Zero-shot Artistic Text Recognition with Multimodal Language Models

Tien Do [1,2], Thuyen Tran [1,2], Duy-Dinh Le[1,2], Thanh Duc Ngo[1,2]
[1] University of Information Technology, VNU-HCM, Vietnam
[2] Vietnam National University, Ho Chi Minh City, Vietnam
{tiendv, thuyentd, duyld, thanhnd}@uit.edu.vn

***Abstract*—Artistic Text (AT), characterized by stylized fonts, geometric distortions, and complex visual effects, frequently appears in real-world contexts such as advertising, signage, and branding. Recognizing such text is crucial for downstream tasks like image understanding and multimodal reasoning. While recent Multimodal Language Models (MLLMs) demonstrate strong general vision-language capabilities, it remains unclear whether they can effectively perceive and interpret artistic text. In this work, we conduct a comprehensive zero-shot evaluation of representative MLLMs on a diverse benchmark of artistic text, revealing key limitations and motivating the need for hybrid approaches that combine visual grounding with robust language modeling.**

I. Introduction

Artistic text (artext), characterized by stylized typography, distortions, and visual effects, is increasingly common in real-world settings such as advertising, branding, and digital media. Recognizing such text is not only challenging for conventional OCR systems but also crucial for downstream applications like visual search, multimodal retrieval, and document understanding. As illustrated in Fig 1, artistic text often appears with diverse styles, distortions, and background noise, making recognition particularly challenging.

Recent advances in MLLMs, such as GPT-4V [1], Gemini [2], or Qwen-VL [3] have demonstrated impressive capabilities in understanding both visual and textual modalities, enabling end-to-end tasks like image captioning, visual question answering (VQA), and document parsing. These models are trained on large-scale image-text pairs and exhibit strong generalization to various domains.

In the domain of scene text recognition (STR), prior works have proposed specialized architectures that explicitly model geometric deformations, such as attention-based decoders, rectification modules, or segmentation-guided pipelines. Recent methods like Corner-Guided Transformers [4] or Skeleton-Guided ATR (SG-ATR) [5] have further improved robustness by incorporating structural priors of characters, showing strong performance on challenging benchmarks like WordArt and Artistic Text-In-The-Wild (ATTW). These OCR-focused methods typically rely on task-specific architectures and synthetic training data, enabling them to achieve high accuracy on curated benchmarks. However, their limited flexibility and lack of broader reasoning capabilities contrast with the general-purpose nature of MLLMs. This raises a key question: Can modern MLLMs recognize and interpret artistic text in the wild without any fine-tuning?

To investigate this, we evaluate a set of representative MLLMs—GPT-4.1, Gemini (1.5–2.5), Qwen2.5-VL, InternVL3-VL, and MiniCPM-o-2.6—on ATTW benchmark, which includes 16,627 artistic text instances across seven diverse subsets. We perform zero-shot recognition on the 4,152-image test set and compare the results against several robust OCR baselines.

Our findings show that the best-performing MLLM (GPT-4.1) achieves 68.01% accuracy, substantially lower than SG-ATR (75.39%). This performance gap highlights the limitations of current MLLMs in perceiving stylized text and underscores the need for dedicated benchmarking and future integration of visual priors into multimodal architectures.

II. Related Works

A. Scene Text Recognition

Scene text recognition (STR) has advanced considerably with deep learning, typically involving stages such as Image Rectification, Feature Extraction, Context Modeling, and Prediction [6]. In particular, STR methods are categorized by their prediction strategies into four types: Encoder–CTC-decoder, Encoder–Attention-decoder, Transformer with MLP-decoder, and Character Segmentation.

Encoder–CTC-decoder methods, originating from the work of [7], use convolutional and recurrent layers with CTC loss to align visual features with character sequences. Although effective, they struggle with irregular text. Rectification-based variants [8]–[10] incorporate spatial transformer networks to address this, with recent improvements deriving transformation parameters from image features for adaptability [10].

Encoder–Attention-decoder approaches enhance prediction via attention mechanisms. SAR [11] introduces attention decoders, further refined by methods such as DAN [12] and SCATTER [13], which integrate visual and linguistic features. To resolve linguistic ambiguities, newer architectures [4], [14] apply multi-head attention and cross-modal encodings leveraging handcrafted features for robust character recognition.

Transformer with MLP-decoder methods build upon Vision Transformers (ViTs), removing CNN inductive biases. Some, like [15], predict characters using encoder-only models, while others like MGP-STR [16] and TrOCR [17] adopt hierarchical

979-8-3315-7206-8/25/$31.00

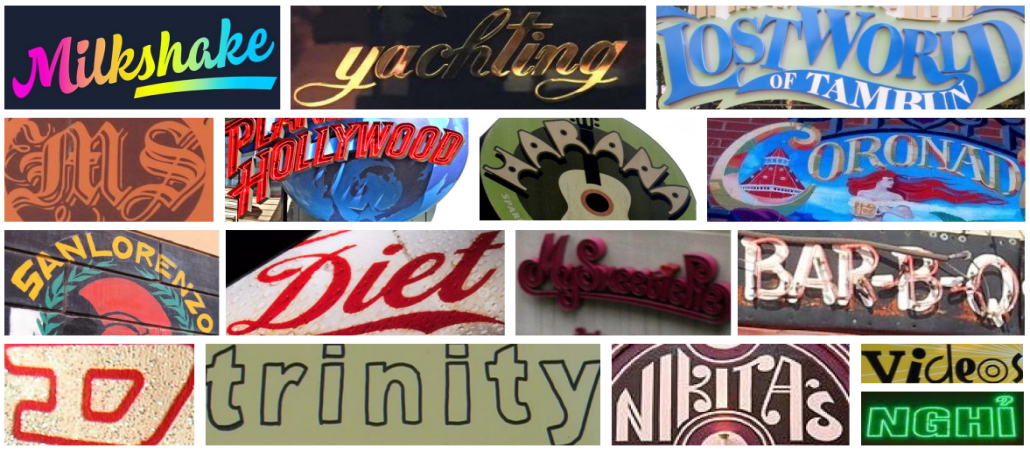

Fig. 1. ATR challenges: Artistic text varies widely in style, effects, and background complexity, posing difficulties for robust recognition.

and pretrained language model-based decoding. PARSeq [18] mitigates directional bias using varied attention masks. To improve robustness, hybrid methods [13], [19], [20] incorporate external language models or cross-modal mechanisms; however, they still face challenges in recognizing out-of-vocabulary (OOV) words [21].

Character Segmentation methods [22]–[25] treat recognition as character segmentation, achieving higher accuracy but requiring dense annotations and complex post-processing.

B. Vision-Language Models for STR

Vision-Language Models (VLMs) have achieved remarkable performance across a wide range of vision tasks [26], largely due to their pretraining on large-scale image-text pairs sourced from the internet. Representative examples of such models include CLIP [27] and BLIP [28].

A growing trend in the field involves directly applying VLMs to downstream visual recognition tasks such as object detection (e.g., DetCLIP [29]) and text-video retrieval (e.g., CLIP4Clip [30]). STR is no exception to this development. Recent studies have explored integrating CLIP into STR pipelines, as demonstrated in approaches such as CLIPTER [31], CLIP-OCR [32], and most recently, CLIP4STR [33]. Specifically, CLIPTER enhances character recognition by utilizing CLIP features extracted from the global image, while CLIP-OCR leverages both visual and linguistic knowledge from CLIP through feature distillation.

On the heavier end of VLMs are MLLMs, such as GPT-4V [1], Gemini [2], Qwen-VL [3], Intern-VL [34], and MiniCPM [35]. These MLLMs have demonstrated strong performance across a wide range of benchmarks that evaluate multimodal understanding and reasoning capabilities, including tasks related to OCR [36].

However, as noted in Long et al. [37], STR methods generally struggle with irregular text, such as rotated, curved, blurred, or occluded instances. Furthermore, [5] highlights the challenges posed by artistic text styles, which continue to hinder even state-of-the-art STR methods such as ABINet [19] and PARSeq [18]. Therefore, in this work, we investigate the robustness and understanding capabilities of MLLMs under such challenging and visually complex conditions.

III. Methodology

A. Experimental Settings

To investigate the robustness and understanding capabilities of MLLMs, we evaluate the following representative models:

Closed-Source MLLMS:

- **OpenAI GPT** [1]: Including GPT-4.1 and GPT-4.1-mini variants;
- **Gemini** [2]: Including Gemini 1.5-Flash, Gemini 2.0-Flash, Gemini 2.5-Flash, and Gemini 2.5-Flash-Pro variants;

Open-Source MLLMS:

- **Qwen-VL** [3]: Including Qwen2.5-VL 3B and Qwen2.5-VL 7B[1] variants;
- **Intern-VL** [34]: Including InternVL3-VL 1B, InternVL3-VL 8B, and InternVL3-VL 14B[1] variants;
- **MiniCPM** [35]: Including MiniCPM-o-2.6 8B;

B. Dataset

The *Artistic Text-In-The-Wild* (ATTW) [5] benchmark is introduced to address the limitations of existing datasets such as WordArt [4], which predominantly focus on designed text while overlooking naturally occurring artistic text in real-world environments. ATTW captures a broad spectrum of visual complexity, including font deformation, diverse text effects, and environmental noise.

ATTW consists of 16,627 annotated artistic text instances collected from widely used STR datasets, such as **CUTE**, **IC13**, **TotalText**, **WordArt**, **BKAI**, **SignboardText** (ST), and **VinText**, featuring characteristics such as intricate fonts, geometric deformations, diverse visual effects (e.g., 3D, neon), overlapping characters, and background interference. ATTW includes two subsets:

[1]#B denotes the number of billions of model parameters.

TABLE I

EXPERIMENTAL RESULTS OF **CLOSED-SOURCE MLLMS** ARE REPORTED IN TERMS OF ACCURACY (HIGHER IS BETTER) ON $\text{ATTW}_{\text{TEST}}$. THE **BOLD NUMBERS** INDICATE THE HIGHEST ACCURACY ACHIEVED FOR EACH RESPECTIVE DATASET.

	Methods	Test datasets and # of samples							
		CUTE 116	**IC13** 251	**Total** 696	**WordArt** 1511	**BKAI** 336	**ST** 209	**VinText** 1033	**$\text{ATTW}_{\text{test}}$** 4152
OpenAI	GPT-4.1	**77.59**	**84.86**	**74.14**	**68.12**	71.04	**54.07**	**60.41**	**68.01**
	GPT-4.1-mini	71.55	74.90	63.94	66.04	63.28	45.45	50.34	61.19
Gemini	Gemini 2.5-Pro	72.41	75.70	58.33	53.56	**73.13**	48.33	55.95	58.16
	Gemini 2.5-Flash	70.69	73.71	61.78	57.11	68.66	49.28	57.79	60.00
	Gemini 2.0-Flash	**77.59**	76.89	66.81	60.40	71.04	51.20	56.53	62.40
	Gemini 1.5-Flash	70.69	62.95	55.60	54.16	44.18	29.19	34.75	48.47

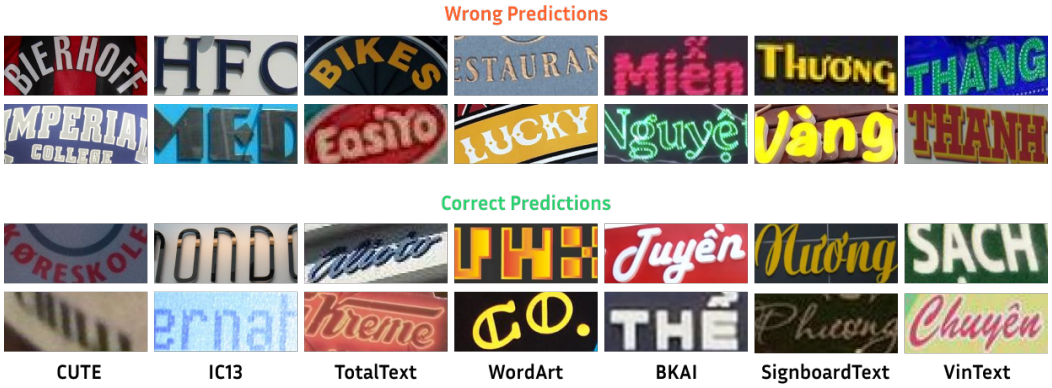

Fig. 2. Example predictions from GPT-4.1 on ATTW test images. Top: incorrect predictions. Bottom: correct predictions. Challenges include distortion, background clutter, and language-specific typography.

- A training set consisting of 12,475 artistic images, referred to as $\text{ATTW}_{\text{train}}$.
- A test set consisting of 4,152 artistic images, referred to as $\text{ATTW}_{\text{test}}$.

In this setting, we do not perform retraining on the $\text{ATTW}_{\text{train}}$ subset but evaluate only on $\text{ATTW}_{\text{test}}$. This decision is motivated by the computational demands of MLLMs, which typically comprise billions of parameters. According to the InternVL documentation [34], fine-tuning InternVL3-8B requires 8 A100-80GB GPUs, even for modest batch sizes. Moreover, training on a limited number of STR images (12,475 from $\text{ATTW}_{\text{train}}$) would be highly susceptible to overfitting. To focus on zero-shot evaluation, we omit the training process and directly query MLLMs on $\text{ATTW}_{\text{test}}$ using prompt-based interaction.

C. Evaluation Tasks

To evaluate the capabilities of Multimodal Language Models (MLLMs) in recognizing artistic text, we define three complementary tasks, each targeting a distinct aspect of model performance.

The first task examines zero-shot recognition, where MLLMs are prompted to predict textual content from ATTW images without any fine-tuning. This task evaluates the models' inherent ability to handle geometric distortions, complex backgrounds, and stylized typography using prompt-based queries (e.g., "read the text").

The second task investigates the impact of model family and scale by comparing open- and closed-source MLLMs (e.g., Qwen2.5-VL, GPT-4.1, Gemini) across multiple ATTW subsets such as CUTE, IC13, and VinText. This analysis highlights how architectural design and model size influence robustness across different textual domains.

The third task benchmarks MLLMs against specialized STR models such as SG-ATR, which incorporate explicit visual priors and structure-aware decoding. This comparison quantifies the performance gap and examines whether MLLMs can serve as viable OCR alternatives or benefit from hybrid integration.

Collectively, these tasks form a structured evaluation framework for analyzing the limitations of current MLLMs and guiding future advances in multimodal artistic text understanding.

TABLE II

EXPERIMENTAL RESULTS OF **OPEN-SOURCE MLLMS** (# DENOTES THE NUMBER OF BILLIONS OF PARAMETERS) ARE REPORTED IN TERMS OF ACCURACY (HIGHER IS BETTER) ON $\text{ATTW}_{\text{TEST}}$. THE **BOLD NUMBERS** INDICATE THE HIGHEST ACCURACY ACHIEVED FOR EACH RESPECTIVE DATASET.

Methods		Test datasets and # of samples							
		CUTE 116	**IC13** 251	**Total** 696	**WordArt** 1511	**BKAI** 336	**ST** 209	**VinText** 1033	**$\text{ATTW}_{\text{test}}$** 4152
Qwen2.5-VL	3B	76.72	62.95	62.36	61.95	**72.24**	**43.54**	**54.11**	60.44
	7B	75.00	68.53	66.67	64.23	70.45	40.19	52.86	**61.65**
InternVL3-VL	1B	68.97	72.91	64.08	59.46	39.70	34.45	30.59	51.23
	8B	**76.72**	79.28	68.53	63.89	50.15	42.58	40.08	57.82
	14B	68.97	73.71	63.20	63.22	48.96	41.15	39.11	55.71
MiniCPM-o-2.6	8B	**76.72**	**82.07**	**70.98**	**68.93**	41.19	33.97	32.04	57.05

TABLE III

EXPERIMENTAL RESULTS COMPARING THE BEST **OPEN-SOURCE MLLMS**, THE BEST **CLOSED-SOURCE MLLMS**, AND SG-ATR ARE REPORTED IN TERMS OF ACCURACY (HIGHER IS BETTER) ON $\text{ATTW}_{\text{TEST}}$. THE **BOLD NUMBERS** INDICATE THE HIGHEST ACCURACY ACHIEVED FOR EACH RESPECTIVE DATASET.

Methods	Test datasets and # of samples							
	CUTE 116	**IC13** 251	**Total** 696	**WordArt** 1511	**BKAI** 336	**ST** 209	**VinText** 1033	**$\text{ATTW}_{\text{test}}$** 4152
SG-ATR [5]	**83.62**	**84.86**	70.40	**80.95**	**80.86**	**75.51**	**73.46**	**75.39**
Qwen2.5-VL	75.00	68.53	66.67	64.23	70.45	40.19	52.86	61.65
OpenAI GPT-4.1	77.59	**84.86**	**74.14**	68.12	71.04	54.07	60.41	68.01

IV. EXPERIMENTS AND RESULTS

In this section, we present the experimental results of zero-shot recognition experiments using MLLMs on the artistic texts in ATTW benchmark. The results are summarized in Table I, Table II, and Table III, which report recognition accuracy across seven challenging ATTW subsets: CUTE, IC13, TotalText, WordArt, BKAI, ST, and VinText.

A. Zero-shot Artistic Text Recognition

The first evaluation task assesses the zero-shot capability of MLLMs to recognize artistic text without fine-tuning. Table I and Table II show the results of closed-source and open-source models, respectively. GPT-4.1 leads among MLLMs with 68.01% overall accuracy and shows strong results on IC13 (84.86%) and TotalText (74.14%), though it remains behind STR-focused models like SG-ATR. Smaller variants like GPT-4.1-mini (61.19%) and Gemini 1.5-Flash (48.47%) show degraded results, highlighting the role of model scale and pretraining diversity.

Among open-source models, Qwen2.5-VL (7B) achieves the highest overall accuracy of 61.65%, performing well on BKAI (70.45%) and IC13 (68.53%) but dropping notably on ST (40.19%) and VinText (52.86%).

MiniCPM-o-2.6, another 8B open-source model, performs well on TotalText (70.98%) and WordArt (68.93%), but its overall accuracy (57.05%) remains lower than Qwen2.5-VL and InternVL3-VL. Its weak performance on ST (33.97%) and VinText (32.04%) suggests limitations in handling background clutter and underrepresented scripts.

Overall, while MLLMs show potential for zero-shot recognition, their performance remains limited under complex styles, noisy scenes, or language-specific distortions.

B. Comparison Across Models and Scales

The second evaluation task explores the variability in performance across different model families and sizes. From Table II, we observe that increasing the number of parameters in Qwen2.5-VL from 3B to 7B leads to consistent improvements across most subsets. For example, on the TotalText subset, accuracy improves from 62.36% (3B) to 66.67% (7B). While MiniCPM-o-2.6 shares a similar parameter size with InternVL3-VL (8B) and Qwen2.5-VL (7B), it demonstrates a more uneven performance profile. Despite strong scores on TotalText and IC13, the model underperforms significantly on subsets with Vietnamese sub-sets of ATTW such as VinText or ST, highlighting potential limitations in its OCR grounding

or pretraining diversity. These results suggest that model scale alone does not guarantee generalization in stylized or culturally diverse text settings.

In Table I, a similar trend is observed among the Gemini family, although not strictly monotonic. Gemini 2.0-Flash achieves a higher average accuracy (62.40%) than Gemini 1.5-Flash (48.47%), with marked improvements on IC13, WordArt, and BKAI. However, performance gains diminish beyond a certain scale, and certain subsets (e.g., VinText and ST) remain difficult for all Gemini variants, regardless of size.

These observations suggest that while model scale contributes to robustness, architectural choices and pretraining corpus composition are also crucial, especially in handling linguistic and stylistic diversity present in artistic text.

C. Comparison with STR Baselines

We compare the performance of MLLMs with that of dedicated STR models, as shown in Table III. SG-ATR, a skeleton-guided model tailored for artistic text recognition, achieves the highest overall accuracy of 75.39%, surpassing all MLLMs by a considerable margin. Notably, SG-ATR maintains strong accuracy even on highly stylized subsets such as WordArt (80.95%) and VinText (73.46%).

GPT-4.1 achieves the highest accuracy among MLLMs (68.01%), performing well on IC13 and TotalText, but struggling with subsets featuring heavy distortions or or underrepresented languages. Qwen2.5-VL (7B) ranks third overall with 61.65%, showing moderate performance but a wider variance across subsets.

This performance gap reinforces the value of explicit visual priors in STR-specific architectures. MLLMs, though general and powerful, currently lack mechanisms for structured visual representation and fine-grained character localization, both of which are critical for high-precision text recognition in artistic settings.

D. Qualitative Observations from GPT-4.1 Predictions

To further illustrate model behavior, Fig. 2 shows several correct and incorrect predictions made by GPT-4.1 on ATTW test images. The model performs well when the text is high-contrast, well-aligned, or has clear stroke structure, as seen in examples like "SÁCH" or "PHỞ". Conversely, failures occur in the presence of distortion, cluttered backgrounds, or decorative effects, such as in "LUCKY" or "Nguyễn". Notably, Vietnamese diacritics are often misread or omitted, indicating limited robustness to language-specific visual cues.

These observations align with our quantitative findings, reaffirming that while GPT-4.1 handles moderately stylized text, it struggles with visually complex or culturally specific scenarios.

V. Conclusions

In this study, we evaluated the capability of Multimodal Language Models (MLLMs) to recognize and interpret artistic text, one of the most visually complex and underexplored forms of scene text in a zero-shot setting. Using the ATTW benchmark, we conducted zero-shot experiments across a diverse set of MLLMs, including GPT-4.1, Gemini, Qwen2.5-VL, InternVL3-VL and MiniCPM-o-2.6. Our results reveal that even the strongest MLLMs underperform compared to dedicated OCR models like SG-ATR, especially on subsets with heavy background interference or stylized typography.

These findings suggest that current MLLMs, while powerful in general multimodal reasoning, still lack the structural visual priors necessary for robust artistic text recognition. Our work highlights the need for hybrid solutions that combine strong visual representations with language understanding, and positions ATTW as a valuable benchmark for evaluating future progress in this direction.

References

[1] Z. Yang *et al.*, "The dawn of lmms: Preliminary explorations with gpt-4v (ision)," *arXiv preprint arXiv:2309.17421*, vol. 9, no. 1, p. 1, 2023.

[2] G. Team *et al.*, "Gemini: A family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2023.

[3] S. Bai *et al.*, *Qwen2.5-vl technical report*, 2025. arXiv: 2502.13923 [cs.CV]. [Online]. Available: https://arxiv.org/abs/2502.13923.

[4] X. Xie, L. Fu, Z. Zhang, Z. Wang, and X. Bai, "Toward understanding wordart: Corner-guided transformer for scene text recognition," in *ECCV*, 2022.

[5] T. Do, T. Tran, K. Le, D.-D. Le, and T. D. Ngo, "Skeleton-guided artistic text recognition," in *International Conference on Document Analysis and Recognition*, Springer, 2025.

[6] J. Baek *et al.*, "What is wrong with scene text recognition model comparisons? dataset and model analysis," in *International Conference on Computer Vision (ICCV)*, 2019, published.

[7] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," *IEEE transactions on pattern analysis and machine intelligence*, 2016.

[8] W. Liu, C. Chen, K.-Y. K. Wong, Z. Su, and J. Han, "Star-net: A spatial attention residue network for scene text recognition.," in *BMVC*, vol. 2, 2016, p. 7.

[9] C. Luo, L. Jin, and Z. Sun, "Moran: A multi-object rectified attention network for scene text recognition," *Pattern Recognition*, vol. 90, pp. 109–118, 2019.

[10] T. Zheng, Z. Chen, J. Bai, H. Xie, and Y.-G. Jiang, "Tps++: Attention-enhanced thin-plate spline for scene text recognition," *arXiv preprint arXiv:2305.05322*, 2023.

[11] H. Li, P. Wang, C. Shen, and G. Zhang, "Show, attend and read: A simple and strong baseline for irregular text recognition," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, 2019, pp. 8610–8617.

[12] T. Wang *et al.*, "Decoupled attention network for text recognition," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, 2020, pp. 12 216–12 224.

[13] R. Litman, O. Anschel, S. Tsiper, R. Litman, S. Mazor, and R. Manmatha, "Scatter: Selective context attentional scene text recognizer," in *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 962–11 972.

[14] J. Lee, S. Park, J. Baek, S. J. Oh, S. Kim, and H. Lee, "On recognizing texts of arbitrary shapes with 2d self-attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 546–547.

[15] R. Atienza, "Vision transformer for fast and efficient scene text recognition," in *International Conference on Document Analysis and Recognition*, Springer, 2021, pp. 319–334.

[16] C. D. Peng Wang and C. Yao, "Multi-granularity prediction for scene text recognition," in *European Conference on Computer Vision*, 2022.

[17] M. Li *et al.*, "Trocr: Transformer-based optical character recognition with pre-trained models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, 2023, pp. 13 094–13 102.

[18] D. Bautista and R. Atienza, "Scene text recognition with permuted autoregressive sequence models," in *European Conference on Computer Vision*, Cham: Springer Nature Switzerland, Oct. 2022, pp. 178–196. DOI: 10.1007/978-3-031-19815-1_11.

[19] S. Fang, H. Xie, Y. Wang, Z. Mao, and Y. Zhang, "Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7098–7107.

[20] S. Fang, Z. Mao, H. Xie, Y. Wang, C. Yan, and Y. Zhang, "Abinet++: Autonomous, bidirectional and iterative language modeling for scene text spotting," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–18, 2022. DOI: 10.1109/TPAMI.2022.3223908.

[21] S. Garcia-Bordils *et al.*, "Out-of-vocabulary challenge report," in *European Conference on Computer Vision*, Springer, 2022, pp. 359–375.

[22] P. Lyu, M. Liao, C. Yao, W. Wu, and X. Bai, "Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 67–83.

[23] M. Liao *et al.*, "Scene text recognition from two-dimensional perspective," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, 2019, pp. 8714–8721.

[24] L. Xing, Z. Tian, W. Huang, and M. R. Scott, "Convolutional character networks," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[25] J. Tang, Z. Yang, Y. Wang, Q. Zheng, Y. Xu, and X. Bai, "Seglink++: Detecting dense and arbitrary-shaped scene text by instance-aware component grouping," *Pattern recognition*, vol. 96, p. 106 954, 2019.

[26] J. Zhang, J. Huang, S. Jin, and S. Lu, "Vision-language models for vision tasks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[27] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*, 2021.

[28] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International conference on machine learning*, PMLR, 2022, pp. 12 888–12 900.

[29] L. Yao *et al.*, "Detclip: Dictionary-enriched visual-concept paralleled pre-training for open-world detection," *Advances in Neural Information Processing Systems*, vol. 35, pp. 9125–9138, 2022.

[30] H. Luo *et al.*, "Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning," *Neurocomputing*, vol. 508, pp. 293–304, 2022.

[31] A. Aberdam *et al.*, "Clipter: Looking at the bigger picture in scene text recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 21 706–21 717.

[32] Z. Wang, H. Xie, Y. Wang, J. Xu, B. Zhang, and Y. Zhang, "Symmetrical linguistic feature distillation with clip for scene text recognition," in *Proceedings of the 31st ACM international conference on multimedia*, 2023, pp. 509–518.

[33] S. Zhao, R. Quan, L. Zhu, and Y. Yang, "Clip4str: A simple baseline for scene text recognition with pre-trained vision-language model," *IEEE Transactions on Image Processing*, 2024.

[34] Z. Chen *et al.*, "Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 24 185–24 198.

[35] Y. Yao *et al.*, "Minicpm-v: A gpt-4v level mllm on your phone," *arXiv preprint arXiv:2408.01800*, 2024.

[36] X. Yue *et al.*, "Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9556–9567.

[37] S. Long, X. He, and C. Yao, "Scene text detection and recognition: The deep learning era," *International Journal of Computer Vision*, vol. 129, no. 1, pp. 161–184, 2021.