Supplementary Materials: VL-Reader: Vision and Language Reconstructor is an Effective Scene Text Recognizer

Anonymous Authors

1 RECONSTRUCTION *VS.* RECOGNITION

In this section, we further explore the connection between the quality of reconstruction and the final accuracy of recognition. Specifically, we set different hyper-parameters (*e.g.*, visual masking ratio r_v and patch size (p_h, p_w)) during the first training phase to observe different reconstruction qualities. In the second training phase, we train our recognizer with the same hyper-parameters across different runs. The remaining settings are fixed during both training phases.

1.1 Effect of Visual Masking Ratio r_v

We have conducted an ablation study on how different visual masking ratios r_v may affect the final recognition accuracy in our main manuscript. In this supplementary material, we further explore the potential impact of varying visual masking ratios, denoted as r_v, on the reconstruction quality. We adopt the SSIM [2] and PSNR [1] scores, which were proposed in the field of image quality assessment, as our evaluation metrics.

The results are presented in Table. 1 and are depicted in Fig. 1. As indicated in the table, a gradual decrease in the visual masking ratio from 0.75 to 0.5 results in a notable improvement in reconstruction quality (see Fig. 1). However, despite the improvement in reconstruction quality, the final recognition accuracy does not increase but, in fact, experiences a significant decline. This observation indicates that improved reconstruction quality does not necessarily result in better recognition accuracy. This could be due to the fact that decreasing the visual masking ratio r_v may lead to a significant decrease in training difficulties. Therefore, models trained with a lower visual masking ratio r_v can achieve good reconstruction quality using only the surrounding pixels, without the need to develop robust cross-modal feature representations, which are often crucial for a strong text recognizer.

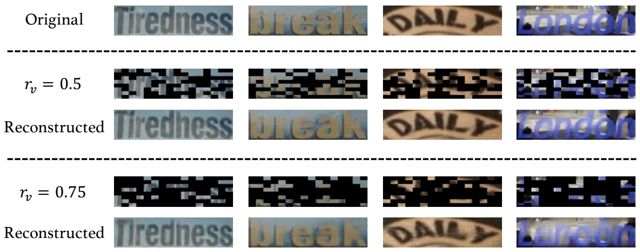

Figure 1: Reconstruction results on images of benchmark datasets (not utilized in training) with varying visual masking ratio r_v. For each column, we show the original image (top), the masked and reconstructed image with $r_v = 0.5$ (middle), and the masked and reconstructed image with $r_v = 0.75$ (bottom).

Table 1: Effect of varying visual masking ratio r_v. "SSIM" and "PSNR" denote the average SSIM and PSNR scores across six standard datasets. "Weighted Avg." indicates the weighted average recognition accuracy across six standard datasets (we use the larger version of IC13 and IC15).

Methods	r_v	SSIM	PSNR	Weighted Avg.
VL-Reader	0.5	**0.936**	**29.4**	96.32
VL-Reader	0.6	0.872	26.1	96.68
VL-Reader	0.7	0.871	26.0	96.89
VL-Reader	0.75	0.839	24.9	**96.90**
VL-Reader	0.8	0.799	23.7	96.79

1.2 Effect of Patch Size

Likewise, we also explore how different patch sizes (p_h, p_w) may impact reconstruction quality and recognition accuracy. The results are presented in Table. 2. As shown in the table, using a smaller patch size $(4, 4)$ can significantly improve the reconstruction quality (SSIM +0.038 and PSNR +1.5). However, we do not observe a significant improvement in recognition accuracy. The reason for this might be attributed to the varying training difficulties associated with different patch sizes during the initial training phase. Concretely, employing a larger patch size $(4, 8)$ increases the likelihood of masking an entire character of a word (see Fig. 2), thereby resulting in greater training difficulty. Conversely, utilizing a smaller patch size $(4, 4)$ indicates a reduced likelihood of masking a critical region, resulting in less training difficulty. Therefore, enhancing reconstruction quality by reducing training difficulty does not automatically ensure improved recognition results.

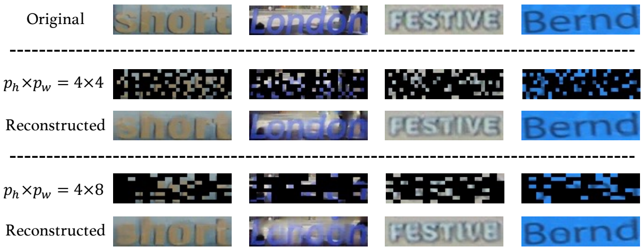

Figure 2: Reconstruction results on images of benchmark datasets (not used in training) with various patch sizes (p_h, p_w). For each column, we show the original image (top), the masked and reconstructed image with $(p_h, p_w) = (4, 4)$ (middle), and the masked and reconstructed image with $(p_h, p_w) = (4, 8)$ (bottom).

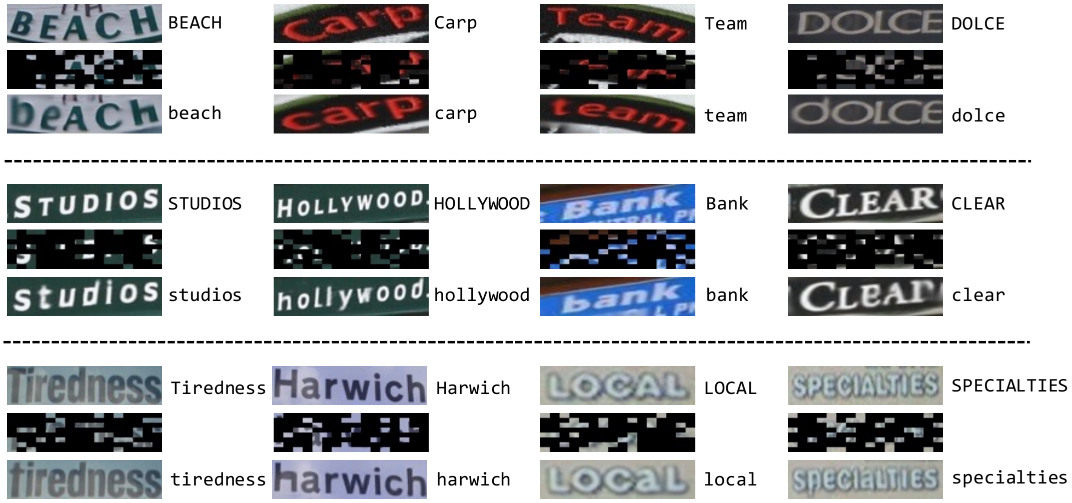

Figure 3: Reconstruction results on images of benchmark datasets (not used in training) with lower-case content prompt (bottom text). For each column, we show the original image (top), the masked image (middle), and the reconstructed image (bottom). Texts alongside the images indicate the original text (top) and lower-case text used for prompting image reconstruction (bottom). The default visual masking ratio r_v is set to 0.75.

Table 2: Effect of different patch size (p_h, p_w). "SSIM" and "PSNR" indicate the average SSIM and PSNR scores across six standard datasets. "Weighted Avg." indicates the weighted average recognition accuracy across six standard datasets (we use the larger version of IC13 and IC15).

Methods	(p_h, p_w)	SSIM	PSNR	Weighted Avg.
VL-Reader	(4, 4)	**0.877**	**26.4**	96.82
VL-Reader	(4, 8)	0.839	24.9	**96.90**

2 RECONSTRUCTION WITH CONTENT PROMPT

We conduct an additional experiment to validate the reconstruction quality with the content prompt by making a minor modification to our MVLD. Concretely, we disable the query tokens and directly use image features and text tokens for bi-modal feature interaction during the first training phase. During reconstruction, an additional modified text sequence, along with the masked image, is fed into the model to produce the reconstructed image. We illustrate several reconstructed images in Fig. 3. As can be seen in the figure, an additional modified text sequence can serve as the content prompt during reconstruction. For example, by changing the original text sequence into lower-case text (*e.g.*, "HOLLYWOOD" to "hollywood"), the reconstructed characters are transformed from the original upper-case ones into lower-case ones (*e.g.*, character "b", "e", "h" in "beach", "h", "l", "y", "d" in "hollywood", see Fig. 3). This result demonstrates that our training objective is capable of learning visual features from linguistic context, resulting in robust cross-modal feature representations.

REFERENCES

[1] Hamid R Sheikh, Alan C Bovik, and Gustavo De Veciana. 2005. An information fidelity criterion for image quality assessment using natural scene statistics. *IEEE Transactions on image processing* 14, 12 (2005), 2117–2128.

[2] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13, 4 (2004), 600–612.