Towards Models that Can See and Read
Supplementary Material

Roy Ganz*
Technion, Israel
ganz@cs.technion.ac.il

Oren Nuriel
AWS AI Labs
onurielamazon.com

Aviad Aberdam
AWS AI Labs
aaberdam@amazon.com

Yair Kittenplon
AWS AI Labs
yairk@amazon.com

Shai Mazor
AWS AI Labs
smazor@amazon.com

Ron Litman
AWS AI Labs
litmanr@amazon.com

A. Implementation Details

In this section, we provide full implementation specifics of UniTNT and divide it into three parts – (1) architecture; (2) training procedure; (3) Scene-text information.

A.1. Architecture

We harness the model agnosticism of UniTNT and apply it to two top-performing VL models. Specifically, we utilize the publicly-available code bases of ALBEF [12][1] and BLIP[2] [11] and apply our method to them. We design our approach in a modular way enabling simple integration into existing models. Below we list the architectural specifics for both $\text{UniTNT}_{\text{ALBEF}}$ and $\text{UniTNT}_{\text{BLIP}}$.

OCR Encoder We use a pretrained BERT-base[3] [6] as our encoder and introduce it with 2-dimensional information, as can be seen in Equation 1. Specifically, we use three separate embedding layers (*i.e.,* `torch.nn.Embedding`)– for the word token and its x and y axis positions for both the OCR and the question. In particular, we define the minimal and the maximal spatial position as 0 and 1000, respectively, and set these values for the question tokens (referred to as "pseudo-2D information" in the main paper). We restrict the number of OCR and question token lengths to 128 and 35, respectively. Next, we sum the 2D-related embeddings and pass them in a 2-layer MLP with a hidden dimension of 768 for additional processing. Finally, we multiply it by α (set to 0.1) and sum it with the token representation to obtain the final one fed into the encoder.

*Work done during an Amazon internship.
[1] https://github.com/salesforce/ALBEF
[2] https://github.com/salesforce/BLIP
[3] https://huggingface.co/docs/transformers/model_doc/bert

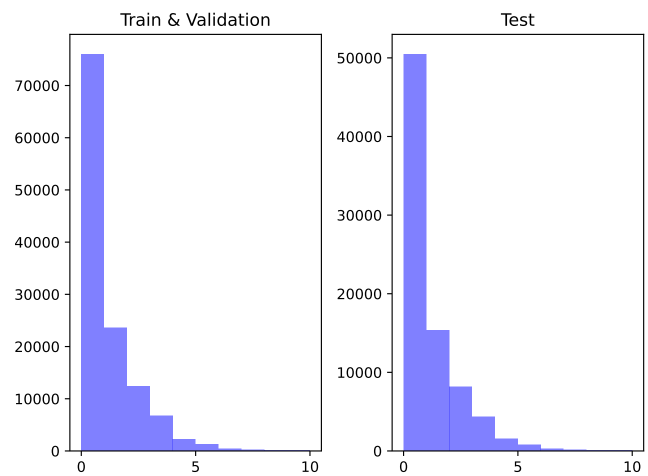


Figure 1: **OCR prevelance in VQAv2.** Histogram of the number of OCR instances per-image in VQAv2 dataset.

VL-OCR Decoder In order to introduce the pretrained decoder with scene-text information, we create new OCR Cross Attention (OCR-CA) blocks and place them in parallel to the existing VL ones. Such newly added components are identical to the existing ones and initialized with the pretrained weights of the latters'. To fuse the outputs of the OCR CA and the VL CA, $\mathcal{F}_{\text{OCR}}$ and $\mathcal{F}_{\text{VL}}$, we concatenate them along the channel dimension and pass them via attention based 2-layers MLP with a hidden size of 768 to obtain $\mathcal{F}_{\text{attn}}$, an attention map that multiplies $\mathcal{F}_{\text{OCR}}$ ($\mathcal{F}_{\text{OCR}} \odot \mathcal{F}_{\text{attn}}$). Namely, this mechanism highlights the important and meaningful features in $\mathcal{F}_{\text{OCR}}$ and masks the less relevant ones. Then, we pass the multiplication output via a learnable gating module (by multiplying it by $tanh(\beta)$, where β is learnable and initialized to 0), aimed to gradually blend the OCR features into the existing VL one.

A.2. Training Procedure

We train all of our models to minimize $\mathcal{L}_{\text{UniTNT}} = \mathcal{L}_{\text{base}} + \alpha_1 \mathcal{L}_{\text{OCR-LM}} + \alpha_2 \mathcal{L}_{\text{OCR-BC}}$ using 8 A100 GPUs, where α_1 and α_2 are hyperparameters.

Visual Question Answering We train both UniTNT$_{\text{ALBEF}}$ and UniTNT$_{\text{BLIP}}$ on a unified Text-Non-Text VQA dataset, containing VQAv2 [1], TextVQA [16] and ST-VQA [3] for 10 epochs using a batch size of 8 and 16 for ALBEF and BLIP, respectively. Moreover, we set $\alpha_1 = \alpha_2 = 1$ and keep the other training-related hyperparameters as in the original papers.

Image Captioning We train UniTNT$_{\text{BLIP}}$ on a the unified Text-Non-Text CAP dataset, comprised of COCO Captions [4] and TextCaps [15], for 5 epochs with batch size of 32. We set $\alpha_1 = 0.05$ and $\alpha_2 = 0$ since contrary to VQA, CAP does not contain textual information available both in training and inference time, making it infeasible to implement OCR-BC. Moreover, we keep the rest of the hyperparameters as in BLIP.

A.3. Scene-text information

As specified in the paper, we extract the scene-text information (word tokens and 2-dimensional position) for all the VQA and CAP datasets (both the general and scene-text counterparts) using Amazon Text-in-Image. To better understand the prevalence of OCR in the non-scene-text datasets, we plot the statistics of OCR in VQAv2 in Fig. 1 (same images are in COCO Captions as well). While a large portion of the images does not contain text in them, there is a large amount of such with OCR (38.36% and 38.03% of train and test images contain OCR). Since OCR conveys meaningful information, it sheds light on the significant improvement of UniTNT up his baselines (ALBEF and BLIP).

B. Datasets

B.1. Visual Question Answering

VQAv2 contains 204,721 images (82,783, 40,504, and 81,434) from COCO [13], 1,105,904 questions (443,757, 214,354, and 447,793), and 6,581,110 answers (4,437,570, 2,143,540, and the test answers are held-out). Answering the questions requires vision-language understanding and commonsense knowledge. Each question has ten ground-truth answers.

TextVQA contains 28,408 images from OpenImages [10], 45,336 questions and 453,360 ground-truth answers. The annotators were instructed to formulate questions that require reasoning from the text in the image. As in VQAv2, each question has 10 ground-truth answers.

ST-VQA is a fusion of computer-vision datasets – ImageNet [5], VizWiz [2], Visual Genome [9], IIIT Scene Text Retrieval [14], ICDAR 2013 [8], ICDAR 2015 [7] and COCO Text [17]. It contains 31K questions, split into training (26K) and testing (5K), requiring scene-text understanding.

B.2. Image Captioning

COCO Captions contains over one and a half million captions describing over 330,000 images from the COCO dataset. Each image has five human-generated captions.

TextCaps is composed of 28,408 images and 142,040 captions (5 captions per image). The images are from the TextVQA dataset, and the captions are based on the text in the image. Specifically, models have to reason over the scene-text information to generate correct captions.

C. The Impact of Training Data

In this section, we study the effect of the different combinations of training datasets and report our findings in Tab. 1. In particular, we experiment with UniTNT and BLIP in Visual Question Answering and Image Captioning using separate training on vision-oriented and OCR-oriented datasets and combined training. In VQA, using both dataset types leads to the best standalone and average performance in the tested benchmarks. This attests to the symbiosis between general and scene-text-oriented VQA, encouraging avoidance of the common practice of separate finetuning.

However, using a unified training set in CAP leads to the best COCO Captions and average results, but not in TextCaps. Specifically, separate finetuning on TextCaps achieves a CIDEr score of 130.5, compared to 119.1 in the combined training. It corresponds with [15], which shows that combining COCO Captions with an upsampled version of TextCaps reduces the model's performance on the former. It is because while training on TextCaps encourages the model to insert OCR into the caption, training on COCO Captions which barely contains OCR in its captions, penalizes such behavior, leading to an intrinsic tradeoff. To better understand the effects of training models solely on TextCaps, we qualitatively test them on COCO Captions. Notably, we finetune both BLIP and UniTNT of TextCaps and demonstrate their performance on COCO Captions in Fig. 2. Our analysis shows that as TextCaps contains OCR in all its captions, separate finetuning causes models to fixate on OCR, regardless of their importance. Moreover, in images without an OCR signal, the models sometimes hallucinate text in the image. While both models showcase similar behavior, since UniTNT has better scene-text understanding, it is more prone to such phenomena. It is also expressed in Tab. 1, where BLIP and UniTNT trained

Method	Vision-oriented dataset	OCR-oriented dataset	VQA test-dev	TextVQA val	Avg.	COCO Caps val	TextCaps val	Avg.
BLIP	✗	✓	40.16	30.12	35.14	84.8	112.7	98.8
$\text{UniTNT}_{\text{BLIP}}$			37.01	50.19	43.60	70.4	**130.5**	100.5
BLIP	✓	✗	76.39	20.50	48.45	133.3	59.4	96.4
$\text{UniTNT}_{\text{BLIP}}$			79.68	36.33	58.01	133.7	59.6	96.7
BLIP	✓	✓	77.40	32.43	54.92	133.4	101.4	117.4
$\text{UniTNT}_{\text{BLIP}}$			**79.90**	**55.21**	**67.56**	**134.0**	119.1	**126.6**

Table 1: **The impact of training data.** We show the effect of each dataset configuration for training UniTNT and BLIP.

on TextCaps obtain 84.8 and 70.4 on COCO Captions, respectively. Despite the improved performance on TextCaps when performing separate finetuning on it, our findings highlight its drawbacks. Thus, we claim that also in CAP, combined training should be applied.

From a general view, we hypothesize that since numerous valid captions exist for a given image, both with and without OCR, the model struggles to decide whether to use the OCR in its caption. Due to the datasets' sizes in combined training that favors the vision-oriented ones, the model opts to reduce its use of OCR, not fully maximizing its performance on TextCaps. It is contrary to VQA, where the conditioning over the question makes it easier for the model to decide whether to use OCR or not (*e.g.*, `"What is written in the sign?"` versus `"What color is this shirt?"`).

D. Qualitative Analysis

Visual Question Answering We provide an additional qualitative demonstration of UniTNT and compare it to BLIP and M4C on both TextVQA validation set (Fig. 3) and VQAv2 test set (Fig. 4). We depict in the four leftmost columns success-cases and the rightmost, fail cases, and color in green the correct answers and red, incorrect ones. Moreover, we divide the figures such that the upper part corresponds with the benchmark's goal (VQAv2 – see, TextVQA – read) and the lower one with the counterpart goal (VQAv2 – read, TextVQA – see). These results further demonstrate that UniTNT is capable of reasoning over both visual and scene-text information, while other competing methods perform unsatisfactorily on at least one of the benchmarks. Moreover, as the visualizations in Fig. 4 testify, granting scene-text understanding also benefit VQAv2, corresponding with the quantitative evidence in the main paper. It is demonstrated in the bottom part of the figure, where the OCR is crucial for answering the questions or providing meaningful information that facilitates answering them.

Image Captioning Similar to the VQA demonstration, we present a visualization of UniTNT performance on TextCaps (Fig. 5 and COCO Captions (Fig. 6) and compare the performance to M4C and BLIP. On the left columns, we show images where our method outperforms the other methods, and on the right, its failure cases. Moreover, we list the CIDEr scores of each prediction and color in green the highest one. These findings attest that BLIP is incapable of incorporating scene-text information, which results in relatively low CIDEr results. Interestingly, M4C is too overfitted for TextCaps, causing it to fail completely on COCO Captions where OCR is scarce. Specifically, it focuses on the OCR regardless of their importance (*e.g.*, the third example in the last row of Fig. 6) and thus provides an irrelevant caption. Despite the intrinsic tradeoff described in the paper between TextCaps and COCO Captions, UniTNT is capable of providing adequate captions for both benchmarks. Specifically, our method is the only one to cope satisfactorily on both benchmarks altogether and is capable of harnessing both scene-text and visual information.

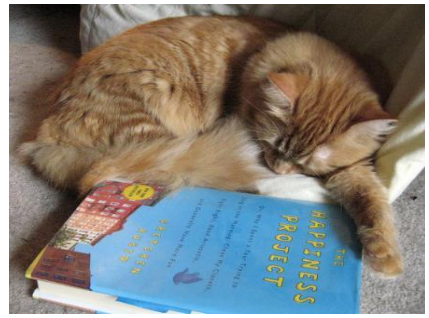
Hallucinating OCR | **Over-fixation on OCR** | **OCR is useful**

BLIP: a young boy is eating a piece of cake with a yellow frosting on it (54.5)

Ours: a young boy is eating a cake with the word cake on it (47.9)

BLIP: a man is surfing in the ocean and is wearing a swim suit (18.9)

Ours: a man is surfing in the ocean with the name jimmy bravo (8.8)



BLIP: a cat sleeping on top of a book that has the word paris on it (103.1)

Ours: a cat sleeping on a book titled happiness project (184.0)

BLIP: a traffic light has a red light on it (42.5)

Ours: a traffic light has a red light that says red on it (28.1)

BLIP: a display of donuts with a coca cola can in the background (24.3)

Ours: a coca cola box is behind some donuts (14.6)

BLIP: a boy wearing a green and yellow jersey with the word fell on it (82.8)

Ours: a boy in a jerlin baseball uniform holds a bat (127.2)

BLIP: two women are decorating a cake with a pepsi logo on it (96.9)

Ours: two women are decorating a cake on a counter (197.4)

BLIP: a seagull is flying over a body of water with the words mr nicholas (26.5)

Ours: a seagull is flying over the water with the words sharklady adventures (71.4)

BLIP: a poster with a baseball player and the words baseball memories (34.5)

Ours: a picture of baseball items and the words baseball memoribilia (93.2)

Figure 2: **Qualitative demonstration of the effects of finetuning on TextCaps.** BLIP and UniTNT results of COCO Captions when finetuned solely on TextCaps. In some cases, scene-text understanding helps the models, but it also leads to over-reliance on the OCR signal and even to the hallucination of OCR. While such phenomena occur in both models, it is more prevalent in UniTNT due to its better scene-text understanding.

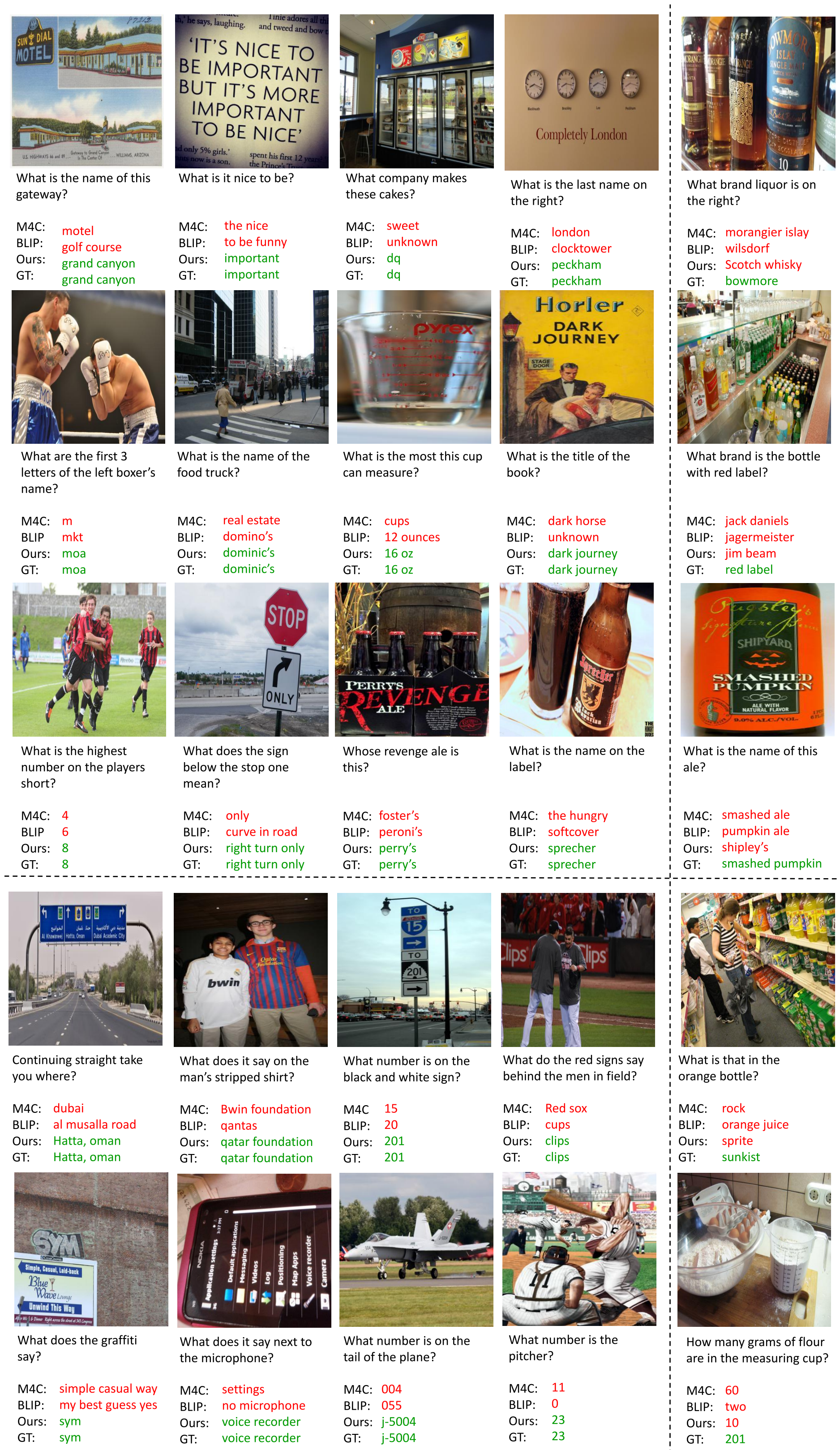


Figure 3: **Qualitative demonstration on TextVQA validation.** UniTNT, M4C, and BLIP answers, containing both success (left) and fail (right) cases of our method on image-question pairs that require mainly reading (top) and ones that require also visual reasoning (bottom).

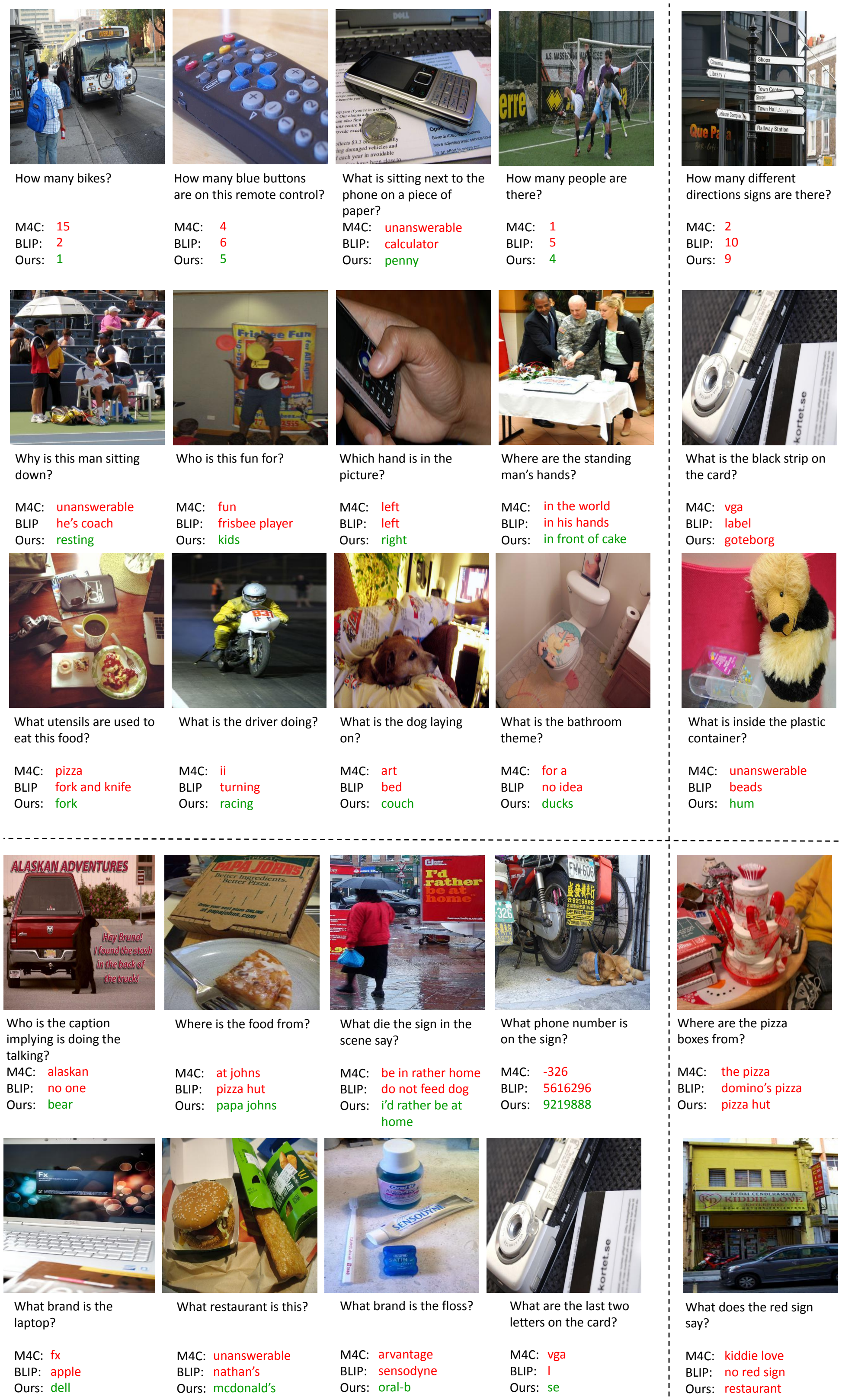


Figure 4: **Qualitative demonstration on VQAv2 test.** UniTNT, M4C, and BLIP answers, containing both success (left) and fail (right) cases of our method on image-question pairs that require mainly vision (top) and ones that require also scene-text understanding (bottom).

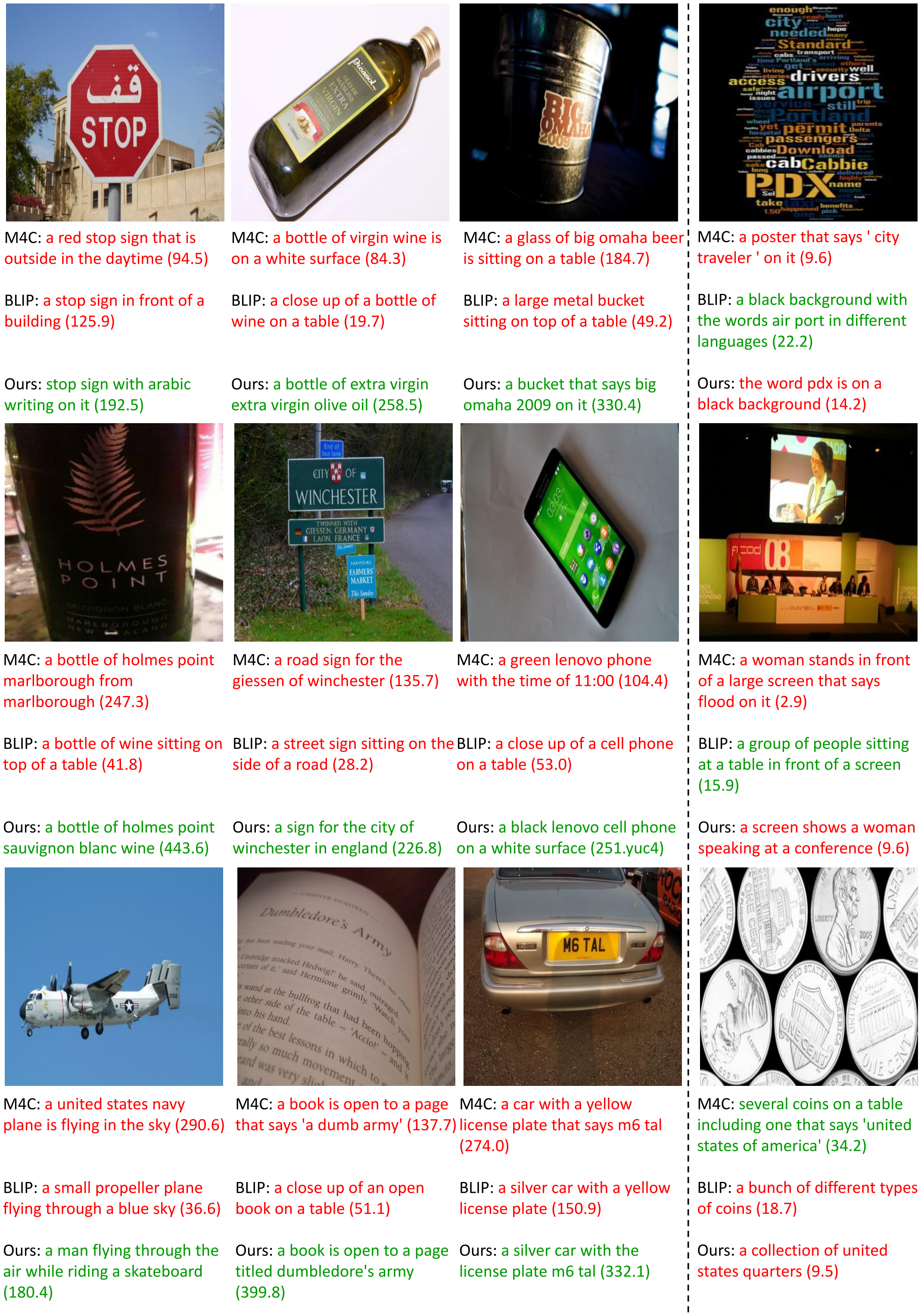


Figure 5: **Qualitative demonstration on TextCaps.** UniTNT, M4C-Captioner, and BLIP answers, containing both success (left) and fail (right) cases of our method alongside the per-caption CIDEr score.

M4C: a white car with the number 3 on it (0.5)

BLIP: a group of people sitting on top of a sandy beach (59.2)

Ours: a group of people on the beach under an umbrella (162.6)

M4C: a large white and red sign that says 'd' on it (0.4)

BLIP: a man sitting at a table with a plate of food (68.5)

Ours: a man in a tie is smiling for the camera (116.7)

M4C: a white car with the word "no" on it (2.5)

BLIP: a man riding a wave on top of a surfboard (123.4)

Ours: a man riding a surfboard on top of a river (181.8)

M4C: a small white sign that says "w" on it (1.5)

BLIP: a group of people standing in a room (27.7)

Ours: a man and a woman standing next to each other (17.3)

M4C: a sign that says vote & laduke on it (0.1)

BLIP: a man sitting at a table with a plate of food (42.8)

Ours: a man in a green shirt holding a glass of wine (186.2)

M4C: a large white sign that says "no parking" on it (10.9)

BLIP: a man riding a skateboard up the side of a ramp (69.0)

Ours: a man flying through the air while riding a skateboard (180.4)

M4C: a picture of a woman in a suit with a sign that says "say say cheese!" (0)

BLIP: two stuffed animals sitting next to each other on a chair (52.3)

Ours: two stuffed animals are sitting next to a book (99.3)

M4C: a picture of a man with the number 3 on it (1.7)

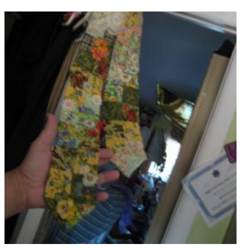
BLIP: a bathroom with a washer and a window (23.9)

Ours: a bathroom with a washer and dryer in it (20.9)

M4C: a picture of a woman and a yellow and white dress with the word "middle middle" on it (0.2)

BLIP: a couple of people that are holding a skateboard (7.8)

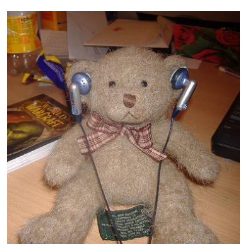
Ours: a man holding a snowboard next to another man (196.4)



M4C: a large number of a train is on the ground with a red and white sign that says "sample" (0.4)

BLIP: a person holding a piece of fabric in their hand (41.7)

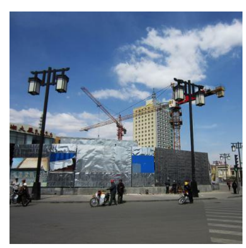
Ours: a person holding a tie in their hand (205.5)



M4C: a book called warcraft is on a table with a picture of a person in the background (5.0)

BLIP: a brown teddy bear sitting on top of a desk (156.7)

Ours: teddy bear wearing headphones sitting on a desk (189.5)



M4C: a green sign that says ' no parking ' on it (6.7)

BLIP: a group of people walking across a street next to a tall building (16.4)

Ours: a group of people walking across a street (14.7)

Figure 6: **Qualitative demonstration on COCO Captions.** UniTNT, M4C-Captioner, and BLIP answers, containing both success (left) and fail (right) cases of our method alongside the per-caption CIDEr score.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 2

[2] Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, et al. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pages 333–342, 2010. 2

[3] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4291–4301, 2019. 2

[4] 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 preprint arXiv:1504.00325*, 2015. 2

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019. 1

[7] Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, et al. Icdar 2015 competition on robust reading. In *2015 13th international conference on document analysis and recognition (ICDAR)*, pages 1156–1160. IEEE, 2015. 2

[8] Dimosthenis Karatzas, Faisal Shafait, Seiichi Uchida, Masakazu Iwamura, Lluis Gomez i Bigorda, Sergi Robles Mestre, Joan Mas, David Fernandez Mota, Jon Almazan Almazan, and Lluis Pere De Las Heras. Icdar 2013 robust reading competition. In *2013 12th international conference on document analysis and recognition*, pages 1484–1493. IEEE, 2013. 2

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

[10] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4. *International Journal of Computer Vision*, 128(7):1956–1981, 2020. 2

[11] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022. 1

[12] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 1

[13] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2

[14] A. Mishra, K. Alahari, and C. V. Jawahar. Image retrieval using textual cues. In *ICCV*, 2013. 2

[15] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *European conference on computer vision*, pages 742–758. Springer, 2020. 2

[16] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019. 2

[17] Andreas Veit, Tomas Matera, Lukas Neumann, Jiri Matas, and Serge Belongie. Coco-text: Dataset and benchmark for text detection and recognition in natural images. *arXiv preprint arXiv:1601.07140*, 2016. 2