2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) | 979-8-3503-1892-0/24/$31.00 ©2024 IEEE | DOI: 10.1109/WACV57701.2024.00564

Simple Token-Level Confidence Improves Caption Correctness

Suzanne Petryk[1,2] Spencer Whitehead[2] Joseph E. Gonzalez[1]
Trevor Darrell[1] Anna Rohrbach[1] Marcus Rohrbach[2]

[1] UC Berkeley [2] Meta

Abstract

The ability to judge whether a caption correctly describes an image is a critical part of vision-language understanding. However, state-of-the-art models often misinterpret the correctness of fine-grained details, leading to errors in outputs such as hallucinating objects in generated captions or poor compositional reasoning. In this work, we explore Token-Level Confidence, or TLC, as a simple yet surprisingly effective method to assess caption correctness. Specifically, we fine-tune a vision-language model on image captioning, input an image and proposed caption to the model, and aggregate either algebraic or learned token confidences over words or sequences to estimate image-caption consistency. Compared to sequence-level scores from pretrained models, TLC with algebraic confidence measures achieves a relative improvement in accuracy by 10% on verb understanding in SVO-Probes and more than doubles image and group scores for compositional reasoning in Winoground. When training data are available, a learned confidence estimator provides further improved performance, reducing object hallucination rates in MS COCO Captions by a relative 30% over the original model and setting a new state-of-the-art.

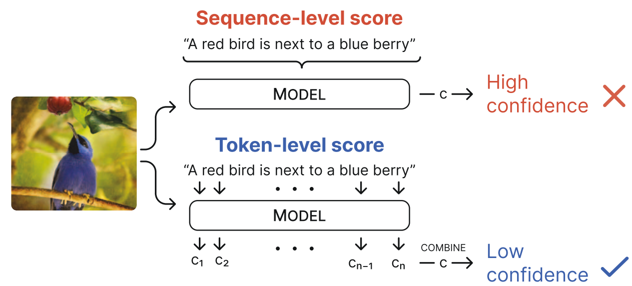

Figure 1. Judging caption correctness is still a challenge for large-scale models that operate at a sequence-level (e.g., with image-text matching scores). For instance, these models may erroneously assign high confidence to "A red bird is next to a blue berry" for an image with the fine-grained difference of a blue bird next to a red berry. We show that both algebraic and learned confidences at a *token-level* from a finetuned image captioning model improve fine-grained estimates of caption correctness.

1. Introduction

For vision-and-language models, grounding and the ability to assess the correctness of a caption with respect to an image is critical for vision-language understanding. When models have difficulties with these, the outputs can be error prone [45] or rely on biases [1, 21]. State-of-the-art models, like CLIP [42] or OFA [60], demonstrate impressive capabilities in a variety of settings, in part, thanks to these properties. While these models have had much success, recent efforts for probing state-of-the-art models have revealed some weaknesses in these areas. For instance, the recent Winoground task [52] illustrates that these models, including large-scale pre-trained ones, can struggle to correctly associate image-caption pairs when the captions have differences in word order. Similarly, SVO-Probes [22] has shown that models can fail in situations that require understanding verbs compared to other parts of speech. The observations from these probing tasks suggest that existing models have difficulties discerning fine-grained details that can appear in multimodal data. This may hinder their accuracy and reliability when used in real settings, which presents significant issues in scenarios that require highly correct outputs, such as assisting people with visual impairments [20, 65].

We conjecture that these weaknesses may be related to the granularity with which models perform image-text matching (ITM). As shown in Fig. 1, many existing models often operate at a sequence-level, pooling the representations of the image and caption to assess whether the text correctly describes the image. This pretext task relies on sequence-level supervision and data with sufficient scale to learn finer-grained concepts, such as the difference between "a cat jumping over a box" and "a box with a cat inside". Typical generative image captioning methods, on the other hand, generate words token-by-token and produce confidences for each one. They are supervised at a token-level rather than sequence-level, which may emphasize the consistency of each token in a sequence more explicitly.

2642-9381/24/$31.00 ©2024 IEEE
DOI 10.1109/WACV57701.2024.00564

Leveraging this observation, we explore Token-Level Confidence, or TLC, for assessing image-caption correctness. We input an image and proposed caption into a fine-tuned captioning model, which produces a distribution over the vocabulary at each time step. The base TLC method, TLC-A, uses algebraic confidence measures (*e.g.*, softmax score) to compute confidence for a given token. To produce a single score for image-caption correctness, we either aggregate token confidences over the sequence (*e.g.*, by taking the average value), or over particular words, such as verbs or objects. Next, we further investigate whether learned confidences can outperform algebraic ones. We propose a Learned confidence estimator, TLC-L, for use in the caption generation setting where training data is available. We use existing annotations to model the likelihood that a predicted token matches reference tokens, and an additional validation set to calibrate our estimated confidence to actual correct and incorrect concepts. Using TLC-L to re-rank candidate captions, we reduce hallucination rates in the final output captions.

Both TLC-A and TLC-L are simple to implement and can be applied on top of any autoregressive image captioning model with an encoder and decoder, an architecture found to scale well with data and multimodal tasks [9, 58, 59, 64]. In this work, we demonstrate the effectiveness of token-level confidence across multiple model sizes of OFA [59], a recent Transformer-based model [55] with strong performance on many vision-language tasks. On the challenging Winoground [52] benchmark evaluating compositional reasoning, we show that TLC-A more than doubles accuracy over pretrained ITM scores, *e.g.*, from 10.25% to 27% on image score (Sec. 4.2). TLC-A also outperforms ITM on a fine-grained verb understanding task [22] by a relative 10% (Sec. 4.3). When using TLC-L to re-rank candidate captions on MS COCO [8], we achieve a 30% relative reduction in object hallucination rate over the original captions and set a new state-of-the-art on a hallucination benchmark [45] (Sec. 4.4). These results demonstrate that token-level confidence, whether algebraic (TLC-A) or learned (TLC-L) are a powerful yet simple resource for improving multimodal reliability.

2. Related Work

Caption correctness. One of the desired properties of a good caption is correctness, *i.e.*, being faithful to an image. [22, 41, 52] propose benchmarks to probe for sensitivity to hard negatives of different types, such as compositional reasoning or action understanding. We use probing benchmarks in our work to demonstrate the effectiveness of TLC-A. Within caption generation, [45] notes that in practice, image captioning models suffer from object hallucination [45], driven by visual misclassification and over-reliance on language priors. Several recent works addressed the issue of object hallucination [7, 34], in some cases relying on causal inference-based approaches [35, 68, 69]. [12] propose a pretraining loss to reduce object hallucination; in our work, we reduce hallucinations without retraining the captioning model. Other recent works pose a slightly distinct problem of correcting errors in a caption provided for a given image (i.e., not as part of the caption generation process) [46, 47, 63]. Some works propose caption decoding methods to target criteria such as correctness [3, 15, 66]. However, the original formulation of beam search remains the dominant decoding method used in modern multimodal architectures [9, 30, 59, 62]. We apply our approach on top of captions generated with beam search and demonstrate that simply re-ranking beams based on token confidences can reduce hallucinations.

Correctness estimation in language models. Similar issues around correctness and hallucination are also relevant for many language-only tasks that require autoregressive prediction. Hallucination in particular has been studied for tasks like abstractive summarization [37], e.g., one work performs token-level hallucination detection [73]. A number of works study model uncertainty and aim to improve model calibration for machine translation [16, 18, 61], dialog [38], question answering [72] and spoken language understanding [49], to name a few tasks. While our focus on image captioning is similarly a conditional generation task, estimating confidence in the multimodal setting can be challenging as errors are driven by factors from both modalities [65].

Image captioning. Image captioning has seen significant progress since the arrival of deep learning as a dominant methodology [5, 14, 26, 27, 44, 57]. In recent years Transformer-based architectures have gained particular prominence [11, 30, 32, 48, 71, 74]. Many papers take the approach of pretraining large vision-and-language models and then adapting them to downstream tasks, including captioning [31, 64]. Recent efforts focus on further scaling these pretraining-based methods [2, 11, 25, 58, 70, 74], while many also aim to unify multiple vision-and-language tasks during pretraining [9, 10, 59, 62]. Despite steady improvements in image caption quality over the past years, even the best models still make mistakes. Here, we study the reliability of vision-language models, with the goal of assessing caption correctness.

Reliability in multimodal models. With the adoption of Large Language/Vision/Vision-and-Language Models (LLMs, LVMs, LVLMs), it is increasingly important to study their limitations and outline expectations regarding their *reliability*. One of the first efforts in doing that for LLMs and LVMs (unimodally) is [53], whose broad definition of reliability includes aspects from modeling uncertainty to robust generalization and adaptation. A recent work in multimodal learning outlines reliability of visual question answering [65], defining it as a model's ability to ensure a low risk of error by means of abstaining from answering. In our work, we approach reliability by improving assessments of caption correctness, and incorporating these estimates to reduce rates of error in generated captions.

3. TLC: Token-Level Confidence for Caption Correctness

Overview. Given an image and a caption, TLC produces a confidence score for each token and aggregates these scores to produce an estimate of caption correctness, *i.e.*, semantic consistency with the image. First, we describe two forms of confidences: algebraic (TLC-A, Sec. 3.1) and learned (TLC-L, Sec. 3.2). Next, in Sec. 3.3, we describe how to combine token confidences to measure caption correctness and use token confidences to re-rank captions during generation. In our experiments, we will then verify TLC-A primarily on out-of-domain probing benchmarks (Sections 4.2 and 4.3). We then evaluate TLC-L in a setting where in-domain training data is available (Sec. 4.4).

Preliminaries. Let f_{pre} be a vision-language model pretrained on a large multimodal dataset, and f_{cap} be a model initialized with f_{pre} and subsequently finetuned for autoregressive image captioning. Given an image x, a caption consists of a sequence of n tokens $t_{1:n}$ describing the image. At each decoding time step $k \in \{1...n\}$, f_{cap} produces a distribution of token likelihoods $\vec{z}_k \in \mathbb{R}^{|V|}$ for a vocabulary V, conditioned on previous outputs $\vec{z}_{1:k-1}$. Autoregressive captioning models are typically trained with a token-level cross-entropy loss on $\vec{z}_k$, often followed by self-critical sequence training [44]. Decoding methods such as sampling or beam search can then be used to select tokens at inference time, typically aiming to maximize the image-conditional sequence likelihood.

3.1. TLC-A: Algebraic Confidences

A simple method for measuring token-level confidence is to use an algebraic function of the distribution $\vec{z}_k$ directly, such as taking the logit or softmax value at the selected token index. We refer to token confidences derived from algebraic functions of $\vec{z}_k$ as TLC-Algebraic, or TLC-A. Prior works find simple measures such as softmax to be unreliable in both vision and vision-language "one-of-K" classification tasks [19, 65]. In contrast, we find that softmax scores from autoregressively-generated tokens perform surprisingly well, even on data that is out-of-distribution from the image captioning training set used by f_{cap}. This is aligned with findings in the language-only setting [13, 51, 54], suggesting that token-level language modeling may be key for reliable confidence measures.

3.2. TLC-L: Learned Domain-Specific Confidences

Although we observe that TLC-A performs well on evaluation benchmarks out-of-distribution from the image captioning training data (Sections 4.2 and 4.3), we would like to see whether *learning* a confidence estimator on in-distribution training data could improve estimates of correctness, similar to [65]. However, we do not have direct supervision to measure the correctness of a specific token in an arbitrary predicted caption with an image, aside from human evaluation. Instead, we leverage existing reference captions to learn a binary classification task, measuring whether a predicted token matches one or more reference tokens at the same time step. Fig. 2 presents an overview of this method, which we refer to as TLC-Learned, or TLC-L.

Forming the training set. We use a frozen f_{cap} and held-out dataset $\mathcal{X}$ to train a confidence estimator g. In practice, we use the f_{cap} validation set. For each image in $\mathcal{X}$, we select one of its reference captions $t_{1:n}$ and a random time step k within it. We first input the *prefix* $t_{1:k-1}$ into the f_{cap} decoder to predict the next token $\hat{t}_k$. We assign a binary label c to $\hat{t}_k$: $c = 1$ if it matches t_k or any token at k from other references with the same prefix, else $c = 0$. For example, in Fig. 2, if t_k is "sleeping", then "asleep" and "laying" are also considered correct due to sharing the same prefix "a dog". $\hat{t}_k$ "standing" is thus labeled as incorrect. These labels provide a noisy yet effective proxy for image consistency. At each epoch, we re-sample a reference caption and a time step k for each image in order to leverage all available ground-truth tokens.

Training a confidence estimator. The output of g is a scalar $\hat{c}$, trained with binary cross-entropy loss with c as supervision. As input, g receives image features from the model, such as those output by an encoder. It also receives token-level features from the decoder (*e.g.*, just before decoder features are projected into the vocabulary space). We find that including the reference *postfix*, or $t_{k+1:n}$, in addition to the prefix $t_{1:k}$ and predicted token $\hat{t}_k$ improves the confidence estimation. We pass the encoder features and position-encoded decoder sequence into a Transformer encoder [55], and pass the output embedding of token $\hat{t}_k$ into a small feed-forward network to produce $\hat{c}$. We provide details on our specific choice of architecture in Sec. 4.1. At inference time, we run our confidence estimator once per time step within a predicted caption $\hat{t}_{1:n}$.

A bidirectional confidence. We use the full caption context as input when supervising confidence for a single token $\hat{t}_k$. Due to self-attention in the Transformer encoder within g, the final prediction $\hat{c}$ represents a bidirectional confidence estimate, in contrast to the original autoregressive token predictions. This enables a useful combination: generating tokens autoregressively scales well with data and model size [9, 59], whereas estimating token confidence bidirectionally uses future context to inform correctness.

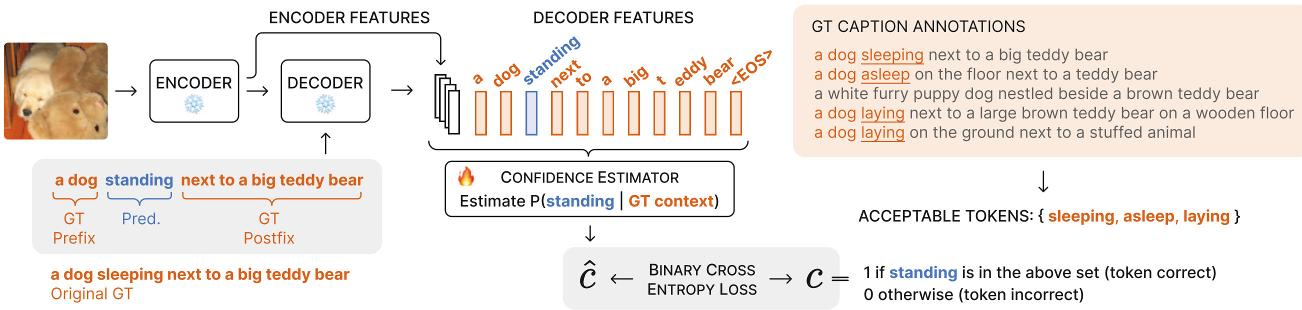

Figure 2. TLC-L: A framework to learn token-level confidence for an autoregressive encoder-decoder captioning model. We first use the captioning model to predict the next token (*e.g.*, "standing") after a partial reference caption (*e.g.*, "a dog"), shown in the bottom left. We input this sequence along with the image and the rest of the reference caption to the model to get corresponding embeddings, which are the inputs to our confidence estimator. We supervise correctness with a binary classification task: whether or not the predicted token matched any reference token at the same time step with the same prefix.

3.3. From Confidence to Caption Correctness

3.3.1 Combining Confidences

In practice, we would like to measure correctness over an entire caption or particular span, such as a word or phrase. To obtain such a score from token-level confidences, we can simply aggregate the confidences over a specific span of tokens $t_{i:j}$ or the full sequence $t_{1:n}$ by taking, *e.g.*, the minimum or average confidence value. We exclude the end-of-sentence (EOS) token, as its confidence is often poorly calibrated relative to previous tokens [28]. In our experiments, we compare correctness between image-caption pairs by aggregating over the full sequence (Sec. 4.2) or specific words (Sections 4.3 and 4.4).

3.3.2 Confidence During Caption Generation

We can use token-level confidences to not only estimate correctness between an image and an *existing* caption but also between a *proposed* caption candidate during generation. By re-ranking candidates relative to estimated correctness, we can reduce errors in the final selected captions.

When generating a caption, it is common to first predict a set of B candidate captions using an autoregressive decoding method such as beam search. Initially, the beams are ranked according to their cumulative token log likelihoods from the captioning model. However, token likelihood can fail to rank captions that are fully correct above those that contain an error. For example, a fluent and detailed sentence with a single-word hallucination may rank above a simpler, yet correct, caption. This is observed in [45], where captions with higher CIDEr [56] could also have higher hallucination rates.

To alleviate this, we first define a set of words or concepts $\mathcal{S}$ that we estimate correctness for. For example, in our experiments, we consider only the tokens that correspond to MS COCO [8] object categories, as we have annotations for their correctness during validation and evaluation. Beginning from the highest-likelihood beam, we estimate confidence $\hat{c}$ for each set of words in $\mathcal{S}$ that appear in the beam (*e.g.*, each MS COCO object that is mentioned). If any $\hat{c}$ are less than a threshold γ, we reject the beam, and continue to the next one until we reach a beam where all relevant tokens are predicted to be correct ($\hat{c} \geq \gamma$), or where there are no tokens from $\mathcal{S}$. If none of the beams satisfy these criteria, we output the original (highest-likelihood) caption. In that setting, we could alternatively choose to abstain from providing a caption in order to avoid misleading a user, similar to [65]. However, we instead choose the original caption in our experiments to simplify the comparison between methods. We choose the threshold γ on a validation set to maximize the recall of non-hallucinated objects at 99% precision.

4. Experiments

After discussing the experimental setup (Sec. 4.1), we demonstrate the effectiveness of TLC-A for identifying correct image-caption pairs that test understanding of compositionality (Sec. 4.2) and verbs (Sec. 4.3). We then evaluate both TLC-A and TLC-L on reducing object hallucinations in generated captions (Sec. 4.4).

4.1. Experimental Setup

As a captioning model, we choose to experiment with OFA [59], a recent open-source sequence-to-sequence multimodal transformer that achieves state-of-the-art captioning performance. OFA has a simple encoder-decoder architecture designed to unify multimodal tasks conditioned on an image and specific input instruction (e.g., "What does the image describe?" prompts the model to output a sequence of tokens for captioning). We use the official implementation and checkpoints (f_{pre}) for $\text{OFA}_{\text{Large}}$, OFA_{Base}, and

Model	Conf.	Text	Image	Group
MTurk Human [52]	-	89.50	88.50	85.50
Random Chance [52]	-	25.00	25.00	16.67
UNITER$_{\text{Large}}$ [52]	ITM	38.00	14.00	10.50
VinVL [52]	ITM	37.75	17.75	14.50
CACR$_{\text{Base}}$ [39]	CACR	39.25	17.75	14.25
IAIS$_{\text{Large}}$ [39]	IAIS	42.50	19.75	16.00
OFA$_{\text{Large}}$	ITM	**30.75**	10.25	7.25
	TLC-A	29.25	**27.00**	**17.50**
	(Δ)	(−1.5)	(+16.75)	(+10.25)
OFA$_{\text{Base}}$	ITM	**26.75**	10.75	6.50
	TLC-A	24.50	**23.50**	**13.75**
	(Δ)	(−2.25)	(+12.75)	(+7.25)
OFA$_{\text{Tiny}}$	ITM	**22.75**	7.75	4.50
	TLC-A	16.50	**15.75**	**6.75**
	(Δ)	(−6.25)	(+8.00)	(+2.25)

Table 1. Accuracy on text, image, and group score for the Winoground evaluation dataset [52]. Citations indicate where scores are reported.

OFA$_{\text{Tiny}}$, pretrained on a dataset with 20M publicly available image-text pairs. As image-text matching was included as a task in OFA pretraining, we use **ITM** in our results to denote the image-text matching score from f_{pre}. For f_{cap}, we finetune each scale of OFA model on MS COCO Captions [8], which has about 80k training images. We split the validation set of 40k images into three parts for training, validation, and testing of g, following [65]. Additional dataset details are in Appendix E.

For TLC-A, we use the softmax score at the selected token index. We experiment with several other choices of algebraic function and report results in Appendix C. For TLC-L, as input to the learned confidence estimator g, we use multimodal image and instruction features output from the OFA encoder, as well as token embeddings from the decoder just before they are projected onto the logit space by a linear layer. g itself is a 4-layer Transformer encoder [55], followed by a 2-layer MLP. Additional details are in Appendix F.

4.2. Correctness Around Compositional Reasoning

First, we assess the ability of TLC-A to select corresponding image-caption pairs. We use Winoground [52], a dataset curated to test the compositionality of vision-language models. Each of the 400 examples contains two image-caption pairs (I_0, C_0) and (I_1, C_1). Captions C_0 and C_1 contain the same words and/or morphemes, yet differ in order; for example, "there is a mug in some grass" and "there is some grass in a mug". There are three evaluations per example: text score (given an image, select the correct caption), image score (given a caption, select the correct image), and group score (all text and image scores for an example must be correct). A pairing is considered correct if the image-caption matching score for the correct pair is greater than that of the incorrect pair (*i.e.*, $c_{POS} > c_{NEG}$). [52] find that the task is surprisingly difficult, with all models they test performing below random chance for image and group score.

As correctness estimates, [52] use image-text matching scores (ITM) from a range of pretrained vision-language models. Other works [39, 43] design training losses specifically targeting relation alignment. Using TLC-A, we produce a correctness estimate c by simply averaging token-level softmax scores for each proposed image-caption pair. We present results in Tab. 1.

TLC-A outperforms ITM in image and group scores by over 2x. Compared to ITM across OFA model sizes, TLC-A more than doubles the image and group scores in all but one case (OFA$_{\text{Tiny}}$ group). Despite its simplicity, with no additional training beyond standard image captioning, TLC-A also outperforms IAIS in image and group scores (proposed in [43]), a training method optimized for multi-modal attention alignment.

Confidence	Model		
	OFA$_{\text{Large}}$	OFA$_{\text{Base}}$	OFA$_{\text{Tiny}}$
ITM	81.23	78.44	65.25
TLC-A	**89.47**	**89.64**	**81.34**
(Δ)	(+8.24)	(+11.20)	(+16.09)

Table 2. Image-caption matching accuracy for verb understanding with a subset of SVO-Probes [22]. TLC-A uses token-level softmax scores aggregated over the verb in each example.

4.3. Correctness Around Verb Understanding

Next, we consider caption correctness when aggregating token confidences over a single word, rather than over a full sequence as in Sec. 4.2. To evaluate this, we use SVO-Probes, a dataset designed by Hendricks and Nematzadeh [22] to test the verb understanding of vision-language transformers. Each example contains an image and a caption describing a ⟨subject, verb, object⟩ relation in the scene. It also contains a negative image, where only one part of the relation is different, such as ⟨person, swim, water⟩ and ⟨person, walk, water⟩. We use a publicly available subset of about 6,500 examples for verb understanding, and use a parser [24] to annotate the location of the verb in each caption. We aggregate token confidences over the verb tokens for TLC-A. Tab. 2 presents image-caption accuracy, where a score is 1 if the confidence is greater for the correct image (again, if $c_{POS} > c_{NEG}$).

TLC-A outperforms image-text matching scores. From Tab. 2, we see that TLC-A reaches higher image-caption matching accuracy compared to the ITM scores from pretrained models, across a range of model sizes (*e.g.*, 8.24% and 11.20% improvement for $\text{OFA}_{\text{Large}}$ and OFA_{Base} respectively). Therefore, when localized word or token positions are available, they can be leveraged for a finer-grained matching score than ITM operating on the full sequence.

4.4. Reducing Object Hallucinations

We now test our approach from Sec. 3.3.2, where we select a caption from a set of candidates to lower the likelihood of error. We also evaluate learned confidences from TLC-L, now that we can use domain-specific training data for g with the image captioning validation set.

Prior work [45] provides a framework for measuring object hallucination on MS COCO, comparing objects mentioned in references with those in a prediction. We add part-of-speech taggers [6, 24] to exclude predicted words that are not nouns; however, when comparing directly to prior work, we use the original implementation. We report standard captioning metrics [4, 56] as well as hallucination metrics CHAIRs and CHAIRi (or CHs and CHi) from [45]:

$$\text{CHAIR}_s = \frac{\text{\# captions with} \geq 1 \text{ hallucination}}{\text{\# captions}} \quad (1)$$

$$\text{CHAIR}_i = \frac{\text{\# objects hallucinated}}{\text{\# objects mentioned}} \quad (2)$$

We also report several caption diversity measures [50, 67] to examine whether captions with lower hallucination rates reduce caption diversity: *Vocab Size* measures unique unigrams across predictions, *% Novel* measures the percentage of generated captions which do not appear in the training set annotations, *Div-2* measures the ratio of unique bigrams to the number of generated words, and *Re-4* measures the repetition of four-grams. Other prior work proposes a polling-based approach to measure object hallucination [33]; however, we do not include this here as it evaluates a model via question-answering rather than evaluating a specific generated caption.

We show results from the following methods. **Standard** uses the original top caption from f_{cap}. **ITM** uses f_{pre} to re-rank the B candidate captions from Standard based on their image-text matching score, and selects the highest-ranked caption as output. **TLC-A** and **TLC-L** use the respective algebraic or learned confidences over the MS COCO object words to re-rank captions as described in Sec. 3.3.2. We find a threshold γ for TLC-A and TLC-L on the validation set (see Appendix A for details). We use a large beam size for all methods ($B = 25$) to observe the behavior of our caption selection method when given many possible candidates.

Learned confidences lead to the least hallucinations. From Tab. 3, we can see that both TLC-A and TLC-L lower the CHs and CHi hallucination rates across all model sizes compared to the original (Standard) captions. TLC-L reaches the lowest rates in each case; for example, it lowers CHs and CHi for $\text{OFA}_{\text{Large}}$ by a relative 37.6% and 34.3% respectively. Using ITM scores slightly lessens hallucination rates over Standard, yet at the cost of large degradation in CIDEr and SPICE, and underperforms TLC in all metrics. In Tab. 5, we further evaluate hallucination rates on the subset of images where the top beam from Standard was *not* selected by TLC-L with $\text{OFA}_{\text{Large}}$– in other words, samples where using TLC-L made a difference. This occurred in almost a quarter of the captions. Standard hallucination rates are much higher on this subset (*e.g.*, 6.78% CHs), whereas TLC-L reduces this by at least half.

Model	Confidence	Hallucination		Quality	
		CHs (↓)	CHi (↓)	CIDEr (↑)	SPICE (↑)
$\text{OFA}_{\text{Large}}$	Standard	2.79	1.78	**144.4**	**25.8**
	ITM	2.57	1.76	126.5	24.4
	TLC-A	1.81	1.24	140.7	25.5
	TLC-L	**1.74**	**1.17**	141.8	25.4
OFA_{Base}	Standard	3.78	2.39	**142.9**	**25.6**
	ITM	3.22	2.15	127.1	24.3
	TLC-A	2.47	1.75	137.5	25.2
	TLC-L	**2.05**	**1.48**	137.5	24.9
OFA_{Tiny}	Standard	11.01	7.23	**117.4**	**21.7**
	ITM	9.42	6.51	106.6	20.6
	TLC-A	9.87	6.86	115.8	21.5
	TLC-L	**8.79**	**6.43**	113.9	21.3

Table 3. Hallucination rates and captioning metrics on our test set when generating captions with a beam size of 25.

Standard metrics overlook hallucinations. CIDEr and SPICE decrease across all TLC approaches, despite having dramatic reductions in hallucination rates. This effect was also observed in prior works [12, 45], which describe how standard metrics can often fail to penalize hallucinations. For instance, the majority of a sentence might overlap with a reference caption, yet still, misclassify an object. [36] nevertheless find that some visually-impaired users of captioning systems prefer correctness above possibly-wrong detail, motivating the drive for low hallucination rates.

TLC improves caption diversity. From Tab. 4, our method achieves better diversity performance across all model sizes. For instance, TLC-A increases bigram uniqueness score *Div-2* and decreases the repetition measure *Re-4*. Incorporating confidence into caption selection may help overcome language priors, where co-occurrence statistics from training influence token likelihoods. Diversity can improve as a result, where captions are driven more by consistency with the image rather than language. For example, the top center sample in Fig. 3 shows the baseline hallucinating a “metal chair”, compared to the correct yet uncommon words “wrought iron fence” described by TLC-L.

TLC-L generalizes beyond beam search. We test TLC-L with nucleus sampling [23] instead of beam search to gen-

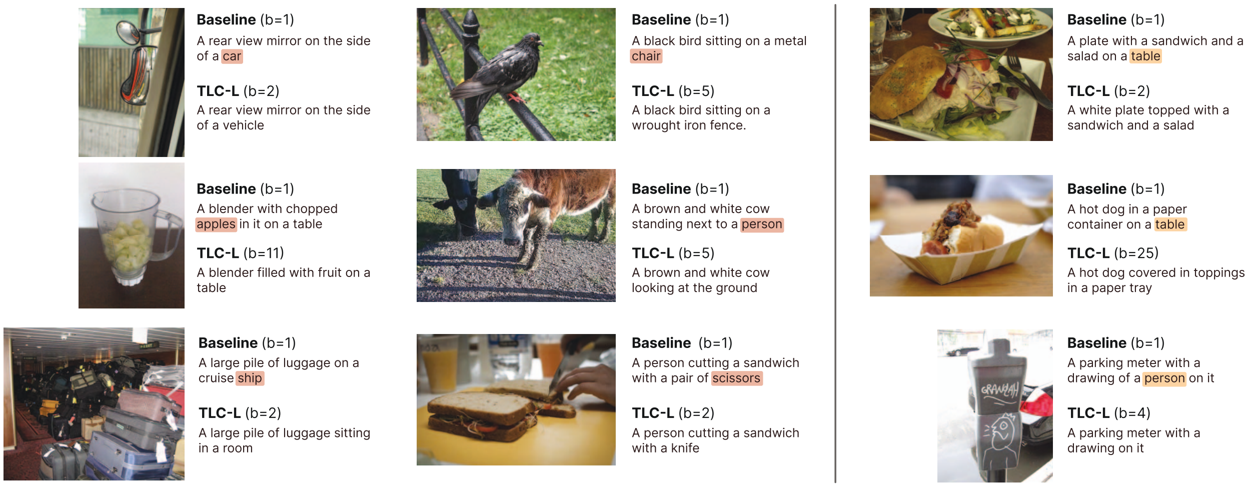

Figure 3. Qualitative examples from our test set in which TLC-L avoided hallucinations in the original (Baseline) captions. In the rightmost column, we show cases where the MS COCO object annotations did not exhaustively include all objects present. Captions are generated with $\text{OFA}_{\text{Large}}$ and a beam size of 25, and $(b = i)$ refers to the index i of the beam as ranked by the Baseline.

Model	Conf.	Vocab Size (↑)	% Novel (↑)	Div-2 (↑)	Re-4 (↓)
$\text{OFA}_{\text{Large}}$	Std.	2822	77.07	6.97	66.34
	TLC-A	**2980**	**78.97**	**7.37**	**64.74**
	TLC-L	2915	77.70	7.13	65.54
OFA_{Base}	Std.	2272	75.43	5.68	71.14
	TLC-A	**2453**	**78.49**	**6.13**	69.28
	TLC-L	2452	77.53	6.03	**69.76**
OFA_{Tiny}	Std.	1130	74.80	2.73	83.29
	TLC-A	1211	75.71	2.91	82.68
	TLC-L	**1243**	**77.05**	**3.01**	**82.12**

Table 4. Caption diversity metrics, evaluated on our test set.

Subset	# I	Method	CHs (↓)	CHi (↓)
Full test set	20,252	Standard	2.79	1.78
		TLC-L	**1.74**	**1.17**
TLC-L, $b > 1$	5,401	Standard	6.78	3.22
		TLC-L	**2.81**	**1.61**

Table 5. Top: Results on the full test set reported in Tab. 3. Bottom: Hallucination rates on a subset of images where TLC-L did not choose the top beam. # I denotes the number of images in each set. Results are shown for $\text{OFA}_{\text{Large}}$.

erate captions. We sample 25 candidates with a top-p of 0.6, and apply the same re-ranking procedure for TLC-L as in Sec. 3.3.2. In Tab. 6, we find a similar trend to experiments with beam search: TLC-L lowers hallucination rates, with a decrease in standard captioning metrics.

Qualitative analysis. We show several qualitative examples in Fig. 3. In the left column, we see two cases where TLC-L "backed-off" to a more general concept, whereas the baseline was specific, yet the image did not contain enough information to determine whether the specificity was indeed correct (*e.g.*, "car" vs. "vehicle" and "apples" vs. "fruit"). A prior work [17] explicitly optimized for this hierarchical generalization of unknown concepts, whereas here it emerges when considering confidence. TLC-L also avoids misclassification errors, such as "person" or "scissors" in the middle column. On the right column, we show cases influenced by incomplete object annotations. For example, the reference segmentations and captions might overlook the object "table". TLC-L rejects captions that mention "table" in some of these cases, reflecting its training objective where correctness was judged based on faithfulness to the reference distribution. We include additional examples, including several failure cases, in Appendix D.

Confidence	CHs (↓)	CHi (↓)	CIDEr (↑)	SPICE (↑)
Standard	3.00	1.89	**142.4**	**25.8**
TLC-L	**2.03**	**1.37**	138.6	25.3

Table 6. Generating $B = 25$ candidates with nucleus sampling (top-p: 0.6), using $\text{OFA}_{\text{Large}}$.

Improving the captioning model improves TLC-A. In Tab. 8, we test TLC when augmenting the $\text{OFA}_{\text{Large}}$ training set with the training data for g. Both Standard-Aug and TLC-A-Aug outperform their non-augmented counterparts, with TLC-A-Aug reaching the lowest hallucination rates. Unsurprisingly, training TLC-L-Aug on the Standard-Aug training set overfits. When keeping the captioning model frozen, TLC-L remains the most effective.

Reported in	Method	Beam Size	XE Loss B@4	S	M	C	CHs (↓)	CHi (↓)	SC Loss B@4	S	M	C	CHs (↓)	CHi (↓)
[69] CVPR 2021	Transformer	unk	-	-	-	-	-	-	38.6	22.0	28.5	128.5	12.1	8.1
[69] CVPR 2021	Transformer+CATT	unk	-	-	-	-	-	-	39.4	22.8	29.3	131.7	9.7	6.5
[68] PAMI 2021	UD-DICv1.0	5	-	-	-	-	-	-	38.7	21.9	28.4	128.2	10.2	6.7
[7] WACV 2022	UD-L	no	34.4	20.7	27.3	112.7	6.4	4.1	37.7	22.1	28.6	124.7	5.9	3.7
[7] WACV 2022	UD-L + Occ	no	33.9	20.3	27.0	110.7	5.9	3.8	37.7	22.2	28.7	125.2	5.8	3.7
[35] CVPR 2022	CIIC$_G$	3	37.3	21.5	28.5	119.0	5.3	3.6	40.2	23.2	29.5	133.1	7.7	4.5
[34] CVPR 2022	COS-Net	3	39.1	22.7	29.7	127.4	4.7	3.2	42.0	24.6	30.6	141.1	6.8	4.2
This work	OFA$_{\text{Large}}$ [59]	5	**41.8**	**24.4**	**31.3**	**140.7**	3.1	2.0	**42.3**	**25.5**	**31.6**	**145.0**	3.1	2.0
This work	OFA$_{\text{Large}}$ + TLC-L	5	41.2	24.1	30.9	138.4	***2.0**	***1.4**	42.0	25.2	31.4	143.8	**2.3**	**1.5**

Table 7. Comparison to prior work for hallucination in image captioning on the MS COCO Karpathy test split. Although we add a noun parser for our results in Tables 3, 4, 5, and 6, we remove this step here and use the original evaluation provided by [45] to be consistent with prior work. We show captioning metrics B@4 (BLEU [40]), S (SPICE [4]), M (METEOR [29]), and C (CIDEr [56]). * indicates state-of-the-art for hallucination rates.

Confidence	Train set f_{cap}	Train set g	CHs (↓)	CHi (↓)	C (↑)	S (↑)
Standard	Train	–	2.79	1.78	**144.4**	**25.8**
TLC-A	Train	–	1.81	1.24	140.7	25.5
TLC-L	Train	Val	**1.74**	**1.17**	141.8	25.4
Standard-Aug	Train + Val	–	2.20	1.38	**153.3**	**26.7**
TLC-A-Aug	Train + Val	–	**1.49**	**1.00**	149.8	26.2
TLC-L-Aug	Train + Val	Train + Val	2.02	1.28	152.7	26.6

Table 8. Results when augmenting the OFA$_{\text{Large}}$ train set with validation data ("-Aug"). TLC-L-Aug thus overfits confidence to data the captioning model was already trained on.

TLC-L with OFA$_{\text{Large}}$ sets a new state-of-the-art. We compare to previous results on MS COCO object hallucination in Tab. 7. We re-train our captioning models and confidence estimators on a dataset split that does not overlap with the Karpathy test split used for evaluation [27]. [45] show that training with a self-critical (SC) loss after training with cross-entropy (XE) [44] can improve captioning metrics, yet worsen hallucination rates compared to training with XE alone. We find that the baseline OFA$_{\text{Large}}$ has similar hallucination rates for XE and SC, yet TLC-L indeed produces the least hallucinations on top of the XE model. This sets a new state-of-the-art of **2.0%** and **1.4%** for CHs and CHi respectively. Notably, TLC-L reduces hallucination without requiring any architecture changes to its captioning model, in contrast to the recent COS-Net, where specific modules were introduced to capture image semantics.

5. Discussion and Limitations

While TLC-L provides effective confidence estimates for caption generation, it requires domain-specific training data for learning a confidence estimator from scratch on top of captioning model features. TLC-A, on the other hand, uses the captioning model outputs directly, which leverages generalization ability from large-scale pretraining. Thus, TLC-A can be effectively applied in settings where in-domain training data for captioning is not available. To combine these advantages, future research could explore unsupervised methods for learning correctness. Additionally, we use algebraic confidence estimates from uncalibrated output distributions, where output probabilities do not necessarily match actual probabilities of correctness. Potential future work may apply calibration methods to token-level confidence for improving caption correctness.

6. Conclusion

In this work, we have explored a simple method using Token-Level Confidence (TLC) for determining whether a caption correctly describes an image, a critical part of vision-language understanding. We find that judging caption correctness at a finer granularity than existing approaches leads to improvements in several settings, such as evaluating compositional reasoning with image-caption pairs or reducing object hallucinations in generated captions. To do so, TLC uses a vision-language model finetuned on image captioning to produce token confidences, and then aggregates either algebraic (TLC-A) or learned token confidences (TLC-L) over words or sequences to estimate image-caption consistency. Increasing the confidence granularity with TLC-A raises image and group scores over ITM on Winoground [52] by over 2x, and improves accuracy in verb understanding on SVO-Probes [22] by a relative 10%. When training data are available to learn and calibrate confidences with TLC-L, we reduce object hallucination rates on COCO Captions by a relative 30%, setting a new state-of-the-art. Overall, our results demonstrate that token-level confidence, whether algebraic or learned, can be a powerful yet simple resource for reducing errors in captioning output and assessing image-caption consistency.

References

[1] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4971–4980, 2018. 1

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022. 2

[3] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Guided open vocabulary image captioning with constrained beam search. *arXiv preprint arXiv:1612.00576*, 2016. 2

[4] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *European conference on computer vision*, pages 382–398. Springer, 2016. 6, 8

[5] 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*, pages 6077–6086, 2018. 2

[6] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009. 6

[7] Ali Furkan Biten, Lluís Gómez, and Dimosthenis Karatzas. Let there be a clock on the beach: Reducing object hallucination in image captioning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1381–1390, 2022. 2, 8

[8] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *arXiv:1504.00325*, 2015. 2, 4, 5

[9] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv:2209.06794*, 2022. 2, 3

[10] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*, pages 1931–1942. PMLR, 2021. 2

[11] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 2

[12] Wenliang Dai, Zihan Liu, Ziwei Ji, Dan Su, and Pascale Fung. Plausible may not be faithful: Probing object hallucination in vision-language pre-training. *arXiv preprint arXiv:2210.07688*, 2022. 2, 6

[13] Shrey Desai and Greg Durrett. Calibration of pre-trained transformers. *arXiv preprint arXiv:2003.07892*, 2020. 3

[14] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015. 2

[15] Zhengcong Fei. Efficient modeling of future context for image captioning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5026–5035, 2022. 2

[16] Taisiya Glushkova, Chrysoula Zerva, Ricardo Rei, and André FT Martins. Uncertainty-aware machine translation evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021. 2

[17] Sergio Guadarrama, Niveda Krishnamoorthy, Girish Malkarnenkar, Subhashini Venugopalan, Raymond Mooney, Trevor Darrell, and Kate Saenko. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 2712–2719, 2013. 7

[18] Nuno M Guerreiro, Elena Voita, and André FT Martins. Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation. *arXiv preprint arXiv:2208.05309*, 2022. 2

[19] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017. 3

[20] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3608–3617, 2018. 1

[21] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *Proceedings of the European conference on computer vision (ECCV)*, pages 771–787, 2018. 1

[22] Lisa Anne Hendricks and Aida Nematzadeh. Probing image-language transformers for verb understanding. *arXiv preprint arXiv:2106.09141*, 2021. 1, 2, 5, 8

[23] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019. 6

[24] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spacy: Industrial-strength natural language processing in python. 2020. 5, 6

[25] Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. Scaling up vision-language pre-training for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17980–17989, 2022. 2

[26] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4634–4643, 2019. 2

[27] 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. 2, 8

[28] Aviral Kumar and Sunita Sarawagi. Calibration of encoder decoder models for neural machine translation. *arXiv preprint arXiv:1903.00802*, 2019. 4

[29] Alon Lavie and Abhaya Agarwal. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the second workshop on statistical machine translation*, pages 228–231, 2007. 8

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

[31] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 2021. 2

[32] 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 (ECCV)*, pages 121–137. Springer, 2020. 2

[33] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. 6

[34] Yehao Li, Yingwei Pan, Ting Yao, and Tao Mei. Comprehending and ordering semantics for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17990–17999, 2022. 2, 8

[35] Bing Liu, Dong Wang, Xu Yang, Yong Zhou, Rui Yao, Zhiwen Shao, and Jiaqi Zhao. Show, deconfound and tell: Image captioning with causal inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18041–18050, 2022. 2, 8

[36] Haley MacLeod, Cynthia L Bennett, Meredith Ringel Morris, and Edward Cutrell. Understanding blind people's experiences with computer-generated captions of social media images. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 5988–5999, 2017. 6

[37] Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020. 2

[38] Sabrina J Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. Reducing conversational agents' overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics*, 10:857–872, 2022. 2

[39] Rohan Pandey, Rulin Shao, Paul Pu Liang, Ruslan Salakhutdinov, and Louis-Philippe Morency. Cross-modal attention congruence regularization for vision-language relation alignment. *arXiv preprint arXiv:2212.10549*, 2022. 5

[40] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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, 2002. 8

[41] Jae Sung Park, Sheng Shen, Ali Farhadi, Trevor Darrell, Yejin Choi, and Anna Rohrbach. Exposing the limits of video-text models through contrast sets. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3574–3586, 2022. 2

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. 1

[43] Shuhuai Ren, Junyang Lin, Guangxiang Zhao, Rui Men, An Yang, Jingren Zhou, Xu Sun, and Hongxia Yang. Learning relation alignment for calibrated cross-modal retrieval. *arXiv preprint arXiv:2105.13868*, 2021. 5

[44] 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*, pages 7008–7024, 2017. 2, 3, 8

[45] 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*, 2018. 1, 2, 4, 6, 8

[46] Fawaz Sammani and Mahmoud Elsayed. Look and modify: Modification networks for image captioning. In *British Machine Vision Conference (BMVC)*, 2019. 2

[47] Fawaz Sammani and Luke Melas-Kyriazi. Show, edit and tell: a framework for editing image captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4808–4816, 2020. 2

[48] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How much can clip benefit vision-and-language tasks? In *International Conference on Learning Representations (ICLR)*, 2022. 2

[49] Yilin Shen, Wenhu Chen, and Hongxia Jin. Modeling token-level uncertainty to learn unknown concepts in slu via calibrated dirichlet prior rnn. *arXiv preprint arXiv:2010.08101*, 2020. 2

[50] Rakshith Shetty, Marcus Rohrbach, Lisa Anne Hendricks, Mario Fritz, and Bernt Schiele. Speaking the same language: Matching machine to human captions by adversarial training. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4135–4144, 2017. 6

[51] Elias Stengel-Eskin and Benjamin Van Durme. Calibrated interpretation: Confidence estimation in semantic parsing. *arXiv preprint arXiv:2211.07443*, 2022. 3

[52] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross.

Winoground: Probing vision and language models for visio-linguistic compositionality. In *CVPR*, 2022. 1, 2, 5, 8

[53] Dustin Tran, Jeremiah Liu, Michael W Dusenberry, Du Phan, Mark Collier, Jie Ren, Kehang Han, Zi Wang, Zelda Mariet, Huiyi Hu, et al. Plex: Towards reliability using pretrained large model extensions. *arXiv preprint arXiv:2207.07411*, 2022. 2

[54] Neeraj Varshney, Swaroop Mishra, and Chitta Baral. Investigating selective prediction approaches across several tasks in iid, ood, and adversarial settings. *arXiv preprint arXiv:2203.00211*, 2022. 3

[55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010, 2017. 2, 3, 5

[56] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 4, 6, 8

[57] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015. 2

[58] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022. 2

[59] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *Proceedings of the 39th International Conference on Machine Learning*, 2022. 2, 3, 4, 8

[60] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *arXiv preprint arXiv:2202.03052*, 2022. 1

[61] Shuo Wang, Zhaopeng Tu, Shuming Shi, and Yang Liu. On the inference calibration of neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020. 2

[62] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022. 2

[63] Zhen Wang, Long Chen, Wenbo Ma, Guangxing Han, Yulei Niu, Jian Shao, and Jun Xiao. Explicit image caption editing. In *European Conference on Computer Vision (ECCV)*, pages 113–129. Springer, 2022. 2

[64] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. In *International Conference on Learning Representations (ICLR)*, 2022. 2

[65] Spencer Whitehead, Suzanne Petryk, Vedaad Shakib, Joseph Gonzalez, Trevor Darrell, Anna Rohrbach, and Marcus Rohrbach. Reliable visual question answering: Abstain rather than answer incorrectly. In *ECCV*, 2022. 1, 2, 3, 4, 5

[66] Yijun Xiao and William Yang Wang. On hallucination and predictive uncertainty in conditional language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, 2021. 2

[67] Yilei Xiong, Bo Dai, and Dahua Lin. Move forward and tell: A progressive generator of video descriptions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 468–483, 2018. 6

[68] Xu Yang, Hanwang Zhang, and Jianfei Cai. Deconfounded image captioning: A causal retrospect. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2, 8

[69] Xu Yang, Hanwang Zhang, Guojun Qi, and Jianfei Cai. Causal attention for vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9847–9857, 2021. 2, 8

[70] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 2

[71] 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*, pages 5579–5588, 2021. 2

[72] Shujian Zhang, Chengyue Gong, and Eunsol Choi. Knowing more about questions can help: Improving calibration in question answering. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021. 2

[73] Chunting Zhou, Graham Neubig, Jiatao Gu, Mona Diab, Paco Guzman, Luke Zettlemoyer, and Marjan Ghazvininejad. Detecting hallucinated content in conditional neural sequence generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021. 2

[74] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 2