Appendix for: "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

A. Overview

Appendix B provides details on how the confidence threshold γ is chosen. Appendix C 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 D presents additional qualitative examples of both success and failure cases, comparing TLC-L to the Baseline model. Appendix E and Appendix F provide further details on datasets and models respectively.

B. Choosing the threshold γ

As mentioned in Sec. 4.4, we provide more details about how the threshold γ is chosen, used at test time to make binary decisions on the correctness of a given object in a predicted caption. For both TLC-A and TLC-L, we choose γ on the validation set. Note that 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. 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. 3), 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.

C. 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 [10] as well as a penalty in image caption decoding [13]. Finally, we consider the **Energy** score [6], 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. 9). 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.

D. 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. 3. 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,

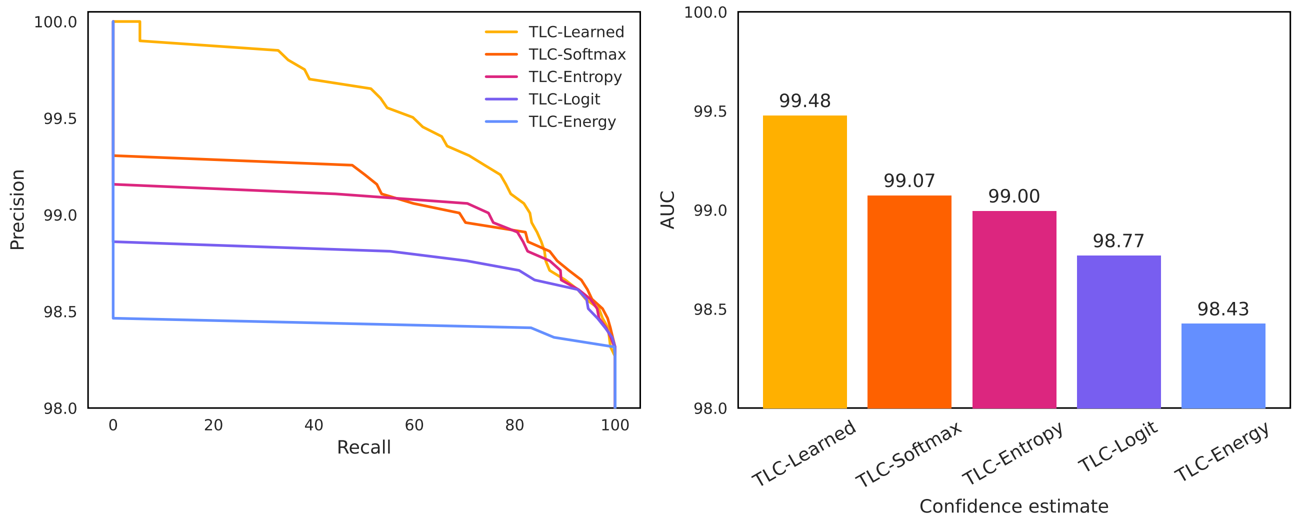

Figure 4. Precision-recall curve (left) and AUC (right) with different confidence estimates for separating correct and hallucinated objects. Results are shown on our validation set using $\mathrm{OFA}_{\mathrm{Large}}$.

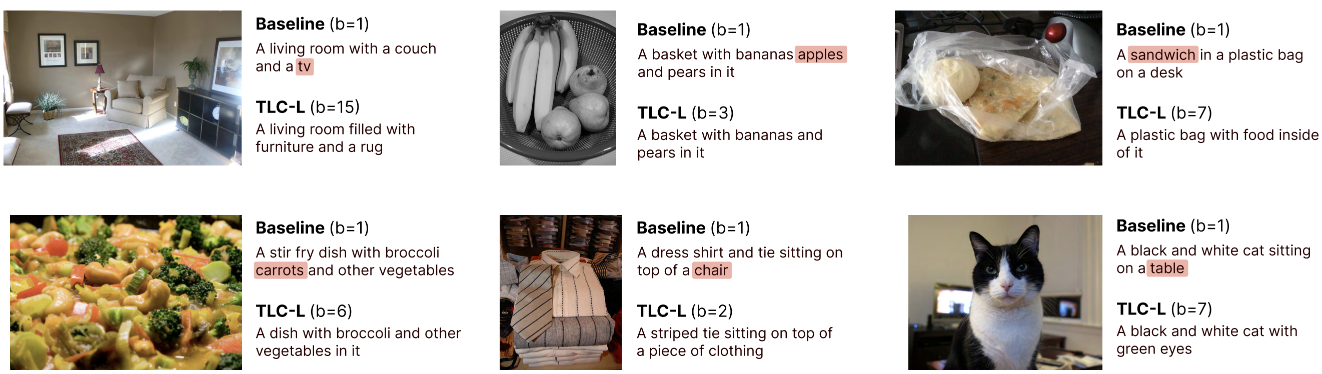

Figure 5. Additional qualitative examples on our test set for TLC-L on $\mathrm{OFA}_{\mathrm{Large}}$, where the Baseline model caption contained a hallucination, yet the caption selected by TLC-L did not.

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.

E. Dataset details

MS COCO Captions. We use the same dataset splits as [12] for training and validating the captioning model f_{cap} and confidence estimator g, as [12] 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 in Tab. 8, we include the training set for g as part of the training set for f_{cap}. In Tab. 9, we refer to these splits as MSC-Main (for MS COCO Main), and use them for results in Tabs. 3, 4, 5, 6, and 8, 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. 9.

Winoground. We use the original data and evaluation setup for Winoground as in the original paper [8], which consisted of 800 unique images and captions. This leads to 400 examples, each consisting of two image-caption pairs, where the captions contain the same words and/or morphemes yet a different word order.

SVO-Probes. For SVO-Probes [3], 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 $\langle$subject, verb, object$\rangle$ relation, *e.g.*, $\langle$girl, sit, shore$\rangle$ relation. We take the available data that contrasts two verbs, *e.g.*, a "positive" or image-consistent relation $\langle$girl,

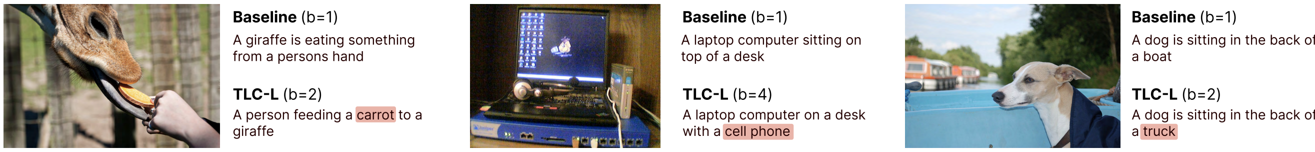

Figure 6. Failure cases on our test set for TLC-L on $\text{OFA}_{\text{Large}}$, where TLC-L selected a caption with a hallucination, yet the Baseline did not.

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 9. Overview of datasets used in our work. MSC indicates MS COCO Captions [2].

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 [4] 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. 9), half of which are correct pairs. This evaluation is not directly comparable to prior work [3], 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 [11]. Winoground was hand-curated from the Getty Images API [1, 8], which is not used by OFA pretraining. Data from SVO-Probes was collected via the Google Image Search API and de-duplicated against Conceptual Captions [3, 7]. As OFA models used Conceptual Captions during pretraining, we assume there is no further overlap.

F. 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 [11]. Parameter counts are 930M for $\text{OFA}_{\text{Large}}$, 180M for OFA_{Base}, and 33M for OFA_{Tiny} [11]. To finetune the pretrained models on MS COCO Captions, we follow the same settings from [11], 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 [9] 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 [5] and clip gradients over 1.0. For aggregating tokens over objects for caption generation (Sec. 4.4), we use the minimum score for softmax and average for TLC-L, found on our validation set.

References

[1] Getty images api. https://www.gettyimages.com/. 3

[2] 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. 3

[3] Lisa Anne Hendricks and Aida Nematzadeh. Probing image-language transformers for verb understanding. *arXiv preprint arXiv:2106.09141*, 2021. 2, 3

[4] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spacy: Industrial-strength natural language processing in python. 2020. 3

[5] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 3

[6] 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. 1

[7] 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. 3

[8] 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. 2, 3

[9] 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. 3

[10] 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. 1

[11] 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. 3

[12] 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. 2

[13] 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. 1