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 AI

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 outperforms prior state-of-the-art in image and group scores for compositional reasoning in Winoground by a relative 37% and 9%, respectively. 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.

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 [2, 20]. State-of-the-art models, like CLIP [42] or OFA [62], 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 [53] 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 [21] 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 [19, 67].

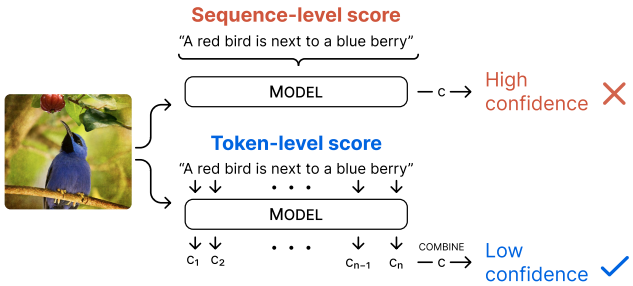


Figure 1: Judging caption correctness is still a challenge for large-scale models that operate at a sequence-level. We show that both algebraic and learned confidences at a token-level from a finetuned image captioning model improve fine-grained estimates of caption correctness.

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.

Leveraging this observation, we explore Token-Level

Dataset & Task	Winoground [53]	SVO-Probes [21]	Hallucination in Captioning [45]
Metric	Acc Image (↑)	Accuracy (↑)	CHi (↓)
Prior SoTA	19.75 [43]	–	3.2 [32]
Baseline (ours)	10.25	81.23	2.0
Ours	**27.00**	**89.47**	**1.4**
Rel. Improvement	37%	10%	30%

Table 1: Summary of results. Despite its simplicity, the relative improvement over the next best approach highlights the significance of TLC for caption correctness.

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 [10, 61, 66, 60]. In this work, we demonstrate the effectiveness of token-level confidence across multiple model sizes of OFA [61], a recent Transformer-based model [56] with strong performance on many vision-language tasks. As summarized in Tab. 1, on the challenging Winoground [53] 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 additionally shows a relative improvement of image and group scores of 37% and 9%, respectively, over the prior state-of-the-art on Winoground [43], which used a regularization tailored for multimodal alignment. TLC-A also outperforms ITM on a fine-grained verb understanding task [21] by a relative 10% (Sec. 4.3). When using TLC-L to re-rank candidate captions on MS COCO [9], 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. [21, 41, 53] 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 [8, 32], in some cases relying on causal inference-based approaches [33, 71, 70]. 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, 65]. Some works propose caption decoding methods such as constrained beam search [4], an uncertainty-aware beam search using prediction entropy [68], or a non-autoregressive caption decoding method [14] to target criteria such as correctness. However, the original formulation of beam search remains the dominant decoding method used in modern multimodal architectures [10, 29, 61, 64]. 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 [75]. A number of works study model uncertainty and aim to improve model calibration for machine translation [15, 17, 63], dialog [38], question answering [74] and spoken language understanding [50], 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 [67].

Image captioning. Image captioning has seen significant progress since the arrival of deep learning as a dominant methodology [6, 13, 24, 25, 44, 58]. In recent years Transformer-based architectures have gained particular prominence [31, 49, 73]. Many papers take the approach of pretraining large vision-and-language models and then adapting them to downstream tasks, including captioning [30, 66]. Recent efforts focus on further scaling these

pretraining-based methods [3, 23, 60, 72], while many also aim to unify multiple vision-and-language tasks during pretraining [10, 11, 61, 64]. 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 [54], 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 [67], 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 [18, 67]. 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 [12, 52, 55], 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 [67]. 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 begin with a trained and frozen f_{cap} and use a heldout dataset $\mathcal{X}$ for training a confidence estimator g. Compared to the training set for f_{cap}, $\mathcal{X}$ provides a better estimate of the captioning model's performance on test data. In this work, we simply use the f_{cap} validation set. For each image in $\mathcal{X}$, paired with one or more references, we select one of the reference captions $t_{1:n}$ and time step k within the caption. We first input the *prefix*, or $t_{1:k-1}$, into the f_{cap} decoder to predict the next token, $\hat{t}_k$. We assign a binary label c to $\hat{t}_k$ – it is **correct** $(c = 1)$ if it matches the reference token t_k or any token at k from other reference captions with the same prefix. Otherwise, $\hat{t}_k$ is labeled as **incorrect** $(c = 0)$. For example, in Fig. 2, the original reference token t_k is "sleeping", yet "asleep" and "laying" are also considered correct, given that they share the same prefix "a dog". The predicted token $\hat{t}_k$ "standing" is therefore labeled as incorrect. This provides proxy for true consistency with the image, which may be noisy; for example, "resting" would be considered incorrect in Fig. 2. Nevertheless, these labels enable TLC-L to learn effective in-domain confidences (Sec. 4.4). At each epoch, we resample a reference caption and a time step k for each image in order to leverage all available ground-truth tokens.

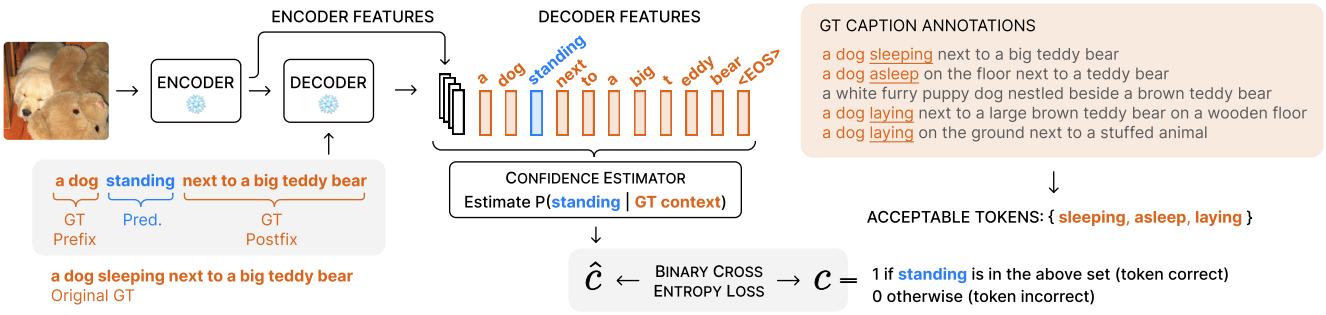


Figure 2: TLC-L: A framework to learn token-level confidence for a pretrained autoregressive encoder-decoder captioning model. We first use the captioning model to predict the next token (*e.g.*, "standing") after a partial reference caption (*e.g.*, "a dog"), shown in the bottom left. We input this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

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 [56], 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. Although we supervise confidence for a single token $\hat{t}_k$ at a time, the full caption context is given as input. 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 [61, 10], whereas estimating token confidence bidirectionally uses future context to inform correctness.

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 [27]. 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:

$$\mathbb{P}(t_{1:n}) = \sum_{k=1}^{n} \log p(t_k \mid t_{1:k-1}, x) \tag{1}$$

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 [57] could also have higher hallucination rates. It is also similar to prior work in machine translation [17], which noted that errors can be "bad luck" from generation rather than inherent model failure.

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 [9] 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 [67]. 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 control the rate of false positives. This is captured by the precision: *"out of all samples predicted as correct, what fraction are* actually *correct?"* We define a target precision α, such as 99%, and select γ such that the binary decisions $\hat{c} \geq \gamma$ maximize the recall of correct samples in $\mathcal{S}$ on the validation set.

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 [61], 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 OFA_{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 [9], 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 [67]. Additional dataset details are in Appendix D.

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 B. 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 [56], followed by a 2-layer MLP. We add a learned positional encoding to the token features, and train g for 200 epochs on 8 V100 GPUs. Additional details are in Appendix E.

Model	Conf.	Text	Image	Group
MTurk Human [53]	-	89.50	88.50	85.50
Random Chance [53]	-	25.00	25.00	16.67
$\text{UNITER}_{\text{Large}}$ [53]	ITM	38.00	14.00	10.50
VinVL [53]	ITM	37.75	17.75	14.50
$\text{CACR}_{\text{Base}}$ [39]	CACR	39.25	17.75	14.25
$\text{IAIS}_{\text{Large}}$ [39]	IAIS	*42.50	19.75	16.00
$\text{OFA}_{\text{Large}}$	ITM	**30.75**	10.25	7.25
	TLC-A	29.25	***27.00**	***17.50**
	(Δ)	(−1.5)	(+16.75)	(+10.25)
OFA_{Base}	ITM	**26.75**	10.75	6.50
	TLC-A	24.50	**23.50**	**13.75**
	(Δ)	(−2.25)	(+12.75)	(+7.25)
OFA_{Tiny}	ITM	**22.75**	7.75	4.50
	TLC-A	16.50	**15.75**	**6.75**
	(Δ)	(−6.25)	(+8.00)	(+2.25)

Table 2: Accuracy on text, image, and group score for the Winoground evaluation dataset [53]. Citations indicate where scores are reported, and * indicates state-of-the-art.

4.2. Correctness Around Compositional Reasoning

First, we assess the ability of TLC-A to select corresponding image-caption pairs. We use Winoground [53], 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.*, $c_{POS} > c_{NEG}$). [53] find that the task is surprisingly difficult, with all models they test performing below random chance for image and group score.

As correctness estimates, [53] 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. 2.

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

Table 3: Image-caption matching accuracy for verb understanding with a subset of SVO-Probes [21]. TLC-A uses token-level softmax scores aggregated over the verb in each example.

TLC-A outperforms prior SOTA image and group performance. TLC-A with $\text{OFA}_{\text{Large}}$ reaches above random chance for both image and group score, improving over prior state-of-the-art. Despite its simplicity, with no additional training beyond standard image captioning, TLC-A outperforms IAIS (proposed in [43]), a training method optimized for multimodal attention alignment. Compared to ITM across OFA model sizes, TLC-A more than doubles the image and group scores in all but one case (OFA_{Tiny} group).

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 [21] 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 [22] to annotate the location of the verb in each caption. We aggregate token confidences over the verb tokens for TLC-A. Tab. 3 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. 3, 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 described in 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 data. [45] provides a method to enumerate MS COCO objects mentioned in references for a given image and enumerate objects mentioned in an arbitrary, predicted caption. We also add part-of-speech taggers [7, 22] to exclude predicted words that are not nouns; however, when comparing directly to prior work, we use the original implementation. A hallucination is flagged when a prediction mentioned an object not present in the reference set. This is evaluated by sentence-level and object instance-level CHAIRs and CHAIRi metrics [45]:

$$\text{CHAIR}_s = \frac{\text{\# captions with} \geq 1 \text{ hallucination}}{\text{\# captions}} \quad (2)$$

$$\text{CHAIR}_i = \frac{\text{\# objects hallucinated}}{\text{\# objects mentioned}} \quad (3)$$

We report standard captioning metrics [5, 57] as well as CHAIRs and CHAIRi (or CHs and CHi). We also report several caption diversity measures [51, 69] 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.

For both TLC-A and TLC-L, we choose a threshold γ on the validation set. This threshold is used at test time to make binary decisions on the correctness of a given object in a predicted caption. We extract all objects from the validation set predictions, as well as corresponding token confidences and ground-truth hallucination scores. Then, we choose a confidence level γ that reaches at least 99% precision when separating correct vs. hallucinated objects. This precision is intentionally very high; the OFA captioning models have fairly low rates of hallucination on MS COCO already (as seen in Tab. 4), yet we are interested in pushing the caption reliability as far as possible. When aggregating token confidences over object words, we select the minimum value for TLC-A and the average value for TLC-L based on the validation set recall. We use a large beam size of $B = 25$ to observe the behavior of our caption selection method when given many possible candidates.

We show results from the following methods. **Standard** uses the original top caption, that is, the caption from the beam ranked highest by f_{cap}. **Standard-Aug** uses the top caption from a captioning model f'_{cap}, where its training set is augmented by the training set for g. This tests whether

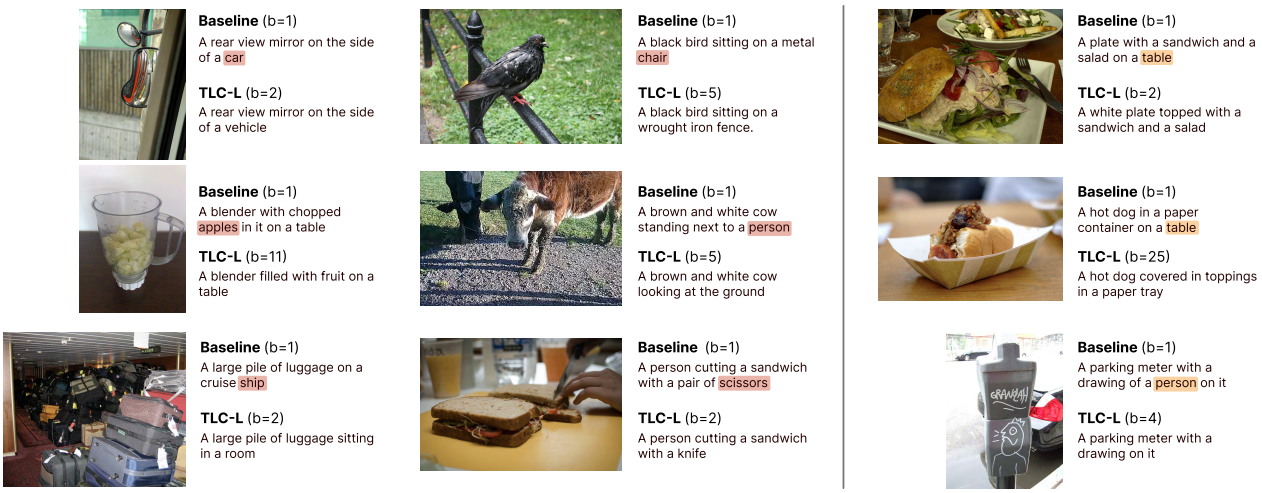


Figure 3: Qualitative examples from our test set in which TLC-L avoided hallucinations in the original (Baseline) captions. In the rightmost column, we show cases where the MS COCO object annotations did not exhaustively include all objects present. Captions are generated with $\text{OFA}_{\text{Large}}$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

the improvements from TLC-L result from using token confidence itself or from additional training data. More details on Standard-Aug are in Appendix E. **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.

Learned confidences lead to the least hallucinations. From Tab. 4, 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. Additionally, TLC-L lowers hallucination rates compared to Standard-Aug as well (*e.g.*, a relative 20.9% lower CHs for $\text{OFA}_{\text{Large}}$). This indicates that reserving a portion of data to train g can have a bigger impact on reducing hallucinations than does using the data for augmentation. 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. 6, 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.

Captioning metrics do not capture hallucinations. CIDEr and SPICE decrease across all TLC-based approaches, despite having dramatic reductions in hallucination rates. This effect was also observed by [45], which described 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.

Model	Confidence	Hallucination		Quality	
		CHs (↓)	CHi (↓)	CIDEr (↑)	SPICE (↑)
$\text{OFA}_{\text{Large}}$	Standard-Aug	2.20	1.38	153.3	26.7
	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-Aug	3.00	1.89	148.8	26.1
	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-Aug	10.58	6.83	119.8	22.1
	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 4: Hallucination rates and captioning metrics on our test set when generating captions with a beam size of 25.

TLC improves caption diversity. From Tab. 5, our method achieves higher performance on diversity metrics across all model sizes. For instance, TLC-A consistently increases bigram uniqueness score *Div-2*, and decreases the repeti-

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 5: 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 6: Top: Results on the full test set reported in Tab. 4. 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 $\text{OFA}_{\text{Large}}$.

tion 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.

Qualitative analysis. We show several qualitative examples in Fig. 3. In the left column, we see two examples 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 [16] 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 examples 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 C.

TLC-L with $\text{OFA}_{\text{Large}}$ 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 [25]. [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 $\text{OFA}_{\text{Large}}$ has similar hallucination rates for XE and SC, yet TLC-L indeed produces the least hallucinations on top of the XE model. This leads to 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 prior SOTA of COSNet, 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. Finally, learned confidences may also be incorporated into decoding methods that are not autoregressive.

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 fine-tuned 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 esti-

Reported in	Method	Beam Size	XE Loss						SC Loss					
			B@4	S	M	C	CHs (↓)	CHi (↓)	B@4	S	M	C	CHs (↓)	CHi (↓)
[45] EMNLP 2018	NBT [35]	5	-	19.4	26.2	105.1	7.4	5.4	-	-	-	-	-	-
[45] EMNLP 2018	TopDown [6] (no Boxes)	5	-	19.9	26.7	107.6	8.4	6.1	-	20.4	27.0	117.2	13.6	8.8
[45] EMNLP 2018	TopDown [6]	5	-	20.4	27.1	113.7	8.3	5.9	-	21.4	27.7	120.6	10.4	6.9
[71] CVPR 2021	Transformer	unk	-	-	-	-	-	-	38.6	22.0	28.5	128.5	12.1	8.1
[71] CVPR 2021	Transformer+CATT	unk	-	-	-	-	-	-	39.4	22.8	29.3	131.7	9.7	6.5
[70] PAMI 2021	UD-DICv1.0	5	-	-	-	-	-	-	38.7	21.9	28.4	128.2	10.2	6.7
[8] 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
[8] 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
[33] 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
[32] 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$_{\text{Large}}$ [61]	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 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 [5]), M (METEOR [28]), and C (CIDEr [57]). * indicates state-of-the-art for hallucination rates.

mate image-caption consistency. Increasing the confidence granularity with TLC-A improves over prior state-of-the-art image and group scores on Winoground [53] by a relative 37% and 9%, respectively, and improves accuracy in verb understanding on SVO-Probes [21] 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.

Acknowledgements. We thank David Chan, Kate Saenko, and Anastasios Angelopoulos for helpful discussions and feedback. Authors, as part of their affiliation with UC Berkeley, were supported in part by the NSF CISE Expeditions Award CCF-1730628, DoD, including DARPA's LwLL, PTG, and/or SemaFor programs, the Berkeley Artificial Intelligence Research (BAIR) industrial alliance program, as well as gifts from Astronomer, Google, IBM, Intel, Lacework, Microsoft, Mohamed Bin Zayed University of Artificial Intelligence, Nexla, Samsung SDS, Uber, and VMware.

References

[1] Getty images api. https://www.gettyimages.com/. 14

[2] 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

[3] 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. 3

[4] 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

[5] 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, 9

[6] 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, 9

[7] 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

[8] 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, 9

[9] 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, 13

[10] 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, 4

[11] 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. 3

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

[13] 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

[14] 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

[15] 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

[16] 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. 8

[17] 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, 4

[18] 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

[19] 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

[20] 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

[21] Lisa Anne Hendricks and Aida Nematzadeh. Probing image-language transformers for verb understanding. *arXiv preprint arXiv:2106.09141*, 2021. 1, 2, 6, 9, 14, 15

[22] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spacy: Industrial-strength natural language processing in python. 2020. 6, 14

[23] 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. 3

[24] 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

[25] 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

[26] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 15

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

[28] 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. 9

[29] 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

[30] 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

[31] 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

[32] 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, 9

[33] 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, 9

[34] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems*, 33:21464–21475, 2020. 13

[35] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Neural baby talk. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 9

[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. 7

[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. 9

[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. 2, 5, 6

[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, 7, 8, 9

[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] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 15

[49] 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

[50] 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

[51] 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

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

[53] 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, 9, 13, 14

[54] 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. 3

[55] 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

[56] 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, 4, 5, 15

[57] 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, 9

[58] 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

[59] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020. 13

[60] 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, 3

[61] 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, 5, 9, 14, 15

[62] 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

[63] 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

[64] 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, 3

[65] 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

[66] 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

[67] 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, 5, 13

[68] 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, 13

[69] 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

[70] Xu Yang, Hanwang Zhang, and Jianfei Cai. Deconfounded image captioning: A causal retrospect. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2, 9

[71] 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, 9

[72] 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. 3

[73] 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

[74] 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

[75] 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

Appendix

A. Overview

Appendix B presents an ablation showing several alternative algebraic confidence estimates, and compares the precision-recall curve for the learned TLC-L to that of algebraic confidences when separating correct and hallucinated objects. Appendix C presents additional qualitative examples of both success and failure cases, comparing TLC-L to the Baseline model. Appendix D and Appendix E provide further details on datasets and models respectively.

B. Alternative Confidence Estimates

We compare several other choices of algebraic confidence estimates for TLC-A besides softmax score used in the main paper. All are derived from the likelihood (logit) distribution $\vec{z}_k$, as mentioned in Sec. 3.1. **Logit** is the logit value for the selected token directly from $\vec{z}_k$, whereas **Softmax** is the corresponding value after a softmax function. Again, in our main paper, TLC-A is based on this softmax score confidence. **Entropy** is the negative entropy of the log-softmax distribution, as a higher entropy should indicate higher uncertainty. Entropy has been previously used as a direct estimate of model uncertainty [59] as well as a penalty in image caption decoding [68]. Finally, we consider the **Energy** score [34], originally proposed as a measure for OOD detection that theoretically correlates with the probability density of the in-domain samples. We use a temperature of 1, and negate the energy score so positive values indicate confident samples.

In Fig. 4, we show the precision-recall curve for various confidence estimates to separate correct and hallucinated objects. We compute these results on our MSC-Main validation set for g (see Tab. 8). Specifically, we are not interested in the exact values of confidence estimates themselves, but rather how well they can *rank* correct objects over those that are hallucinated. When using confidence estimates in practice, we need a threshold to make a binary decision about whether an object in a caption is considered hallucinated or not (Sec. 3.3.2). We choose this threshold for a specific precision level, above the accuracy that the model achieves on its own. For instance, on the validation set for g, about 98.3% of the captioning model's predicted objects are correct (and the rest hallucinated). To push reliability further, we choose a threshold γ for each method that achieves a precision of 99%. In Fig. 4 (left), we therefore only show recall rates above 98% precision, yet show the overall area-under-the-curve (AUC) in Fig. 4 (right).

From Fig. 4, we can see that TLC-Learned (*i.e.*, TLC-L) achieves the highest AUC of 99.48%, and TLC-Softmax achieves the second-highest of 99.07%. The precision-recall plot shows that all algebraic confidences reach 0% recall before 99.5% precision, whereas TLC-L still retains about 60% recall at this high precision rate. In our main paper, we use TLC-A to denote TLC-Softmax, as it performed the best among the algebraic confidences.

Dataset	Use Case	# Images	# Captions
MSC - Main	Train f_{cap} and f'_{cap}	82,783	414,113
	Validate f_{cap}, Train g and f'_{cap}	16,202	81,065
	Validate g and f'_{cap}, Select g thresholds	4,050	20,268
	Evaluation	20,252	101,321
MSC - Prior	Train f_{cap}	82,783	414,113
	Validate f_{cap}, Train g	28,403	142,120
	Validate g, Select g thresholds	7,101	35,524
	Evaluation	5,000	25,010
Winoground	Evaluation	800	800
SVO-Probes	Evaluation	12,958	6,479

Table 8: Overview of datasets used in our work. MSC indicates MS COCO Captions [9].

C. Additional Qualitative Examples

In Fig. 5, we present qualitative examples (in addition to those in Fig. 3) where the Baseline model caption contained a hallucination, yet the caption selected by TLC-L did not. Note that "Baseline" refers to "Standard" as in Tab. 4. In Fig. 6, we show several failure cases of TLC-L. On the left is a case where the Baseline model selects a more general caption, whereas TLC-L erroneously rejects it for one with a hallucinated "carrot". On the middle and right, TLC-L selects captions that include other hallucinations of objects. Nevertheless, TLC-L corrected 44.5% $(252/566)$ of captions that contained a hallucination from the Baseline model, whereas TLC-L introduced a hallucination in only 0.2% $(38/19,686)$ of captions that did not contain a hallucination from the Baseline model.

D. Dataset details

MS COCO Captions. We use the same dataset splits as [67] for training and validating the captioning model f_{cap} and confidence estimator g, as [67] similarly reserves validation data in MS COCO for training a confidence estimator (yet for the visual question answering task, rather than image captioning). For the Standard-Aug model f'_{cap} in Tab. 4, we include the training set for g as part of the training set for f'_{cap}. In Tab. 8, we refer to these splits as MSC-Main (for MS COCO Main), and use them for results in Tabs. 4, 5, and 6, and Figs. 3, 4, 5, and 6. For comparison to prior work that uses the Karpathy test split (Tab. 7), we re-split the validation set to prevent overlap. These details are presented as MSC-Prior in Tab. 8.

Winoground. We use the original data and evaluation setup for Winoground as in the original paper [53], which consisted of 800 unique images and captions. This leads to 400 examples, each consisting of two image-caption pairs,

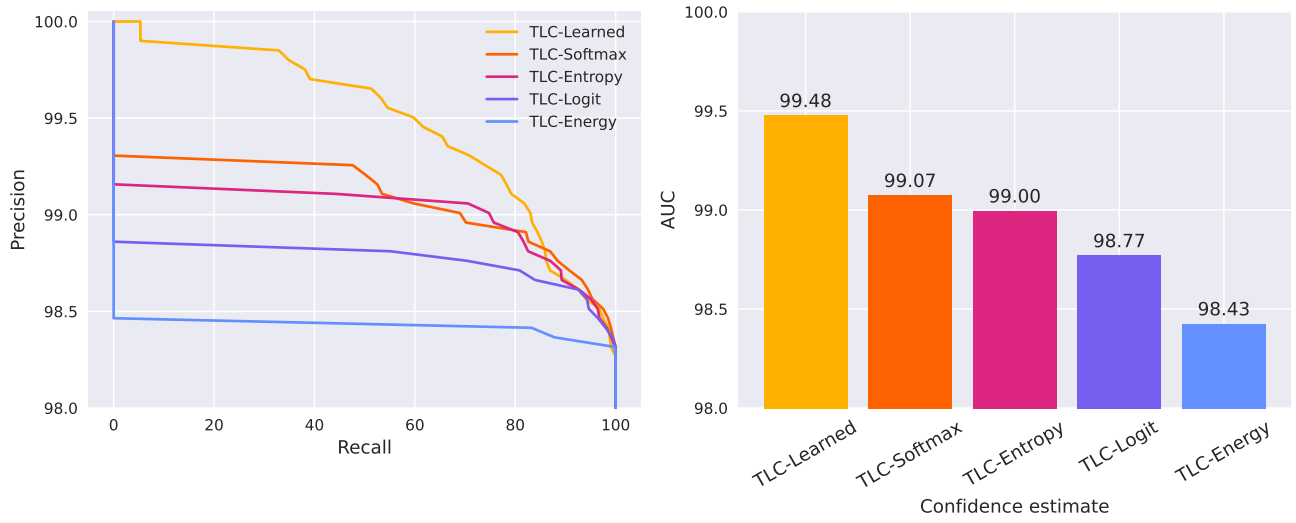


Figure 4: Precision-recall curve (left) and AUC (right) with different confidence estimates for separating correct and hallucinated objects. Results are shown on our validation set using $\text{OFA}_{\text{Large}}$.

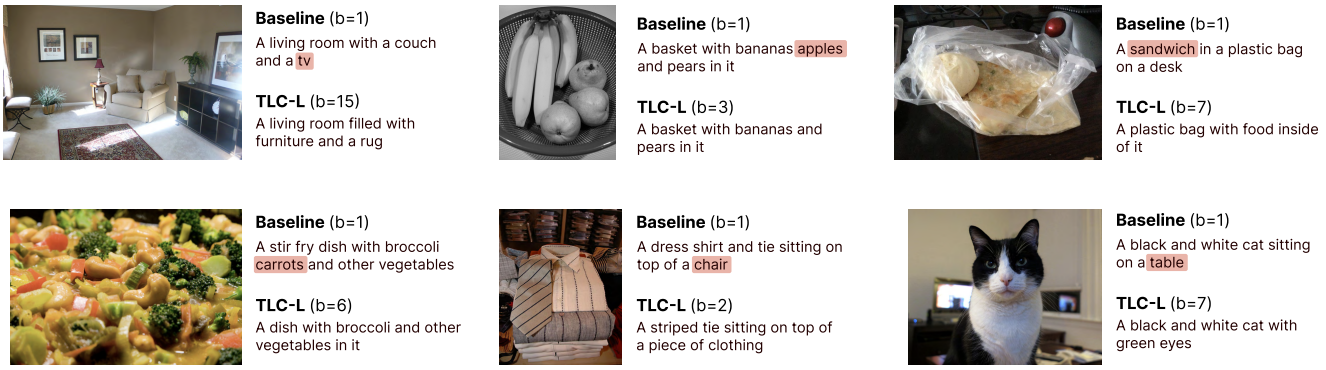


Figure 5: Additional qualitative examples on our test set for TLC-L on $\text{OFA}_{\text{Large}}$, where the Baseline model caption contained a hallucination, yet the caption selected by TLC-L did not.

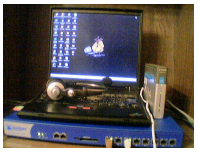
where the captions contain the same words and/or morphemes yet a different word order.

SVO-Probes. For SVO-Probes [21], we use the authors' public code to access a subset of data where the images were available. As discussed in Sec. 4.3, each image is annotated with a ⟨subject, verb, object⟩ relation, *e.g.*, ⟨girl, sit, shore⟩ relation. We take the available data that contrasts two verbs, *e.g.*, a "positive" or image-consistent relation ⟨girl, sit, shore⟩ and a "negative" or inconsistent relation ⟨girl, walk, shore⟩. For each image, we take the provided "positive" caption (*e.g.*, "A girl sits on the shore"), and use a part-of-speech tagger [22] to localize the verb (*e.g.*, "sit") in the sentence. We do not use images where the tagger failed to identify the verb, often in cases where the verb did not appear in the caption itself (*e.g.*, a triplet of ⟨person, wear, glasses⟩ with a caption of "The glasses fogged up"). The final split contains about 6,500 image-caption pairs (Tab. 8), half of which are correct pairs. This evaluation is not directly comparable to prior work [21], which used the full set of data, chose a threshold of 0.5 to indicate whether or not an individual sample matched an image, and was performed at a sequence-level rather than word-level. In our work, we contrast a positive and negative image for a given caption, and label a sample as correct if the confidence for the positive pair is larger than the confidence for the negative pair, similar to Winoground.

Overlap with training data. All OFA models were not exposed to any MS COCO validation or test data during pretraining [61]. Winoground was hand-curated from the Getty Images API [1, 53], which is not used by OFA pretraining. Data from SVO-Probes was collected via the Google Image Search API and de-duplicated against Conceptual Cap-

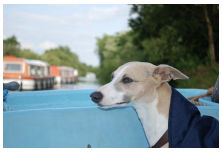
Baseline (b=1)
A giraffe is eating something from a persons hand

TLC-L (b=2)
A person feeding a carrot to a giraffe



Baseline (b=1)
A laptop computer sitting on top of a desk

TLC-L (b=4)
A laptop computer on a desk with a cell phone



Baseline (b=1)
A dog is sitting in the back of a boat

TLC-L (b=2)
A dog is sitting in the back of a truck

Figure 6: Failure cases on our test set for TLC-L on $\text{OFA}_{\text{Large}}$, where TLC-L selected a caption with a hallucination, yet the Baseline did not.

tions [21, 48]. As OFA models used Conceptual Captions during pretraining, we assume there is no further overlap.

E. Model details

Captioning. To complement the details in Sec. 4.1, we provide additional experimental details for the captioning models. We use publicly available checkpoints for pretrained models provided by [61]. Parameter counts are 930M for $\text{OFA}_{\text{Large}}$, 180M for OFA_{Base}, and 33M for OFA_{Tiny} [61]. To finetune the pretrained models on MS COCO Captions, we follow the same settings from [61], where we train with cross entropy loss for 2 epochs for $\text{OFA}_{\text{Large}}$, and 5 epochs for OFA_{Base} and OFA_{Tiny}. We then train with CIDEr optimization for 3 epochs.

TLC-L. In addition to details in Sec. 4.1, we provide further information on the learned confidence estimator g. We use a 4-layer Transformer encoder [56] with 4 attention heads each. The embedded output corresponding to the token of interest t_k (Sec. 3.2) is passed to a 2-layer MLP, with hidden dimensions of size 512. The embedding dimension is 1024 for $\text{OFA}_{\text{Large}}$, 768 for OFA_{Base}, and 512 for OFA_{Tiny}. We train g for 200 epochs, with a batch size of 256, starting learning rate of 0.001, warm up ratio of 0.06 and polynomial learning rate decay to $2e$-7. We use the Adam optimizer [26] and clip gradients over 1.0. For aggregating tokens over objects for caption generation (Sec. 3.3.2), we use the minimum score for softmax and average for TLC-L, found on our validation set.