Towards Addressing the Misalignment of Object Proposal Evaluation for Vision-Language Tasks via Semantic Grounding

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

Abstract

Object proposal generation serves as a standard pre-processing step in Vision-Language (VL) tasks (image captioning, visual question answering, etc.). The performance of object proposals generated for VL tasks is currently evaluated across all available annotations, a protocol that we show is "misaligned" - higher scores do not necessarily correspond to improved performance on downstream VL tasks. Our work serves as a study of this phenomenon and explores the effectiveness of semantic grounding to mitigate its effects. To this end, we propose evaluating object proposals against only a subset of available annotations, selected by thresholding an annotation importance score. Importance of object annotations to VL tasks is quantified by extracting relevant semantic information from text describing the image. We show that our method is consistent and demonstrates greatly improved alignment with annotations selected by image captioning metrics and human annotation when compared against existing techniques. Lastly, we compare current detectors used in the Scene Graph Generation (SGG) benchmark as a use case, which serves as an example of when traditional object proposal evaluation techniques are misaligned[1].

1. Introduction

Vision-Language (VL) tasks are a growing topic in both the Natural Language Processing (NLP) and Computer Vision communities with the majority of techniques relying on object proposal generation for pre-processing [2, 51]. Object proposals are a set of regions or bounding boxes deemed likely to contain the object specified by a detector. Object proposal generation offers an explainable, efficient, and highly effective bridge between raw images and VL tasks.

However, current evaluation techniques of object proposal generation are poorly aligned with the VL use-case, resulting in adverse effects like "gameability" [8]. While [8] claim that this misalignment is caused by missing annotations, we theorize that the inclusion of superfluous object annotations not relevant to VL tasks in evaluation is also a contributing factor. Contrary to the prevailing attitude that models should be evaluated across all available annotations, we postulate that models only need information about a few critical objects to understand a scene. This intuition aligns well with the idea that not all test examples are equally

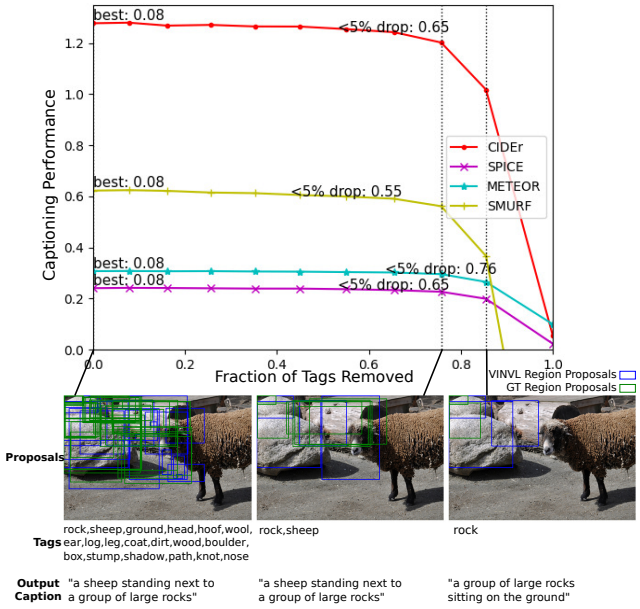

Figure 1. A system-level plot showing the performance of OSCAR [21] as tags and their corresponding image features are removed based on our proposed annotation importance scores. Captions are evaluated using standard metrics [1, 4, 12, 42] with all punctuation removed. The text on the plot shows at what fraction of tag and corresponding bounding box feature removal the metrics achieve their best score and the highest fraction of tags that can be removed before the performance drops by more than 5% for each metric. The results suggest that model performance depends on a critical subset of object regions.

[1]Source codes, data, and surveys will be released at https://github.com/JoshuaFeinglass/VL-detector-eval.

important for evaluation [33], which is rapidly gaining traction in NLP benchmarks and benefits not just evaluation but data annotation as well. Thus, we propose selecting ground truth annotations for use in evaluation based on a semantic grounding signal, specifically image captions or region descriptions. To measure the importance of a given object, we extract relevant semantic information using typicality analysis [12, 25] and propagate this importance to adjacent objects using graph signal processing techniques [10]. This importance score is then used to select only the objects most relevant to VL tasks for evaluation. Exploring image captioning as a case study of this phenomenon, we observe that *high image captioning performance can be maintained with only 24%-44% of object tags and their corresponding features from regions of interest depending on the performance metric* as shown in Figure 1. Furthermore, the preservation of a high SMURF [12] score suggests that removing these annotations does not significantly impact the detail/diversity of the generated captions.

Due to the scarcity of relevant detectors and lack of related benchmarks, we opt for a holistic approach when validating our metric. We perform three independent studies, each of which provides unique insight into the effectiveness and advantages of our approach. We begin with an empirical analysis and find that annotations selected using our importance scoring result in the highest alignment with improved image captioning performance for a widely adopted captioning pipeline when compared to an area-based baseline. To get an example-level view, we perform three human surveys using Amazon Mechanical Turk (AMT) and find that our proposed metric adjustments are highly aligned with human judgement while most of the existing metrics exhibit little to no alignment with VL bias. We then show that our selections are consistent across human-annotated text descriptions from different datasets, in particular, COCO captions and Visual Genome region descriptions. Our findings support the existence of a critical annotation set which remains consistent when using different semantic grounding sources. Furthermore, in our last experiment, we explore a Scene Graph Generation (SGG) use case where our approach provides information about model performance not captured in previous benchmarks. More specifically, we observe an instance where the standard evaluation approach fails to capture poor precision performance on VL task essential objects due to its misalignment.

Contributions: We create a theoretical formulation of misalignment in object proposal evaluation and develop an object importance score which can be used to mitigate the effect of this phenomenon and enhance the feedback provided when designing VL detectors. To support these insights, we perform 4 experiments: an analysis of the alignment between our importance score and performance on a downstream VL task, 3 human surveys, a study of the consistency between selected object regions from different annotation sets, and a demonstration of mitigated misalignment on a SGG benchmark.

2. Related Work

Object Proposal Evaluation relies primarily on variations of mean Average Precision (mAP) [9, 11, 26, 32], although Average Recall (AR) and mAR (mean Average Recall) are employed for evaluation benchmarks of related tasks like SGG [44]. These methods are considered to be intuitive and are not validated against human judgement. There are no existing benchmarks or metrics for VL task related object detection, despite the existence of benchmarks for other subtasks like salient object detection [6]. Thus, our work is the first to introduce such a benchmark. There are however numerous scene-oriented object detectors developed via pretraining [2, 14, 39, 48, 50, 51], with Visual Genome (VG) [19] serving as the standard dataset.

Scene Understanding tasks including scene representation and scene recognition rely largely on supervisory signals such as object segmentation and labels, which can be erroneous or incomplete [23]. Previous works have also shown that human captions and text alone can serve as a strong supervisory signal for object detection [17, 45, 46] and Visual Question Answering (VQA) [3]. [29] sought to find the minimum set of objects needed for the task of scene recognition. Objects relations have also been shown to be important for scene representation and recognition [37], with graph-based methods achieving significant success [5]. Dataset filtering [7, 47] has also explored the use of supervisory signals for data example selection from an inference perspective. Our work combines these scattered concepts into a single coherent formulation of scene-oriented bias for evaluation.

Annotation Weighting is gaining popularity with [33] and [24] asserting that each test example is not equally informative for evaluation benchmarks and that quantifying this importance can improve annotation and help detect overfitting. In particular, Item Response Theory (IRT) is a test example selection and weighting mechanism gaining popularity in Natural Language Processing benchmarks [20, 35, 41] which seeks to provide greater rewards for more difficult text examples and has been shown to be more reliable and representative than standard accuracy. Rather than selecting and weighting examples based on difficulty, our work instead focuses on selecting test examples based on their relevance to a task of interest.

3. Our Approach

3.1. Vision-Language Task Background

An object proposal based approach to VL tasks consists of a vision module **V** and cross-modal understanding module

VL

$$\{v_d\}_{d\in\mathcal{D}} = \mathbf{V}(image), \quad y = \mathbf{VL}(w_{task}, \{v_d\}_{d\in\mathcal{D}}), \quad (1)$$

where the pre-preprocessed image information $v_d = (b_d, f_d, c_d)$ consists of a region or bounding box b_d and extracted features corresponding to the region f_d along with a category label c_d for each object detector proposal $d \in \mathcal{D}$. The text prompt w_{task} and output y are VL task-specific, corresponding to a question and answer in VQA, text and matching score in text-image retrieval, an empty prompt and an object predicate graph in SGG, and an empty prompt and output caption in image captioning.

3.2. Object Proposal Evaluation Background

For a specific object category, an object region proposal b_d provided by a detector is typically deemed correct or incorrect based on the largest intersection over union (IOU) it is able to achieve with a ground truth human region annotation b_a from their category as shown

$$\text{IOU}(b_d, b_a) = \frac{area(b_d \cap b_a)}{area(b_d \cup b_a)}. \quad (2)$$

IOU is an extension of the Jaccard Index applied to a region's pixels. A threshold γ is then applied to the IOU scores to obtain a set of correct detections. Precision is the most commonly used performance measure of an object detector proposal set $\mathcal{D}$ and is calculated over the ground truth region annotation set $\mathcal{A}$ for a specific category $\mathcal{A}_c \in \mathcal{A}$ as

$$P(\mathcal{D}, \mathcal{A}) = \frac{|\{max_{a\in\mathcal{A}_c}[\text{IOU}(b_d, b_a)] \geq \gamma\}_{d\in\mathcal{D}}|}{|\mathcal{D}|}. \quad (3)$$

Further reading on object proposal evaluation is in [27].

3.3. Object Importance and Misalignment

We now propose a novel formulation of object importance and explain how this results in metric misalignment. We first assume that the information relevant to task output y provided by the output from the vision module $\mathbf{V}$ is limited to a critical subset of ground truth object annotations $\mathcal{I} \in \mathcal{A}$. This implies that the increase in alignment, represented as the mutual information MI, between the output and y provided by additional annotations is limited to an arbitrarily small constant δ such that

$$MI(\mathbf{V}_{\mathcal{I}}(image); y) = MI(\mathbf{V}_{\mathcal{A}}(image); y) + \delta. \quad (4)$$

Thus, at a fixed number of detector proposals $|\mathcal{D}|$, a precision metric is misaligned when the ranking of detectors evaluated using the critical objects does not match the ranking of detectors evaluated using all the objects as shown

$$P(\mathcal{D}_1, \mathcal{I}) > P(\mathcal{D}_2, \mathcal{I}) \implies P(\mathcal{D}_1, \mathcal{A}) > P(\mathcal{D}_2, \mathcal{A}). \quad (5)$$

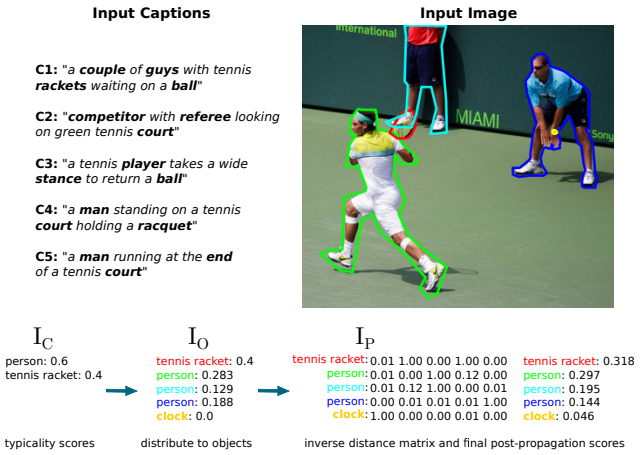

Figure 2. An example illustrating our processing pipeline. Words used for object typicality are shown in bold font. Object annotations are color-coded to their corresponding label in the I_O and I_P processing stages. If we set $T = 0.2$, only the tennis racket and adjacent player would be selected based on their high I_P scores.

Here $\mathcal{D}_1$ and $\mathcal{D}_2$ are object region proposals from two different detectors. More specifically, this condition is violated when ground truth region annotations from the outside the critical subset $\overline{\mathcal{I}}$ impact the ranking of the detectors by skewing the size of the correct detection set as shown

$$P(\mathcal{D}_1, \mathcal{I}) + P(\mathcal{D}_1, \overline{\mathcal{I}}) < P(\mathcal{D}_2, \mathcal{I}) + P(\mathcal{D}_2, \overline{\mathcal{I}}). \quad (6)$$

This misalignment is more severe in tasks with a larger number of superfluous ground truth region annotations. By removing annotations that are unlikely to be critical to VL tasks, we reduce the size of $\overline{\mathcal{I}}$, thereby mitigating the risk of the condition from Eq. 5 being violated. Thus, we define an object's importance I as the likelihood it is a member of $\mathcal{I}$.

3.4. Estimating Object Importance

We estimate an object's importance using semantic grounding from human annotated captions for each image. Our methodology consists of 3 steps: characterization of the underlying semantic process in order obtain importance scores for each object category present in the captions (I_C), distributing these importance scores to each object from the category based on the area of its region annotation (I_O), and propagating object importance to adjacent objects to reduce sparsity (I_P). Critical objects proposals to be used for evaluation are then selected based on a threshold T. An example is shown in Figure 2.

Typicality Scores (I_C): We utilize typicality [12, 25] to characterize the underlying semantic process generating the object instances present in the ground truth captions. To estimate the semantic typicality for our application, we extract the object-specific concepts from the caption using a Parts of Speech (POS) tagger [15]. The prevalence of each object in

the ground truth caption set $\mathcal{S}$ is then determined by taking its document frequency df where each caption is treated as a separate document. The typicality is

$$I_C(c_s) = \frac{df_{\mathcal{S}}(c_s)}{|\mathcal{S}|}, \tag{7}$$

where c_s is an object category present in the caption sentence based on mappings from object-specific concepts to the most similar object class by ConceptNet [38], $|\mathcal{S}|$ is the number of captions present in the ground truth human caption annotations, $df_{\mathcal{S}}$ is a function that counts the number of captions in which the object-specific concept is present, and $I_C(c_s)$ is the estimated importance of the object category. In the rare case that no importance is assigned to any object categories, the data example has poor alignment between its captions and object annotations and is skipped during the evaluation.

Distribute to Objects (I_O): To begin quantifying the importance of each object in the category c_s, the importance is then distributed to each of the ground truth object annotations from the given category $\mathcal{A}_{c_s}$ based on its area $area(b_a)$

$$I_O(b_a) = \frac{area(b_a) * I_C(c_s)}{\sum_{a \in \mathcal{A}_{c_s}} area(b_a)}, \tag{8}$$

where $I_O(b_a)$ is the importance of the object b_a.

Propogate Scores (I_P): We have now identified objects of importance to VL tasks in the image. However, larger scores are likely to be sparse among the objects since most objects are not in a category with a high I_C score. We instead infer that VL task importance is highly dependent on regional interactions (e.g. person holding tennis rack in Figure 2) and utilize heat kernel based dispersion modeling techniques [49] from graph signal processing in order to capture inter-objects interactions. To this end, we model the objects in the image as a graph with adjacency matrix, W, and construct a heat kernel using the PyGSP toolkit [10]. The values of W are based on the inverse of the minimum Euclidean distance (d) between the ground truth object region annotations in the image as shown

$$W_{ij} = \frac{1}{max(d(b_i, b_j), 1)}, if\ i \neq j, \tag{9}$$

where W_{ij} (shown in bottom right of Figure 2) is transpose invariant and $W_{ij} = 0$ if $i = j$.

The heat kernel, $H_t(W)$, is a function of graph connectedness and can be used to smooth the values of each node on a graph over time, t. The heat kernel is defined in the spectral domain as $g_t(\lambda) = \exp(-t\lambda)$, where $\lambda \in [0, 1]$ are the normalized eigenvalues of the graph Laplacian (formed by W). Since the kernel is applied to the graph eigenvalues λ, which can be interpreted as squared frequencies, and as a generalization of the Gaussian kernel on graphs.

We apply this kernel to the object importance vector to propagate the importance of the objects based on their proximity to one another as shown

$$I_P(j) = \sum_{i=1}^{|V|} I_O(b_i) \times H_t(i, j), \tag{10}$$

where $|V|$ is the number of connections a node has on the graph and t represents the dispersion time parameter (as $t \to \infty$, all importance scores I_P become uniform) set to the standard value of 1 [10]. We then normalize the sum of I_P to 1 as a post-processing step.

Once we have determined the final importance weight after propagation, denoted as I_P, of each object, we apply a threshold T in order to select a suitable subset of objects (with $I_P > T$) to be used for object evaluation. Therefore, increasing T allows for more focus on the central aspects of the scene while decreasing T incorporates more details from the annotations at the risk of including noisy or irrelevant annotations.

3.5. Extending Existing Metrics

After our proposed approach of selecting critical objects, standard evaluation procedures are then employed. In practice, precision performance is aggregated across all object categories. Region proposals are selected based on either a confidence threshold or by taking a set number of the most confident predictions. There are 6 evaluation metrics currently employed in object detection: $mAP_{IOU=0.50:0.05:0.95}$, the primary COCO evaluation metric, $mAP_{IOU=0.50}$, the primary metric for PASCAL VOC and VL task object detection, $mAP_{IOU=0.75}$, the precision metric most attune to localization, and 3 variations of $mAR_{IOU=0.50:0.05:0.95}$ where either 1, 10, or 100 annotations are used as the ground truth set. Additional detail regarding these metrics can be found in [27]. For convenience, we will refer to these 6 COCO metrics as mAP, mAP_{50}, mAP_{75}, mAR^1, mAR^{10}, and mAR^{100}. We also include an adjusted recall metric with an IOU of 0.50 $mAR^1_{IOU=0.50}$ which when combined with mAP_{50} via harmonic mean, creates a proposed F1 score, $F1_{IOU=0.50}$. For convenience, we will refer to our proposed metrics as mAR^1_{50} and $F1$. All proposed metrics are developed based on the implementation from [22]. By combining the perspectives of precision and recall with the importance threshold of object annotations, our approach provides insight into detector comparison and improvement.

4. Experiments

4.1. Alignment with Captioning Metrics

We first measure our method's agreement with downstream captioning metrics when the importance scores are used to select annotations and proposals from VINVL [51]

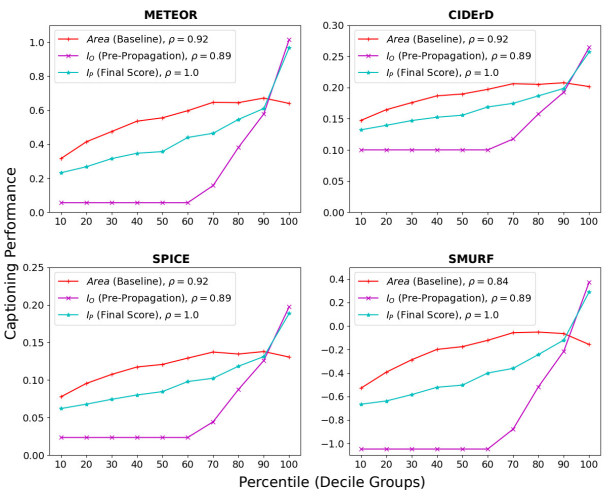

Figure 3. Plots of the captioning performance for each importance score decile group for image captioning on the COCO "Karpathy" test split [16, 22] evaluated using CIDEr (C) [42], Meteor (M) [4], SPICE (SP) [1], and SMURF (SM) [12]. The Spearman's ρ rank correlation is used to measure the alignment between our importance score selections and image captioning results and is shown for each method in the legend.

to perform the task of image captioning. VINVL uses the 1594 most frequent object classes and 524 most frequent object attributes from VG for their label prediction set. Their work uses OSCAR [21] as a downstream captioner, which takes category tags and features from regions of interest as input and uses the CIDEr optimization methodology [31]. Annotations and proposals deemed more important by our algorithm should result in higher captioning scores, while annotations and proposals deemed less important by our algorithm should result in lower captioning scores. We select image captioning as the representative task of the VL domain since it directly incorporates text-image alignment in a consistent manner.

For our experiment, we first select the 2109 images from the 5000 images in the Karpathy test split [16] that have at least 1 VG annotation. We then remove 38 examples with poor image-caption alignment, leaving us with 2071 images for use as a benchmark. We use the provided pre-trained models with default settings and do not perform any fine-tuning. OSCAR is provided with regional information sourced from both ground truth annotations and VINVL [51] and tag information sourced from ground truth annotations. To measure agreement, we split the annotations into the 10 decile groups based on the I_P importance scores to generate adjusted tag and corresponding bounding box feature sets as input to the OSCAR captioner. Figure 3 shows the Spearman's ρ rank correlation between the mean captioning metric scores and our importance score, along with the correlation for a developed area-based baseline. We select

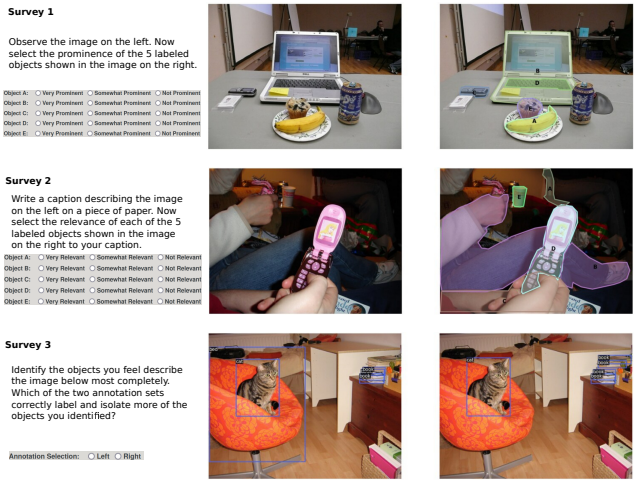

Figure 4. A visualization of the 3 AMT surveys performed with the instructions and input interface shown on the left and example images shown on the right.

the Spearman's ρ rank correlation since we expect the rank of the mean captioning score for each percentile range to increase in a monotonic, linear fashion if our importance scores are well aligned with VINVL and the human annotations. Figure 1 follows a similar procedure, but instead removes the lowest importance decile group at each iteration and shows the result for a data example at 3 different tag and feature removal points.

Our results show that the I_P importance score is highly aligned with captioning metrics. Although the max captioning scores from 90th-100th percentile of the pre-propagation and final scores are comparable, performance in the lower percentiles is very low and for the most part constant for the pre-propagation scores. This agrees with our expectations since the pre-propagation scores are sparse and provide no information about objects not mentioned in the captions. The area-based baseline is roughly correlated with improved captioning performance but has a drastically lower performance in the higher percentile groups.

4.2. Human Surveys

We perform 3 human surveys using AMT in order to provide an example-level view of our approach and collect information about object importance based on human perception. We start with a set of 225 randomly selected images from COCO that contain at least 5 annotated objects. We then use the importance score of each object to select 5 objects from each image to be visually annotated using the provided ground truth regions and labelled for our first two surveys. Letter labels {'A','B','C','D','E'} are assigned to objects randomly and images with confusing labels due to object overlap were removed, leaving 198 labelled images that were used in our first two surveys. The first two surveys

Method	Prominent	Caption Aligned
$Area$ (Baseline)	0.089	0.064
I_O (Pre-Propagation)	0.154	0.083
I_P (Final Score)	0.152	0.160

Table 1. Kendall Tau rank correlation between object scoring algorithms and rating-based survey results. 'Prominent' corresponds to the survey 1 responses while 'Caption Aligned' corresponds to the survey 2 responses. The correlation between the two surveys was 0.25. A ranking based on the area of the annotations is used as a baseline along with an ablation study.

ask Turkers to rate objects based on two separate criteria in order to reduce survey bias and provide a more diverse view of object importance. For the first survey, Turkers are asked rate each object's prominence on a scale of 1 to 3, 1 being "not prominent" and 3 being "very prominent". In the second survey, we follow a similar procedure except we ask Turkers to first write a caption describing the image, then rate the objects based on their relevance to that caption. For our third survey, two state-of-the-art anchor-free detectors, FoveaBox [18] and fcos [40], are used to automatically annotate our previously selected 225 images with bounding boxes and class labels. We include only the top 5 most confident predictions of each model. The 142 images with inconsistent class labels were used for the third survey, which asks Turkers to choose which image of the automatically annotated images includes and correctly labels more of the objects most important for understanding the scene. Each detector's annotated image was placed randomly on either the left or right side of the survey page for the selection process. The majority decision from 3 different Turkers is used as the final selection. The selections were quite consistent with all 3 annotators choosing the same image 81% of the time. The full survey is shown in Figure 4. We publicly release the AMT responses (822 in total), survey templates, and labelled images for all the surveys.

Table 1 shows the rank correlation of each object scoring methodology with the human responses from our first two surveys, as well as an additional ablation study showing the importance of propagating the importance over regions yielding results consistent with Figure 3. As a baseline, we use object area as a method for importance scoring. For the purposes of this comparison, the importance scores are mapped to the discrete set {1,2,3} such that the frequency of each value matches that of the human survey response distribution. We use Kendall Tau due to its focus on concordant and discordant pairs, making it more robust to ties and survey noise and more appropriate for experiments not fitting to a linear representation. Our importance scores demonstrate a dramatic increase in human object rating alignment over the more naïve area-based approach for both survey prompts, despite the inter-correlation between the surveys being only 25%. Based on the results from the third survey, we measure

Metric	Acc
mAP	0.489
mAP_{50}	0.692
mAP_{75}	0.600
mAR^1	0.559
mAR^{10}	0.425
mAR^{100}	0.425
mAR^1_{50} (Ours)	0.750
$F1$ (Ours)	0.737

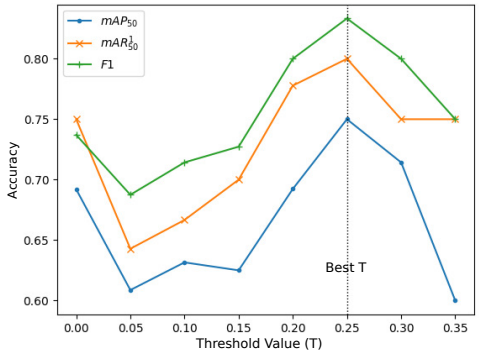

Figure 5. The table on the top shows the example-level accuracy with human judgement from Survey 3 for different detection metrics using all ground truth annotations (T=0). The plot on the bottom shows how this accuracy improves for the best performing metrics by selecting ground truth object annotations based on importance with T=0.25 (92% of annotations removed) yielding the best results.

the agreement between our proposed detection evaluation metrics along with existing metrics in the table in Figure 5. We observe that mAP_{50} along with our proposed recall and F1 scores have the greatest alignment with human judgement when compared with other existing metrics. We are able to further improve alignment with human judgement by using our annotation selection methodology, which can be seen in the plot shown in Figure 5. The initial dip in alignment between human and metric selections is largely caused by the forced selection of object rankings in Survey 1 and 2 rather than allowing for ties. This necessary limitation on the human survey's granularity simplifies the response process but forces arbitrary selections for lower importance objects. One such example can be observed in Figure 6. This selection should have instead been considered a tie by the metric since most scene-essential objects have been accounted for by both models and any image selection is likely to be arbitrary.

4.3. Consistency Study

Although COCO is a crucial benchmark for object detection, VG is one of the primary benchmarks for VL tasks. Therefore, we assess whether our approach is consistent across the COCO and VG annotation formats. By showing the consistency of selected objects across datasets, we further support the existence of a critical subset of ground truth object annotations and the effectiveness of our approach.

In place of captions, VG utilizes region descriptions

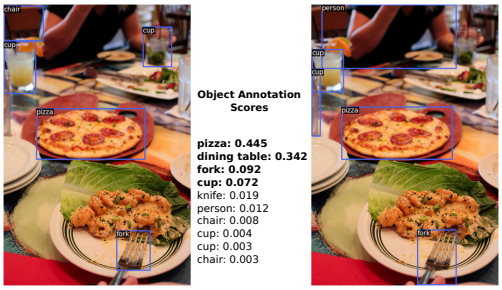

Figure 6. A failure case for mAP_{50} for T=0.05. Object annotations used in the evaluation are shown in bold.

where a human annotator describes a section of the image. These descriptions are very similar to captions and at least 10 descriptions are provided for each image, so we postulate they can be used in place of captions for our method. We use the preprocessed VG split from [44], commonly utilized for SGG tasks, which contains information on only the 150 most frequent object categories, to perform a study. We first acquire the annotations of 11,597 training split images that appear in both the VG and COCO dataset. Then we perform our score-based selection of objects on the COCO dataset with a threshold value of 0.25. We perform the same selection on VG, but sweep the threshold value from 0 to 0.35 in increments of 0.05. Here we use the Intersection-over-Union (IOU) between the selected annotations of the two datasets to show that our method is gradually removing excessive and noisy annotations and instead focusing on essential regions, like those annotated in COCO, as supported by the results in Figure 7. Visualizations of selected annotations for two images show that selected VG annotations become very similar to the COCO annotations in terms of quality and focus, but still include some additional detail. The gradual increase in IOU is highly significant since it occurs despite the large amount of overlap between the annotations of VG. Our results suggest that some of the additional annotations found in VG actually lead to worse scene coverage and that there is a subset of objects recognized by annotators as capturing the essence of the scene, which we are able to identify with our method. Based this study, we determine T=0.075 (61% of annotations removed) to be a threshold of interest for the VG dataset since an inflection point occurs around this value, meaning many of the most irrelevant annotations have been removed and the rate of IOU increase has begun to slow. We also determine T=0.30 (96% of annotations removed) to be a threshold of interest since this selection has the greatest IOU with the selected COCO annotations.

4.4. Importance in a SGG Benchmark

Current VL works typically report the mAP_{50} score of a proposed detector as a intermediate validation of their

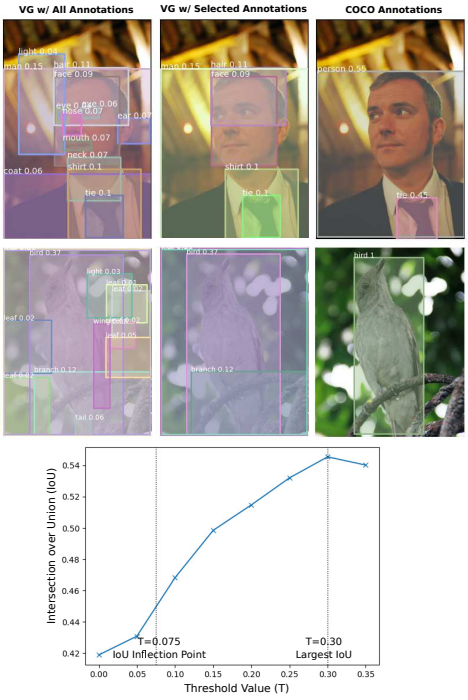

Figure 7. Examples of the annotation selection process for Visual Genome using a threshold value of 0.075 are shown on the top along with a comparison to COCO. The plot on the bottom shows the average IOU between COCO and VG annotations as the importance threshold is raised.

methodologies. While mAP_{50} is certainly correlated with the bias needed for improved VL performance, cases of disagreement between the intermediate measure and downstream task performance measures are very common. Another issue is that since specialized detectors like VINVL [51] and BUTD [2] use unique category sets and training procedures, it is difficult to compare them with other detectors. Scene Graph Generation (SGG) benchmarks, on the other hand, have a strict category set and training procedure used by proposed detectors, allowing for direct and meaningful comparison. Authors have also claimed that SGG can serve as an important benchmark for connecting detection and scene-understanding tasks. However, there is little to no evidence to support that SGG provides any more feedback on VL-oriented performance than mAP_{50}, nor is there any evidence that techniques improving SGG performance lead to improved performance in other VL tasks. On the contrary, the novel portion of SGG, the object relations, have been found to be highly imbalanced and ambiguous [39]. In addition, it is difficult to determine how results reported in Scene Graph Generation works can be related to other VL tasks.

In this experiment, we present a use case that highlights the flaws of these competing methodologies in a highly explainable manner. We follow the progress of detectors

Model	Previous	Ours					
	$T=0$	$T=0.075$			$T=0.30$		
	P	P	R	F	P	R	F
2018 [48]	20.4	18.0	37.7	24.3	5.9	46.7	10.6
2020 [39]	22.9	18.7	40.0	25.5	5.2	47.2	9.4
2021 [14]	24.5	20.0	41.7	27.0	5.7	50.9	10.2

Table 2. Detector evaluation based on the Visual Genome experimental procedure from [44]. The "Previous" column represents the relevant information provided by the unmodified mAP_{50} metric, while the remaining columns correspond to the feedback provided by our proposed modifications using thresholds determined in the consistency study.

utilized in Scene Graph Generation (SGG) by extensively evaluating 3 selected detectors. All detection models in SGG currently use the Faster RCNN architecture [30]. However, the 2018 model from "Neural Motifs" [48] incorporates a VGG backbone [36] and both the 2020 model from [39] and the 2021 model [14] incorporate the ResNeXt101-FPN as the backbone as originally done in [50]. Since VG has approximately 20 object annotations per image, we use the top 20 most confident proposals from each detector in our evaluation. A comparison of the performance of these detectors is shown in Table 2.

Based on this analysis, it is apparent that the information provided by previous measures is very limited. Questions like which objects were detected and whether these objects were worth detecting are ignored. Using our more detailed evaluation, we see the 2021 detector consistently achieves higher recall than the other two detectors, but the 2018 model has higher precision on objects critical to VL tasks. This leads to the 2018 model having a larger F1 score at the higher importance threshold and shows that newer models may not be adequately prioritizing the correct classification of essential objects, a phenomenon not captured by existing evaluation methods due to their misalignment. While our measures demonstrate a degree of agreement with previous methods, our method is able to capture a much broader and more nuanced story about a given detector's performance.

5. Discussion and Limitations

It should be noted that selection of critical objects is by no means a "one size fits all solution" to metric misalignment. Our method is intended to be an enhancement of existing VL evaluation metrics like caption evaluation by providing more detailed feedback on object prioritization by the vision module based on downstream semantic information. Furthermore, proper alignment between object proposals and evaluation is a task-specific problem where the severity of misalignment will vary depending on how few of the annotated objects are relevant to a given downstream task.

Another challenge facing VL evaluation metrics is degrading performance when there are fewer or noisier captions provided by human annotation [1]. For our method, fewer captions would result in more data examples being skipped during the evaluation due to no importance being assigned to any of the objects in the image. We also find that while the mappings performed by ConceptNet are very reasonable, there are a number of failure cases. An example of such a case is the word "player", which is mapped to "sports ball" instead of the more appropriate "person". This can lead to sports players being given less importance at the I_O stage, especially if all the annotators only use the term "player" to describe the person in question. Importance propagation helps compensate for this issue as long as the "player" is in close proximity to the "sports ball".

We also observe some limitations in the human study in the form of annotation noise. We attempt to reduce influence of this noise by using two different prompts for the first two surveys, using the majority decision of 3 responses in the third survey, and removing select responses. To avoid bias in our selections, we do not keep any intermediate data from the survey process and instead make removal selections based on survey completion in an unreasonable amount of time, with too little variation in scoring, or with excessively repeated patterns. In addition, there is a clear trade-off between instruction specificity and bias when conducting surveys [28]. We appeal to the AMT annotators innate understanding of importance with unique and open-ended tasks, but this can potentially lead to less consistent responses.

6. Conclusion and Broader Impact

We present a formulation of misalignment in object proposal metrics along with an importance score that can be used to select objects critical to VL tasks in order to address this phenomenon. Our object proposal evaluation methodology is the first to be validated with both human judgement and empirical results. The current lack of a VL specific detector evaluation benchmark has contributed to a shift towards embedding-based approaches for VL tasks [43, 52]. This shift represents a dangerous trend in VL pipelines, which are known to be sensitive to language [34] and dataset [13] priors. To avoid such issues, vision and language components of these pipelines should be evaluated independently in an explainable manner. Our approach could help revitalize research in detectors and enable transparent and explainable approaches and analyses in VL tasks. Future work could focus on applying our method to other elements of scene understanding such as activities or attributes by shifting the focus of the concept extraction to the element of interest.

Acknowledgments: The authors acknowledge support from the NSF RI Project VR-K #1750082, CPS Project #2038666 and and a grant from Meta AI Learning Alliance. Any opinions, findings, and conclusions in this publication are those of the authors and do not necessarily reflect the view of the funding agencies.

References

[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *ECCV*, 2016. 1, 5, 8

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 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)*, June 2018. 1, 2, 7

[3] Pratyay Banerjee, Tejas Gokhale, Yezhou Yang, and Chitta Baral. Weaqa: Weak supervision via captions for visual question answering. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3420–3435, 2021. 2

[4] Satanjeev Banerjee and Alon Lavie. 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, 2005. 1, 5

[5] Daniel M. Bear, Chaofei Fan, Damian Mrowca, Yunzhu Li, Seth Alter, Aran Nayebi, Jeremy Schwartz, Li Fei-Fei, Jiajun Wu, Joshua B. Tenenbaum, and Daniel L. K. Yamins. Learning physical graph representations from visual scenes, 2020. 2

[6] Ali Borji, Ming-Ming Cheng, Huaizu Jiang, and Jia Li. Salient object detection: A benchmark. *IEEE Transactions on Image Processing*, 24(12):5706–5722, Dec 2015. 2

[7] Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew Peters, Ashish Sabharwal, and Yejin Choi. Adversarial filters of dataset biases. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1078–1088. PMLR, 13–18 Jul 2020. 2

[8] Neelima Chavali, Harsh Agrawal, Aroma Mahendru, and Dhruv Batra. Object-proposal evaluation protocol is 'gameable'. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 1

[9] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server, 2015. 2

[10] Michaël Defferrard, Lionel Martin, Rodrigo Pena, and Nathanaël Perraudin. Pygsp: Graph signal processing in python, Oct. 2017. 2, 4

[11] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015. 2

[12] Joshua Feinglass and Yezhou Yang. 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, Aug. 2021. Association for Computational Linguistics. 1, 2, 3, 5

[13] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 8

[14] Xiaotian Han, Jianwei Yang, Houdong Hu, Lei Zhang, Jianfeng Gao, and Pengchuan Zhang. Image scene graph generation (sgg) benchmark, 2021. 2, 8

[15] Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017. 3

[16] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015. 5

[17] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 2

[18] Tao Kong, Fuchun Sun, Huaping Liu, Yuning Jiang, Lei Li, and Jianbo Shi. Foveabox: Beyound anchor-based object detection. *IEEE Transactions on Image Processing*, 29:7389–7398, 2020. 6

[19] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017. 2

[20] John P Lalor, Hao Wu, and Hong Yu. Building an evaluation scale using item response theory. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2016, page 648. NIH Public Access, 2016. 2

[21] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020. 1, 5

[22] 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*, pages 740–755. Springer, 2014. 4, 5

[23] Man Luo, Shailaja Keyur Sampat, Riley Tallman, Yankai Zeng, Manuha Vancha, Akarshan Sajja, and Chitta Baral. 'just because you are right, doesn't mean i am wrong': Overcoming a bottleneck in the development and evaluation of open-ended visual question answering (vqa) tasks. *arXiv preprint arXiv:2103.15022*, 2021. 2

[24] Tom McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages

3428–3448, Florence, Italy, July 2019. Association for Computational Linguistics. 2

[25] Daniel Osherson and Edward E Smith. On typicality and vagueness. *Cognition*, 64(2):189–206, 1997. 2, 3

[26] Rafael Padilla, Sergio L Netto, and Eduardo AB da Silva. A survey on performance metrics for object-detection algorithms. In *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*, pages 237–242. IEEE, 2020. 2

[27] Rafael Padilla, Sergio L Netto, and Eduardo AB Da Silva. A survey on performance metrics for object-detection algorithms. In *2020 international conference on systems, signals and image processing (IWSSIP)*, pages 237–242. IEEE, 2020. 3, 4

[28] Mihir Parmar, Swaroop Mishra, Mor Geva, and Chitta Baral. Don't blame the annotator: Bias already starts in the annotation instructions. In *Conference of the European Chapter of the Association for Computational Linguistics*, 2022. 8

[29] Jiayan Qiu, Yiding Yang, Xinchao Wang, and Dacheng Tao. Scene essence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8322–8333, June 2021. 2

[30] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. 8

[31] Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 5

[32] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 658–666, 2019. 2

[33] Pedro Rodriguez, Joe Barrow, Alexander Miserlis Hoyle, John P. Lalor, Robin Jia, and Jordan Boyd-Graber. Evaluation examples are not equally informative: How should that change NLP leaderboards? 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 4486–4503, Online, Aug. 2021. Association for Computational Linguistics. 2

[34] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics. 8

[35] João Sedoc and Lyle Ungar. Item response theory for efficient human evaluation of chatbots. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 21–33, 2020. 2

[36] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015. 8

[37] Xinhang Song, Shuqiang Jiang, Bohan Wang, Chengpeng Chen, and Gongwei Chen. Image representations with spatial object-to-object relations for rgb-d scene recognition. *IEEE Transactions on Image Processing*, 29:525–537, 2020. 2

[38] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge, 2018. 4

[39] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *Conference on Computer Vision and Pattern Recognition*, 2020. 2, 7, 8

[40] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9627–9636, 2019. 6

[41] Clara Vania, Phu Mon Htut, William Huang, Dhara Mungra, Richard Yuanzhe Pang, Jason Phang, Haokun Liu, Kyunghyun Cho, and Samuel R. Bowman. Comparing test sets with item response theory. 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 1141–1158, Online, Aug. 2021. Association for Computational Linguistics. 2

[42] R. Vedantam, C. L. Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575, 2015. 1, 5

[43] 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, 2022. 8

[44] Danfei Xu, Yuke Zhu, Christopher B. Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2, 7, 8

[45] Keren Ye and Adriana Kovashka. Linguistic structures as weak supervision for visual scene graph generation. *ArXiv*, abs/2105.13994, 2021. 2

[46] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *CVPR*, 2021. 2

[47] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[48] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5831–5840, 2018. 2, 8

[49] Fan Zhang and Edwin R Hancock. Graph spectral image smoothing using the heat kernel. *Pattern recognition*, 41(11):3328–3342, 2008. 4

[50] Ji Zhang, Kevin J Shih, Ahmed Elgammal, Andrew Tao, and Bryan Catanzaro. Graphical contrastive losses for scene graph parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11535–11543, 2019. 2, 8

[51] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 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, June 2021. 1, 2, 4, 5, 7

[52] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):13041–13049, Apr. 2020. 8