A Token-level Text Image Foundation Model for Document Understanding

Tongkun Guan [1 *] **Zining Wang** [2 *] **Pei Fu** [2] **Zhengtao Guo** [3] **Wei Shen** [1] **Kai Zhou** [2] **Tiezhu Yue** [2] **Chen Duan** [2]
Hao Sun [4] **Qianyi Jiang** [2] **Junfeng Luo** [2] **Xiaokang Yang** [1]

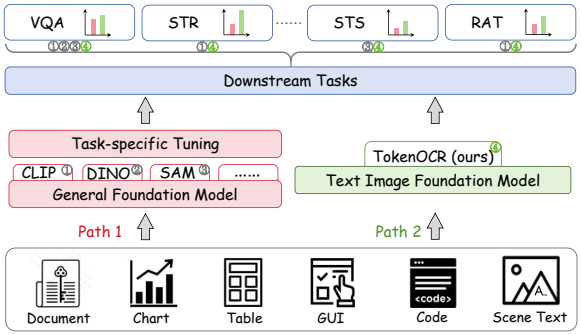
Abstract

In recent years, general visual foundation models (VFMs) have witnessed increasing adoption, particularly as image encoders for popular multi-modal large language models (MLLMs). However, without semantically fine-grained supervision, these models still encounter fundamental prediction errors in the context of downstream text-image-related tasks, *i.e.*, perception, understanding and reasoning with images containing small and dense texts. To bridge this gap, we develop **TokenOCR**, the first token-level visual foundation model specifically tailored for text-image-related tasks, designed to support a variety of traditional downstream applications. To facilitate the pretraining of TokenOCR, we also devise a high-quality data production pipeline that constructs the first token-level image text dataset, **TokenIT**, comprising 20 million images and 1.8 billion token-mask pairs. Furthermore, leveraging this foundation with exceptional image-as-text capability, we seamlessly replace previous VFMs with TokenOCR to construct a document-level MLLM, **TokenVL**, for VQA-based document understanding tasks. Finally, extensive experiments demonstrate the effectiveness of TokenOCR and TokenVL. Code, datasets, and weights will be available at https://token-family.github.io/TokenOCR_project/.

1. Introduction

Text image acts as a crucial medium for information transmission in everyday life. The precise interpretation of these images significantly enhances the automation of information processes, including text recognition, retrieval, segmentation, and understanding.

With the trend towards these tasks unification and the advancement of multi-modal large language models (MLLMs), visual foundation models (VFMs) have garnered considerable attention due to their broad capabilities in providing visual understanding for these downstream vision tasks (Covert et al., 2024). For instance, popular general models CLIP (Radford et al., 2021), DINO (Caron et al., 2021), and SAM (Kirillov et al., 2023) are widely adapted for text-image-related tasks to achieve performance gains through LoRA/adapter tuning (Ye et al., 2024), prompt learning (Yu et al., 2023), and learnable position interpolation technology. Additionally, CLIP and SigLIP (Zhai et al., 2023) have also proven effective as visual encoders for MLLMs in concurrent studies (McKinzie et al., 2024; Tong et al., 2024).



VFM	Granularity	Dataset	#Image	#Pairs
CLIP (Radford et al., 2021)	image-level	WIT400M	400M	0.4B
DINO (Caron et al., 2021)	image-level	ImageNet	14M	-
SAM (Kirillov et al., 2023)	pixel-level	SA1B	11M	1.1B
TokenOCR	token-level	TokenIT	20M	1.8B

Figure 1: For different tasks, previous works select different VFMs from general foundation models (path 1). In contrast, we develop a unified token-level foundation model, **TokenOCR**, specifically tailored for text-image-related tasks (path 2). TokenOCR is trained on a substantial self-built dataset, **TokenIT**, comprising 20 million images and 1.8 billion token-mask pairs. This well-learned model is capable of supplanting other VFMs in related downstream tasks.

However, these VFMs, trained with image-level supervision, are not optimal for processing fine-grained dense prediction tasks (Yu et al., 2024), such as document understanding with densely packed and small visual texts. Although several works attempt to incorporate SAM as an additional high-resolution encoder (Wei et al., 2025; Fan et al., 2024) or combine other expert models (Lin et al., 2023b), these dual or more complex VFM combinations result in a doubling of the number of tokens, which is costly and lack flexibility.

Furthermore, to the best of our knowledge, there is currently almost no fine-grained text image foundation model with token granularity, specifically tailored for extracting robust and general visual text semantic feature representations.

In this work, we close the gap and explore the potential of the text image foundation model at a large scale. Leveraging the vast amounts of publicly available data, we develop a high-quality data production pipeline that constructs the first token-level image text dataset, named **TokenIT**, comprising 20 million images and 1.8 billion token-mask pairs. Specifically, we begin by extracting text transcriptions and text masks for each sample. Subsequently, we split each text transcription into several tokens (BPE-level subwords) using a tokenizer (Chen et al., 2024d) and obtain their corresponding BPE token masks. The number of token-mask pairs ultimately constructed is 4.5 times that of CLIP and 0.7B more than SAM as summarized in Figure 1.

Leveraging the self-constructed TokenIT dataset, we further propose the first token-level text image foundation model, named **TokenOCR**, designed to support a wide array of text-image-related downstream tasks. To achieve *image-as-text* semantic alignment, token-level visual embeddings are aligned with token-level language embeddings for positive token-mask pairs, meanwhile ensuring that negative pairs remain distinct within the embedding space. Specifically, each token-level visual embedding is derived through a mean pooling operation applied to the visual image features within a corresponding token mask; each token-level language embedding is produced via a straightforward token embedding layer, obviating the need for a complex text encoder like CLIP.

The *image-as-text* semantic attributes, aligned at the VFM level, effectively bridge the gaps between visual and language modalities. This approach creates a unified sequence representation that can be seamlessly integrated into any large language model (LLM) for popular MLLM tasks. Building upon this foundation, we propose a document-level MLLM, named **TokenVL**, which further enhances spatially visual-language token alignment at the LLM level for document understanding in Visual Question Answering (VQA) tasks. Additionally, we freeze the weights of the TokenOCR model to facilitate other downstream applications, including text segmentation, text retrieval, and end-to-end text recognition tasks.

Overall, the main contributions are summarized as follows:

1) The first token-level image text dataset (TokenIT) is proposed, which consists of 20M images and 1.8B high-quality token-mask pairs.

2) The first token-level text image foundation model, TokenOCR, is proposed to support various downstream tasks, including text segmentation, text retrial, and text understanding.

3) The *image-as-text* semantic capability inspires us to develop TokenVL, a VQA-based MLLM tailored for document perception, understanding, and reasoning.

4) Extensive experiments demonstrate the effectiveness of our proposed TokenOCR and TokenVL. Specifically, TokenOCR shows exceptional "zero-shot" capabilities and flexibility compared to other VFMs, such as CLIP, SAM, and InternViT2.5. TokenVL with 8B parameters, incorporating TokenOCR as the VFM, achieves performance gains of 38 on the OCRBench task and an average of 8.8% across ten document VQA tasks. Similarly, TokenVL with 2B parameters results in performance gains of 17 on the OCRBench task and an average of 13.34% on the ten VQA tasks.

2. Related Work

Visual Foundation Models. Visual foundation models (VFMs) are a vitally important component, which serves various downstream tasks, such as semantic segmentation (Shen et al., 2023), optical character recognition (Guan et al., 2022; 2025c;a), object detection (Liu et al., 2025), and remote sensing (Hong et al., 2024). Notably, Radford et al. (Radford et al., 2021) introduce CLIP to align visual and language modalities through contrastive learning from large-scale image-text pairs. SigLIP (Zhai et al., 2023) demonstrate that a simple sigmoid loss can be more effective than a contrastive loss. Caron et al. (Caron et al., 2021) propose DINO, a method for self-supervised learning of image features without labeled data, utilizing self-distillation. However, several studies have observed that these image-level supervised paradigms often encounter basic perceptual errors and fail to capture localized features necessary for dense prediction tasks. Kirillov et al. (Kirillov et al., 2023) introduce the pixel-level SAM, ushering in a new era of segmenting virtually anything. Despite the model's prominence in segmentation tasks, its limited semantic capabilities constrain its applicability to tasks requiring deeper understanding and reasoning. Recently, with the advancement of multimodal large language models (MLLMs) and the trend towards task unification, building more suitable VFMs has become increasingly important.

MLLMs for Document Understanding. Multimodal Large Language Models (MLLMs) connect the powerful Visual Foundation Model and Large Language Model to facilitate perception, understanding, and reasoning, which generate coherent texts through visual question answering. Recent advancements have empowered MLLMs to extract meaningful information from text images for Visual Document Understanding (VDU) tasks. Specifically, these

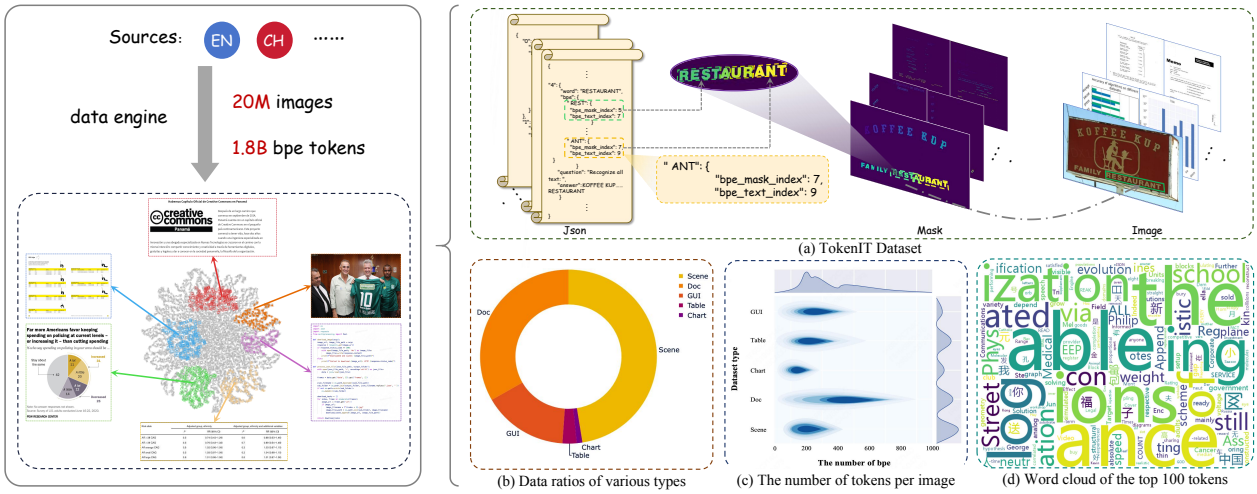


Figure 2: An overview of the self-constructed token-level TokenIT dataset, comprising 20 million images and 1.8 billion text-mask pairs. (a) provides a detailed description of each sample, including the raw image, a mask, and a JSON file that records BPE token information. We also count (b) the data distribution, (c) the number of selected BPE tokens, and (d) a word cloud map highlighting the top 100 BPE tokens.

methods can be roughly categorized into two types: OCR-dependent MLLMs (Lu et al., 2024; Lee et al., 2024; Kim et al., 2023a; Tanaka et al., 2024; Liao et al., 2024a) and OCR-free MLLMs (Chen et al., 2024d; Zhu et al., 2023; Huang et al., 2024b; Li et al., 2024e; Hu et al., 2024e; Ye et al., 2023a). OCR-dependent MLLMs utilize an external OCR engine to extract text information and merge the generated results into MLLMs, which brings excessive auxiliary tokens. In contrast, OCR-free MLLMs have sought to simplify this process by predicting question-driven outputs directly. They incorporate task-specific modules for enhancing the capabilities of Document MLLMs, including high-resolution image processing (Li et al., 2024e; Ye et al., 2023a; Hu et al., 2024c; Feng et al., 2023a), efficient token compression (Zhang et al., 2024d; Hu et al., 2024e; Yu et al., 2024), and refined attention mechanisms (Huang et al., 2024b; Shao et al., 2024). Despite these achievements, existing OCR-free models still struggle to capture fine-grained textual content within images. We speculate that this limitation is caused by the VFMs utilized in large multimodal models. Therefore, we propose the first token-level text image foundation model for visual document understanding tasks. This model aims to bridge the visual-language modality gap by ensuring that the semantic descriptions of each BPE token of visual texts in an image correspond accurately to those of language texts.

3. TokenIT Dataset

In the OCR community, there are almost no datasets of image-text pairs with token granularity, where each language token (split by the BPE tokenizer) aligns precisely with its corresponding image location. However, this type of dataset could effectively enhance the fine-grained perception of VFMs and assist MLLMs in bridging the modality gap between visual and language embeddings. To fill this gap, we curate a Token-level Image Text dataset, TokenIT.

Specifically, to construct a robust and comprehensive TokenIT dataset, we collect various types of data, including natural scene text images, documents (PDF, receipt, letter, note, report, *etc.*), tables, charts, code, and GUI. Finally, three rounds of inspections are conducted to minimize labeling errors, a process that took four months to develop the first token-level image text dataset (TokenIT), which includes 20 million images and 1.8 billion token-mask pairs.

As depicted in Figure 2 (a), each sample in this dataset includes a raw image, a mask image, and a JSON file. The JSON file provides the question-answer pairs and several BPE tokens randomly selected from the answer, along with the ordinal number of each BPE token in the answer and its corresponding pixel value on the mask image. Consequently, each BPE token corresponds one-to-one with a pixel-level mask. The data ratios are summarized in Figure 2 (b). Figure 2 (c) and (d) further provide the number distribution of tokens per image type and a word cloud of the top 100 tokens, respectively. More specific details are introduced in Supplementary Material.

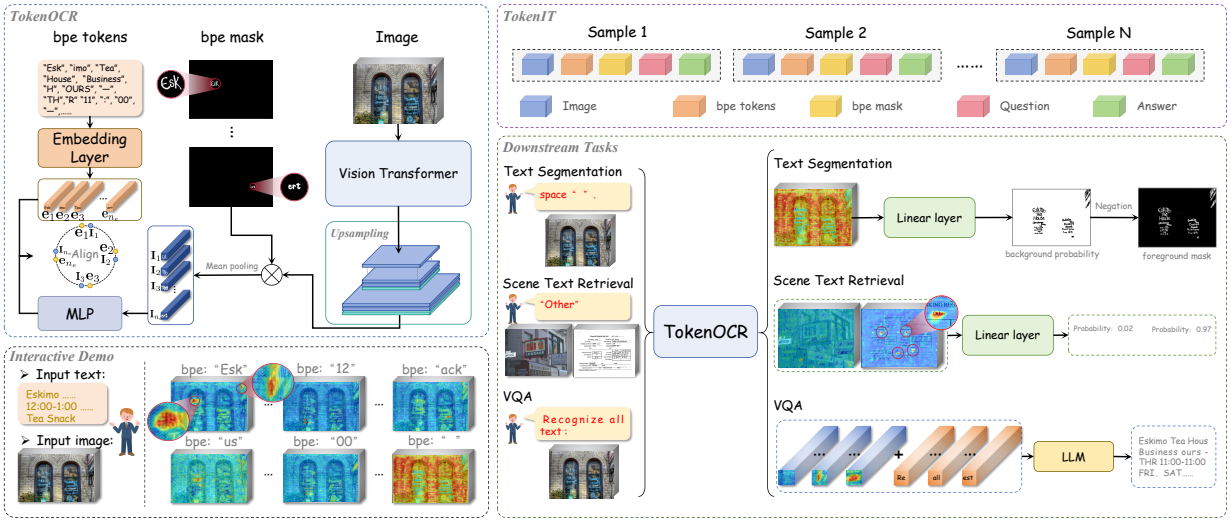


Figure 3: An overview of the proposed TokenOCR, where the token-level image features and token-level language features are aligned within the same semantic space. This "image-as-text" alignment seamlessly facilitates user-interactive applications, including text segmentation, retrieval, and visual question answering.

4. Methodology

Overall. To better describe our method, we define each sample $\mathcal{S}$ of our TokenIT dataset:

$$\begin{cases} \mathcal{S} = \{\mathbf{X}, \mathbf{M}, \mathcal{E}, \mathcal{Q}, \mathcal{A}\}, \\ \mathbf{M} \Rightarrow \{\mathbf{M}_1, ..., \mathbf{M}_{n_e}\}, \\ \mathcal{E} = \{e_1, ..., e_{n_e}\}, \\ \mathcal{Q} = \{q_1, ..., q_{n_q}\}, \\ \mathcal{A} = \{a_1, ..., a_{n_a}\}, \end{cases} \tag{1}$$

where $\mathbf{X}$ is an input image. $\mathcal{Q}$ and $\mathcal{A}$ denote the tokenized question and answer, respectively, processed using a BPE tokenizer (Chen et al., 2024d). $\mathbf{M}$ refers to the mask image, which is divided into n_e BPE token masks $\{\mathbf{M}_1, ..., \mathbf{M}_{n_e}\}$, according to the pixel value (recorded in the JSON file) of each BPE token on the mask image. Consequently, for any BPE token (e_i in $\mathcal{E}$), the pixel value at its specific position in the mask image $\mathbf{M}_i$ is set to 1, with all other positions set to 0. Notably, $\mathcal{E}$ is a subset consisting of n_e BPE tokens, which are randomly selected from $\mathcal{A}$.

Utilizing the TokenIT dataset with 1.8B token-mask pairs, we construct the first token-level OCR foundation model (TokenOCR) by token-level *image-as-text* alignment. For VQA-based document understanding downstream tasks, we employ the well-learned foundation model to construct an MLLM (TokenVL), which includes the following stages: 1) LLM-guided Token Alignment; 2) Supervised Instruction Tuning. Besides, we also freeze the foundation model (unless otherwise stated) to conduct other text-related downstream tasks, including text segmentation, text retrieval, and text understanding.

4.1. TokenOCR

Although existing VFMs produce good representations for zero-shot or fine-tuning tasks, they still encounter significant challenges in processing fine-grained tasks, such as document scenarios with densely packed small texts. Thus a suitable VFM that is tailored for text images is in demand. In light of this, we construct the first token-level VFM, which fills the gap in the field. Concretely, the pre-training process is formulated as follows:

The input image $\mathbf{X} \in \mathbb{R}^{H\times W\times 3}$ is first fed into a ViT-based visual encoder $f(\cdot)$ to extract image features $\mathbf{F} \in \mathbb{R}^{\frac{H}{p}\times\frac{W}{p}\times C}$, where p is the patch size, set to 14 by default. A simple two-layer deconvolution is then applied to the image feature $\mathbf{F}$ to enlarge the feature resolution. Subsequently, a linear layer ($\mathbb{R}^C \rightarrow \mathbb{R}^D$) is applied to expand to the same embedding dimension as the language embedding layer. The processed image feature is denoted as $\tilde{\mathbf{F}} \in \mathbb{R}^{\frac{4\times H}{p}\times\frac{4\times W}{p}\times D}$.

Next, given all BPE token-mask pairs $\mathcal{B} = \{(e_1, \mathbf{M}_1), (e_2, \mathbf{M}_2), ..., (e_{n_e}, \mathbf{M}_{n_e})\}$ corresponding to the input image, the pre-training objective encourages embeddings of matching pairs $\{(\mathbf{e}_1, \mathbf{t}_1), (\mathbf{e}_2, \mathbf{t}_2), ..., (\mathbf{e}_{n_e}, \mathbf{t}_{n_e})\}$ to align with each other, where $\mathbf{e}_i \in \mathbb{R}^D$ is the token embeddings of e_i. The associate token-level visual features $\mathbf{t}_i \in \mathbb{R}^D$ are yielded by a mean-pooling operation:

$$\mathbf{t}_i = \frac{1}{\sum_{x,y} \texttt{BI}(\mathrm{M_i})^{(x,y)}} \sum_{x,y} \texttt{BI}(\mathrm{M_i})^{(x,y)} \tilde{\mathbf{F}}^{(x,y)}, \tag{2}$$

where $\texttt{BI}(\cdot)$ refers to the bilinear interpolation operation to match the feature resolution of $\tilde{\mathbf{F}}$. The coordinate (x, y) indicates a point on the x-axis and y-axis, respectively. Finally, without requiring a complex text encoder like CLIP-Text, we adopt a simple and learnable token embedding layer to align

the visual-language modality at the token level. Specifically, following the previous works (Zhai et al., 2023; Zhang et al., 2023a; Guan et al., 2023b), the objectives are to minimize:

$$\begin{cases} \mathcal{L}_{dis} = \frac{1}{|\mathcal{B}|}\frac{1}{D}\sum_{i=1}^{|\mathcal{B}|}\sum_{j=1}^{D}|e_i^j - \mathrm{t}_i^j|, \\ \mathcal{L}_{sim} = \frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\left(1 - \frac{\mathbf{e}_i \cdot \mathbf{t}_i}{\|\mathbf{e}_i\|\|\mathbf{t}_i\|}\right), \\ \mathcal{L}_{sig} = -\frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\sum_{j=1}^{|\mathcal{B}|}\underbrace{\log\frac{1}{1+e^{z_{ij}(-k\mathbf{e}_i\cdot\mathbf{t}_j+b)}}}_{\mathcal{L}_{sig}^{ij}}, \end{cases} \tag{3}$$

where k and b are learnable parameters, and we initialize k and b to `log 10` and `-10`, respectively. The label z_{ij} indicates whether the token-level visual feature $\mathbf{t}_i$ and token embedding $\mathbf{e}_j$ are a pair, being `1` if they are paired and `-1` otherwise.

After pre-training, the input image's visual embeddings and corresponding text embeddings share the same feature space, achieving *image-as-text* semantic alignment. This alignment facilitates seamless image-text interaction, *i.e.,* inputting text to highlight the corresponding area in the image (as illustrated in the "Interactive Demo" area of Figure 3), along with other derivative downstream tasks. More visualization examples are presented in Supplementary Materials.

4.2. TokenVL

The *image-as-text* semantic attributes inherently bridge the gaps between visual and language modalities, creating a unified sequence representation that LLM can effectively understand. Inspired by this, we employ the TokenOCR as the visual foundation model and further develop an MLLM, named TokenVL, tailored for document understanding. Following the previous training paradigm (Chen et al., 2024d; Lv et al., 2023; Wei et al., 2025; Hu et al., 2024c), TokenVL also includes two stages: 1) LLM-guided Token Alignment Training for text parsing tasks and 2) Supervised Instruction Tuning for VQA tasks.

Specifically, adopting the widely-used multi-scale adaptive cropping strategy (Ye et al., 2023b), the input image $\mathbf{X} \in \mathbb{R}^{H\times W\times 3}$ is initially divided into several non-overlapping sub-images $\{\mathbf{X}_i \in \mathbb{R}^{\iota\times\iota\times 3} | i \in \{1,2,...,N\}\}$. By default, ι is set to 448 and N does not exceed 6. Additionally, the original image $\mathbf{X}$ is resized to a global image $\mathbf{X}_g$ with the same size to preserve the overall layout. Subsequently, our proposed TokenOCR processes these images $\mathcal{X} = \{\mathbf{X}_g, \mathbf{X}_1, ..., \mathbf{X}_N\}$ to produce their corresponding visual embeddings, denoted as $\mathcal{F} = \{\tilde{\mathbf{F}}_i \in \mathbb{R}^{\frac{4\times\iota}{p}\times\frac{4\times\iota}{p}\times D} | i \in \{g,1,2,...,N\}\}$.

After that, for each visual image features $\tilde{\mathbf{F}}_i$ (global image and sub-images), we apply a token abstractor $\xi : \mathbb{R}^{\frac{4\times\iota}{p}\times\frac{4\times\iota}{p}\times D} \rightarrow \mathbb{R}^{\frac{\iota}{p\times\frac{s}{4}}\times\frac{\iota}{p\times\frac{s}{4}}\times D}$ to adaptively extract a meaningful visual embedding within each window of shape $s \times s$, where s is set to 4 in our experiment. Specifically, in addition to the original dictionary of the tokenizer, we define a special token `<text>` to obtain a learnable token embedding $\mathbf{e}_s \in \mathbb{R}^{1\times 1\times D}$. Benefiting from the priors of the TokenOCR, the special token embedding can easily learn robust representations to identify the most suitable visual embeddings within each window. Concretely, for each sub-image and global image, we first re-organize the shape of its visual embeddings $\tilde{\mathbf{F}}_i$ from $\frac{4\times\iota}{p} \times \frac{4\times\iota}{p} \times D$ to $(\frac{\iota}{p\times\frac{s}{4}})^2 \times D \times s^2$. $\xi(\cdot)$ is then implemented as follows:

$$\begin{cases} \alpha_i = \mathtt{softmax}(\mathbf{e}_s\tilde{\mathbf{F}}_i), \alpha_i \in \mathbb{R}^{(\frac{\iota}{p\times\frac{s}{4}})^2\times 1\times s^2} \\ \mathring{\mathbf{F}}_i = \mathtt{sum}(\alpha_i \circ \tilde{\mathbf{F}}_i), \mathring{\mathbf{F}}_i \in \mathbb{R}^{\frac{\iota}{p\times\frac{s}{4}}\times\frac{\iota}{p\times\frac{s}{4}}\times D} \end{cases} \tag{4}$$

where the `softmax` and `sum` operations are conducted on the last dimension. $\circ$ denotes the Hadamard product. After the token abstractor, we flatten these compressed features $\{\mathring{\mathbf{F}}_g, \mathring{\mathbf{F}}_1, ..., \mathring{\mathbf{F}}_N\}$ to get the final visual embeddings $\mathcal{V} = \{\mathbf{v}_1, ..., \mathbf{v}_{n_v}\}$, which will be fed into LLM. Here, $n_v = \frac{\iota}{p\times\frac{s}{4}} \times \frac{\iota}{p\times\frac{s}{4}} \times (N+1)$ denotes the number of image tokens.

1) LLM-guided Token Alignment Training.

In the pre-training stage, we use the compressed visual embeddings $\mathcal{V}$ as the visual inputs, and $\mathcal{Q}$ and $\mathcal{A}$ from the Eq.1 as the language inputs to simultaneously conduct VQA-based text parsing tasks (*implicitly semantic alignment*) and token alignment (*explicitly spatial alignment*) tasks, as illustrated in Figure 4. It is important to note that the Token Alignment (TA) branch is just introduced during LLM-guided Token Alignment Training, as all answers appear directly in the image.

Text parsing tasks include recognizing full text, recognizing partial text within localization, visual text grounding, converting formulas into LaTeX, converting tables into markdown or LaTeX, and converting charts into CSV or markdown formats. More specific details are introduced in Supplementary Materials. Concretely, the visual and language inputs are concatenated together to be fed into the LLM, which predicts answers step-by-step by $\mathtt{LLM}([\mathcal{V}_{1:n_v}; \mathcal{Q}_{1:n_q}; \mathcal{A}_{1:m-1}]), \forall m \in \{2,...,n_a\}$. The cross-entropy loss is formulated as:

$$\mathcal{L}_{cel} = -\sum_{m=2}^{n_a}\sum_{i=1}^{N}\mathbf{a}_m \log \hat{\mathbf{a}}_m, \tag{5}$$

where $\hat{\mathbf{a}}_m \in \mathbb{R}^Z$ refers to the probability distribution, $\mathbf{a}_m$ is the one-hot vector of a_m, and Z denotes the dictionary size of the tokenizer.

The auto-regressive training task above allows only language inputs to implicitly interact with visual inputs. Without explicitly spatially-aware supervision, the outputs may

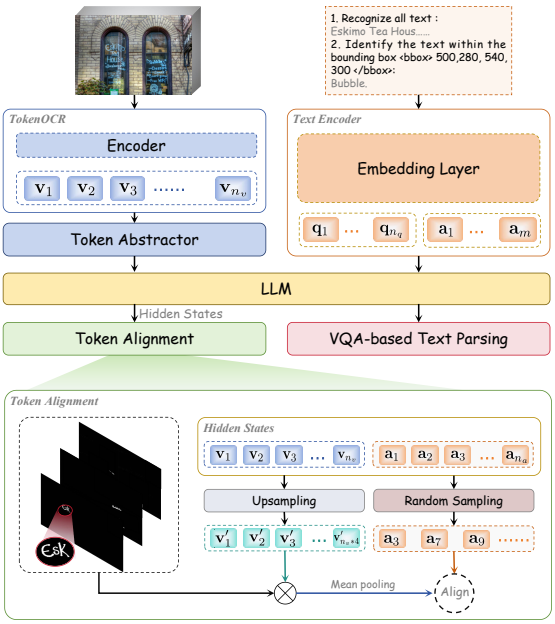


Figure 4: The framework of LLM-guided Token Alignment Training. Existing MLLMs primarily enhance spatial-wise text perception capabilities by integrating localization prompts to predict coordinates. However, this implicit method makes it difficult for these models to have a precise understanding. In contrast, the proposed token alignment uses BPE token masks to directly and explicitly align text with corresponding pixels in the input image, enhancing the MLLM's localization awareness.

depend more on the LLM's robust semantic context capabilities rather than the VFM's image feature representations. To explicitly facilitate spatial-wise visual-language alignment at the LLM level, we conduct a fine-grained alignment task with token granularity by leveraging the BPE token-mask pairs $\{(e_1, \mathbf{M}_1), (e_2, \mathbf{M}_2), ..., (e_{n_e}, \mathbf{M}_{n_e})\}$. Specifically, given that the outputs of the k-th hidden layer of the LLM as $\{\mathcal{V}^k, \mathcal{Q}^k, \mathcal{A}^k\} = \{\mathbf{v}_1^k, ..., \mathbf{v}_{n_v}^k, \mathbf{q}_1^k, ..., \mathbf{q}_{n_q}^k, \mathbf{a}_1^k, ..., \mathbf{a}_{n_a}^k\}$, we extract the visual features and language features corresponding to each BPE token. Taking the BPE token e_i as an example, we first compute its index location in $\{\mathcal{V}^k, \mathcal{Q}^k, \mathcal{A}^k\}$ as $|\mathcal{V}^k| + |\mathcal{Q}^k| + \zeta(e_i, \mathcal{A})$, where $\zeta(e_i, \mathcal{A})$ finds the position of e_i in $\mathcal{A}$ according to the relation $e_i \in \mathcal{E}$ and $\mathcal{E} \in \mathcal{A}$. For easy reference, the position has been recorded in our JSON file, which corresponds to the value for the keyword `index_in_text`. Consequently, the selected language features can be easily obtained through indexing operations. Then, to extract the selected visual features corresponding to the BPE token e_i, we exclude the global visual features (global image) and reorganize the remaining visual features (all sub-images) in $\mathcal{V}^k$ to recover a complete feature map, denoted as $\mathbf{F}^k$. A mean-pooling operation yields the associated token-level visual features:

$$\texttt{average}(\mathbf{M}_i \circ \texttt{BI}(\mathbf{F}^k)), \tag{6}$$

where $\texttt{BI}(\cdot)$ refers to the bilinear interpolation operation to match the feature resolution of $\mathbf{M}_i$. `average` means performing a global pooling operation on the features. Finally, the visual-language modality at the token level is aligned by minimizing the objectives following Eq.3.

Building on this, we assist the LLM in achieving fine-grained semantic perception for document understanding. This enables the visual semantics of each image patch with text to be consistent with the language semantics of its corresponding BPE token at the LLM level.

2) Supervised Instruction Tuning.

Following the final stage of the previous MLLMs, we collect the existing VQA datasets to conduct supervised instruction tuning. These datasets cover a wide range of scenarios, including Documents (DocVQA, InfoVQA, DeepForm, KLC, DocMatix, AI2D, KIE, DocReason25K), Tables (TabFact, WTQ, TableBench, TabMWP, TableVQA), Charts (ChartQA, FigureQA, DVQA, PlotQA, UniChart, GeoQA+, Sujet-Finance), Formulas (UniMER, HME100k), and Scene Texts (TextVQA, ST-VQA, OCR-VQA, IAM, EST-VQA, SynthDoG).

During the Supervised Instruction Tuning stage, we cancel the token alignment branch as answers may not appear in the image for some reasoning tasks (*e.g., How much taller is the red bar compared to the green bar?*). This also ensures no computational overhead during inference to improve the document understanding capability. Finally, we inherit the remaining weights from the LLM-guided Token Alignment and unfreeze all parameters to facilitate comprehensive parameter updates.

5. Experiments

Implementation Details. To pre-train the TokenOCR foundation model, we employ the AdamW optimizer alongside a cosine learning rate schedule, with a base learning rate set at 5e-4. The model undergoes pre-training for two epochs on the TokenIT dataset. For TokenVL, we leverage the well-trained TokenOCR as the visual foundation model and InternLM (Cai et al., 2024) as the language model. Specifically, during the LLM-guided token alignment stage, InternLM remains frozen while we train the TokenOCR and newly introduced token abstractor. This stage involves training for one epoch on the TokenIT dataset, utilizing a base learning rate of 2e-4. In the subsequent supervised instruction tuning stage, all parameters are fully trainable, with a base learning rate of 1e-5. These experiments are executed on 64 H800 GPUs. Additional implementation details about other downstream experiments are provided within each respective subtask and **Supplementary Material**.

Tasks	Method	#Param	TextSeg	TotalText	HierText	average
ZS	CLIP-L-336px	304M	19.71	13.56	13.39	15.55
	CLIP-L-448px	304M	20.50	13.91	13.19	15.86
	CLIP-L-1024px	304M	21.35	14.33	11.77	15.81
	TokenOCR-448px	323M	38.27	33.10	26.46	32.61
	TokenOCR-1024px	323M	**38.28**	**33.54**	**31.95**	**34.59**
LP	SAM-H	632M	40.82	36.83	25.87	34.51
	InternViT2.5	300M	49.77	42.54	34.31	42.21
	TokenOCR	323M	**55.66**	**47.53**	**43.11**	**48.77**

Table 1: Text segmentation experiments of various visual foundation models. "ZS" refers to the zero-shot experiment. "LP" denotes the linear probe experiment.

Method	#Param	DocVQA	InfoVQA	TextVQA	ChartQA	average
SAM-H	632M	17.0	23.1	33.1	30.1	25.82
CLIP-L	304M	64.9	38.6	80.7	65.2	62.36
InternViT2.5	300M	77.3	49.3	84.4	74.0	71.25
TokenOCR	323M	**78.9**	**50.0**	**85.6**	**74.4**	**72.21**

Table 2: The ANLS results of various visual foundation models on VQA tasks.

5.1. Effectiveness of TokenOCR

Our work focuses on developing a high-performing dataset-agnostic foundation model. Fine-tuning, because it adapts representations to each dataset during the fine-tuning phase, can compensate for and potentially mask failures to learn general and robust representations. As a result, employing zero-shot transfer or fitting a linear classifier on representations extracted from the model, and then measuring its performance across various datasets, is a common approach (Radford et al., 2021). This method provides a clearer assessment of VFMs' ability to generalize without relying on dataset-specific tuning.

Text Segmentation: 1) Zero-shot Segmentation: We compute the similarity between visual and language features to get the segmentation results. For CLIP, in line with prior work, we select "text" as the language prompt, which has been proven to be the most effective (Yu et al., 2023). In our method, we use a space " " as the language prompt and then apply a negation operation to derive the foreground similarity map. 2) Linear Probe: We keep the VFM frozen and train a linear layer to perform segmentation. Based on the results shown in Table 1, TokenOCR demonstrates significant average performance improvement across various text segmentation tasks. In the zero-shot setting, TokenOCR-1024px achieves the highest average score of 34.59%, significantly outperforming CLIP-L by 18.78%. In the linear probe setting, TokenOCR again leads with an average score of 48.77%, showing considerable improvements over SAM-H and InternViT2.5.

Visual Question Answering: To further explore the representation learning capabilities of VFMs, we keep them frozen and fine-tune Vicuna-7B (Zheng et al., 2023) as the language model to conduct the text-related VQA tasks. All comparison methods employ the same configuration—training data, test benchmarks, learnable parameters, and optimizer—to ensure a fair evaluation. As seen in Table 2, TokenOCR achieves the highest scores on popular benchmarks, outperforming SAM-H, CLIP-L, and InternViT2.5 by 46.39%, 9.85%, and 0.96%, respectively.

Tasks	Methods	#Param	CTR (EN)	CSVTRv2 (CH)	average
LP	CLIP-L	304M	1.21	6.03	3.62
	InternViT2.5	300M	4.21	22.37	13.29
	TokenOCR	323M	**43.04**	**84.19**	**63.62**

Table 3: Linear probe experiments of various VFMs on text retrieval tasks. All VFMs are frozen.

Text Retrieval: We select representative models, CLIP and InternViT2.5, to compare with our proposed TokenOCR on a Chinese dataset and an English dataset. Specifically, all VFMs are frozen. We calculate the similarity maps between the visual embeddings (extracted from the VFM) of all retrieval images and the language embeddings of all queries. For linear probe experiments, we use the same training data and train a simple linear classifier to score each similarity map, assigning a 1 if the similarity score is greater than 0.5, and a 0 otherwise. Finally, mean Average Precision (mAP) is employed to evaluate the performance of each VFM. The comparison results show that using only a few parameters, TokenOCR can perform well. Specifically, our proposed TokenOCR can achieve an average score of 63.62% on bilingual tasks. Additionally, there remains significant room for improvement through specific designs and components in the future.

5.2. Effectiveness of TokenVL

OCRBench results: OCRBench is a widely recognized and comprehensive benchmark comprising 29 tasks, commonly utilized to assess the OCR capabilities of MLLMs. As illustrated in Table 4, we compare the performance of our TokenVL against previously existing MLLMs. TokenVL achieves the highest score of 860 among the 8B-Model group, significantly outperforming models like general-MLLM InternVL2.5 (↑ 38) and expert TextHawk2 (↑ 76). In the 2B-Model group, our method achieves the top score of 821, surpassing competitors such as MiniMonkey (↑ 19) and InternVL2.5 (↑ 17).

Document Benchmarks results: To demonstrate the perception, understanding, and reasoning capabilities of our TokenVL, we collect existing evaluation benchmarks across five categories: Document, Chart, Natural Scene, Table, and KIE. The results, presented in Table 5, show a consistent and significant outperformance over other 8B MLLMs. Specifically, for widely used evaluation benchmarks (Doc/Info/Chart/TextVQA), TokenVL-2B achieves an average gain of 2.18% and 1.33% over MiniMonkey and InternVL2.5, respectively. TokenVL-8B obtains gains

8B-Model	ShareGPT4V	Cambrian	MM1.5	POINT1.5	GPT-4o	Gemini-1.5-Pro	GLM-4v	Claude3.5	InternVL2.5
Score	398	614	635	720	736	754	776	788	822

8B-Model	TextMonkey	DocOwl-1.5	TextHawk2	TokenVL(ours)	**2B-Model**	MiniMonkey	InternVL2.5	TokenVL(ours)
Score	561	599	784	**860**	Score	802	804	**821**

Table 4: Comparison results of our TokenVL with other MLLMs on the OCRbench benchmark.

Model	size	Venue	DocVQA	InfoVQA	DeepForm	ChartQA	$\text{TextVQA}_{\text{Val}}$	WTQ	TabFact	FUNSD	SROIE	KLC
MiniCPM-V	3B	COLM'24	71.9	-	-	55.6	74.1	-	-	-	-	
Mini-Monkey	2B	ICLR'25	87.4	60.1	-	76.5	75.7	-	-	42.9	70.3	-
InternVL2.5	2B	arxiv'24	88.7	60.9	15.2	79.2	74.3	38.7	58.1	37.9	68.1	16.1
TokenVL	2B	-	**89.9**	**61.0**	**71.9**	**81.1**	**76.4**	**49.0**	**76.9**	**43.0**	**82.6**	**38.8**
Claude-3.5 Sonnet	Closed-source model		88.5	59.1	31.4	51.8	71.4	47.1	53.5	-	-	24.8
GeminiPro-1.5	Closed-source model		91.2	73.9	32.2	34.7	80.4	50.3	71.2	-	-	24.1
GPT4o 20240806	Closed-source model		92.8	66.4	38.4	85.7	70.5	46.6	81.1	-	-	29.9
DocPeida	7B	arxiv'23	47.1	15.2	-	46.9	60.2	-	-	29.9	21.4	-
DocOwl	7B	arxiv'23	62.2	38.2	42.6	57.4	52.6	26.9	67.6	0.5	1.7	30.3
LLaVA1.5	7B	NeurIPS'23	-	-	-	9.3	-	-	-	0.2	1.7	-
UReader	7B	EMNLP'23	65.4	42.2	49.5	59.3	57.6	29.4	67.6	-	-	32.8
CHOPINLLM	7B	arxiv'24	-	-	-	70.0	-	-	-	-	-	-
TextHawk	7B	arxiv'24	76.4	50.6	-	66.6	-	34.7	71.1	-	-	-
DocKylin	7B	arxiv'24	77.3	46.6	-	66.8	-	32.4	-	-	-	-
MM1.5	7B	arxiv'24	88.1	59.5	-	78.6	76.8	46.0	75.9	-	-	-
DocOwl-1.5	8B	EMNLP'24	81.6	50.4	68.8	70.5	68.8	39.8	80.4	-	-	37.9
DocOwl-1.5-Chat	8B	EMNLP'24	82.2	50.7	68.8	70.2	68.6	40.6	80.2	-	-	38.7
CogAgent	17B	CVPR'24	81.6	44.5	-	68.4	76.1	-	-	-	-	-
Monkey	10B	CVPR'24	66.5	36.1	40.6	65.1	67.6	25.3	-	-	-	-
TextMonkey	8B	arxiv'24	73.0	28.6	-	66.9	65.6	-	-	32.3	47.0	-
HRVDA	7B	CVPR'24	72.1	43.5	63.2	67.6	73.3	31.2	72.3	-	-	37.5
InternVL2	8B	CVPR'24	91.6	74.8	-	-	77.4	-	-	-	-	-
Park et al.	7B	NeurIPS'24	72.7	45.9	53.0	36.7	59.2	34.5	68.2	-	-	36.7
MOAI	7B	ECCV'24	-	-	-	-	67.8	-	-	-	-	-
Vary	7B	ECCV'24	76.3	-	-	66.1	-	-	-	-	-	-
TextHawk2	7B	arxiv'24	89.6	67.8	-	81.4	75.1	46.2	78.1	-	-	-
PDF-WuKong	9B	arxiv'24	76.9	-	-	-	-	-	-	-	-	-
LLaVA-NEXT-7B	7B	arxiv'24	63.5	30.9	1.3	52.1	65.1	20.1	52.8	-	-	5.35
LLama3.2-11B	11B	arxiv'24	82.7	36.6	1.78	23.8	54.3	23.0	58.3	-	-	3.47
Pixtral-12B	12B	arxiv'24	87.7	49.5	27.4	71.8	76.1	45.2	73.5	-	-	24.1
Ovis	9B	arxiv'24	88.8	74.0	45.2	81.4	77.7	50.7	76.7	-	-	23.9
InternVL2.5	8B	arxiv'24	93.0	**77.6**	37.9	84.8	79.1	52.7	74.8	38.26	71.7	22.9
AlignVLM	8B	arxiv'25	81.2	53.8	63.3	75.0	64.6	45.3	83.0	-	-	35.5
TokenVL w/o TA	8B	-	93.8	75.3	72.4	86.5	79.3	57.2	83.6	41.5	79.0	39.6
TokenVL	8B	-	**94.2**	76.5	**72.9**	**86.6**	**79.9**	**61.4**	**85.2**	**42.2**	**81.9**	**39.9**

Table 5: Comparisons on various types of text-rich image understanding tasks. All evaluation benchmarks use the officially designated metrics. "size" refers to the number of parameters in the model, and "Val" refers to the validation set.

Method	TotalText (↓)	IC15 (↓)	IIT (↓)	Docgenome (↓)
w/o token alignment	0.3592	0.2388	0.2388	0.2837
w token alignment	**0.3547**	**0.2324**	**0.1921**	**0.2806**

Table 6: Edit distance for full-image text recognition.

Abstractor	Alignment	DocVQA	InfoVQA	ChartVQA	$\text{TextVQA}_{\text{Val}}$
×	×	93.1	74.7	86.5	79.1
✓	×	93.8	75.3	86.5	79.3
✓	✓	**94.2**	**76.5**	**86.6**	**79.9**

Table 7: Comparison experiments on the VQA tasks.

of 1.2%, 1.8%, and 0.8% on DocVQA, ChartQA, and TextVQA compared to the previous SOTA InternVL2.5. Additionally, TokenVL achieves a larger performance gain on other benchmarks while maintaining these properties.

5.3. Ablation Study

w/o token alignment. Token alignment at the LLM level explicitly facilitates interaction between image embeddings and language embeddings. This method encourages the LLM to reference image content more directly when responding to questions, rather than relying solely on its powerful semantic context capabilities. To verify the effectiveness of this strategy: 1) we perform a text recognition experiment of full-text images, which predicts all texts within a given image from top to bottom and left to right. As shown in Table 6, without fine-tuning on downstream text data, we directly evaluate our model's performance with and without Token Alignment, using document scenes (1000 images extracted from IIT-CDIP and DocGenome respectively) and natural scenes (ICDAR15 and TotalText). Specifically, given

the question "recognize all texts in the image" for MLLMs, we calculate edit distance by comparing the model's outputs with the ground truth answers sorted by spatial position. It was observed that token alignment significantly improves text recognition performance on full images. 2) we also evaluate the final VQA performance of the MLLM on four widely used evaluation benchmarks, both with and without Token Alignment, referring to the last two group results of Table 7. As a result, an average gain of 0.6% is obtained. More details are provided in Supplementary Material.

w/o token abstractor. To reduce the spatial dimensions, we designed a learnable token embedding vector to adaptively capture useful visual information. Without the token abstractor, we use a simple pooling layer instead. The ablation results are shown in the top two groups of Table 7, where an average gain of 0.3% is obtained, even though the token abstractor is not our main contribution.

6. Conclusion

In the paper, we take a step towards constructing a fine-grained visual foundation model, and propose a series of token-level product families: TokenIT, TokenOCR, and TokenVL. We also explore the potential and effectiveness of TokenOCR and TokenVL at a sufficiently large scale. While this approach demonstrates good and consistent performance gains on downstream tasks, there remains significant room for improvement through effective training strategies or additional designs. Therefore, we hope these products will serve as easily reproducible baselines for more downstream tasks in the future.

References

Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., and Zhou, J. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

Biten, A. F., Tito, R., Mafla, A., Gomez, L., Rusinol, M., Valveny, E., Jawahar, C., and Karatzas, D. Scene text visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4291–4301, 2019.

Cai, Z., Cao, M., Chen, H., Chen, K., Chen, K., Chen, X., Chen, X., Chen, Z., Chen, Z., and et al. Internlm2 technical report, 2024.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. Emerging properties in self-supervised vision transformers. In *ICCV*, pp. 9650–9660, 2021.

Chen, W., Wang, H., Chen, J., Zhang, Y., Wang, H., Li, S., Zhou, X., and Wang, W. Y. Tabfact: A large-scale dataset for table-based fact verification. *arXiv preprint arXiv:1909.02164*, 2019.

Chen, X., Djolonga, J., Padlewski, P., Mustafa, B., Changpinyo, S., Wu, J., Ruiz, C. R., Goodman, S., Wang, X., Tay, Y., et al. Pali-x: On scaling up a multilingual vision and language model. *arXiv preprint arXiv:2305.18565*, 2023.

Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024a.

Chen, Z., Wang, W., and et al. Internvl2: Better than the best—expanding performance boundaries of open-source multimodal models with the progressive scaling strategy. 2024b. URL https://internvl.github.io/blog/2024-07-02-InternVL-2.0/.

Chen, Z., Wang, W., Tian, H., Ye, S., Gao, Z., Cui, E., Tong, W., Hu, K., Luo, J., Ma, Z., et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024c.

Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., 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*, pp. 24185–24198, 2024d.

Cheng, K., Sun, Q., Chu, Y., Xu, F., Li, Y., Zhang, J., and Wu, Z. Seeclick: Harnessing gui grounding for advanced visual gui agents. *arXiv preprint arXiv:2401.10935*, 2024.

Cheng, Z., Bai, F., Xu, Y., Zheng, G., Pu, S., and Zhou, S. Focusing attention: Towards accurate text recognition in natural images. In *Proceedings of the IEEE international conference on computer vision*, pp. 5076–5084, 2017.

Ch'ng, C. K. and Chan, C. S. Total-text: A comprehensive dataset for scene text detection and recognition. In *2017 14th IAPR international conference on document analysis and recognition (ICDAR)*, volume 1, pp. 935–942. IEEE, 2017.

Chng, C. K., Liu, Y., Sun, Y., Ng, C. C., Luo, C., Ni, Z., Fang, C., Zhang, S., Han, J., Ding, E., et al. Icdar2019 robust reading challenge on arbitrary-shaped text-rrc-art. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1571–1576. IEEE, 2019.

Co., L. B. A. Z. T. Chinese ocr. 2024. URL https://huggingface.co/datasets/longmaodata/Chinese-OCR.

Covert, I., Sun, T., Zou, J., and Hashimoto, T. Locality alignment improves vision-language models. *arXiv preprint arXiv:2410.11087*, 2024.

Deng, X., Sun, H., Lees, A., Wu, Y., and Yu, C. Turl: Table understanding through representation learning. *ACM SIGMOD Record*, 51(1):33–40, 2022.

Deng, X., Gu, Y., Zheng, B., Chen, S., Stevens, S., Wang, B., Sun, H., and Su, Y. Mind2web: Towards a generalist agent for the web. *Advances in Neural Information Processing Systems*, 36, 2024.

Fan, X., Ji, T., Jiang, C., Li, S., Jin, S., Song, S., Wang, J., Hong, B., Chen, L., Zheng, G., et al. Mousi: Poly-visual-expert vision-language models. *arXiv preprint arXiv:2401.17221*, 2024.

Feng, H., Liu, Q., Liu, H., Zhou, W., Li, H., and Huang, C. Docpedia: Unleashing the power of large multimodal model in the frequency domain for versatile document understanding. *arXiv preprint arXiv:2311.11810*, 2023a.

Feng, H., Wang, Z., Tang, J., Lu, J., Zhou, W., Li, H., and Huang, C. Unidoc: A universal large multimodal model for simultaneous text detection, recognition, spotting and understanding. *arXiv preprint arXiv:2308.11592*, 2023b.

Feng, H., Liu, Q., Liu, H., Tang, J., Zhou, W., Li, H., and Huang, C. DocPedia: unleashing the power of large multimodal model in the frequency domain for versatile document understanding. *arXiv*, 2311.11810, 2024.

Gu, J., Meng, X., Lu, G., Hou, L., Minzhe, N., Liang, X., Yao, L., Huang, R., Zhang, W., Jiang, X., et al. Wukong: A 100 million large-scale chinese cross-modal pre-training benchmark. *Advances in Neural Information Processing Systems*, 35:26418–26431, 2022.

Guan, T., Gu, C., Lu, C., Tu, J., Feng, Q., Wu, K., and Guan, X. Industrial scene text detection with refined feature-attentive network. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(9):6073–6085, 2022.

Guan, T., Shen, W., Yang, X., Feng, Q., Jiang, Z., and Yang, X. Self-supervised character-to-character distillation for text recognition. In *ICCV*, pp. 19473–19484, 2023a.

Guan, T., Shen, W., Yang, X., Feng, Q., Jiang, Z., and Yang, X. Self-supervised character-to-character distillation for text recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19473–19484, 2023b.

Guan, T., Lin, C., Shen, W., and Yang, X. Posformer: recognizing complex handwritten mathematical expression with position forest transformer. In *European Conference on Computer Vision*, pp. 130–147. Springer, 2025a.

Guan, T., Shen, W., and Yang, X. Ccdplus: Towards accurate character to character distillation for text recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025b.

Guan, T., Shen, W., Yang, X., Wang, X., and Yang, X. Bridging synthetic and real worlds for pre-training scene text detectors. In *European Conference on Computer Vision*, pp. 428–446. Springer, 2025c.

Harley, A. W., Ufkes, A., and Derpanis, K. G. Evaluation of deep convolutional nets for document image classification and retrieval. In *International Conference on Document Analysis and Recognition (ICDAR)*.

He, M., Liu, Y., Yang, Z., Zhang, S., Luo, C., Gao, F., Zheng, Q., Wang, Y., Zhang, X., and Jin, L. Icpr2018 contest on robust reading for multi-type web images. In *2018 24th international conference on pattern recognition (ICPR)*, pp. 7–12. IEEE, 2018.

Hong, D., Zhang, B., Li, X., Li, Y., Li, C., Yao, J., Yokoya, N., Li, H., Ghamisi, P., Jia, X., et al. Spectralgpt: Spectral remote sensing foundation model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Hong, W., Wang, W., Lv, Q., Xu, J., Yu, W., Ji, J., Wang, Y., Wang, Z., Dong, Y., Ding, M., and Tang, J. Cogagent: A visual language model for gui agents, 2023.

Hu, A., Xu, H., Ye, J., Yan, M., Zhang, L., Zhang, B., Li, C., Zhang, J., Jin, Q., Huang, F., and Zhou, J. mPLUG-DocOwl 1.5:unified structure learning for OCR-free document understanding. *arXiv*, 2403.12895, 2024a.

Hu, A., Xu, H., Ye, J., Yan, M., Zhang, L., Zhang, B., Li, C., Zhang, J., Jin, Q., Huang, F., et al. mplug-docowl 1.5: Unified structure learning for ocr-free document understanding. *arXiv preprint arXiv:2403.12895*, 2024b.

Hu, A., Xu, H., Ye, J., Yan, M., Zhang, L., Zhang, B., Li, C., Zhang, J., Jin, Q., Huang, F., et al. mplug-docowl 1.5: Unified structure learning for ocr-free document understanding. *arXiv preprint arXiv:2403.12895*, 2024c.

Hu, A., Xu, H., Zhang, L., Ye, J., Yan, M., Zhang, J., Jin, Q., Huang, F., and Zhou, J. mPLUG-DocOwl2: high-resolution compressing for OCR-free multi-page document understanding. *arXiv*, 2409.03420, 2024d.

Hu, A., Xu, H., Zhang, L., Ye, J., Yan, M., Zhang, J., Jin, Q., Huang, F., and Zhou, J. mplug-docowl2: High-resolution compressing for ocr-free multi-page document understanding. *arXiv preprint arXiv:2409.03420*, 2024e.

Huang, M., Liu, Y., Liang, D., Jin, L., and Bai, X. Mini-monkey:alleviating the semantic sawtooth effect for lightweight MLLMs via complementary image pyramid. *arXiv*, 2408.02034, 2024a.

Huang, M., Liu, Y., Liang, D., Jin, L., and Bai, X. Mini-monkey: Alleviate the sawtooth effect by multi-scale adaptive cropping. *arXiv preprint arXiv:2408.02034*, 2024b.

Kafle, K., Price, B., Cohen, S., and Kanan, C. Dvqa: Understanding data visualizations via question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5648–5656, 2018.

Kahou, S. E., Michalski, V., Atkinson, A., Kádár, Á., Trischler, A., and Bengio, Y. Figureqa: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300*, 2017.

Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., and Wu, A. Y. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE TPAMI*, 24(7):881–892, 2002.

Kapoor, R., Butala, Y. P., Russak, M., Koh, J. Y., Kamble, K., AlShikh, W., and Salakhutdinov, R. Omniact: A dataset and benchmark for enabling multimodal generalist autonomous agents for desktop and web. In *European Conference on Computer Vision*, pp. 161–178. Springer, 2024.

Karatzas, D., Shafait, F., Uchida, S., Iwamura, M., i Bigorda, L. G., Mestre, S. R., Mas, J., Mota, D. F., Almazan, J. A., and De Las Heras, L. P. Icdar 2013 robust reading competition. In *2013 12th international conference on document analysis and recognition*, pp. 1484–1493. IEEE, 2013.

Kareem, S. A., Pozos-Parra, P., and Wilson, N. An application of belief merging for the diagnosis of oral cancer. *Applied Soft Computing*, 61:1105–1112, 2017.

Kim, G., Hong, T., Yim, M., Nam, J., Park, J., Yim, J., Hwang, W., Yun, S., Han, D., and Park, S. Ocr-free document understanding transformer. In *European Conference on Computer Vision*, pp. 498–517. Springer, 2022.

Kim, G., Lee, H., Kim, D., Jung, H., Park, S., Kim, Y., Yun, S., Kil, T., Lee, B., and Park, S. Visually-situated natural language understanding with contrastive reading model and frozen large language models. *arXiv preprint arXiv:2305.15080*, 2023a.

Kim, G., Lee, H., Kim, D., Jung, H., Park, S., Kim, Y., Yun, S., Kil, T., Lee, B., and Park, S. Visually-situated natural language understanding with contrastive reading model and frozen large language models. *arXiv preprint arXiv:2305.15080*, 2023b.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al. Segment anything. In *ICCV*, pp. 4015–4026, 2023.

Krylov, I., Nosov, S., and Sovrasov, V. Open images v5 text annotation and yet another mask text spotter. In *Asian Conference on Machine Learning*, pp. 379–389. PMLR, 2021.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Laurençon, H., Marafioti, A., Sanh, V., and Tronchon, L. Building and better understanding vision-language models: insights and future directions. In *Workshop on Responsibly Building the Next Generation of Multimodal Foundational Models*, 2024a.

Laurençon, H., Tronchon, L., and Sanh, V. Unlocking the conversion of web screenshots into html code with the websight dataset. *arXiv preprint arXiv:2403.09029*, 2024b.

Laurençon, H., Tronchon, L., Cord, M., and Sanh, V. What matters when building vision-language models?, 2024a.

Laurençon, H., Tronchon, L., Cord, M., and Sanh, V. What matters when building vision-language models?, 2024b. URL https://arxiv.org/abs/2405.02246.

Lee, B.-K., Park, B., Kim, C. W., and Ro, Y. M. Moai: Mixture of all intelligence for large language and vision models. *ECCV*, 2024.

Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., and Li, C. Llava-onevision: Easy visual task transfer, 2024a. URL https://arxiv.org/abs/2408.03326.

Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024b.

Li, W., Yuan, Y., Liu, J., Tang, D., Wang, S., Qin, J., Zhu, J., and Zhang, L. Tokenpacker: Efficient visual projector for multimodal llm, 2024c. URL https://arxiv.org/abs/2407.02392.

Li, Z., Yang, B., Liu, Q., Ma, Z., Zhang, S., Yang, J., Sun, Y., Liu, Y., and Bai, X. Monkey:image resolution and text label are important things for large multi-modal models. In *Computer Vision and Pattern Recognition (CVPR)*, 2024d.

Li, Z., Yang, B., Liu, Q., Ma, Z., Zhang, S., Yang, J., Sun, Y., Liu, Y., and Bai, X. Monkey: Image resolution and text label are important things for large multi-modal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26763–26773, 2024e.

Liao, W., Wang, J., Li, H., Wang, C., Huang, J., and Jin, L. Doclayllm: An efficient and effective multi-modal extension of large language models for text-rich document understanding. *arXiv preprint arXiv:2408.15045*, 2024a.

Liao, W., Wang, J., Li, H., Wang, C., Huang, J., and Jin, L. Doclayllm: An efficient and effective multi-modal extension of large language models for text-rich document understanding. *arXiv preprint arXiv:2408.15045*, 2024b.

Lin, Z., Liu, C., Zhang, R., Gao, P., Qiu, L., Xiao, H., Qiu, H., Lin, C., Shao, W., Chen, K., et al. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *arXiv preprint arXiv:2311.07575*, 2023a.

Lin, Z., Liu, C., Zhang, R., Gao, P., Qiu, L., Xiao, H., Qiu, H., Lin, C., Shao, W., Chen, K., et al. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *arXiv preprint arXiv:2311.07575*, 2023b.

Liu, C., Yin, K., Cao, H., Jiang, X., Li, X., Liu, Y., Jiang, D., Sun, X., and Xu, L. Hrvda: High-resolution visual document assistant. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15534–15545, 2024a.

Liu, H., Li, C., Li, Y., and Lee, Y. J. Improved baselines with visual instruction tuning, 2023a. URL https://arxiv.org/abs/2310.03744.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning, 2023b.

Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pp. 38–55. Springer, 2025.

Liu, Y., Chen, H., Shen, C., He, T., Jin, L., and Wang, L. Abcnet: Real-time scene text spotting with adaptive bezier-curve network, 2020. URL https://arxiv.org/abs/2002.10200.

Liu, Y., Yang, B., Liu, Q., Li, Z., Ma, Z., Zhang, S., and Bai, X. TextMonkey: an OCR-Free large multimodal model for understanding document. *arXiv*, 2403.04473, 2024b.

Long, S., Qin, S., Panteleev, D., Bissacco, A., Fujii, Y., and Raptis, M. Towards end-to-end unified scene text detection and layout analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1049–1059, 2022.

Lu, J., Yu, H., Wang, Y., Ye, Y., Tang, J., Yang, Z., Wu, B., Liu, Q., Feng, H., Wang, H., et al. A bounding box is worth one token: Interleaving layout and text in a large language model for document understanding. *arXiv preprint arXiv:2407.01976*, 2024.

Lu, P., Qiu, L., Chang, K.-W., Wu, Y. N., Zhu, S.-C., Rajpurohit, T., Clark, P., and Kalyan, A. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. *arXiv preprint arXiv:2209.14610*, 2022.

Luo, C., Shen, Y., Zhu, Z., Zheng, Q., Yu, Z., and Yao, C. Layoutllm: Layout instruction tuning with large language models for document understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15630–15640, 2024.

Lv, T., Huang, Y., Chen, J., Zhao, Y., Jia, Y., Cui, L., Ma, S., Chang, Y., Huang, S., Wang, W., et al. Kosmos-2.5: A multimodal literate model. *arXiv preprint arXiv:2309.11419*, 2023.

Masry, A., Long, D. X., Tan, J. Q., Joty, S., and Hoque, E. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*, 2022.

Mathew, M., Karatzas, D., and Jawahar, C. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2200–2209, 2021.

Mathew, M., Bagal, V., Tito, R., Karatzas, D., Valveny, E., and Jawahar, C. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1697–1706, 2022.

McKinzie, B., Gan, Z., Fauconnier, J., Dodge, S., Zhang, B., Dufter, P., Shah, D., Du, X., Peng, F., Weers, F., et al. Mm1: methods, analysis & insights from multimodal llm pre-training. arxiv. *Preprint posted online on April*, 18, 2024.

Methani, N., Ganguly, P., Khapra, M. M., and Kumar, P. Plotqa: Reasoning over scientific plots. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1527–1536, 2020.

Mishra, A., Shekhar, S., Singh, A. K., and Chakraborty, A. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pp. 947–952. IEEE, 2019.

Nacson, M. S., Aberdam, A., Ganz, R., Avraham, E. B., Golts, A., Kittenplon, Y., Mazor, S., and Litman, R. Docvlm: Make your vlm an efficient reader, 2024. URL https://arxiv.org/abs/2412.08746.

Nayef, N., Yin, F., Bizid, I., Choi, H., Feng, Y., Karatzas, D., Luo, Z., Pal, U., Rigaud, C., Chazalon, J., et al. Icdar2017 robust reading challenge on multi-lingual scene text detection and script identification-rrc-mlt. In *2017 14th IAPR international conference on document analysis and recognition (ICDAR)*, volume 1, pp. 1454–1459. IEEE, 2017.

Pasupat, P. and Liang, P. Compositional semantic parsing on semi-structured tables. *arXiv preprint arXiv:1508.00305*, 2015.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. pp. 8748–8763, 2021.

Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., and Komatsuzaki, A. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

Shao, H., Qian, S., Xiao, H., Song, G., Zong, Z., Wang, L., Liu, Y., and Li, H. Visual cot: Unleashing chain-of-thought reasoning in multi-modal language models. 2024.

Shen, W., Peng, Z., Wang, X., Wang, H., Cen, J., Jiang, D., Xie, L., Yang, X., and Tian, Q. A survey on label-efficient deep image segmentation: Bridging the gap between weak supervision and dense prediction. *IEEE transactions on pattern analysis and machine intelligence*, 45 (8):9284–9305, 2023.

Shi, B., Yao, C., Liao, M., Yang, M., Xu, P., Cui, L., Belongie, S., Lu, S., and Bai, X. Icdar2017 competition on reading chinese text in the wild (rctw-17). In *2017 14th iapr international conference on document analysis and recognition (ICDAR)*, volume 1, pp. 1429–1434. IEEE, 2017.

Shi, M., Liu, F., Wang, S., Liao, S., Radhakrishnan, S., Huang, D.-A., Yin, H., Sapra, K., Yacoob, Y., Shi, H., Catanzaro, B., Tao, A., Kautz, J., Yu, Z., and Liu, G. Eagle: Exploring the design space for multimodal llms with mixture of encoders, 2024. URL https://arxiv.org/abs/2408.15998.

Sidorov, O., Hu, R., Rohrbach, M., and Singh, A. Textcaps: a dataset for image captioning with reading comprehension. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pp. 742–758. Springer, 2020.

Singh, A., Pang, G., Toh, M., Huang, J., Galuba, W., and Hassner, T. Textocr: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8802–8812, 2021.

Stanisławek, T., Graliński, F., Wróblewska, A., Lipiński, D., Kaliska, A., Rosalska, P., Topolski, B., and Biecek, P. Kleister: key information extraction datasets involving long documents with complex layouts. In *International Conference on Document Analysis and Recognition*, pp. 564–579. Springer, 2021.

Sun, N., Yang, X., and Liu, Y. Tableqa: a large-scale chinese text-to-sql dataset for table-aware sql generation. *arXiv preprint arXiv:2006.06434*, 2020.

Sun, Y., Ni, Z., Chng, C.-K., Liu, Y., Luo, C., Ng, C. C., Han, J., Ding, E., Liu, J., Karatzas, D., et al. Icdar 2019 competition on large-scale street view text with partial labeling-rrc-lsvt. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pp. 1557–1562. IEEE, 2019.

Svetlichnaya, S. Deepform: Understand structured documents at scale. 2020.

Tanaka, R., Nishida, K., and Yoshida, S. Visualmrc: Machine reading comprehension on document images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 13878–13888, 2021.

Tanaka, R., Iki, T., Nishida, K., Saito, K., and Suzuki, J. Instructdoc: A dataset for zero-shot generalization of visual document understanding with instructions. In *AAAI*, volume 38, pp. 19071–19079, 2024.

Tong, S., Brown, E., Wu, P., Woo, S., Middepogu, M., Akula, S. C., Yang, J., Yang, S., Iyer, A., Pan, X., et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*, 2024.

Turski, M., Stanisławek, T., Kaczmarek, K., Dyda, P., and Graliński, F. Ccpdf: Building a high quality corpus for visually rich documents from web crawl data. In *International Conference on Document Analysis and Recognition*, pp. 348–365. Springer, 2023.

Veit, A., Matera, T., Neumann, L., Matas, J., and Belongie, S. Coco-text: Dataset and benchmark for text detection and recognition in natural images. *arXiv preprint arXiv:1601.07140*, 2016.

Wang, B., Li, G., Zhou, X., Chen, Z., Grossman, T., and Li, Y. Screen2words: Automatic mobile ui summarization with multimodal learning. In *The 34th Annual ACM Symposium on User Interface Software and Technology*, pp. 498–510, 2021a.

Wang, D., Raman, N., Sibue, M., Ma, Z., Babkin, P., Kaur, S., Pei, Y., Nourbakhsh, A., and Liu, X. DocLLM: a layout-aware generative language model for multimodal document understanding. *arXiv*, 2401.00908, 2023.

Wang, H., Bai, X., Yang, M., Zhu, S., Wang, J., and Liu, W. Scene text retrieval via joint text detection and similarity learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4558–4567, June 2021b.

Wang, W., Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Zhu, J., Zhu, X., Lu, L., Qiao, Y., and Dai, J. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization, 2024. URL https://arxiv.org/abs/2411.10442.

Wei, H., Kong, L., Chen, J., Zhao, L., Ge, Z., Yang, J., Sun, J., Han, C., and Zhang, X. Vary: Scaling up the vision vocabulary for large vision-language model. In *European Conference on Computer Vision*, pp. 408–424. Springer, 2024.

Wei, H., Kong, L., Chen, J., Zhao, L., Ge, Z., Yang, J., Sun, J., Han, C., and Zhang, X. Vary: Scaling up the vision vocabulary for large vision-language model. In *ECCV*, pp. 408–424. Springer, 2025.

Wu, Z., Chen, X., Pan, Z., Liu, X., Liu, W., Dai, D., Gao, H., Ma, Y., Wu, C., Wang, B., et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024.

Xia, R., Mao, S., Yan, X., Zhou, H., Zhang, B., Peng, H., Pi, J., Fu, D., Wu, W., Ye, H., et al. Docgenome: An open large-scale scientific document benchmark for training and testing multi-modal large language models. *arXiv preprint arXiv:2406.11633*, 2024.

Xie, X., Yin, L., Yan, H., Liu, Y., Ding, J., Liao, M., Liu, Y., Chen, W., and Bai, X. PDF-WuKong: a large multimodal model for efficient long PDF reading with end-to-end sparse sampling. *arXiv*, 2410.05970, 2024.

Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., and Zhou, M. LayoutLM: pre-training of text and layout for document image understanding. In *Knowledge Discovery and Data Mining*, 2019.

Ye, J., Hu, A., Xu, H., Ye, Q., Yan, M., Xu, G., Li, C., Tian, J., Qian, Q., Zhang, J., et al. Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model. *arXiv preprint arXiv:2310.05126*, 2023a.

Ye, J., Hu, A., Xu, H., Ye, Q., Yan, M., Xu, G., Li, C., Tian, J., Qian, Q., Zhang, J., et al. Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model. *arXiv preprint arXiv:2310.05126*, 2023b.

Ye, M., Zhang, J., Liu, J., Liu, C., Yin, B., Liu, C., Du, B., and Tao, D. Hi-sam: Marrying segment anything model for hierarchical text segmentation. *arXiv preprint arXiv:2401.17904*, 2024.

Yu, W., Liu, Y., Hua, W., Jiang, D., Ren, B., and Bai, X. Turning a clip model into a scene text detector. In *CVPR*, pp. 6978–6988, 2023.

Yu, Y.-Q., Liao, M., Zhang, J., and Wu, J. Texthawk2: A large vision-language model excels in bilingual ocr and grounding with 16x fewer tokens. *arXiv preprint arXiv:2410.05261*, 2024.

Yuliang, L., Lianwen, J., Shuaitao, Z., and Sheng, Z. Detecting curve text in the wild: New dataset and new solution. *arXiv preprint arXiv:1712.02170*, 2017.

Zhai, X., Mustafa, B., Kolesnikov, A., and Beyer, L. Sigmoid loss for language image pre-training. In *ICCV*, pp. 11975–11986, 2023.

Zhang, C., Han, D., Qiao, Y., Kim, J. U., Bae, S.-H., Lee, S., and Hong, C. S. Faster segment anything: Towards lightweight sam for mobile applications. *arXiv preprint arXiv:2306.14289*, 2023a.

Zhang, H., Liang, L., and Jin, L. Scut-hccdoc: A new benchmark dataset of handwritten chinese text in unconstrained camera-captured documents. *Pattern Recognition*, 108: 107559, 2020.

Zhang, H., Gao, M., Gan, Z., Dufter, P., Wenzel, N., Huang, F., Shah, D., Du, X., Zhang, B., Li, Y., et al. Mm1. 5: Methods, analysis & insights from multimodal llm fine-tuning. *arXiv preprint arXiv:2409.20566*, 2024a.

Zhang, J., Yang, W., Lai, S., Xie, Z., and Jin, L. Dockylin: A large multimodal model for visual document understanding with efficient visual slimming. *arXiv preprint arXiv:2406.19101*, 2024b.

Zhang, L., Hu, A., Xu, H., Yan, M., Xu, Y., Jin, Q., Zhang, J., and Huang, F. Tinychart: Efficient chart understanding with visual token merging and program-of-thoughts learning. *arXiv preprint arXiv:2404.16635*, 2024c.

Zhang, R., Zhou, Y., Jiang, Q., Song, Q., Li, N., Zhou, K., Wang, L., Wang, D., Liao, M., Yang, M., et al. Icdar 2019 robust reading challenge on reading chinese text on signboard. In *2019 international conference on document analysis and recognition (ICDAR)*, pp. 1577–1581. IEEE, 2019.

Zhang, R., Lyu, Y., Shao, R., Chen, G., Guan, W., and Nie, L. Token-level correlation-guided compression for efficient multimodal document understanding. *arXiv preprint arXiv:2407.14439*, 2024d.

Zhang, Y., Zhang, R., Gu, J., Zhou, Y., Lipka, N., Yang, D., and Sun, T. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*, 2023b.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36: 46595–46623, 2023.

Zhong, X., Tang, J., and Yepes, A. J. Publaynet: largest dataset ever for document layout analysis. In *2019 International conference on document analysis and recognition (ICDAR)*, pp. 1015–1022. IEEE, 2019.

Zhong, X., ShafieiBavani, E., and Jimeno Yepes, A. Image-based table recognition: data, model, and evaluation. In *European conference on computer vision*, pp. 564–580. Springer, 2020.

Zhu, D., Chen, J., Shen, X., Li, X., and Elhoseiny, M. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

A. VQA-based Text Parsing Tasks

Modality connectors act as the bridge between the visual foundation model (VFM) and the LLM. Previous MLLMs employ image-text pairs of natural images (*e.g.,* Conceptual Captions, LAION, COYO) to pre-train them. In the work, to endow our MLLM TokenVL with generality and comprehensive document understanding abilities, we follow DocOwl (Hu et al., 2024c) to conduct modality alignment. It involves both structure-aware parsing tasks (recognizing full text, converting formulas into LaTeX, converting tables into markdown or LaTeX, and converting charts into CSV or markdown formats) and multi-grained text localization tasks (recognizing partial text within localization, visual text grounding). Specifically, we present an example to introduce them, as shown in Table 8. In this way, the pre-trained modality connector can understand the visual features of VFM and project them into the same feature space with the language features of LLM.

B. Interactive Demo

As shown in Figure 5 and Figure 6, we provide more interactive examples, including natural scene images, documents, codes, charts, tables, and GUIs. For each scene, we provide two examples. The first column is the original image, the second to fourth columns are visualizations of the corresponding BPE words within the image, and the last column shows the highlighted area of the image when the prompt is a space " ". As we observed,

1) Our foundation model TokenOCR can distinguish text and background areas well. This means that when using the foundation model for downstream tasks, we can remove redundant background features at very low cost;

2) For complex, dense, and small texts, TokenOCR still precisely perceive, such as "picture"(Code), "f"(Code), "19"(Table), "P"(Table), *etc*. Additionally, our TokenOCR still can capture punctuation marks, such as commas, periods, double quotes, *etc*;

3) Our TokenOCR also supports handwritten texts, such as "STE"(Document) and "USA"(Document). We will provide a demo address for users to experience it.

C. TokenIT Dataset

C.1. Data Source

To construct a comprehensive TokenIT dataset, we collect various types of data, including natural scene text images, documents (PDF, receipt, letter, note, report, *etc.*), tables, charts, and GUI. The data sources are summarized in Table 9.

C.2. Data Generation

Next, we elaborate on the data construction pipeline for the TokenIT dataset, which involves four steps:

1) Text Image Segmentation. For natural scene text images, charts and tables, we fine-tune the SAM model (Kirillov et al., 2023) on datasets with character-level mask annotations and leverage the well-learned model to generate text masks, since these images are relatively complex and diverse in color and style. For PDFs and industrial documents, we conduct simple unsupervised clustering (Kanungo et al., 2002) to get their text masks;

2) Text Recognition. We use the previous SOTA method (Guan et al., 2023a) to obtain the recognition results for all types, except for natural scene text images. As these natural scene datasets already provide text transcriptions, we adopt them directly;

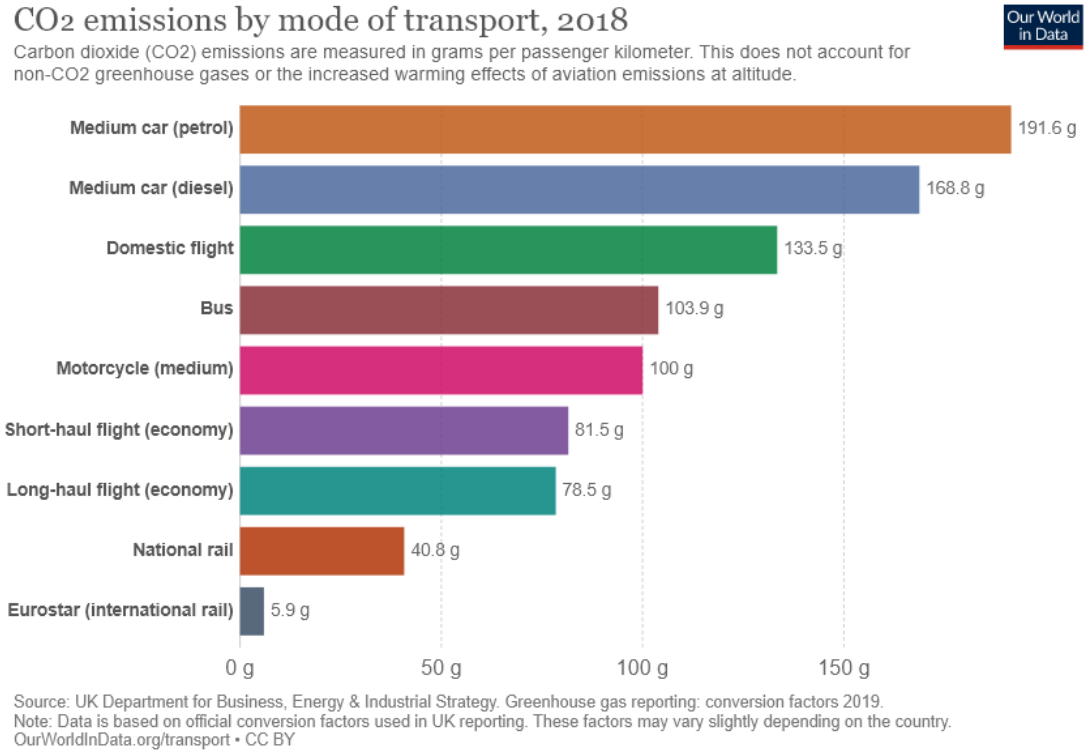
3) Tokenizer. We choose the widely adopted BPE tokenizer (Chen et al., 2024d) to split the language texts into multiple BPE tokens, where each token corresponds to a BPE-level subword;

4) Token-level Image Text Construction. After obtaining the text masks in Step 1, we apply the method (Guan et al., 2025b) to produce character-level segmentation masks. Subsequently, we combine each token's corresponding character-level masks to create a complete token-level segmentation mask.

5) Data Correction. For each image and its generated labels following the above stage, we render the labels onto the images to verify data labeling quality and perform manual relabeling as needed. Finally, *three rounds of inspections are conducted to minimize labeling errors, a process that took four months to develop the first token-level image text dataset (TokenIT).*

Overall, the proposed TokenIT dataset includes 20 million images (including natural scene text images, documents, tables,

VQA-based Text Parsing



Question:	Recognizing full text.
Answer:	CO_2 emissions by mode of transport, 2018. Carbon dioxide (CO2) emissions are measured in grams per passenger kilometer. This does not account for non-CO2 greenhouse gases or the · · · · · · .
Question:	Recognizing the text within the bounding box <bbox>75, 200, 160, 230</bbox>.
Answer:	Medium car (diesel).
Question:	Pblueict the bounding box of the text <ocr>Eurostar (International rall)</ocr>
Answer:	<bbox>25, 520, 160, 550</bbox>.
Question:	Converting the chart into CSV format.
Answer:	Mode of Transport, CO2 Emissions (g CO2 per passenger km) Medium car (petrol), 191.6 Medium car (diesel), 168.8 Domestic flight, 133.5 Bus, 103.9 · · · · · · .
Question:	Converting the chart into Markdown format.
Answer:	|Mode of Transport |CO2 Emissions (g CO2 per passenger km) | |———————|———————| |Medium car (petrol) |191.6 | |Medium car (diesel) |168.8 | |Domestic flight |133.5 | |Bus |103.9 | · · · · · · .

Table 8: The illustration of VQA-based Text Parsing tasks of TokenVL.

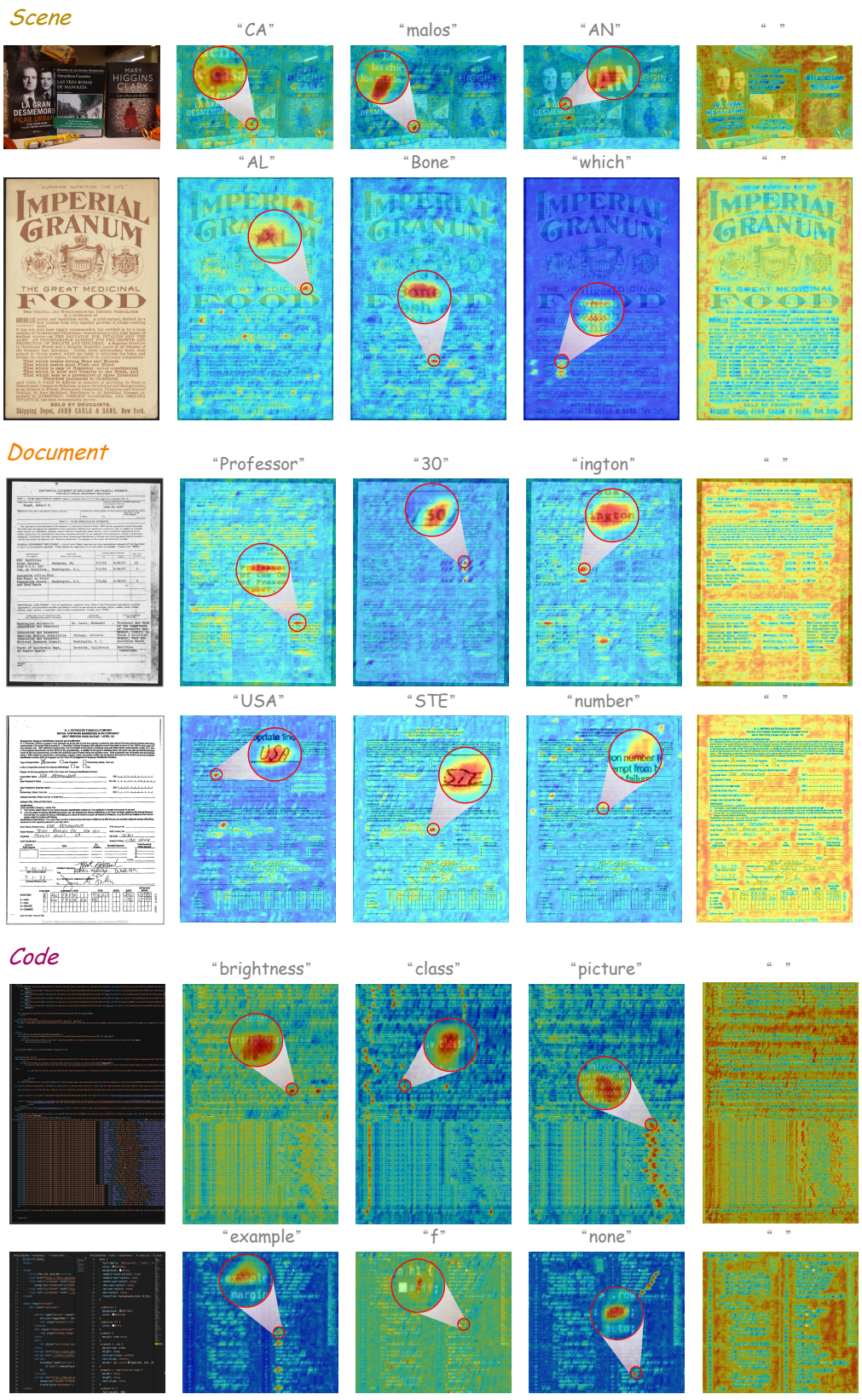


Figure 5: More visualization examples of the natural scene images, document images, and code images.

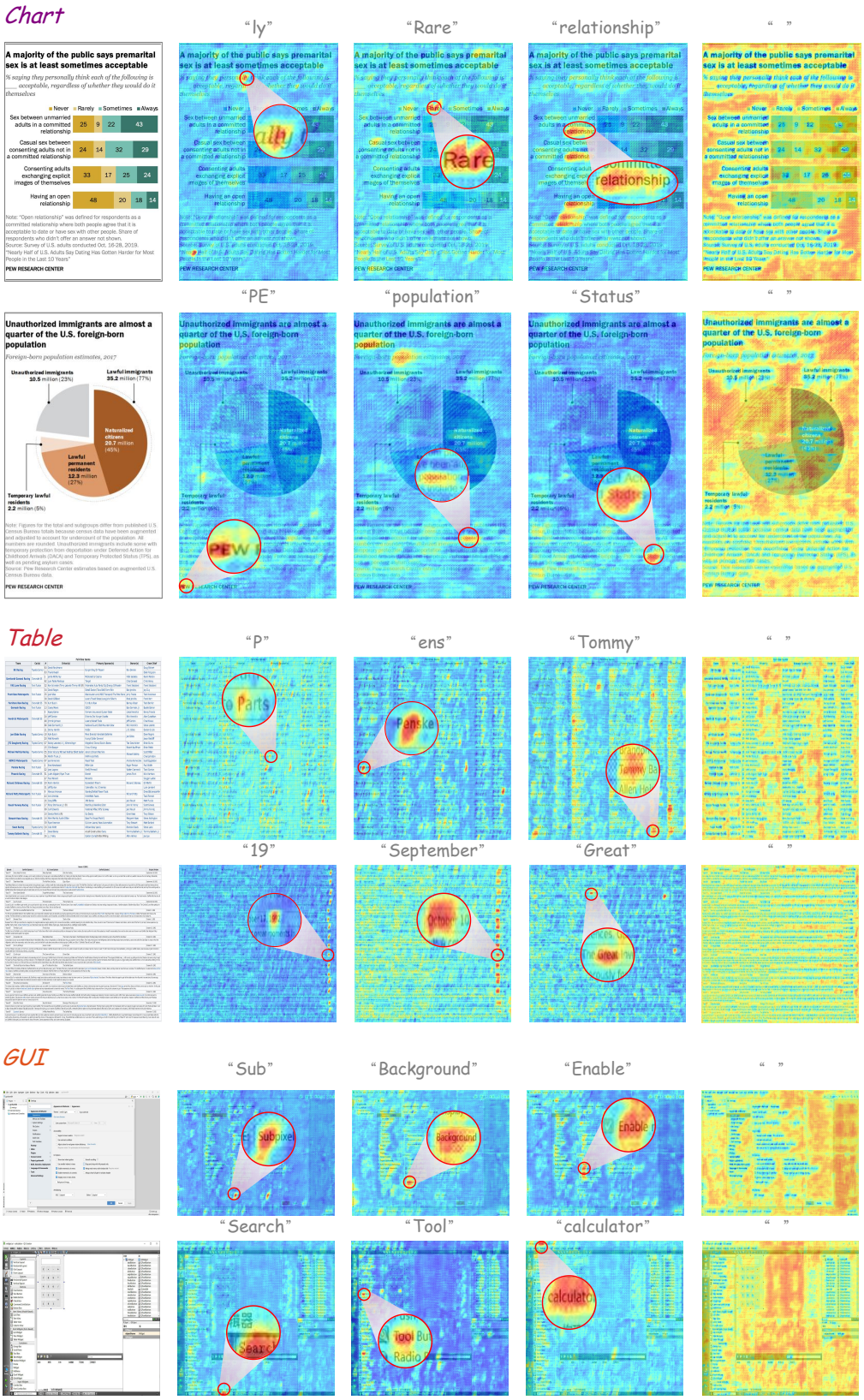


Figure 6: More visualization examples of the chart, table, and GUI images.

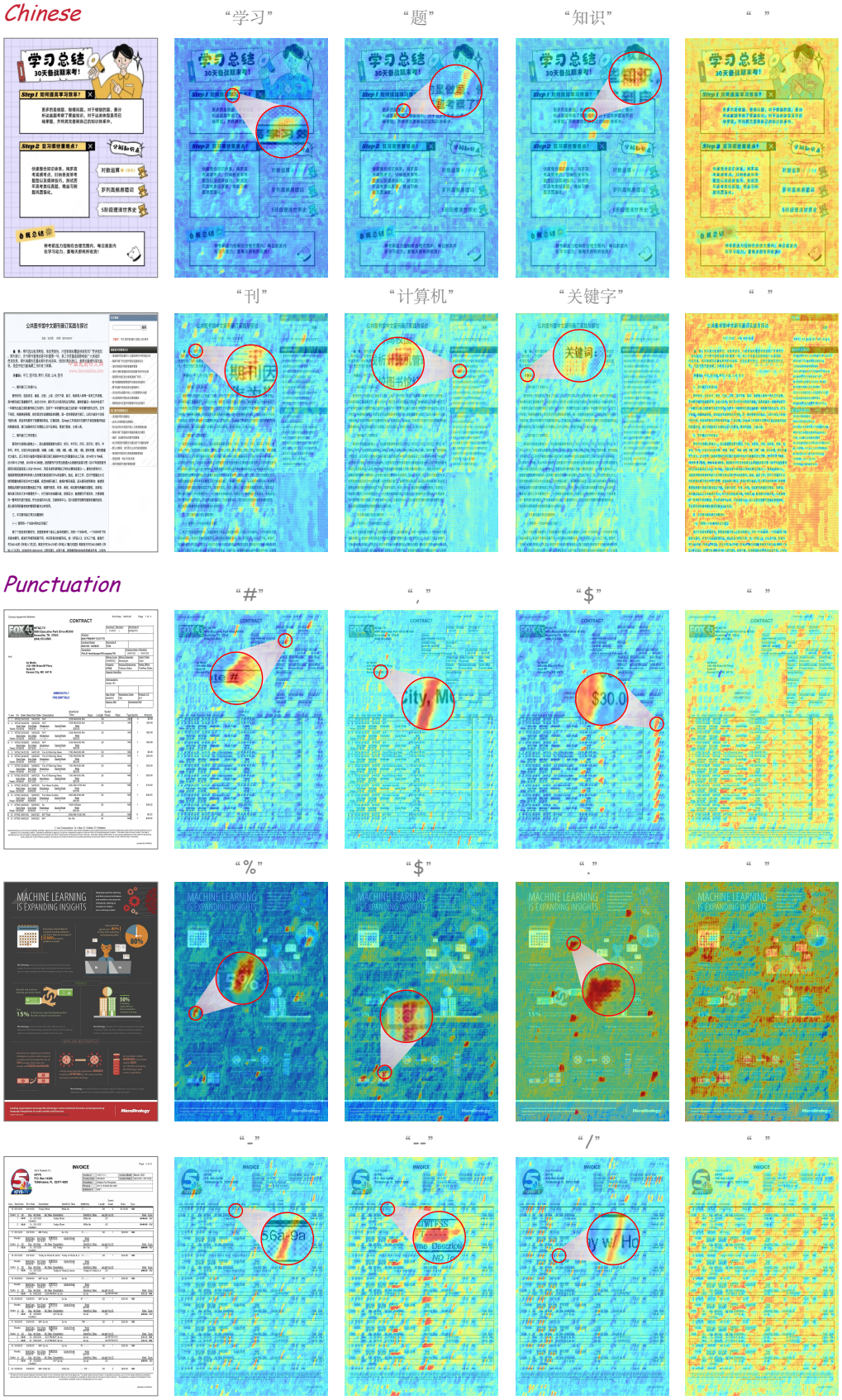


Figure 7: More visualization examples of the Chinese.

Dataset Type	Dataset Name
Natural Scene	ICDAR2013 (Karatzas et al., 2013), COCOText (Veit et al., 2016), CTW1500 (Yuliang et al., 2017), HierText (Long et al., 2022), ICDAR2015 (Cheng et al., 2017), OCRCC (Kareem et al., 2017), OpenImagesV5Text (Krylov et al., 2021), TextCaps (Sidorov et al., 2020), TextOCR (Singh et al., 2021), TotalText (Ch'ng & Chan, 2017), Laion-OCR (Schuhmann et al., 2021), Wukong-OCR (Gu et al., 2022), Mlt2017 (Nayef et al., 2017), ocrvqa (Mishra et al., 2019), ST-VQA (Biten et al., 2019), SynText (Liu et al., 2020), the-cauldron (Laurençon et al., 2024a), ArT (Chng et al., 2019), ChineseOCR (Co., 2024), HCCDoc (Zhang et al., 2020), ICDAR2017rctw (Shi et al., 2017), LSVT (Sun et al., 2019), MTWI (He et al., 2018), and ReCTS (Zhang et al., 2019)
Document	DocVQA (Mathew et al., 2021), InfographicsVQA (Mathew et al., 2022), KIE, Kleister-Charity (Stanisławek et al., 2021), PubTabNet (Zhong et al., 2020), RVL-CDIP (Harley et al.), VisualMRC (Tanaka et al., 2021), Docmatix (Laurençon et al., 2024a), IIT-CDIP (Xu et al., 2019), publaynet (Zhong et al., 2019), Synthdog-en (Kim et al., 2022), DocGenome (Xia et al., 2024), CCpdf (Turski et al., 2023)
Chart	ChartQA (Masry et al., 2022), FigureQA (Kahou et al., 2017), PlotQA (Methani et al., 2020),TabMWP (Lu et al., 2022), DVQA (Kafle et al., 2018)
Table	TableQA (Sun et al., 2020), DeepForm (Svetlichnaya, 2020), TURL (Deng et al., 2022), TabFact (Chen et al., 2019), WikiTableQuestions (Pasupat & Liang, 2015)
GUI	Screen2Words (Wang et al., 2021a), WebSight (Laurençon et al., 2024b), Omni-ACT (Kapoor et al., 2024), SeeCliCK (Cheng et al., 2024), Mind2Web (Deng et al., 2024)

Table 9: Data source of our TokenIT dataset.

charts, Code, and GUI) and 1780679833 (1.8 billion) token-mask pairs. Each BPE token corresponds one-to-one with a pixel-level mask.

D. Training Details

D.1. Text Segmentation

In this section, we evaluate the performance of text segmentation using TextSeg, COCOText, and HierText, which provide pixel-level annotations. The test sets of these datasets are utilized for zero-shot experiments. In the linear probe setting, all methods are trained on the combined three training sets and evaluated separately on each test set. The training configuration includes 70 epochs, a learning rate of 0.0001, a batch size of 6, and the optimizer AdamW.

D.2. Visual Question Answering

In this section, we evaluate the performance of visual document understanding using the test sets of DocVQA, InfoVQA, ChartQA, and TextVQA. The proposed model undergoes a two-phase training process: pre-training and fine-tuning.

During the pre-training phase, we randomly sampled 200,000 images each from the IIT-CDIP and DocMatix document datasets. Full-text recognition was implemented using PaddleOCR to generate ground-truth textual content, which served as target answers. The model was trained with the instructional prompt "Recognize all text:" where only the Multilayer Perceptron (MLP) component received parameter updates. The training configuration included one epoch with a learning rate of 1e-3 and a batch size of 24.

For the fine-tuning phase, we extended the training scope to incorporate Low-Rank Adaptation (LoRA) for Large Language Model (LLM) optimization while continuing MLP updates. The training data comprised the training splits of the aforementioned QA evaluation datasets. This phase maintained single-epoch training with modified hyperparameters, specifically a reduced learning rate of 2e-4 and a batch size of 12 to ensure stable parameter convergence. This hierarchical training paradigm progressively enhances both text recognition accuracy and semantic comprehension capabilities in document understanding tasks.

D.3. Text Retrieval

In this section, we evaluate model performance using the CTR benchmark (English) (Veit et al., 2016) and the CSVTRv2 benchmark (Chinese) (Wang et al., 2021b). For English text retrieval, we employ the training sets from ICDAR2013, ICDAR2015, COCOText, MLT2017, OpenImagesV5Text, CTW1500, TotalText, HierText, and TextOCR. For Chinese text retrieval, we use ArT, ChineseOCR, HCCDoc, icdar2017rctw, LSVT, MTWI, and ReCTS as the training sets. These methods are optimized using the AdamW optimizer. The initial learning rate is 1e-4. We use a batch size of 6 and a number of training epochs of 10. After the first 5 epochs, the initial learning rate is reduced to 1e-5.

E. Mainstream Benchmark Results

General multi-modal large models typically use DocVQA, InfoVQA, ChartQA, and TextVQA to evaluate document understanding capabilities, as these benchmarks encompass diverse and comprehensive scenarios that reflect real-world applications. To compare performance intuitively and clearly, we collected data from nearly all models that reported scores on these four benchmarks and summarized them in Table 10. Specifically, we categorized the existing MLLMs into three types based on model size: "<2B", "<8B", and ">8B". Due to resource constraints, we did not conduct experiments with models exceeding 8B parameters in our TokenVL, providing only two versions: TokenVL-2B and TokenVL-8B. Notably, our TokenVL-2B improves upon the previous state-of-the-art (SOTA) result by 1.32%, and our TokenVL-8B improves by 0.63%. Compared to models with larger parameters, our 8B version slightly surpasses DeepSeek-VL2-16B and InternVL2-40B by 0.3%.

Size	Model	Visual Encoder	LLM Decoder	DocVQA	InfoVQA	ChartQA	TextVQA	Avg.
<2B	DocLLM-1B (Wang et al., 2023)	-	Falcon-1B	61.4	-	-	-	-
	Mini-Monkey (Huang et al., 2024a)	InternViT-300M	InternLLM2-2B	87.4	60.1	76.5	75.7	74.93
	MM1.5-1B (Zhang et al., 2024a)	CLIP-ViT-H	Private	81.0	50.5	67.2	72.5	67.80
	MM1.5-3B (Zhang et al., 2024a)	CLIP-ViT-H	Private	87.7	58.5	74.2	76.5	74.23
	InternVL2-1B (Chen et al., 2024b)	InternViT-300M	Qwen2-0.5B	81.7	50.9	72.9	70.5	69.00
	InternVL2-2B (Chen et al., 2024b)	InternViT-300M	InternLM2-1.8B	86.9	58.9	76.2	73.3	73.83
	InternVL2.5-1B (Chen et al., 2024a)	InternViT-300M	Qwen2.5-0.5B	84.8	56.0	75.9	72.0	72.18
	InternVL2.5-2B (Chen et al., 2024a)	InternViT-300M	InternLM2.5-1.8B	88.7	60.9	79.2	74.3	75.78
	LLaVA-OneVision-0.5B (Li et al., 2024b)	SigLIP	qwen2-0.5B	70.0	41.8	61.4	-	-
	TokenVL-2B	TokenOCR	InternLM2.5-1.8B	**89.9**	**61.0**	**81.1**	**76.4**	**77.10**
<8B	UReader (Ye et al., 2023b)	CLIP-ViT-L/14	LLaMA-7B	65.4	42.2	59.3	57.6	56.13
	DocLLM-7B (Wang et al., 2023)	-	LLaMA2-7B	69.5	-	-	-	-
	Cream (Kim et al., 2023b)	CLIP-ViT-L/14	Vicuna-7B	79.5	43.5	63.0	-	-
	Qwen-VL (Bai et al., 2023)	ViT-bigG	Qwen-7B	65.1	35.4	65.7	63.8	57.50
	LLaVA-1.5-7B (Liu et al., 2023a)	CLIP-ViT-L	Vicuna1.5-7B	-	-	-	58.2	-
	SPHINX (Lin et al., 2023a)	CLIP-ViT+CLIP-ConvNext+DINOv2-ViT	LLaMA2-7B	-	-	-	61.2	-
	LLaVA-OneVision (Li et al., 2024a)	SigLIP	Qwen2-7B	87.5	68.8	80.0	-	-
	Monkey (Li et al., 2024d)	Vit-BigG	Qwen-7B	66.5	36.1	65.1	67.6	58.83
	TextMonkey (Liu et al., 2024b)	Vit-BigG	Qwen-7B	73.0	-	66.9	65.6	-
	IDEFICS2 ((Laurençon et al., 2024b))	SigLIP-SO400M	Mistral-7B	74.0	-	-	73.0	-
	LayoutLLM (Luo et al., 2024)	LayoutLMv3-large	Vicuna1.5-7B	74.25	-	-	-	-
	DocKylin (Zhang et al., 2024b)	Swin	Qwen-7B	77.3	46.6	66.8	-	-
	DocLayLLM (Liao et al., 2024b)	LayoutLMV3	LLaMA3-8B	77.79	42.02	-	-	-
	mPLUG-DocOwl (Hu et al., 2024a)	CLIP-ViT-L/14	LLaMA-7B	62.2	38.2	57.4	52.6	52.60
	mPLUG-DocOwl1.5 (Hu et al., 2024b)	CLIP-ViT-L/14	LLaMA2-7B	82.2	50.7	70.2	68.6	67.93
	mPLUG-DocOwl2 (Hu et al., 2024d)	CLIP-ViT-L/14	LLaMA2-7B	80.7	46.4	70.0	66.7	65.95
	Vary (Wei et al., 2024)	CLIP-ViT-L/14 + SAM	Qwen-7B	76.3	-	66.1	-	-
	Eagle (Shi et al., 2024)	CLIP + ConvNeX + Pix2Struct + EVA2 + SAM	LLaMA3-8B	86.6	-	80.1	77.1	-
	PDF-WuKong (Xie et al., 2024)	CLIP-ViT-L-14	InernLM2-7B	85.1	61.3	80.0	-	-
	TextHawk2 (Yu et al., 2024)	SigLIP	Qwen2-7B	89.6	67.8	81.4	75.1	78.48
	MM1.5-7B (Zhang et al., 2024a)	CLIP-ViT-H	Private	88.1	59.5	78.6	76.5	75.68
	HRVDA (Liu et al., 2024a)	Swin-L	LLaMA2-7B	72.1	43.5	67.6	73.3	64.13
	InternVL2-4B (Chen et al., 2024b)	InternViT-300M	Phi-3-mini	89.2	67.0	81.5	74.4	78.03
	InternVL2-8B (Chen et al., 2024b)	InternViT-300M	InternLM2.5-7B	91.6	74.8	83.3	77.4	81.78
	InternVL2.5-4B (Chen et al., 2024a)	InternViT-300M	Qwen2.5-3B	91.6	72.1	84.0	76.8	81.13
	InternVL2.5-8B (Chen et al., 2024a)	InternViT-300M	InternLM2.5-7B	93.0	**77.6**	84.8	79.1	83.63
	InternVL2.5-8B-mpo(Wang et al., 2024)†	InternViT-300M	InternLM2.5-7B	92.3	76.0	83.8	79.1	82.80
	DeepSeek-VL2-3B (Wu et al., 2024)	SigLIP-SO400M-384	DeepSeekMoE	88.9	66.1	81.0	80.7	79.18
	DocPeida (Feng et al., 2024)	Swin	Vicuna-7B	47.1	15.2	46.9	60.2	42.35
	TokenPacker-7B (Li et al., 2024c)	CLIP-ViT-L/14	Vicuna-7B	60.2	-	-	-	-
	LLaVA-OneVision-7B (Li et al., 2024b)	SigLIP	qwen2-7B	87.5	68.8	80.0	-	-
	DocVLM (Nacson et al., 2024)	CLIP-ViT-G/14 + DocFormerV2	Qwen2-7B	92.8	66.8	-	**82.8**	-
	TokenVL-8B	TokenOCR	InternLM2.5-7B	**94.2**	76.5	**86.6**	79.9	**84.30**
>8B	LLaVA-13B (Liu et al., 2023b)	CLIP-ViT-L/14	Vicuna-13B	6.9	-	-	36.7	-
	PaLI-X (Chen et al., 2023)	ViT-22B	UL2-32B	86.8	54.8	72.3	80.8	73.68
	LLaVAR (Zhang et al., 2023b)	CLIP-ViT-L/14	Vicuna-13B	11.6	-	-	48.5	-
	LLaVA-1.5-13B (Liu et al., 2023a)	CLIP-ViT-L	Vicuna1.5-13B	-	-	-	62.5	-
	CogAgent (Hong et al., 2023)	EVA2-CLIP+CogVLM +Cross Attention	Vicuna-13B	81.6	44.5	68.4	76.1	67.65
	Unidoc (Feng et al., 2023b)	CLIP-ViT-L/14	Vicuna-13B	90.2	36.8	70.5	73.7	67.80
	MM1.5-30B (Zhang et al., 2024a)	CLIP-ViT-H	Private	91.4	67.3	83.6	79.2	80.38
	InternVL1.5-26B (Chen et al., 2024c)	InternViT-6B	InternLM2-20B	90.9	72.5	83.8	80.6	81.95
	InternVL2-26B (Chen et al., 2024b)	InternViT-6B	InternLM2-20B	92.9	75.9	84.9	82.3	84.00
	InternVL2-40B (Chen et al., 2024b)	InternViT-6B	Nous-Hermes-2-Yi-34B	93.9	78.7	86.2	83.0	85.45
	InternVL2.5-26B (Chen et al., 2024a)	InternViT-6B	InternLM2.5-20B	94.0	79.8	87.2	82.4	85.85
	InternVL2.5-38B (Chen et al., 2024a)	InternViT-6B	Qwen2.5-32B	**95.3**	83.6	88.2	82.7	87.45
	InternVL2.5-78B (Chen et al., 2024a)	InternViT-6B	Qwen2.5-72B	95.1	**84.1**	**88.3**	83.4	**87.73**
	TinyChart (Zhang et al., 2024c)	SigLIP	Phi-2	-	-	83.6	-	-
	TokenPacker-13B (Li et al., 2024c)	CLIP-ViT-G/14	Vicuna-13B	70.0	-	-	-	-
	DeepSeek-VL2-16B (Wu et al., 2024)	SigLIP-SO400M-384	DeepSeekMoE	92.3	75.8	84.5	83.4	84.00
	DeepSeek-VL2-27B (Wu et al., 2024)	SigLIP-SO400M-384	DeepSeekMoE	93.3	78.1	86.0	**84.2**	85.40

Table 10: Comparison results on four widely evaluated datasets.