresponding sections.

4 Structured Enhancement

4.1 Overview

As depicted in Figure 2 and outlined in Algorithm 1, TROPE leverages raw object data from an object detector (VinVL) to enhance a base caption y with supplemental text to form y^+. The additions to the base caption include semantic part proposals consisting of an article (if the object is singular), part attributes, and a part descriptor (e.g., "pink and white petals"). These proposals are associated with objects mentioned in the base caption (e.g., birds or flowers) and are integrated using punctuation and connective phrases such as ",", "and", "with", and "in addition to".

4.2 Extract Key Objects from Base Caption (h)

Initially, key objects in the caption are identified. This involves extracting nouns and their corresponding caption indices from the base caption using tokenization and Parts-of-Speech (POS) tagging (Honnibal et al., 2020). For possessive phrases marked by "'s" or "of", only the possessing noun is considered, and the insertion index is set to the end of the phrase. As object detectors primarily recognize unigrams, compound nouns with space separation are not extracted. The identified nouns and indices are then matched with object labels and bounding boxes from VinVL, with plural nouns reverted to their singular form using Inflect (Paul Dyson, 2024) to ensure consistent matching. The bounding box with the highest confidence is used for singular nouns, and the smallest box encompassing all relevant bounding boxes is used for plural nouns, resulting in bounding boxes R^O and and corresponding caption indices I^O for all key objects.

4.3 Generate Semantic Proposals (g)

The next process involves assigning smaller objects (indexed by r in Algorithm 1) as parts of the larger key objects (indexed by k). Objects are sorted by the area of their bounding boxes, from smallest to largest, to prioritize smaller key objects and avoid assigning everything to large background objects like the "sky". Object proposals that overlap (defined as $area(b_k \cap b_r)/area(b_k)$ in Algorithm 1) significantly with a key object (exceeding a predefined threshold $T = 0.5$ based on Table 4) are assigned as parts. If multiple parts share the same label and attribute, the label is pluralized using Inflect. Semantic part proposals P are created by appending the attribute and part labels, prefacing with an article if the part is singular. Proposals in the list for an object S_k are then ranked by adding overlap percentage and detector confidence (based on the ablation in Table 2), determining their inclusion order.

4.4 Aggregate and Add Semantic Proposals to Caption (m)

Redundant proposals, such as "white petals" and "pink petals", are organized coherently. Matching part labels with different attributes are combined using commas or "and" (e.g., "white and pink petals"). These semantic proposals are then inserted into the base caption at the identified object indices I_k^O. The number of proposals included per object is based on a user-defined parameter N. Proposals are introduced by "with", the most frequent semantic indicator from our word frequency study in Section 5.4. If an object's description already includes "with", "in addition to" is used to prepend the new proposals, ensuring a cohesive augmentation of the existing caption (e.g., "a flower with white and pink petals in addition to green leaves").

5 Experiments

To validate the effectiveness of TROPE for enhancing the detail of generated image captions, we conduct experiments on our proposed fine-grained IC benchmark in Section 5.3, where TROPE demonstrates consistent improvement in fine-grained image captioning performance for standard metrics. To further motivate the use of TROPE, we then explore the bias of general domain datasets in a word frequency study in Section 5.4.

Algorithm 1 High-Level TROPE Pseudocode

Input: y, $\{b_r, l_r^o, l_r^a\}_{r \in \mathcal{R}}$,$N$
Output: y^+ ▷ Enhanced caption
  $(R^O, I^O) = h(y)$ ▷ Section 4.2
  for $k \leftarrow 1$ to $length(I^O)$ **do**
    $S_k \leftarrow \emptyset$
    for $r \in (\mathcal{R} - R^O)$ **do**
      $S_k \leftarrow \{\}$
      if $[area(b_k \cap b_r)/area(b_k)] > T$ **then**
        $P \leftarrow g(l_r^o, l_r^a)$ ▷ Section 4.3
        S_k.append(P)
      end if
    end for
  end for
  Sort I^O in reverse order
  Apply the same reordering to S_k
  $y^+ \leftarrow y$
  for $k \leftarrow 1$ to $length(I^O)$ **do**
    if $S_k \neq \emptyset$ **then**
      $y^+ \leftarrow m(S_k, I_k^O, y^+, N)$ ▷ Section 4.4
    end if
  end for

5.1 Datasets

CUB or the Caltech-UCSD Birds (Welinder et al., 2010) dataset is a very popular benchmark for fine-grained classification and contains 200 classes of bird species (bobolink, cardinal, etc.). We use the 5,794 image test set with 10 captions for each image annotated by Reed et al. (2016) for our benchmark.

FLO or the Oxford Flowers (Nilsback and Zisserman, 2008) is another popular benchmark for fine-grained classification and contains 102 classes of flower species (moon orchid, snapdragon, etc.). We use the 6,149 image test set with 10 captions for each image annotated by Reed et al. (2016) for our benchmark.

SC or the Sydney Captions (Zhang et al., 2015) dataset consists of 7 land-use classes (residential, airport, etc.). We use the 58 image test set with 5 captions for each image annotated by Qu et al. (2016) for our benchmark.

UCM or the UC Merced Land Use (Yang and Newsam, 2010) dataset consists of 21 land-use classes (agricultural, harbor, etc.). We use the 210 image test set with 5 captions for each image annotated by Qu et al. (2016) for our benchmark.

MSCOCO or the Microsoft Common Objects in Common Context (Lin et al., 2014) dataset which is comprised of curated images containing

	Method	CUB C	CUB M	CUB SP	CUB SM	FLO C	FLO M	FLO SP	FLO SM	UCM C	UCM M	UCM SP	UCM SM	SC C	SC M	SC SP	SC SM
Domain Gen.	Up-Down	3.70	7.99	14.56	-	9.04	8.07	12.38	-	-	-	-	-	-	-	-	-
	AoANet	4.84	8.58	15.47	-	10.29	7.61	11.92	-	-	-	-	-	-	-	-	-
	M^2Trans.	7.78	8.68	15.17	-	11.12	8.28	13.95	-	-	-	-	-	-	-	-	-
	EISNet	6.83	8.82	15.20	-	11.38	8.62	12.52	-	-	-	-	-	-	-	-	-
	LSML	9.60	10.24	15.72	-	14.35	9.72	**15.23**	-	-	-	-	-	-	-	-	-
Zero-Shot	ZeroCap	0.33	4.75	0.25	-1.14	0.47	5.12	0.34	-1.13	0.59	5.39	1.84	-1.13	0.32	4.31	0.60	-1.14
	ConZIC	10.30	10.17	2.02	0.41	15.07	11.32	3.19	0.78	7.54	6.51	2.88	-0.27	12.33	7.62	3.48	-0.05
	1 part	14.89	12.34	3.27	0.50	23.83	13.26	5.00	0.78	**8.08**	7.12	3.42	-0.22	**13.40**	**7.83**	3.78	**0.00**
	5 parts	7.21	14.31	5.80	0.58	16.21	14.06	5.60	0.81	6.64	7.16	3.36	**-0.21**	13.25	7.82	3.73	**0.00**
	10 parts	5.73	14.10	6.56	0.55	16.16	14.05	5.62	0.81	6.64	7.15	3.35	**-0.21**	13.25	7.82	3.73	**0.00**
	Oscar	29.63	15.52	6.40	0.26	41.56	15.15	7.67	0.32	7.95	7.31	4.85	-0.83	5.12	6.19	2.62	-0.82
	1 part	**50.16**	21.36	10.52	0.72	**68.28**	19.66	12.59	0.84	7.67	8.08	5.39	-0.73	6.71	7.06	2.78	-0.72
	5 parts	11.00	**25.47**	17.39	**1.00**	44.16	**21.26**	**14.03**	**0.99**	4.31	**8.54**	**5.55**	-0.58	5.55	7.69	4.01	-0.41
	10 parts	4.06	24.78	**19.44**	0.95	44.00	21.24	14.02	**0.99**	4.30	8.53	5.53	-0.58	5.55	7.68	**4.08**	-0.38

Table 1: A fine-grained IC benchmark comparing the performance of domain generalization models from Ren et al. (2023) including: Up-Down (Anderson et al., 2018), AoANet (Huang et al., 2019), M^2Transformer (Cornia et al., 2020), EISNet (Wang et al., 2020), and LSML (Ren et al., 2023), zero-shot IC methods including: ZeroCap (Tewel et al., 2022), ConZIC (Zeng et al., 2023), and Oscar (Li et al., 2020), and TROPE based enhancements of select zero-shot IC methods with varying numbers of semantic part proposals. Enhancements provided by TROPE are denoted with ().

	Criteria	CUB M	CUB SP	FLO M	FLO SP
1 part	Score	**21.42**	10.49	19.41	12.20
	Overlap	20.49	9.39	19.62	**12.59**
	Score+Overlap	21.36	**10.52**	**19.66**	**12.59**
5 parts	Score	25.32	17.08	21.24	14.01
	Overlap	24.97	16.53	21.21	14.00
	Score+Overlap	**25.47**	**17.39**	**21.26**	**14.03**

Table 2: An ablation study of the performance of different criteria for selecting proposals. Adding the object score (detector confidence) and overlap (from detector bounding boxes) yields the best captioning results.

	Component	CUB M	CUB SP	FLO M	FLO SP
1 part	Descriptor	18.71	6.8	16.25	7.72
	Part	17.74	9.31	16.44	10.56
	Both	**21.36**	**10.52**	**19.66**	**12.59**
5 parts	Descriptor	22.82	7.03	18.37	8.11
	Part	19.77	10.4	16.89	10.21
	Both	**25.47**	**17.39**	**21.26**	**14.03**

Table 3: An ablation study of the impact on performance when including only the descriptor or part component of the semantic proposal. The results suggest that METEOR is more sensitive to descriptors, while SPICE is more sensitive to object parts.

80 common object classes (like 'human' or 'truck') with 5 human annotated captions for each image. **Flickr8k** (Hodosh et al., 2013) is a popular dataset comprised of 8000 images with 5 human annotated captions for each image. The images were crawled from social media postings and like MS-COCO, are primarily common objects.

5.2 Evaluation Metrics

We utilize the four rule-based caption evaluation specific metrics which exhibit high agreement with human judgement across all commonly reported benchmarks: CIDEr (C) (Vedantam et al., 2015), METEOR (M) (Banerjee and Lavie, 2005), SPICE (SP) (Anderson et al., 2016), and SMURF (SM) (Feinglass and Yang, 2021). For all utilized metrics, a larger value indicators better performance with all metrics aside from SMURF varying within the range 0 to 1. SMURF is standardized to human performance, meaning a value of 0 is on par with human captions and negative or positive values indicate worse or better performance than humans, respectively. We exclude BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) since they exhibit very poor agreement with human judgement in caption evaluation (Anderson et al., 2016; Feinglass and Yang, 2021) and also do not consider metrics like CLIPScore (Hessel et al., 2021) which are exclusively referenceless since they are likely to be sensitive to domain shift.

Threshold	CUB M	CUB SP	FLO M	FLO SP
0.25	21.2	10.35	**19.72**	12.47
0.50	**21.36**	**10.52**	19.66	**12.59**
0.75	21.3	10.43	19.37	**12.59**

Table 4: An ablation study of the threshold T used to assign parts to each object based on overlap for 1 part proposal. Although TROPE's performance does not seem to be strongly impacted by changes to T, a setting of $T = 0.5$ exhibits the highest performance the majority of the time for the tested datasets and parameters.

5.3 Fine-Grained Captioning Benchmark

Table 1 shows a comparison between base captioners enhanced using TROPE and relevant base-

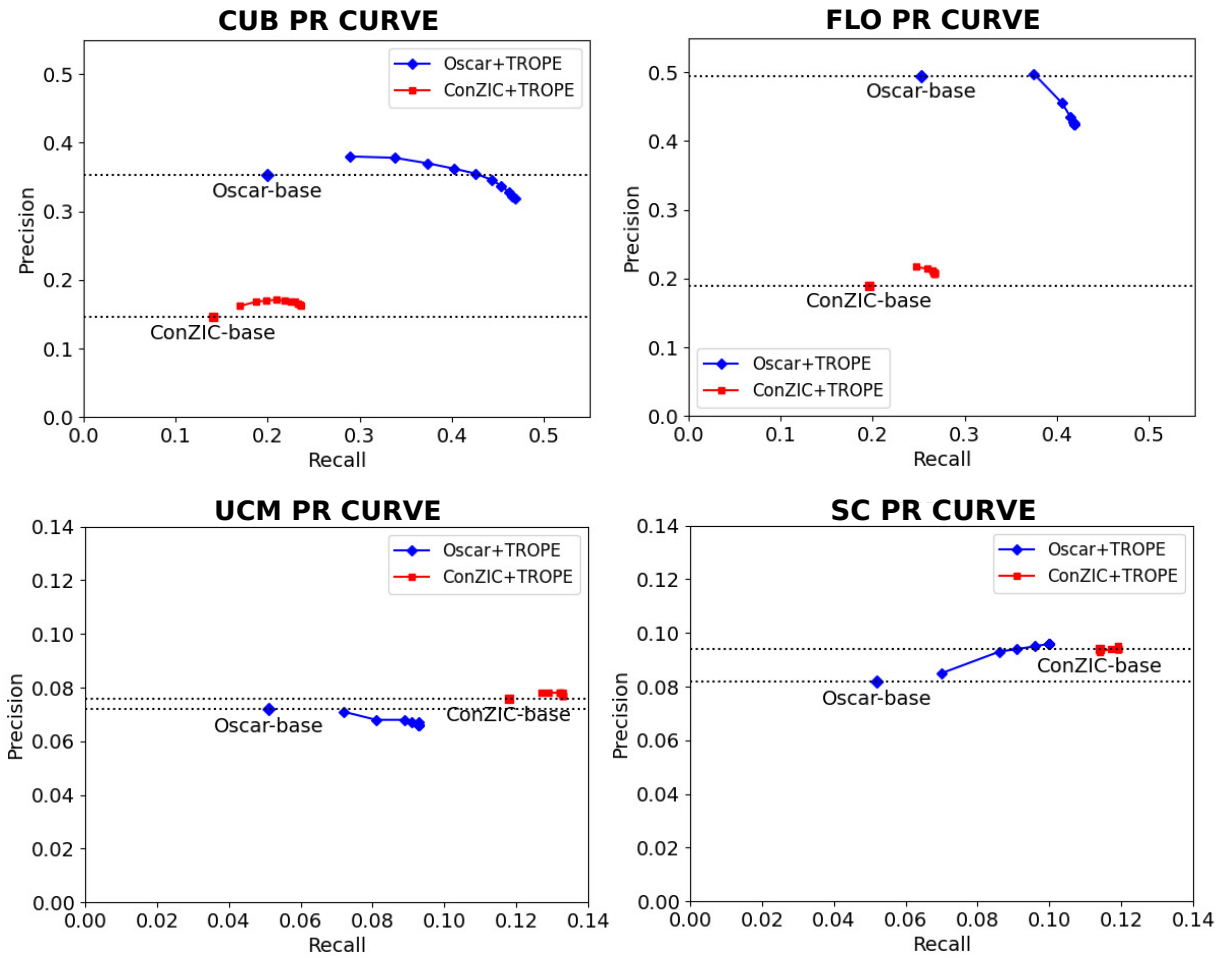

Figure 3: Precision-recall curves generated by sweeping the number of semantic proposals added to the base caption from 1 to 10 for both the Oscar and ConZIC base captions. Horizontal lines represent the base caption precision performance.

lines. Zero-shot IC methods ZeroCap (Tewel et al., 2022) and ConZIC (Zeng et al., 2023) utilizing pre-trained models CLIP (Radford et al., 2021), BERT (Devlin et al., 2019), and GPT2 (Radford et al., 2019) as well as the Oscar (Li et al., 2020) model utilizing VinVL (Zhang et al., 2021) features are included since they are publicly available and achieve state-of-the-art results on MSCOCO (Lin et al., 2014) and Flickr8k (Hodosh et al., 2013). Domain generalization model results reported by Ren et al. (2023) for CUB and FLO are also included in the benchmark because although they are not publicly available and train across four separate captioning sets, they still do not have access to target domain captions, making the setting zero-shot. The two most competitive and publicly available base captioners, Oscar and ConZIC, are selected for enhancement by TROPE. To aid in the design of TROPE, ablation studies utilizing the VinVL+Oscar pipeline shown in Tables 2, 3, and 4 explore the impact of the proposal selection criteria, semantic components, and overlap threshold T on captioning performance, respectively. To better show the trend of TROPE's performance for each additional proposal added to the base caption, we derive a precision and recall metrics from the SPARCS (SMURF's state-of-the-art semantic score) and use these metrics to generate precision and recall curves for each dataset shown in Figure 3.

In general, the two state-of-the-art zero-shot IC methods achieved poor results compared to the rest of captioners. Zero-Cap in particular generated captions with little diversity and almost no relevance to the images. ConZIC performed significantly better, especially on aerial images when compared with Oscar, achieving the highest base model score for SMURF on UCM and the highest base model scores across all standard metrics on SC. Oscar achieved highly competitive results, especially on

Figure 4: Qualitative examples of TROPE applied to captions generated by GPT4 (Achiam et al., 2023) with $N = 5$ semantic part proposals. Minor failures can be observed in the 2nd image caption with erroneous attributes like "green" house and redundant parts like "house" and "building". The 3rd caption is another failure case where no supplemental information from TROPE is added to the caption since the base caption contains no key objects detected by VinVL.

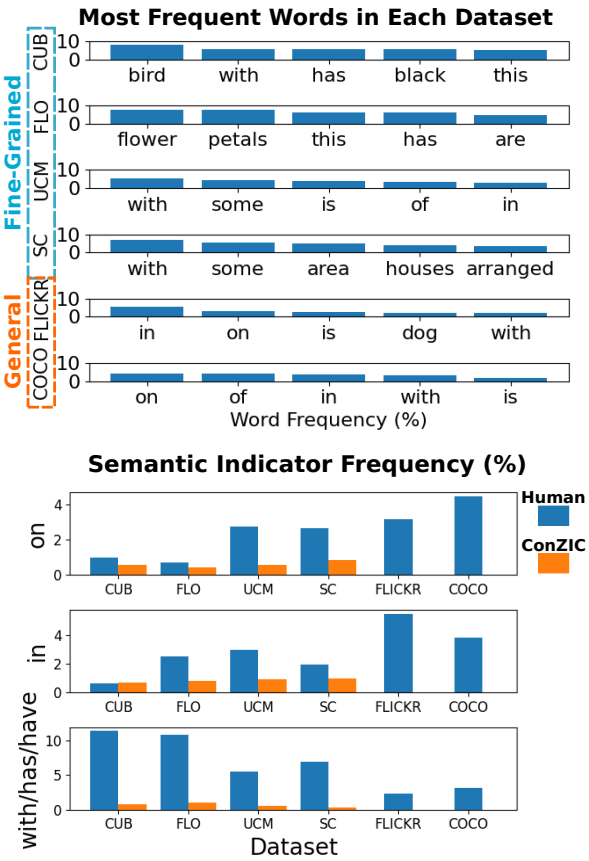

Figure 5: Visualizations showcasing the unique characteristics of fine-grained datasets. The top plot shows the frequency of the 5 most common terms in our selected fine-grained and general domain datasets. The bottom plot shows the frequency of different semantic indicators across our selected datasets for both human annotations and available base captions from ConZIC.

CUB and FLO, where it was the highest performing base model for all reported metrics except SMURF and SPICE. Inference for both ConZIC and ZeroCap is extremely slow, taking more than a day to generate captions for the benchmark compared to the VinVL+Oscar pipeline which took a few hours.

Performance achieved by the Oscar and ConZIC increased significantly across all standard metrics, datasets, and tested models after adding 1 semantic part proposal. This can be attributed to a large jump in recall performance across all datasets in Figure 3, which then increases less significantly with each additional semantic part proposal added. Conversely, precision typically changes slightly with the first proposal, then decreases at an increasing rate with each additional proposal, with SC as a notable exception. These findings are discussed further in Section 6. Although enterprise models like GPT4 (Achiam et al., 2023) are not included in the benchmark due to cost and rate limitations, we show 3 examples of TROPE integrated with GPT4 in Figure 4, with 2 examples demonstrating improvements in caption detail and 1 example demonstrating a failure case with no change to the base caption.

5.4 Word Frequency Study

To explore the distinctive features of fine-grained captions, we analyzed the word frequency statistics of the training sets of our selected fine-grained image captioning datasets alongside general domain datasets such as MSCOCO and Flickr8k. The findings, illustrated in Figure 5, reveal distinct linguistic patterns between these dataset categories. Words that serve as semantic indicators of object-to-object interactions, such as 'on' and 'in', appear with greater frequency in general domain datasets.

In contrast, words that indicate object-part descriptions, like 'with', 'has', and 'have', are more prevalent in fine-grained datasets.

This variation in word usage underscores the unique requirements of fine-grained captioning, which often necessitates detailed descriptions of object parts and attributes. The state-of-the-art zero-shot IC captioning method, ConZIC, exhibits a notable deficiency in incorporating these semantic indicators for object-part descriptions, which likely contributes to its underperformance in fine-grained tasks. This observation supports our hypothesis that effective fine-grained captioning relies heavily on the precise depiction of object parts and attributes.

Moreover, the word frequencies highlight that the granularity of fine-grained properties exists on a spectrum. Datasets like CUB and FLO, which typically feature a single salient object such as a "bird" or "flower", exhibit a high degree of specificity. Aerial datasets like UCM and SC, however, occupy a middle ground between general domain and fine-grained datasets. Although these datasets may include dominant objects like "airport" or "ocean", they lack the intense focus on singular objects characteristic of the most fine-grained datasets. This spectrum of granularity provides further context for tailoring image captioning approaches to suit the specific demands of different dataset types.

6 Discussion

Based on our results and the apparent spectrum of fine-grained dataset characteristics, TROPE's effectiveness appears widely applicable to numerous image datasets. However, its performance varies depending on each dataset's structure. A significant factor influencing TROPE's success is the alignment between the common terminology used by human captioners and the vocabulary of the object detector employed. For instance, while VinVL effectively covers common terms related to bird parts (e.g., head, tail, wing), flower parts (e.g., petal, leaf), and aerial views (e.g., airplane, airport), it lacks specialized terms frequently used in flower descriptions (e.g., stamen, pistil, veins). This gap is notable in our results: TROPE shows state-of-the-art performance on CUB, UCM, and SC, but somewhat underperforms in the FLO dataset compared to domain generalization techniques, particularly as additional part proposals are integrated, which dramatically affects precision.

Furthermore, as the precision of object detectors improves, we anticipate that methods like TROPE will yield even greater improvements in image captioning performance. TROPE's strength lies in significantly boosting recall with minimal reductions in precision. In cases of poor detector performance, the typical outcome is no change to the base caption, whereas a mismatch between the detector's vocabulary and human captions can lead to redundant or irrelevant descriptions, thereby decreasing precision.

Our analysis also indicates that different caption evaluation metrics prioritize different aspects of TROPE's semantic components (see Table 3) and precision-recall performance curve. METEOR, SPICE, and SMURF achieve their highest scores with the incorporation of five additional parts per object, suggesting a preference for detailed content. Conversely, CIDEr peaks with just one additional part, likely because it penalizes excessive wordiness beyond the average reference caption length, which may not suit fine-grained captioning settings where detailed descriptions are crucial.

Considering these insights, the optimal number of semantic part proposals to add to a base caption depends on the specific needs and goals of the research. For applications requiring high accuracy, such as assistive technologies, we recommend adding only a single proposal. Conversely, for purposes like training generative models or enhancing retrieval systems, incorporating multiple proposals may be beneficial as it enhances the discriminative information available, despite the potential for introducing irrelevant details. Researchers should select evaluation metrics that best align with their objectives and tailor their approach accordingly.

7 Conclusion and Broader Impact

We have introduced TROPE, a training-free method for zero-shot captioning that enhances base captions by adding semantic part proposals to key object instances. This approach has demonstrated state-of-the-art performance in fine-grained zero-shot image captioning (IC), consistently improving captions across all tested models, metrics, and datasets. Given the foundational role of IC in a variety of Vision-Language tasks, TROPE holds potential for enhancing fine-grained performance in applications such as text-to-image generation, text-to-image retrieval, and image-to-text retrieval. Future work could also focus on extending the prin-

ciples underlying TROPE to other modalities, such as audio or video. This would involve adapting TROPE to work with relevant pre-trained models tailored to these modalities, potentially opening new avenues for multimodal integration and captioning enhancements.

8 Limitations

Because TROPE relies on inferences from pre-trained IC models, domains where these pre-trained models have little familiarity with the constituent objects, parts, and terminology like medical imagery are likely to yield very poor zero-shot IC results. These limitations are also applicable to the other training-free baselines presented in this work and could possibly be mitigated with domain-specific human annotation as explored in few-shot or text-based training methods. For high-risk applications, practitioners should examine the overlap between the utilized detectors vocabulary and objects commonly present in the target domain. In such applications, including more than a single semantic part proposal should only be considered if this overlap is high, which reduces the risk of decreasing base caption precision.

9 Ethics Statement

Bias in pre-trained IC models (Rohrbach et al., 2018; Mehrabi et al., 2019) is a concerning challenge for researchers that can potentially impact gender and racial inclusion (Hendricks et al., 2018). Zero-shot settings are especially susceptible to carrying over bias from the training dataset since no test set data is available. The use of object detector-based primitives in zero-shot settings could be a promising avenue for mitigating bias in a concise and explainable manner. TROPE has the potential to improve the diversity of generated captions and models trained using those captions. This in turn could improve the inclusion of different genders and races.

10 Acknowledgements

TROPE is supported by NSF Robust Intelligence program grants #1750082 and #2038666. The authors acknowledge technical access (through ASU-OpenAI collaboration) and support from ASU Enterprise Technology. The views and opinions of the authors expressed herein do not necessarily state or reflect those of the funding agencies and employers.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *ECCV*.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

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.

Ali Furkan Biten, Lluis Gomez, Marçal Rusinol, and Dimosthenis Karatzas. 2019. Good news, everyone! context driven entity-aware captioning for news images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12466–12475.

Shizhe Chen, Qin Jin, Peng Wang, and Qi Wu. 2020. Say as you wish: Fine-grained control of image caption generation with abstract scene graphs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9962–9971.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Junjie Fei, Teng Wang, Jinrui Zhang, Zhenyu He, Chengjie Wang, and Feng Zheng. 2023. Transferable decoding with visual entities for zero-shot image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3136–3146.

Joshua Feinglass, Jayaraman J. Thiagarajan, Rushil Anirudh, T.S. Jayram, and Yezhou Yang. 2024. 'eyes of a hawk and ears of a fox': Part prototype network for generalized zero-shot learning. In *Proceedings*

of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pages 7791–7798.

Joshua Feinglass and Yezhou Yang. 2021. SMURF: SeMantic and linguistic UndeRstanding fusion for caption evaluation via typicality analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2250–2260, Online. Association for Computational Linguistics.

Joshua Feinglass and Yezhou Yang. 2024. Towards addressing the misalignment of object proposal evaluation for vision-language tasks via semantic grounding. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4397–4407.

Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. 2018. Women also snowboard: Overcoming bias in captioning models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 771–787.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*.

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Anwen Hu, Shizhe Chen, and Qin Jin. 2020. Icecap: information concentrated entity-aware image captioning. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4217–4225.

Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. 2019. Attention on attention for image captioning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4634–4643.

Yun Jing, Xu Zhiwei, and Gao Guanglai. 2020. Context-driven image caption with global semantic relations of the named entities. *IEEE Access*, 8:143584–143594.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.

Wei Li, Linchao Zhu, Longyin Wen, and Yi Yang. 2023a. Decap: Decoding clip latents for zero-shot captioning via text-only training. *arXiv preprint arXiv:2303.03032*.

Wei Li, Linchao Zhu, Longyin Wen, and Yi Yang. 2023b. Decap: Decoding CLIP latents for zero-shot captioning via text-only training. In *The Eleventh International Conference on Learning Representations*.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

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

Man Liu, Chunjie Zhang, Huihui Bai, and Yao Zhao. 2024. Part-object progressive refinement network for zero-shot learning. *IEEE Transactions on Image Processing*, 33:2032–2043.

Di Lu, Spencer Whitehead, Lifu Huang, Heng Ji, and Shih-Fu Chang. 2018. Entity-aware image caption generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4013–4023, Brussels, Belgium. Association for Computational Linguistics.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*.

Maria-Elena Nilsback and Andrew Zisserman. 2008. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE.

David Nukrai, Ron Mokady, and Amir Globerson. 2022. Text-only training for image captioning using noise-injected CLIP. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4055–4063, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Paul Dyson. 2024. Inflect.

Bo Qu, Xuelong Li, Dacheng Tao, and Xiaoqiang Lu. 2016. Deep semantic understanding of high resolution remote sensing image. In *2016 International Conference on Computer, Information and Telecommunication Systems (CITS)*, pages 1–5.

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

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Vignesh Ramanathan, Anmol Kalia, Vladan Petrovic, Yi Wen, Baixue Zheng, Baishan Guo, Rui Wang, Aaron Marquez, Rama Kovvuri, Abhishek Kadian, et al. 2023. Paco: Parts and attributes of common objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7141–7151.

Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. 2016. Learning deep representations of fine-grained visual descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 49–58.

Yuchen Ren, Zhendong Mao, Shancheng Fang, Yan Lu, Tong He, Hao Du, Yongdong Zhang, and Wanli Ouyang. 2023. Crossing the gap: Domain generalization for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2871–2880.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, Brussels, Belgium. Association for Computational Linguistics.

Yixuan Su, Tian Lan, Yahui Liu, Fangyu Liu, Dani Yogatama, Yan Wang, Lingpeng Kong, and Nigel Collier. 2022a. Language models can see: Plugging visual controls in text generation. *arXiv preprint arXiv:2205.02655*.

Yixuan Su, Tian Lan, Yan Wang, Dani Yogatama, Lingpeng Kong, and Nigel Collier. 2022b. A contrastive framework for neural text generation. In *Advances in Neural Information Processing Systems*.

Yoad Tewel, Yoav Shalev, Idan Schwartz, and Lior Wolf. 2022. Zerocap: Zero-shot image-to-text generation for visual-semantic arithmetic. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17918–17928.

Alasdair Tran, Alexander Mathews, and Lexing Xie. 2020. Transform and tell: Entity-aware news image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13035–13045.

Haoqin Tu, Bowen Yang, and Xianfeng Zhao. 2023. Zerogen: Zero-shot multimodal controllable text generation with multiple oracles. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 494–506. Springer.

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

Shujun Wang, Lequan Yu, Caizi Li, Chi-Wing Fu, and Pheng-Ann Heng. 2020. Learning from extrinsic and intrinsic supervisions for domain generalization. In *European Conference on Computer Vision*, pages 159–176. Springer.

Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. 2010. Caltech-ucsd birds 200. Technical Report CNS-TR-201, Caltech.

Dingyi Yang and Qin Jin. 2023. Attractive storyteller: Stylized visual storytelling with unpaired text. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11053–11066, Toronto, Canada. Association for Computational Linguistics.

Yi Yang and Shawn Newsam. 2010. Bag-of-visual-words and spatial extensions for land-use classification. GIS '10, page 270–279, New York, NY, USA. Association for Computing Machinery.

Zequn Zeng, Hao Zhang, Ruiying Lu, Dongsheng Wang, Bo Chen, and Zhengjue Wang. 2023. Conzic: Controllable zero-shot image captioning by sampling-based polishing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23465–23476.

Chunhui Zhang, Chao Huang, Youhuan Li, Xiangliang Zhang, Yanfang Ye, and Chuxu Zhang. 2022. Look twice as much as you say: Scene graph contrastive learning for self-supervised image caption generation. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, page 2519–2528, New York, NY, USA. Association for Computing Machinery.

Fan Zhang, Bo Du, and Liangpei Zhang. 2015. Saliency-guided unsupervised feature learning for scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 53(4):2175–2184.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5579–5588.

Yang Zhang and Songhe Feng. 2023. Enhancing domain-invariant parts for generalized zero-shot learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6283–6291.

Wentian Zhao, Xinxiao Wu, and Xiaoxun Zhang. 2020. Memcap: Memorizing style knowledge for image captioning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):12984–12992.