Towards Models that Can See and Read

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

Abstract

Visual Question Answering (VQA) and Image Captioning (CAP), which are among the most popular vision-language tasks, have analogous scene-text versions that require reasoning from the text in the image. Despite their obvious resemblance, the two are treated independently and, as we show, yield task-specific methods that can either see or read, but not both. In this work, we conduct an in-depth analysis of this phenomenon and propose UniTNT, a Unified Text-Non-Text approach, which grants existing multimodal architectures scene-text understanding capabilities. Specifically, we treat scene-text information as an additional modality, fusing it with any pretrained encoder-decoder-based architecture via designated modules. Thorough experiments reveal that UniTNT leads to the first single model that successfully handles both task types. Moreover, we show that scene-text understanding capabilities can boost vision-language models' performance on general VQA and CAP by up to 2.69% *and* 0.6 *CIDEr, respectively.*

1. Introduction

In recent years, Vision-Language (VL) tasks, such as Visual Question Answering (VQA) [4, 18] and Image Captioning (CAP) [34, 2], have gained immense research interest [55, 39, 49, 30, 22, 46, 10, 47]. However, despite the remarkable success of VL models on these tasks, it was discovered a few years ago that such models are incapable of reasoning from the text in natural images [41, 8, 40]. This finding raised significant concerns, as understanding scene-text is crucial in almost any real-world application.

To address this issue, designated scene-text datasets were introduced for both VQA [41, 8] and CAP [40], aiming to highlight the importance of utilizing textual information in images. Following the introduction of the above datasets, a new line of research has arisen, focusing on scene-text-oriented tasks, evaluated individually and effectively dis-

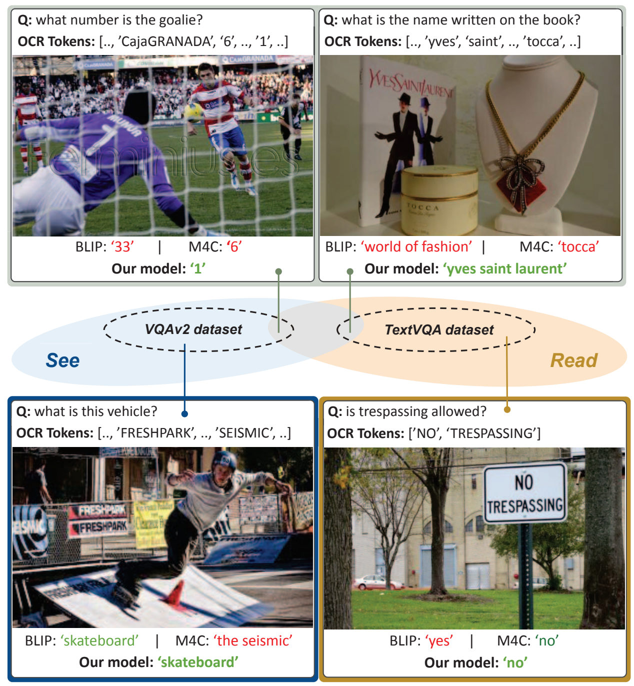

Figure 1: **See and read in VQA.** Illustration of the possible three types of reasoning required in VQA image-question pairs and representative datasets distributions (middle). Samples from the 'see' (bottom left), 'read' (bottom right), and 'see-∩-read' (top) subsets are presented. Each sample includes an image, question, OCR, and model predictions.

sociated from the general one. From a user perspective, this separation is artificial and does not adequately reflect the objective of real-world VQA systems, and as we show, it encourages models to only excel on one task at a time. Therefore, we advocate that VL research should strive towards unified models, and thus, methods should be evaluated accordingly. To this end, we propose conducting *combined evaluation* for VL models on both general and scene-text benchmarks and treating the average results as expressing the "see" and "read" capabilities. We emphasize

*Work done during an Amazon internship.

that even the minority of works that evaluate both types of tasks [46, 11, 3] do it on separate models, which are fine-tuned per task, perpetuating the faulty tasks' segregation.

Apart from being unjustified, this separation introduces biases [7, 48], providing the models with prior knowledge that implies which modality to focus on, which does not exist in real-world scenarios. Namely, it creates a shortcut that encourages models to excel solely on a specific benchmark by acquiring an understanding of either the visual or textual information in the image, but not both. In particular, Biten *et al*. [7] recently showed that SOTA performance on scene-text VQA can be achieved without using the visual modality, and Wang *et al*. [48] revealed that existing scene-text VQA models' success stems from exploiting language priors. Our *combined evaluation* effectively addresses this problem by testing whether models can reason from both types of information, as exploiting such data biases and priors would yield low combined results.

From a more high-level view, three categories span the space of VL data; the first are examples that require reasoning over vision only (dominant in VQA [18] and CAP [9]), the second are instances in which using scene-text information solely is sufficient (dominant in scene-text VQA [8, 41] and scene-text CAP [40]), and the third are ones in which both are essential. We denote the three subsets as **'see'**, **'read'**, and **'see-∩-read'**, respectively. For completion, the whole space is denoted as **'see-∪-read'**, the union of all others. We illustrate this conceptual data distribution for VQA in Fig. 1. Examining the performance of existing VQA approaches over the three types of questions mentioned above, shown in Fig. 2, reveals that while some of the methods [30, 31, 47] perform well on the first subset and some [21, 52] on the second, none are optimal on the entire domain. Moreover, throughout our analysis, we reveal that the 'see-∩-read' subset, in which both visual and textual information are needed for answering, is very challenging and underrepresented, requiring a new dedicated benchmark.

In this work, while striving towards models that excel on the entire space of VL data, we propose UniTNT, a Unified Text-Non-Text model, which provides VL architectures with scene-text understanding capabilities. Specifically, we treat textual information in the image, *i.e.* tokens and positions, as a third modality and introduce it into the pretrained model. Adding a new modality to an already-trained model is challenging and might lead to suboptimal results [16, 43, 54, 3]. To overcome this, we encode such information using a designated encoder and inject it into the existing pretrained decoder via a novel fusing mechanism that gradually shifts between VL features to textual-enriched ones. Moreover, we propose scene-text-related intermediate supervision to encourage the already-trained model to leverage the newly added information. Being both task and model agnostic by its design, our method can be applied to any VL encoder-decoder-based architecture.

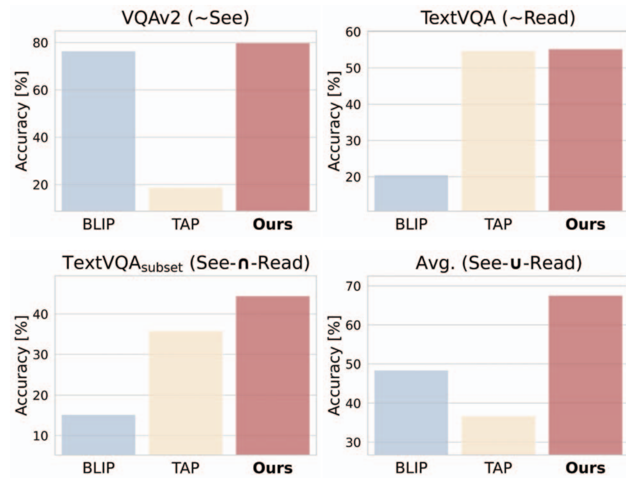

Figure 2: **Models' accuracy on different types of VQA data.** Leading methods and UniTNT performance on different benchmarks. VQAv2 and TextVQA datasets mostly require reasoning from visual information only ('See') and textual information only ('Read'), respectively. 'See-∩-Read' refers to a subset of the TextVQA dataset (Sec. 4.3), in which both modalities are essential for answering each question. 'See-∪-Read' represents the sets' union.

We evaluate UniTNT on both general and scene-text benchmarks of VQA and CAP using *combined evaluation* and show that it leads to the first single model performing well on both tasks. We show that our method can be easily integrated into existing VL models, improving their scene-text understanding substantially by applying it to BLIP [30] and ALBEF [31]. Interestingly, such reasoning abilities boost the base model's VQA results (*e.g.*, improves BLIP [30] by 2.69% on VQAv2 [18]), while achieving state-of-the-art competitive results on scene-text VQA benchmarks. A similar trend exists in captioning, where UniTNT enhances BLIP's performance by 0.6 CIDEr points on COCO Captions [9] while substantially boosting its scene-text CAP performance. These improvements highlight the significance of scene-text comprehension in VL tasks, laying the foundation for future research on general multimodal architectures that can leverage scene-text.

To summarize:

- We thoroughly analyze current methods and reveal that the faulty text-non-text task separation leads to models that either reason from visual or textual information in images, but not both.
- We introduce UniTNT, a model-agnostic method to grant reading capabilities to pretrained VL models by fusing the scene-text information as an additional modality.
- Extensive experiments show that our method not only improves the scene-text benchmarks' results but also significantly enhances the performance of VQA and CAP.

	Method	OCR System	Visual Question Answering						Image Captioning			
			VQA		TextVQA		ST-VQA	Avg.	COCO	TextCaps		Avg.
			test-dev	test-std	val	test	test-ANLS		Karpathy-test	val	test	
Separate	M4C [21, 40]	✓	27.47	27.70	46.53	47.42	0.43	37.56	4.7	95.5	90.1	47.4
	TAP [52]	✓	18.76	18.81	54.71	53.97	0.60	36.39	4.6	109.2	103.2	53.9
	ALBEF [31]	✗	75.22	75.38	11.67	13.88	0.19	44.63	-	-	-	-
	BLIP [30]	✗	76.39	76.59	20.50	23.74	0.34	50.16	133.3	59.4	61.9	97.6
	$\text{OFA}_{\text{Large}}$ [47]	✗	79.70	79.85	22.10	21.47	0.27	50.66	150.7	64.5	66.8	108.8
Comb.	M4C [21]	✓	59.11	59.04	47.22	48.61	0.50	53.83	109.8	102.7	98.0	103.9
	ALBEF [31]	✗	75.61	75.87	16.15	17.04	0.22	46.46	-	-	-	-
	BLIP [30]	✗	77.40	77.39	32.43	31.48	0.44	54.44	133.4	101.4	91.8	112.6

Table 1: **Current status of VQA and CAP models.** The results of leading methods on both scene-text and general VQA and CAP benchmarks reveal that currently, no method performs well on both scene-text and general benchmarks, even when applying combined training. *Separate* and *Comb.* summarize the results described in Sections 2.1, 2.2 and 2.3, respectively.

2. See and Read: Analyzing Methods and Data

In this paper, contrary to the common practice in VL research, we highlight the importance of models to "see" and "read" altogether and start by comprehensively analyzing such capability via a "see-∪-read"-oriented *combined evaluation*. Our analysis reveals that existing models' reasoning abilities over both types of information are lacking, prompting the question of whether this limitation is due to inherent method constraints or biased data. Our evaluation focuses on the performance of leading general and scene-text-oriented models on VQAv2 [18], TextVQA [41], and ST-VQA [8] for VQA, and COCO Captions [9] and TextCaps [40] for captioning.

2.1. Visual Question Answering

General VQA Methods: During the vision-language revolution, numerous methods [33, 39, 31, 30, 47, 49, 46, 3, 10, 55, 53, 13, 22] have been proposed for various multimodal tasks, including VQA, which have advanced the state-of-the-art. These methods can leverage vast online image-caption pairs via vision-language pretraining [32, 12, 39], followed by task-specific fine-tuning. However, a few years ago, such models were shown to be ineffective in reasoning from textual information in the scene, as they primarily focus on the images' visual content [41, 8].

Nevertheless, such models have advanced significantly in the past few years. Thus, to reveal the current status of such models in scene-text understanding, we examine the performance of three leading VQA models, ALBEF [31], BLIP [30], and OFA [47], using unconstrained open-vocabulary generation, on scene-text VQA tasks. As seen in Tab. 1, although such methods perform well on VQA, as expected, their results on the analogous scene-text VQA datasets are unsatisfactory, testifying their incompetence in scene-text understanding. Interestingly, their inability to utilize scene-text information hinders its performance even on VQA, as we later show in Sec. 4.

Scene-Text VQA Methods: Several methods have been proposed to improve the scene-text understanding of VQA models [21, 19, 17, 52, 23, 36, 7]. These models utilize an off-the-shelf OCR system's output alongside the image and question as input to a multimodal transformer. However, some recent studies [48, 7] have indicated that scene-text VQA datasets may have biases discouraging models from relying on the visual modality. To properly test such claims, we evaluate M4C [21] and TAP [52] on general VQA, which requires strong visual understanding and report the results in Tab. 1. As can be seen, M4C and TAP obtain only 27.70% and 18.81%, respectively. When compared to, for example, BLIP's 76.59%, it testifies that, indeed, such methods disregard the visual information. Interestingly, although TAP consistently outperforms M4C on the scene-text benchmarks, it achieves lower results on the general one, implying the data biases in the former datasets.

2.2. Image Captioning

Similar to VQA's analysis, we conduct a captioning *combined evaluation* using TextCaps and COCO Captions for both types of models and report the average CIDEr scores. Our empirical results in Tab. 1 demonstrate that while general models (BLIP and OFA) and scene-text ones (M4C-Captioner and TAP) perform well on their designated benchmarks, they fail to obtain satisfactory results on the analogous one. In particular, BLIP obtains a CIDEr score of only 61.9 on TextCaps, compared to 90.1 of M4C-Captioner. On the other hand, the latter achieves 4.7 on COCO captions, compared to BLIP's 133.3. In addition, like in VQA, while TAP outperforms M4C in TextCaps, it does not occur on COCO Captions. These findings suggest that existing methods exhibit unsatisfactory performance when evaluated on both captioning benchmarks.

2.3. The Role of the Datasets in 'See and Read'

We now examine whether this limitation stems from a lack of representative training data rather than method limitations. Specifically, we test if the inferior performance of

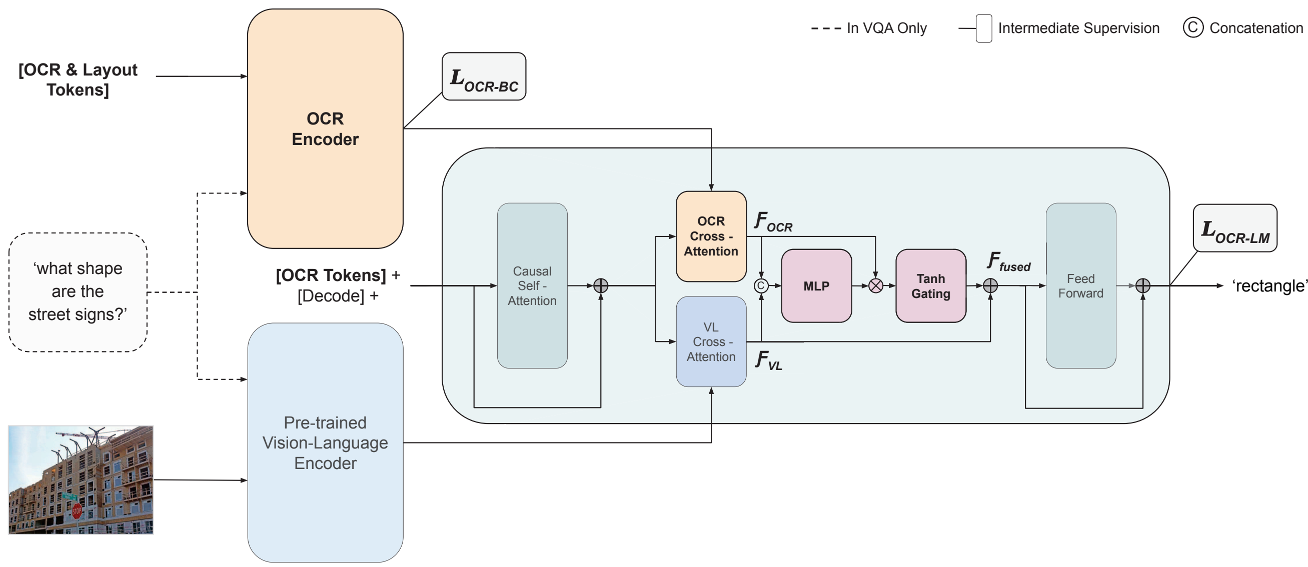

Figure 3: **An overview of UniTNT.** Our method endows existing general VL models with scene-text understanding capability. The OCR information is encoded separately and injected into the decoder via a gated cross-attention-based fusing mechanism as complemental information. $\mathcal{L}_{\text{OCR-BC}}$ and $\mathcal{L}_{\text{OCR-LM}}$ are auxiliary losses, enforcing the model to utilize the scene-text information. UniTNT newly introduced components are presented in bold. 'See', 'Read', and 'Fusing' related modules are in blue, orange, and red, respectively.

scene-text-oriented models on visual tasks and vice versa is solely due to the training data's bias towards reasoning over solely one type of information. To test this claim, we merge two datasets, conduct combined training for both general and scene-text-oriented methods, and report the results in Tab. 1. As can be seen, while unified training leads to improved performance on both types of VL benchmarks, there is a substantial performance gap – scene-text models lag behind general ones on the general benchmarks and vice-versa. Nevertheless, these results indicate that reasoning from text and vision are not at odds and suggest a symbiotic relationship between the two tasks. Furthermore, they provide further motivation for avoiding the common practice of separating the tasks, as done in previous work [3, 11, 46]. To conclude, while joint training is a step forward, it is not enough to achieve our ultimate goal.

3. Method

In this section, we describe UniTNT, a method aimed to obtain our titular goal by granting pretrained general VL models the ability to reason over scene-text information during finetuning while retaining their original reasoning capabilities, depicted in Fig. 3. By doing so, we propose a change of perspective compared to top-performing ST methods, such as [7, 52], that harnesses an OCR-oriented pretrained model but fails to enrich it with visual understanding during finetuning. Adapting pretrained models to consider additional inputs, absent during pretraining, is a non-trivial task tackled by recent literature [43, 3]. On the one hand, we wish to encourage the model to utilize the new stream of information and, on the other hand, to prevent it from neglecting the original stream. To address this, we encode the OCR information via a designated OCR encoder and fuse it residually, retaining the former stream of information and gradually shifting towards an OCR-enriched one. Moreover, we propose auxiliary losses, encouraging the pretrained decoder to utilize this information. Similarly to previous works [21, 52, 36, 7], we utilize an off-the-shelf-OCR system to extract the scene-text information.

3.1. Architecture

We design our architecture in a task-agnostic way – enabling compatibility with both visual question answering and image captioning tasks. In addition, UniTNT is model agnostic and can be applied to any encoder-decoder-based VL model. In this work, we integrate our approach into two top-performing open-source methods – ALBEF [31], and BLIP [30] as a case study, denoted as $\text{UniTNT}_{\text{ALBEF / BLIP}}$.

OCR Encoder Rather than utilizing the pre-existing encoder to process the OCR alongside the visual modality, as in [7, 21, 52], we introduce a dedicated OCR encoder, which maps the scene-text information into features fed into the existing system's decoder. This encoder receives the question alongside OCR information, namely tokens and 2-dimensional (2D) positional information, both extracted by the OCR system. The positional information was proven to be valuable for documents and scene-text understanding tasks [50, 51, 5, 7]. Not only that our approach outperforms the one that utilizes the pre-existing text encoder to process the OCR tokens (demonstrated in Sec. 5), but it also provides flexibility to address tasks that do not utilize a text encoder, such as image captioning.

Formally, each OCR instance is represented by $(t, x_0, y_0, x_1, y_1, w, h)$, namely, its word token, bounding box's top-left, bottom-right, width, and height values, respectively. We embed each value separately using designated embedding layers E (*i.e.*, `torch.nn.Embedding`). Next, we sum the 2D representations, pass them via a 2-layer MLP and add it to the token's representation, yielding the OCR representation:

$$\begin{aligned}\mathbf{e}_{\text{OCR}} = &\mathrm{E}_{\text{OCR}}(\mathrm{t}) + \alpha * \mathrm{MLP}(\mathrm{E_x}(x_0) + \mathrm{E_y}(y_0) + \\ &\mathrm{E_x}(x_1) + \mathrm{E_y}(y_1) + \mathrm{E_w}(w) + \mathrm{E_h}(h))\end{aligned} \tag{1}$$

where α is a predefined hyperparameter. As for the question, we embed its tokens using the same embedding layer. Since both the OCR and the question representations are fed into the same model, we equip the question representations with pseudo-2D information corresponding to the size of the entire image, yielding the final question representation $\mathbf{e}_q$. Finally, we concatenate them to obtain the OCR encoder's input, $\left\{\mathbf{e}_q^1 \dots \mathbf{e}_q^M, \mathbf{e}_{\text{OCR}}^1 \dots \mathbf{e}_{\text{OCR}}^N\right\}$, where M and N are the lengths of the question and OCR, respectively.

VL-OCR Decoder To integrate the OCR information into the decoder, we add a dedicated OCR Cross Attention (CA) and a fusing mechanism, as visualized in Fig. 3. We place the OCR CA block parallel to the pre-existing VL CA module to enrich the decoded features with textual information in the image. This architectural design yields two data streams (visual and scene-text-oriented ones) that need to be merged adequately into a single VL-OCR representation. To this end, we introduce a fusing mechanism composed of a gated cross-attention mechanism, which gradually shifts from VL features to fused, OCR-enriched ones.

Formally, our fusing mechanism merges the output of our new OCR CA with the one of the VL CA, denoted as $\mathcal{F}_{\text{OCR}}$ and $\mathcal{F}_{\text{VL}}$ respectively. Specifically, this module receives two features sequences, $\mathcal{F}_{\text{OCR}}, \mathcal{F}_{\text{VL}} \in \mathbb{R}^{B \times L \times C}$, and outputs $\mathcal{F}_{\text{fused}} \in \mathbb{R}^{B \times L \times C}$, where B, L, C are the batch size, sequence length and the number of channels, respectively. First, we concatenate $\mathcal{F}_{\text{OCR}}$ and $\mathcal{F}_{\text{VL}}$ across the channel dimension and insert them into a simple 2-layer MLP to obtain an attention map $\mathcal{F}_{\text{attn}} \in \mathbb{R}^{B \times L \times C}$. Next, we pass the element-wise product of $\mathcal{F}_{\text{OCR}}$ and $\mathcal{F}_{\text{attn}}$ in a $\tanh$ gating mechanism [20, 3]. The goal of the $\tanh$ gating is to enable gradual OCR blending with the VL one by multiplying its inputs with $\tanh(\beta)$, where β is a learnable parameter initialized to zero. At initialization, it ensures that the added modules are skipped, preserving the model pretraining's data flow. Finally, we sum the output of the $\tanh$ gating with $\mathcal{F}_{\text{VL}}$ to obtain the fused features:

$$\mathcal{F}_{\text{attn}} = \mathrm{MLP}(\mathrm{concat}(\mathcal{F}_{\text{VL}}, \mathcal{F}_{\text{OCR}})), \tag{2}$$

$$\mathcal{F}_{\text{fused}} = \mathcal{F}_{\text{VL}} + \tanh(\beta)(\mathcal{F}_{\text{OCR}} \odot \mathcal{F}_{\text{attn}}), \tag{3}$$

where $\odot$ is the Hadamard product.

3.2. Scene-text Auxiliary Losses

We propose two auxiliary losses, encouraging the model to utilize the scene-text signal rather than ignoring it - OCR Causal Language Modeling (OCR-LM) and OCR Binary Classification (OCR-BC).

OCR Causal Language Modeling To better fuse the scene-text information, we add a causal language modeling supervision over the OCR tokens. Specifically, we prepend the shifted OCR tokens (according to the OCR system reading order) to the inputs of the decoder and train the system to predict the next OCR token based on previous ones,

$$\mathcal{L}_{\text{OCR-LM}} = -\sum_{i=1}^{N} \log\left(\mathbb{P}\left(t^i | t^{<i}\right)\right) \tag{4}$$

where t^i is the i^{th} OCR token. Minimizing such loss enforces the system to account for the scene-text signal, as desired. While variants of such a loss were previously used during pretraining [52, 7], we are the first to utilize it during finetuning. Moreover, inserting the OCR into the decoder at inference has another significant advantage, as it serves as a prefix and enables the model to condition its answers on the OCR. Such behavior is desirable since the OCR can provide meaningful information for general and scene-text VL tasks, as we experimentally demonstrate in Sec. 4.1.

OCR Binary Classification To obtain more meaningful and task-beneficial OCR encodings, we propose a binary classification objective of predicting whether each OCR token is a part of the ground-truth answer. We build a binary linear classifier on top of the outputs of the OCR encoder and train it using a binary cross-entropy loss. More specifically, since most of the OCR tokens are not part of the answer, we employ a weighted version, as such classification task is highly imbalanced. We denote this loss as $\mathcal{L}_{\text{OCR-BC}}$.

3.3. Training Procedure

So far, we have described the main building blocks in our method, and now, as illustrated in Fig. 3, we put it all together. First, we harness a trained general encoder-decoder VL model and modify it as described above in Sec. 3.1. Next, we freeze the VL model's pre-existing image encoder, similarly to [54, 3], and train UniTNT on a unified dataset (*i.e.*, general and scene-text VQA datasets

	Method	OCR System	VQA		TextVQA		ST-VQA	Avg.
			test-dev	test-std	val	test	test-ANLS	
VQA	SimVLM$_{\text{large}}$ [49]	✗	79.32	79.56	-	-	-	-
	GIT$^{\text{VQA}}_{\text{large}}$ [46]	✗	75.51	-	-	-	-	-
	ALBEF [31]	✗	75.22	75.38	11.67	13.88	0.19	44.63
	OFA$_{\text{Large}}$ [47]	✗	79.70	<u>79.85</u>	22.10	21.47	0.27	50.66
	mPLUG$_{\text{ViT-B}}$ [29]	✗	<u>79.79</u>	79.81	-	-	-	-
	BLIP [30]	✗	76.39	76.59	20.50	23.74	0.34	50.17
	UniTNT$_{\text{BLIP}}$	✓	79.68	79.78	36.33	35.90	0.50	57.84
	Δ		↑ 3.28	↑ 3.19	↑ 15.83	↑ 12.16	↑ 0.16	↑ 7.67
TextVQA	GIT$^{\text{TextVQA}}_{\text{large}}$ [46]	✗	-	-	37.47	-	-	-
	SA-M4C[23]	✓	-	-	45.4	44.6	0.50	-
	LOGOS [36]	✓	-	-	51.53	51.08	0.58	-
	M4C [21]	✓	27.47	27.70	46.53	47.42	0.43	37.56
	TAP [52]	✓	18.76	18.81	54.71	53.97	0.60	36.39
	LaTr [7]	✓	-	-	**59.53**	**59.55**	**0.68**	-
	BLIP [30]	✗	40.16	40.39	30.12	27.72	0.36	34.06
	UniTNT$_{\text{BLIP}}$	✓	37.01	37.24	50.19	47.39	0.59	42.32
	Δ		↓ 3.15	↓ 3.15	↑ 20.07	↑ 19.67	↑ 0.23	↑ 8.26
Comb.	M4C [21]	✓	59.11	59.04	47.22	48.61	0.50	53.83
	ALBEF [31]	✗	75.61	75.87	16.15	17.04	0.22	46.46
	UniTNT$_{\text{ALBEF}}$	✓	77.60	77.80	43.73	44.13	0.58	<u>60.97</u>
	Δ		↑ 1.99	↑ 1.93	↑ 27.58	↑ 27.09	↑ 0.36	↑ 14.51
	BLIP [30]	✗	77.40	77.39	32.43	31.48	0.44	54.44
	UniTNT$_{\text{BLIP}}$	✓	**79.90**	**80.08**	<u>55.21</u>	<u>55.35</u>	<u>0.66</u>	**67.72**
	Δ		↑ 2.50	↑ 2.69	↑ 22.77	↑ 23.87	↑ 0.22	↑ 13.28

Table 2: **VQA results.** Accuracy of general, scene-text oriented VQA methods and UniTNT using three training regimes – separate VQA and TextVQA and combined training, where non-open vocabulary methods results are in gray. Δ indicates improvement over the base architecture in the same regime. These results highlight our method's effectiveness, significantly improving the general VQA results by enriching VL models with scene-text understanding.

	Method	OCR System	COCO	TextCaps		Avg.
			Karpathy-test	val	test	
Caps	VinVL [55]	✗	129.3	-	-	-
	LEMON$_{\text{base}}$ [22]	✗	133.3	-	-	-
	GIT$^{\text{Cap}}_{\text{large}}$ [46]	✗	138.5	-	-	-
	SimVLM$_{\text{large}}$ [49]	✗	<u>142.6</u>	-	-	-
	OFA$_{\text{Large}}$ [47]	✗	**150.7**	64.5	66.8	108.8
	BLIP [30]	✗	133.3	59.4	61.9	97.6
	UniTNT$_{\text{BLIP}}$	✓	133.7	59.6	62.8	98.3
	Δ		↑ 0.4	↑ 0.2	↑ 0.9	↑ 0.7
TextCaps	GIT$^{\text{TextCap}}_{\text{large}}$ [46]	✗	-	106.3	-	-
	MMA-SR [45]	✓	-	98.0	88.0	-
	CNMT [42]	✓	-	-	93.0	-
	M4C-Captioner [40]	✓	4.7	95.5	90.1	47.4
	TAP [52]	✓	4.6	109.2	103.2	53.9
	BLIP [30]	✗	84.8	112.7	103.7	94.3
	UniTNT$_{\text{BLIP}}$	✓	70.4	**130.5**	**123.1**	96.8
	Δ		↓ 14.4	↑ 17.8	↑ 19.4	↑ 2.5
Comb.	M4C-Captioner [40]	✓	109.8	102.7	98.0	103.9
	BLIP [30]	✗	133.4	101.4	91.8	<u>112.6</u>
	UniTNT$_{\text{BLIP}}$	✓	134.0	<u>119.1</u>	<u>109.4</u>	**121.7**
	Δ		↑ 0.6	↑ 17.7	↑ 17.6	↑ 9.1

Table 3: **CAP results.** CIDEr scores of general, scene-text oriented CAP methods and UniTNT using three training regimes – separate Caps and TextCaps and combined training. Δ indicates improvement over the base architecture in the same regime. These results highlight our method's effectiveness, significantly improving the general CAP results by enriching VL models with scene-text understanding.

or general and scene-text captioning datasets). Specifically, $\mathcal{L}_{\text{UniTNT}} = \mathcal{L}_{\text{base}} + \alpha_1 \mathcal{L}_{\text{OCR-LM}} + \alpha_2 \mathcal{L}_{\text{OCR-BC}}$ is minimized, where $\mathcal{L}_{\text{base}}$ is the base task-dependent loss term used in our base architecture, and α_1, α_2 are tunable hyperparameters.

4. Experiments

In this section, we experimentally examine UniTNT, comparing its performance with state-of-the-art methods with a similar capacity on both VQA and CAP tasks, using separate and combined training. In particular, to better study the effects of our method, we test it and the baselines in three distinct training regimes; (i) separate training on the general datasets, (ii) separate training on the scene-text ones, and (iii) combined training approach, denoted as Comb. As we focus on models' see and read capabilities, we emphasize the combined training regime and view it as the most crucial one. However, the separate training regimes can provide insights into the impact of scene-text understanding on the general benchmarks and the biases within the scene-text datasets. As in Sec. 2, for each of the regimes, we consider three standard benchmarks for VQA: VQAv2 [18], TextVQA [41] and ST-VQA [8], and two for CAP: COCO Captions [9] and TextCaps [40]. We report the performance on each benchmark and the non-weighted averaged one (*combined evaluation*) to quantify the models' reasoning capabilities from both visual and textual information as a single number. For VQA, we calculate this score only on VQAv2 and TextVQA test sets. Lastly, in Sec. 4.3, we present a new subset evaluation setting for scene-text VQA to measure the model's ability to answer questions requiring reasoning over all modalities simultaneously. For all datasets, we extract OCR information using Amazon Text-in-Image[1] [35, 38, 1, 26]. The supplementary materials list the implementation details, additional dataset information and for completeness, a comparison with other methods, disregarding the models' size.

4.1. Visual Question Answering Experiments

We integrate our approach to two models, ALBEF and BLIP, denoted as UniTNT$_{\text{ALBEF}}$ and UniTNT$_{\text{BLIP}}$, respectively, and report their performance using three training regimes: (i) VQA, (ii) TextVQA, and (iii) *Comb.*, as shown in Tab. 2. In the first regime, training UniTNT$_{\text{BLIP}}$ exclusively on VQAv2 results in performance improvements of $+3.19\%$ and $+12.16\%$ on VQA, and TextVQA, respectively, leading to a significant boost of $+7.67\%$ in the average score. Even though VQAv2 mainly focuses on reasoning from visual information, these results stress the importance of scene-text understanding in this benchmark and the effectiveness of our method. Interestingly, despite the marginal presence of OCR in VQAv2, UniTNT$_{\text{BLIP}}$ manages to effectively harness it and obtain 35.90% on

[1] https://docs.aws.amazon.com/rekognition/latest/dg/text-detection.html

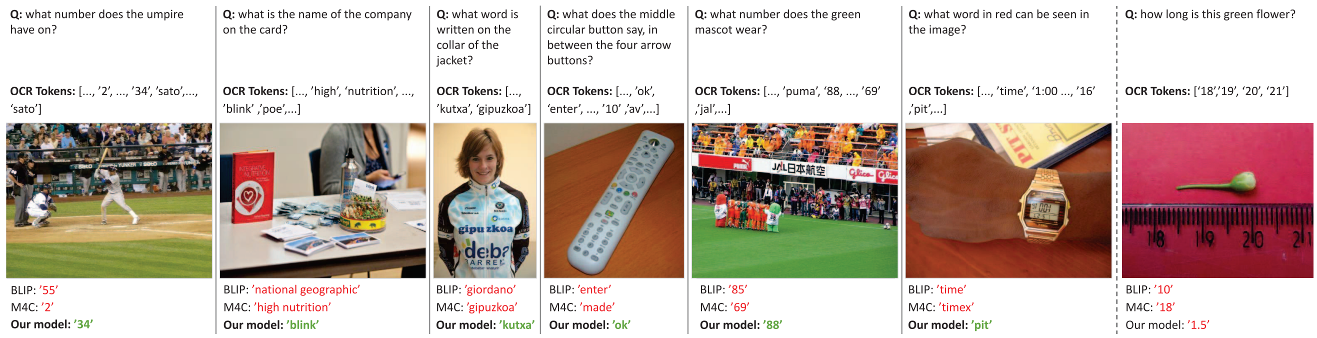

Figure 4: **Reasoning over all modalities.** We curate a subset out of TextVQA [41] validation set, containing only the samples which require reasoning over both vision and scene-text in the same question. Presented are representative examples from this subset, each includes an image, question, OCR input tokens, and model predictions. Green and red stand for correct and wrong predictions, respectively.

Method	TextVQA	$\text{TextVQA}_{\text{Read}}$	$\text{TextVQA}_{\text{See}\cap\text{Read}}$	Gap↓
M4C [21]	46.53	47.94	35.69	12.25
TAP [52]	54.71	56.24	35.83	20.41
$\text{UniTNT}_{\text{BLIP}}$	**55.21**	**56.32**	**44.44**	**11.88**

Table 4: **TextVQA splits.** Accuracy of leading scene-text VQA methods on the two non-overlapping subsets of TextVQA validation data, and the gap between them. 'See-∩-Read' refers to our subset, in which reasoning over all modalities is needed for each sample. 'Read' stands for the rest of the TextVQA validation set.

TextVQA, outperforming BLIP that trained solely on TextVQA itself (27.72%). In the scene-text configuration, performance improves by $+19.67\%$ on TextVQA; however, it decreases by -3.15% on VQA. This reinforces previous findings [48, 7], suggesting that scene-text VQA datasets contain biases encouraging models to over-rely on the OCR and disregard the visual information. As BLIP's scene-text understanding is very restricted, it cannot fully exploit such biases and retains its visual understanding better, expressed via better VQAv2 results. In the final combined training configuration, we showcase the versatility of our approach by presenting results for both $\text{UniTNT}_{\text{ALBEF}}$ and $\text{UniTNT}_{\text{BLIP}}$, highlighting its model-agnostic nature. When trained on both types of datasets, $\text{UniTNT}_{\text{BLIP}}$ improves BLIP by $+2.5\%$, $+22.77\%$, and $+13.28\%$ on VQA, TextVQA, and on average, respectively, achieving the highest average score. The results indicate that despite the biases in scene-text VQA datasets, UniTNT can harness them without sacrificing the visual reasoning capability. Moreover, $\text{UniTNT}_{\text{BLIP}}$ trained on the combined dataset outperforms models trained on each task separately, attesting to the tasks' mutually beneficial relationship, motivating the community to strive towards models that can see and read.

To gain a better understanding of the enhancements achieved by UniTNT across both "see" and "read" datasets, we conducted a qualitative analysis of our method, comparing it to BLIP and M4C in Fig. 1, Fig. 4 and in the supplementary materials. Our analysis indicates that the improvements observed in VQA are due to questions that necessitate reading and those that become easier to answer with OCR information. Regarding scene-text VQA, M4C struggles to reason from visual information, while UniTNT excels in this regard, resulting in significant performance improvements on both TextVQA and ST-VQA.

4.2. Image Captioning Experiments

Similar to our VQA experiments, we evaluate the performance of UniTNT on CAP by comparing it to top-performing methods using the same three training regimes (Caps, TextCaps, and *Comb.*). We only integrate our approach to BLIP ($\text{UniTNT}_{\text{BLIP}}$), as ALBEF was not applied to captioning. As the results in Tab. 3 indicate, in the Caps regime, our approach slightly improves both the per-task and the average results. Like in VQA, in the TextCaps training regime, UniTNT results in significant gains over TextCaps ($+19.4$ CIDEr points) but a decline in COCO (-14.4 CIDEr points), compared to BLIP. Moreover, while combined training leads to the best COCO results, the best TextCaps performance is achieved via designated TextCaps finetuning. This phenomenon aligns with the earlier findings by [40], attributing it to the different nature of ground truth captions in the scene-text and general benchmarks (additional analysis appears in the supplementary materials). Nevertheless, the combined trained UniTNT leads to the best average score across all methods and regimes.

4.3. A subset for Reasoning Over All Modalities

As illustrated in Fig. 1, VQA data is composed of three categories. Some questions can be answered using just vision ('see'), some by reasoning over the scene-text information only ('read'), and some require reasoning over both modalities at once ('see-∩-read'). Since most of the questions in current benchmarks fall either in the 'see' or

OCR-Sys	OCR-Enc	Fuse	$\mathcal{L}_{OCR-LM}$	$\mathcal{L}_{OCR-BC}$	2-D	VQA VQAv2	VQA TextVQA	CAP COCO	CAP TextCap
✗	✗	✗	✗	✗	✗	77.40	32.43	133.4	101.4
✓	✗	✗	✗	✗	✗	77.66	43.02	133.5	109.7
✓	✓	✗	✗	✗	✗	78.41	46.13	133.5	110.4
✓	✓	✓	✗	✗	✗	78.65	47.38	133.5	118.3
✓	✓	✓	✓	✗	✗	79.86	52.66	133.8	118.9
✓	✓	✓	✓	✓	✗	79.81	52.66	-	-
✓	✓	✓	✓	✗	✓	79.74	52.93	134.0	119.1
✓	✓	✓	✓	✓	✓	79.90	55.21	-	-

Table 5: **UniTNT design choices.** UniTNT$_{\text{BLIP}}$ results on VQA and CAP w.r.t its different building blocks.

Method	Freeze VE	VQA test-dev	TextVQA val	Avg.
UniTNT$_{\text{ALBEF}}$	✗	75.55	40.60	58.08
UniTNT$_{\text{ALBEF}}$	✓	77.60	43.73	60.67
UniTNT$_{\text{BLIP}}$	✗	78.77	52.45	65.61
UniTNT$_{\text{BLIP}}$	✓	79.90	55.21	67.56

Table 6: **Visual encoder freezing.** VQA accuracy of UniTNT$_{\text{BLIP}}$ and UniTNT$_{\text{ALBEF}}$, with and without freezing the visual encoder, attesting to the freezing's importance.

'read' category, unifying them is beneficial for testing methods' performance on the whole space, denoted by 'see-∪-read', eliminating the model's prior on whether a question is of type 'see' or 'read'. However, the more challenging and intriguing questions are the ones that require reasoning over scene-text and visual information altogether, denoted as 'see-∩-read'. To provide a more reliable way to evaluate VQA models on this questions' category, we manually curate all such image-question pairs from the TextVQA [41] validation set, producing an evaluation subset of 480 image-question pairs out of the total 5000 (±10%).

This subset can serve as a foundation for measuring models' capabilities on what we believe are the more challenging questions that the research community should tackle. In Fig. 4, we depict examples from this subset alongside the prediction of M4C [21], BLIP [30], and UniTNT. This qualitative analysis confirms that both scene-text and general VQA models struggle to cope with this type of questions, while UniTNT is substantially better. Moreover, in Tab. 4 we report the quantitative results of leading scene-text-oriented methods and UniTNT on the non-overlapping subsets of TextVQA validation set, *i.e.,* the 'TextVQA$_{\text{See}\cap\text{Read}}$' subset and its complementary set, 'TextVQA$_{\text{Read}}$', exposing the performance degradation that occurred on the former. As these findings suggest, our method leads to the best performance, affirming that it is indeed better at reasoning on scene-text and visual information simultaneously. Nevertheless, as can be seen, while UniTNT is a step forward, there is still a big room to improve on these types of challenging questions.

5. Ablation Studies

In this section, we study the effect of our key contributions and test the impact of freezing the vision encoder.

Design Choices: We ablate UniTNT's components on both the general and scene-text-oriented datasets in Tab. 5, where all numbers are reported under the *Comb.* settings. Since the trends in CAP results are similar to VQA, we will focus on analyzing the latter. First, we report the added performance of a naive approach – simply inserting the OCR tokens as an additional input to BLIP's existing text encoder, similar to [21, 52, 7]. As seen in Tab. 5, the accuracy on TextVQA improves by $+10.59\%$ (from 32.43% to 43.02%) while improving VQA results by $+0.22\%$. Our designated OCR encoder increases TextVQA performance to 46.13% ($+3.11\%$) while obtaining an additional $+0.75\%$ gain in VQA. Introducing our VL-OCR decoding scheme (denoted as "Fuse") boosts us to 47.37% on TextVQA and an extra $+0.24\%$ on VQA. Furthermore, using $\mathcal{L}_{\text{OCR-LM}}$ significantly improves TextVQA performance by $+5.22\%$ (from 47.38% to 52.66%) while gaining an extra $+1.21\%$ on VQA. Finally, the combination of $\mathcal{L}_{\text{OCR-BC}}$ with the 2-D information gets us to 55.21% and 79.9% on TextVQA and VQA. Overall, UniTNT leads to significant $+22.78\%$ and $+2.50\%$ improvements on TextVQA and VQA over the combined trained BLIP.

The Effect of Freezing the Visual Encoder: Recently, a few works [43, 3, 54] have examined different freezing configurations to avoid knowledge forgetness when combining pretrained models. Inspired by these works, we examine the effect of freezing the Visual Encoder (VE) weights while applying UniTNT, preserving its valuable knowledge acquired in pretraining, and summarize the results in Tab. 6. As our findings suggest, freezing the VE significantly improves the results on VQA for both UniTNT$_{\text{ALBEF}}$ and UniTNT$_{\text{BLIP}}$ by $+2.05\%$ and $+1.13\%$, and on TextVQA by $+3.13\%$ and $+2.75\%$, respectively.

6. Discussion and Conclusions

We wish to convey a few take-home messages to the VL research community. First, current SOTA methods cannot adequately reason over both scene-text and vision information. Our experiments demonstrate that this occurs even when combining training datasets, suggesting a fundamental limitation of existing methods. Second, our findings discover the symbiotic nature of these two types of reasoning capabilities, as performance on both tasks can be improved jointly. Moreover, by proposing UniTNT, we present the first single model that successfully handles both task types. Finally, we argue that the VL research community should strive to develop models that can simultaneously reason over vision, language, and scene-text. To facilitate this, we curate a suitable subset to serve as a benchmark foundation.

References

[1] Aviad Aberdam, Roy Ganz, Shai Mazor, and Ron Litman. Multimodal semi-supervised learning for text recognition. *arXiv preprint arXiv:2205.03873*, 2022. 6

[2] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8948–8957, 2019. 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. 2, 3, 4, 5, 8

[4] 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. 1, 12

[5] Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R Manmatha. Docformer: End-to-end transformer for document understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 993–1003, 2021. 5

[6] 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. 13

[7] Ali Furkan Biten, Ron Litman, Yusheng Xie, Srikar Appalaraju, and R Manmatha. Latr: Layout-aware transformer for scene-text vqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16548–16558, 2022. 2, 3, 4, 5, 6, 7, 8

[8] 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. 1, 2, 3, 6, 12

[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 preprint arXiv:1504.00325*, 2015. 2, 3, 6, 12

[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 preprint arXiv:2209.06794*, 2022. 1, 3

[11] 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 preprint arXiv:2209.06794*, 2022. 2, 4

[12] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: universal image-text representation learning. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX*, volume 12375 of *Lecture Notes in Computer Science*, pages 104–120. Springer, 2020. 3

[13] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022. 3

[14] 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. 13

[15] 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. 12

[16] Robert M French. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135, 1999. 2

[17] Chenyu Gao, Qi Zhu, Peng Wang, Hui Li, Yuliang Liu, Anton Van den Hengel, and Qi Wu. Structured multimodal attentions for textvqa. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 3

[18] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 1, 2, 3, 6

[19] Wei Han, Hantao Huang, and Tao Han. Finding the evidence: Localization-aware answer prediction for text visual question answering. *arXiv preprint arXiv:2010.02582*, 2020. 3

[20] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Computation*, 1997. 5

[21] Ronghang Hu, Amanpreet Singh, Trevor Darrell, and Marcus Rohrbach. Iterative answer prediction with pointer-augmented multimodal transformers for textvqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9992–10002, 2020. 2, 3, 4, 6, 7, 8

[22] 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. 1, 3, 6

[23] Yash Kant, Dhruv Batra, Peter Anderson, Alexander Schwing, Devi Parikh, Jiasen Lu, and Harsh Agrawal. Spatially aware multimodal transformers for textvqa. In *European Conference on Computer Vision*, pages 715–732. Springer, 2020. 3, 6

[24] 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. 13

[25] 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. 13

[26] Yair Kittenplon, Inbal Lavi, Sharon Fogel, Yarin Bar, R. Manmatha, and Pietro Perona. Towards weakly-supervised text spotting using a multi-task transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4604–4613, June 2022. 6

[27] 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. 13

[28] 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. 13

[29] Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, Hehong Chen, Guohai Xu, Zheng Cao, et al. mplug: Effective and efficient vision-language learning by cross-modal skip-connections. *arXiv preprint arXiv:2205.12005*, 2022. 6

[30] 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, 2, 3, 4, 6, 8, 12

[31] 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. 2, 3, 4, 6, 12

[32] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Gotmare, Shafiq R. Joty, Caiming Xiong, and Steven Chu-Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 9694–9705, 2021. 3

[33] 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. *arXiv preprint arXiv:2012.15409*, 2020. 3

[34] 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. 1, 13

[35] Ron Litman, Oron Anschel, Shahar Tsiper, Roee Litman, Shai Mazor, and R Manmatha. Scatter: selective context attentional scene text recognizer. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11962–11972, 2020. 6

[36] Xiaopeng Lu, Zhen Fan, Yansen Wang, Jean Oh, and Carolyn P Rosé. Localize, group, and select: Boosting text-vqa by scene text modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2631–2639, 2021. 3, 4, 6

[37] A. Mishra, K. Alahari, and C. V. Jawahar. Image retrieval using textual cues. In *ICCV*, 2013. 13

[38] Oren Nuriel, Sharon Fogel, and Ron Litman. Textadain: Fine-grained adain for robust text recognition. *arXiv preprint arXiv:2105.03906*, 2021. 6

[39] 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. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1, 3

[40] 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. 1, 2, 3, 6, 7, 12, 13

[41] 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. 1, 2, 3, 6, 7, 8, 12

[42] Anonymous submission. Cnmt, textcaps challenge 2020 top entry. *https://evalai.cloudcv.org/web/challenges/ challenge-page/573/overview*, 2020. 6

[43] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212, 2021. 2, 4, 8

[44] 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. 13

[45] Jing Wang, Jinhui Tang, and Jiebo Luo. Multimodal attention with image text spatial relationship for ocr-based image captioning. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4337–4345, 2020. 6

[46] 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. 1, 2, 3, 4, 6

[47] 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 Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 23318–23340. PMLR, 2022. 1, 2, 3, 6

[48] Qingqing Wang, Liqiang Xiao, Yue Lu, Yaohui Jin, and Hao He. Towards reasoning ability in scene text visual question answering. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2281–2289, 2021. 2, 3, 7

[49] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021. 1, 3, 6

[50] Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. Layoutlm: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1192–1200, 2020. 5

[51] Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, et al. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. *arXiv preprint arXiv:2012.14740*, 2020. 5

[52] Zhengyuan Yang, Yijuan Lu, Jianfeng Wang, Xi Yin, Dinei Florencio, Lijuan Wang, Cha Zhang, Lei Zhang, and Jiebo Luo. Tap: Text-aware pre-training for text-vqa and text-caption. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8751–8761, 2021. 2, 3, 4, 5, 6, 7, 8

[53] 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

[54] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18123–18133, 2022. 2, 5, 8

[55] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Making visual representations matter in vision-language models. 2021. 1, 3, 6