# SEA-Vision: A Multilingual Benchmark for Comprehensive Document and Scene Text Understanding in Southeast Asia

Supplementary Material

Table 1. Per-language statistics for the Document Parsing subset of SEA-Vision. We report the number of pages, average number of blocks per page, average text-area ratio, and the proportions of pages with tables and formulas.

| Language | #Pages | Avg. blocks/page | Avg. text-area ratio | Pages with tables (%) | Pages with formulas (%) |
|---|---|---|---|---|---|
| EN | 1585 | 25.71 | 63.08 | 15.65 | 6.56 |
| ZH | 636 | 28.11 | 62.59 | 23.27 | 16.50 |
| VI | 1678 | 19.69 | 60.13 | 11.50 | 5.78 |
| TH | 1506 | 10.53 | 54.38 | 6.57 | 1.26 |
| FIL | 1265 | 12.35 | 49.40 | 8.06 | 0.70 |
| MS | 1580 | 14.12 | 49.50 | 9.68 | 0.89 |
| ID | 1532 | 17.59 | 53.79 | 18.73 | 6.33 |
| LO | 914 | 8.86 | 49.85 | 3.71 | 0.11 |
| KM | 1331 | 8.49 | 52.50 | 3.38 | 0.38 |
| MY | 1575 | 9.68 | 52.56 | 5.14 | 0.76 |
| PT | 1632 | 22.79 | 60.90 | 12.81 | 2.80 |

Table 2. Page-type statistics for the Document Parsing subset. Each entry denotes the number of pages for a given page type and language.

| Page Type | EN | ZH | VI | TH | FIL | MS | ID | LO | KM | MY | PT |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Academic Literature | 164 | 178 | 185 | 179 | 161 | 168 | 140 | 61 | 192 | 199 | 176 |
| Book | 187 | 64 | 186 | 193 | 187 | 187 | 187 | 200 | 199 | 171 | 193 |
| Textbook | 183 | 47 | 184 | 179 | 177 | 187 | 177 | 135 | 188 | 192 | 182 |
| Exam Paper | 133 | 9 | 184 | 157 | 141 | 147 | 146 | 129 | 148 | 166 | 180 |
| Magazine | 189 | 56 | 200 | 181 | 65 | 196 | 194 | 4 | 20 | 92 | 196 |
| Newspaper | 175 | 124 | 185 | 49 | 55 | 187 | 186 | 3 | 6 | 188 | 171 |
| Note | 193 | 39 | 190 | 196 | 190 | 190 | 198 | 125 | 198 | 198 | 191 |
| Slide | 188 | 22 | 196 | 196 | 132 | 160 | 169 | 70 | 196 | 195 | 184 |
| Research Report | 173 | 97 | 168 | 176 | 157 | 158 | 135 | 187 | 184 | 174 | 159 |

## A. Dataset Statistics

As outlined in Section 4 of the main text, this appendix expands the statistical analysis of SEA-Vision. It provides additional details for the Document Parsing subset, the distribution of capability labels in the TEC-VQA portion, and a co-occurrence analysis of reasoning skills.

### A.1. Document Parsing Statistics

**Language-wise distribution.** The Document Parsing subset spans 11 languages, with page counts ranging from 636 to 1,678 (Table 1). The comparatively small Chinese portion (636 pages) results from strict safety filtering: many candidate documents contained politically sensitive content and were removed. Lao and Khmer also remain smaller due to the scarcity of high-quality public materials in these languages. In contrast, English, Vietnamese, Malay, Indonesian, Myanmar, and Portuguese each contribute 1.5k–1.7k pages, reflecting richer public sources. Layout complexity varies substantially across languages: English and Chinese pages contain dense structures with over 25 blocks on average, whereas Lao, Khmer, Thai, and Myanmar typically exhibit simpler layouts of 8–10 blocks. Content density is also uneven: Chinese pages show the highest formula proportion (16.5%), consistent with their larger share of technical and educational materials.

**Page-type-wise distribution.** We categorize pages into nine representative document types covering diverse real-world formats, including exam papers, academic articles, books, magazines, newspapers, handwritten notes, research reports, slides, and textbooks (Table 2). Academic, textbook, and book pages constitute the core of the dataset and remain relatively well balanced across most languages. By contrast, Chinese exam papers, magazines, and newspapers appear less frequently due to the removal of safety-sensitive materials. Magazine and newspaper pages are heavily concentrated in high-resource languages such as English, Vietnamese, Malay, Indonesian, and Portuguese, with minimal representation for Lao and Khmer. Handwritten notes and slides are included for all languages but remain notably sparse in low-resource ones, making cross-lingual evaluation on informal or noisy layouts more challenging.

### A.2. TEC-VQA Statistics

**Per-language question–answer distribution.** The Text-Centric Visual Question Answering (TEC-VQA) subset contains 1,839 images and 7,496 question–answer (QA) pairs, spanning the same 11 languages as the Document Parsing subset. As shown in Table 3, we report, for each language, the number of images, the number of QA pairs, and the average lengths of questions and answers in tokens. It can be seen that, for each low-resource language, TEC-VQA still maintain non-trivial coverage to support multilingual evaluation.

**Capability distribution.** Each TEC-VQA question is labeled with one or more of five reasoning skill categories: text recognition (TR), numerical calculation (NC), comparative analysis (CA), logical reasoning (LR), and spatial understanding (SU). These labels indicate the primary capa-

Table 3. Per-language statistics for the TEC-VQA subset, including the number of images, the number of QA pairs, and the average question and answer lengths.

| Language | #Images | #QA pairs | Avg Q length | Avg A length |
|---|---|---|---|---|
| EN | 267 | 1335 | 72.74 | 31.00 |
| ZH | 100 | 500 | 21.72 | 12.81 |
| VI | 225 | 854 | 68.35 | 47.19 |
| TH | 272 | 1064 | 57.01 | 35.31 |
| FIL | 115 | 403 | 82.61 | 86.71 |
| MS | 242 | 924 | 73.54 | 38.29 |
| ID | 200 | 767 | 71.07 | 37.95 |
| LO | 89 | 371 | 58.33 | 43.65 |
| KM | 85 | 325 | 66.37 | 28.80 |
| MY | 242 | 358 | 68.08 | 32.96 |
| PT | 152 | 595 | 73.37 | 46.09 |

Table 4. Distribution of TEC-VQA capability categories in SEA-Vision, reported as the number and proportion of QA pairs attributed to each capability.

| Capability category | #QA pairs | Proportion (%) |
|---|---|---|
| Text recognition | 5573 | 74.35 |
| Numerical calculation | 5065 | 67.57 |
| Comparative analysis | 4306 | 57.44 |
| Logical reasoning | 459 | 6.12 |
| Spatial understanding | 209 | 2.79 |

bilities required to answer the question. Text recognition questions involve directly reading a text snippet from the image. Numerical calculation questions require arithmetic or counting based on textual content. Comparative analysis questions compare two or more values (for example, selecting the larger number or the earlier date). Logical reasoning questions refers to those requiring logical judgment, such as truth and falsehood, and AND/OR/NOT relationships. Spatial understanding questions depend on layout or positional relations of text elements (for example, "Which item is at the top of the page?"). The overall distribution of these capability labels is summarized in Table 4.

**Capability co-occurrence analysis.** Many TEC-VQA questions require multiple capabilities simultaneously (for example, reading several numbers and comparing them after a simple calculation). To analyze such combinations in TEC-VQA, we compute the co-occurrence matrix over the capability labels. Each off-diagonal entry counts QA pairs in Fig. 1 annotated with the corresponding capability combination (for example, "2430" represents the number of questions requiring text recognition and comparative analysis simultaneously). We observe that numerical calculation frequently co-occurs with text recognition and comparative analysis, reflecting that images contain a wealth of statistical information, such as tabulated bills, receipts, or timetables, which often require reading, comparing values, and further calculation. Logical reasoning and spatial understanding appear less frequently overall but tend

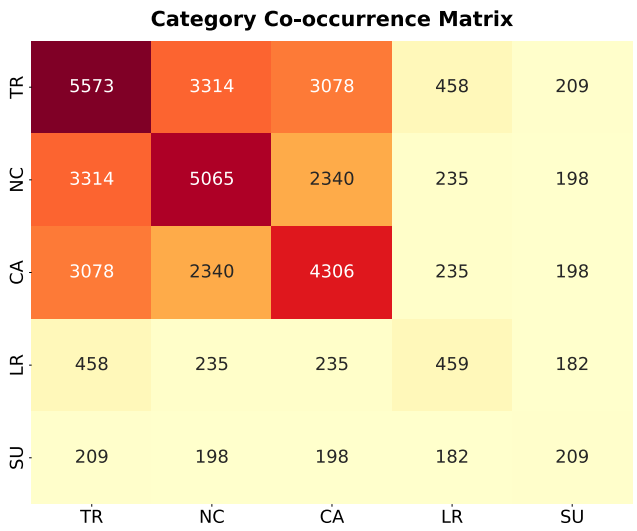



Figure 1. Heatmap of the TEC-VQA capability category co-occurrence matrix. Color intensity indicates the frequency of QA pairs annotated with each capability combination.

to co-occur with text recognition in more complex layout-dependent questions.

## B. Additional Quantitative Results

This appendix provides additional quantitative analyses that decompose the overall results in Section 5 into finer-grained sub-tasks and data subsets. We first present Document Parsing performance on pure text, tables, formulas, and reading order, and then provide a detailed breakdown of TEC-VQA by domain and language/resource level. We use the same metrics as in the main paper: Normalized Edit Distance (NED, Normalized Edit Distance), Tree Edit Distance-based Similarity (TEDS, Tree Edit Distance-based Similarity), Character Detection Matching metric (CDM, Character Detection Matching metric), and Bilingual Evaluation Understudy (BLEU, Bilingual Evaluation Understudy).

### B.1. Document Parsing Sub-Task Results

#### B.1.1. Pure Text

Table 5 shows clear cross-language trends across model families. Latin- and Chinese-based languages exhibit relatively low NED—strong models typically stay around 0.05–0.12 on EN, FIL, ID, and ZH—whereas Khmer, Lao, Thai, and Burmese remain considerably more challenging, with many baselines exceeding 0.80 and some pipeline systems reaching 0.96–0.98. Pipeline and expert models perform well on familiar scripts (e.g., POINTS-Reader at 0.05 on EN and 0.04 on ID) but degrade sharply on low-resource ones (e.g., 0.975 on LO and 0.981 on KM). General models show more stable cross-lingual performance: Qwen3-VL-32B maintains low NED across multiple high-resource lan-

guages, and Gemini 2.5-Pro achieves the best overall average (0.129) while reducing errors on difficult scripts such as Khmer (0.312), Lao (0.150), and Burmese (0.214). These results reflect pronounced and persistent performance gaps across languages in pure-text recognition.

#### B.1.2. Tables

Table 6 shows that table parsing performance varies substantially across languages and model families. For high-resource Latin and Chinese scripts, strong models achieve low text errors (e.g., NED around 0.05–0.12 for EN, ID, MS, ZH) and high structural fidelity (TEDS above 0.90 for several baselines). In contrast, table regions in Khmer and Lao remain challenging for nearly all systems, with NED often exceeding 0.60 and TEDS dropping below 0.35, reflecting frequent structural mismatches and cell-level OCR mistakes. Pipeline and expert models maintain strong structure-text consistency on familiar scripts—for instance, PaddleOCR-VL reaches TEDS 0.91–0.93 on EN/ID and NED around 0.06—but degrade sharply on low-resource scripts (e.g., TEDS 0.23–0.32 on KM/LO). General models exhibit more stable behavior across languages, with Qwen3-VL-32B and Gemini 2.5-Pro delivering the most balanced results: Gemini achieves TEDS above 0.92 on several high-resource languages and lowers NED to 0.40–0.41 on Khmer and Lao, outperforming other general models. Overall, the table results show a pronounced gap between high-resource and low-resource scripts, driven jointly by structural ambiguity and noise in complex table regions.

#### B.1.3. Formulas

Table 7 shows that formula parsing performance is relatively consistent across model families, with moderate variation between systems. Among pipeline models, PaddleOCR-VL achieves the strongest results (NED 0.230, BLEU 0.586), outperforming Dolphin-1.5 and MinerU2.5. Expert models exhibit comparable performance—dots.ocr and DeepSeek-OCR reach BLEU scores around 0.56–0.58 with NED near 0.27–0.29. General models show the widest spread: Qwen2.5-VL-72B and Qwen3-VL-32B deliver the best overall formula recognition (NED 0.223/0.212, BLEU 0.626/0.633), whereas InternVL3.5-38B lags notably behind. Overall, formula reconstruction remains challenging but stable across most baselines, with top-performing general models showing clear advantages in both structural and token-level accuracy.

#### B.1.4. Reading Order

Table 8 and Table 9 indicate that most models handle reading-order reconstruction reliably across high-resource scripts and standard page types. Languages such as EN, FIL, ID, and ZH typically show low NED—often around 0.07–0.15—demonstrating that sequential flow can be recovered accurately when script and document conventions are well supported. Larger errors are concentrated in a few low-resource languages, especially Khmer and Lao, where NED may rise to 0.45–0.65 for several baselines. A similar trend appears across page types: single-column formats such as books, exam papers, and slides are consistently easier, while multi-column newspapers and magazines introduce moderate increases in error. Overall, reading-order performance is strong for most languages and page types, with difficulty mainly arising in scripts and structures that deviate from common training distributions.

### B.2. TEC-VQA Detailed Breakdown

To highlight the impact of language-resource disparity, we group the 11 languages into high-resource and low-resource sets, following the discussion in the main text. High-resource languages include, for example, English, Chinese, Indonesian, and Malay, which have stronger OCR support and richer pre-training corpora. Low-resource languages include Lao, Khmer, and Burmese, among others.

Table 10 summarizes the group-level average TEC-VQA accuracy for these two sets. The average accuracies of MLLMs across the four languages in the high-resource group rank 1st, 2nd, 4th, and 5th among all languages, and their average accuracy (45.78%) is 2.5 times that of the low-resource group (17.45%). This reflects a severe lack of attention to low-resource languages among mainstream MLLMs, and a deficiency in multi-scenario recognition and comprehension capabilities for the text of low-resource languages.

## C. Experimental Details

This section describes evaluation metrics, baseline configurations, and implementation details used in our experiments.

### C.1. Baseline Configurations

#### C.1.1. Document parsing baselines

We evaluate three categories of document parsing baselines—pipeline models, expert models, and general models—under a unified protocol. Pipeline systems (MinerU2.5, Dolphin-1.5, MonkeyOCR-pro-1.2B/3B, PaddleOCR-VL) are run with their official checkpoints and default configurations. Expert models (POINTS-Reader, DeepSeek-OCR, dots.ocr) follow their official inference settings, using deterministic decoding when generative components are involved. General multimodal models (InternVL3.5-38B, Qwen2.5-VL-72B-Instruct, Qwen3-VL-32B-Instruct, GPT4o, Gemini2.5-Pro) also adopt deterministic, non-sampling inference (do_sample=False, temperature = 0.0). Maximum generation length is fixed per model—e.g., 32k tokens for Qwen2.5-VL-72B and 4096 tokens for InternVL—while closed-source models are evaluated with their default decoding settings.

| Method Type | Methods | EN | FIL | ID | KM | LO | MS | MY | PT | TH | VI | ZH | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Pipeline Models | **MinerU2.5** | 0.234 | 0.186 | 0.248 | 0.969 | 0.975 | 0.223 | 0.858 | 0.255 | 0.972 | 0.469 | 0.307 | 0.503 |
| | **Dolphin-1.5** | 0.061 | 0.065 | 0.063 | 0.962 | 0.932 | 0.068 | 0.688 | 0.039 | 0.701 | 0.108 | 0.097 | 0.321 |
| | **MonkeyOCR-pro-1.2B** | 0.059 | 0.07 | 0.241 | 0.985 | 0.985 | 0.248 | 0.806 | 0.24 | 0.987 | 0.878 | 0.091 | 0.51 |
| | **MonkeyOCR-pro-3B** | 0.054 | 0.057 | 0.121 | 0.971 | 0.974 | 0.097 | 0.756 | 0.059 | 0.972 | 0.721 | 0.087 | 0.435 |
| | **PaddleOCR-VL** | 0.044 | 0.046 | 0.036 | 0.966 | 0.964 | 0.041 | 0.787 | 0.039 | 0.081 | 0.172 | 0.083 | 0.268 |
| Expert Models | **POINTS-Reader** | 0.05 | 0.022 | 0.04 | 0.975 | 0.981 | 0.032 | 0.909 | 0.051 | 0.857 | 0.202 | 0.117 | 0.363 |
| | **DeepSeek-OCR** | 0.118 | 0.037 | 0.056 | 0.656 | 0.163 | 0.053 | 0.499 | 0.082 | 0.244 | 0.171 | 0.147 | 0.2 |
| | **dots.ocr** | 0.05 | 0.033 | 0.05 | 0.333 | 0.335 | 0.04 | 0.244 | 0.039 | 0.084 | 0.1 | 0.103 | 0.118 |
| General Models | **InternVL3.5-38B** | 0.264 | 0.413 | 0.414 | 0.962 | 0.97 | 0.419 | 0.787 | 0.453 | 0.943 | 0.577 | 0.761 | 0.605 |
| | **Qwen2.5-VL-72B-Instruct** | 0.064 | 0.057 | 0.057 | 0.922 | 0.897 | 0.054 | 0.761 | 0.043 | 0.094 | 0.069 | 0.118 | 0.256 |
| | **Qwen3-VL-32B-Instruct** | 0.055 | 0.053 | 0.056 | 0.925 | 0.505 | 0.057 | 0.65 | 0.04 | 0.113 | 0.076 | 0.086 | 0.223 |
| | **GPT4o** | 0.124 | 0.091 | 0.115 | 0.798 | 0.797 | 0.11 | 0.577 | 0.147 | 0.323 | 0.214 | 0.595 | 0.303 |
| | **Gemini2.5-Pro** | 0.068 | 0.086 | 0.079 | 0.312 | 0.15 | 0.072 | 0.336 | 0.051 | 0.102 | 0.098 | 0.13 | 0.129 |
| **Avg.** | | 0.096 | 0.094 | 0.121 | 0.826 | 0.741 | 0.116 | 0.666 | 0.118 | 0.498 | 0.297 | 0.209 | 0.326 |

Table 5. Pure-text region performance on SEA-Vision, reported as Normalized Edit Distance (NED↓) across 11 languages.

| Method Type | Methods | EN | | FIL | | ID | | KM | | LO | | MS | | MY | | PT | | TH | | VI | | ZH | | Avg. | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | NED | TEDS | NED | TEDS | NED | TEDS | NED | TEDS | NED | TEDS | NED | TEDS | NED | TEDS | NED | TEDS | NED | TEDS | NED | TEDS | NED | TEDS | NED | TEDS |
| Pipeline Models | **MinerU2.5** | 0.071 | 0.913 | 0.061 | 0.918 | 0.051 | 0.932 | 0.54 | 0.394 | 0.648 | 0.366 | 0.079 | 0.908 | 0.308 | 0.666 | 0.059 | 0.922 | 0.416 | 0.515 | 0.109 | 0.872 | 0.055 | 0.929 | 0.128 | 0.851 |
| | **Dolphin-1.5** | 0.13 | 0.816 | 0.139 | 0.798 | 0.141 | 0.803 | 0.767 | 0.196 | 0.866 | 0.12 | 0.156 | 0.796 | 0.476 | 0.467 | 0.127 | 0.82 | 0.656 | 0.305 | 0.143 | 0.821 | 0.241 | 0.71 | 0.23 | 0.72 |
| | **MonkeyOCR-pro-1.2B** | 0.083 | 0.896 | 0.119 | 0.863 | 0.106 | 0.874 | 0.71 | 0.254 | 0.764 | 0.281 | 0.14 | 0.836 | 0.468 | 0.524 | 0.13 | 0.856 | 0.511 | 0.52 | 0.188 | 0.808 | 0.128 | 0.835 | 0.196 | 0.789 |
| | **MonkeyOCR-pro-3B** | 0.095 | 0.895 | 0.132 | 0.855 | 0.094 | 0.893 | 0.676 | 0.272 | 0.763 | 0.237 | 0.133 | 0.842 | 0.431 | 0.553 | 0.121 | 0.851 | 0.541 | 0.432 | 0.18 | 0.806 | 0.093 | 0.886 | 0.189 | 0.792 |
| | **PaddleOCR-VL** | 0.059 | 0.924 | 0.076 | 0.911 | 0.056 | 0.922 | 0.69 | 0.232 | 0.671 | 0.321 | 0.083 | 0.899 | 0.413 | 0.504 | 0.064 | 0.914 | 0.132 | 0.83 | 0.059 | 0.927 | 0.059 | 0.924 | 0.116 | 0.86 |
| Expert Models | **POINTS-Reader** | 0.076 | 0.905 | 0.073 | 0.915 | 0.061 | 0.921 | 0.909 | 0.09 | 0.863 | 0.15 | 0.078 | 0.914 | 0.737 | 0.245 | 0.09 | 0.887 | 0.79 | 0.204 | 0.154 | 0.837 | 0.056 | 0.932 | 0.201 | 0.785 |
| | **DeepSeek-OCR** | 0.192 | 0.778 | 0.113 | 0.862 | 0.092 | 0.881 | 0.911 | 0.079 | 0.489 | 0.54 | 0.108 | 0.874 | 0.437 | 0.546 | 0.12 | 0.842 | 0.386 | 0.606 | 0.221 | 0.763 | 0.106 | 0.866 | 0.198 | 0.779 |
| | **dots.ocr** | 0.126 | 0.852 | 0.09 | 0.874 | 0.085 | 0.884 | 0.5 | 0.443 | 0.622 | 0.364 | 0.113 | 0.874 | 0.397 | 0.557 | 0.112 | 0.857 | 0.22 | 0.732 | 0.136 | 0.851 | 0.078 | 0.908 | 0.151 | 0.823 |
| General Models | **InternVL3.5-38B** | 0.612 | 0.626 | 0.638 | 0.53 | 0.688 | 0.442 | 0.847 | 0.121 | 0.852 | 0.121 | 0.666 | 0.473 | 0.68 | 0.387 | 0.734 | 0.372 | 0.832 | 0.122 | 0.755 | 0.307 | 0.71 | 0.389 | 0.703 | 0.41 |
| | **Qwen2.5-VL-72B-Instruct** | 0.167 | 0.873 | 0.097 | 0.871 | 0.099 | 0.897 | 0.816 | 0.186 | 0.634 | 0.413 | 0.095 | 0.898 | 0.52 | 0.493 | 0.133 | 0.89 | 0.168 | 0.856 | 0.12 | 0.905 | 0.152 | 0.872 | 0.178 | 0.836 |
| | **Qwen3-VL-32B-Instruct** | 0.112 | 0.895 | 0.085 | 0.9 | 0.078 | 0.912 | 0.76 | 0.253 | 0.511 | 0.586 | 0.081 | 0.911 | 0.39 | 0.599 | 0.097 | 0.905 | 0.131 | 0.87 | 0.086 | 0.921 | 0.093 | 0.911 | 0.136 | 0.864 |
| | **GPT4o** | 0.199 | 0.761 | 0.125 | 0.85 | 0.172 | 0.77 | 0.534 | 0.358 | 0.577 | 0.376 | 0.176 | 0.823 | 0.332 | 0.57 | 0.311 | 0.65 | 0.271 | 0.667 | 0.248 | 0.744 | 0.275 | 0.603 | 0.248 | 0.702 |
| | **Gemini2.5-Pro** | 0.121 | 0.885 | 0.072 | 0.93 | 0.072 | 0.933 | 0.404 | 0.552 | 0.37 | 0.567 | 0.154 | 0.892 | 0.238 | 0.751 | 0.086 | 0.872 | 0.12 | 0.815 | 0.066 | 0.941 | 0.073 | 0.922 | 0.12 | 0.872 |
| **Avg.** | | 0.157 | 0.848 | 0.140 | 0.852 | 0.138 | 0.851 | 0.697 | 0.264 | 0.664 | 0.342 | 0.159 | 0.842 | 0.448 | 0.528 | 0.168 | 0.818 | 0.398 | 0.575 | 0.190 | 0.808 | 0.163 | 0.822 | 0.215 | 0.776 |

Table 6. Table-region performance (TEDS↑ / NED↓) across 11 languages.

| Method Type | Methods | NED | BLEU |
|---|---|---|---|
| Pipeline Models | **MinerU2.5** | 0.314 | 0.551 |
| | **Dolphin-1.5** | 0.332 | 0.491 |
| | **MonkeyOCR-pro-1.2B** | 0.276 | 0.538 |
| | **MonkeyOCR-pro-3B** | 0.277 | 0.539 |
| | **PaddleOCR-VL** | 0.230 | 0.586 |
| Expert Models | **POINTS-Reader** | 0.285 | 0.560 |
| | **DeepSeek-OCR** | 0.291 | 0.581 |
| | **dots.ocr** | 0.266 | 0.567 |
| General Models | **InternVL3.5-38B** | 0.521 | 0.329 |
| | **Qwen2.5-VL-72B-Instruct** | 0.223 | 0.626 |
| | **Qwen3-VL-32B-Instruct** | 0.212 | 0.633 |
| | **GPT4o** | 0.379 | 0.451 |
| | **Gemini2.5-Pro** | 0.267 | 0.578 |
| **Avg.** | | 0.298 | 0.541 |

Table 7. Formula-region performance (NED↓ / BLEU↑) across 11 languages.

### C.1.2. TEC-VQA baselines

For TEC-VQA, we compare both non-generative baselines (for example, classification or retrieval-based models) and generative MLLM-based baselines.

**Prompt template.** All MLLM-based TEC-VQA baselines share a unified prompt template. We use a short system instruction plus an image-conditioned user query. The core user-side template is:

> *You are a helpful AI assistant that answers questions based on the given image. Please follow these guidelines:*
> *1. Answer the question directly and accurately based on what you see in the image, without adding any extra information or explanations.*
> *2. If the question is in a specific language, answer in the same language.*
> *3. Keep your answer concise and relevant to the question.*
> *4. If you cannot find the answer in the image, respond with "No relevant information can be found from the*

| Method Type | Methods | EN | FIL | ID | KM | LO | MS | MY | PT | TH | VI | ZH | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Pipeline Models | **MinerU2.5** | 0.146 | 0.072 | 0.136 | 0.571 | 0.566 | 0.122 | 0.32 | 0.133 | 0.638 | 0.197 | 0.208 | 0.27 |
| | **Dolphin-1.5** | 0.124 | 0.063 | 0.106 | 0.337 | 0.409 | 0.105 | 0.29 | 0.096 | 0.409 | 0.097 | 0.119 | 0.188 |
| | **MonkeyOCR-pro-1.2B** | 0.161 | 0.092 | 0.15 | 0.309 | 0.342 | 0.139 | 0.28 | 0.151 | 0.447 | 0.461 | 0.151 | 0.247 |
| | **MonkeyOCR-pro-3B** | 0.161 | 0.087 | 0.123 | 0.331 | 0.327 | 0.11 | 0.255 | 0.135 | 0.494 | 0.271 | 0.15 | 0.221 |
| | **PaddleOCR-VL** | 0.118 | 0.074 | 0.087 | 0.245 | 0.309 | 0.085 | 0.213 | 0.097 | 0.071 | 0.114 | 0.102 | 0.13 |
| Expert Models | **POINTS-Reader** | 0.118 | 0.044 | 0.092 | 0.65 | 0.601 | 0.088 | 0.503 | 0.111 | 0.552 | 0.164 | 0.14 | 0.265 |
| | **DeepSeek-OCR** | 0.169 | 0.069 | 0.117 | 0.452 | 0.105 | 0.121 | 0.251 | 0.141 | 0.222 | 0.19 | 0.131 | 0.182 |
| | **dots.ocr** | 0.142 | 0.072 | 0.132 | 0.099 | 0.201 | 0.115 | 0.094 | 0.127 | 0.086 | 0.142 | 0.176 | 0.122 |
| General Models | **InternVL3.5-38B** | 0.304 | 0.299 | 0.395 | 0.393 | 0.435 | 0.362 | 0.291 | 0.443 | 0.59 | 0.49 | 0.607 | 0.411 |
| | **Qwen2.5-VL-72B-Instruct** | 0.189 | 0.086 | 0.139 | 0.61 | 0.306 | 0.121 | 0.366 | 0.133 | 0.08 | 0.12 | 0.187 | 0.204 |
| | **Qwen3-VL-32B-Instruct** | 0.142 | 0.064 | 0.12 | 0.497 | 0.203 | 0.112 | 0.302 | 0.107 | 0.075 | 0.109 | 0.126 | 0.165 |
| | **GPT4o** | 0.19 | 0.076 | 0.165 | 0.5 | 0.455 | 0.144 | 0.239 | 0.161 | 0.151 | 0.185 | 0.4 | 0.217 |
| | **Gemini2.5-Pro** | 0.183 | 0.092 | 0.165 | 0.117 | 0.064 | 0.133 | 0.115 | 0.107 | 0.089 | 0.128 | 0.202 | 0.126 |
| **Avg.** | | 0.165 | 0.092 | 0.148 | 0.393 | 0.333 | 0.135 | 0.271 | 0.149 | 0.300 | 0.205 | 0.208 | 0.211 |

Table 8. Reading-order NED by language.

| Method Type | Methods | Academic paper | Book | Exam paper | Magazine | Newspaper | Note | Research report | Slide | Textbook | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Pipeline Models | **MinerU2.5** | 0.290 | 0.306 | 0.268 | 0.235 | 0.233 | 0.218 | 0.344 | 0.205 | 0.307 | 0.267 |
| | **Dolphin-1.5** | 0.145 | 0.179 | 0.231 | 0.187 | 0.199 | 0.158 | 0.203 | 0.175 | 0.221 | 0.189 |
| | **MonkeyOCR-pro-1.2B** | 0.208 | 0.233 | 0.273 | 0.277 | 0.336 | 0.198 | 0.271 | 0.182 | 0.282 | 0.251 |
| | **MonkeyOCR-pro-3B** | 0.173 | 0.204 | 0.235 | 0.248 | 0.303 | 0.176 | 0.248 | 0.171 | 0.265 | 0.225 |
| | **PaddleOCR-VL** | 0.076 | 0.118 | 0.149 | 0.157 | 0.187 | 0.096 | 0.145 | 0.105 | 0.167 | 0.133 |
| Expert Models | **POINTS-Reader** | 0.233 | 0.304 | 0.237 | 0.233 | 0.345 | 0.225 | 0.346 | 0.163 | 0.302 | 0.265 |
| | **DeepSeek-OCR** | 0.155 | 0.189 | 0.182 | 0.212 | 0.285 | 0.127 | 0.173 | 0.144 | 0.211 | 0.186 |
| | **dots.ocr** | 0.102 | 0.095 | 0.109 | 0.198 | 0.230 | 0.050 | 0.101 | 0.110 | 0.154 | 0.127 |
| General Models | **InternVL3.5-38B** | 0.497 | 0.470 | 0.328 | 0.488 | 0.612 | 0.268 | 0.484 | 0.179 | 0.421 | 0.416 |
| | **Qwen2.5-VL-72B-Instruct** | 0.189 | 0.202 | 0.190 | 0.227 | 0.289 | 0.151 | 0.222 | 0.149 | 0.242 | 0.207 |
| | **Qwen3-VL-32B-Instruct** | 0.124 | 0.151 | 0.189 | 0.174 | 0.239 | 0.123 | 0.185 | 0.128 | 0.200 | 0.168 |
| | **GPT4o** | 0.199 | 0.232 | 0.191 | 0.264 | 0.368 | 0.136 | 0.263 | 0.133 | 0.227 | 0.223 |
| | **Gemini2.5-Pro** | 0.094 | 0.095 | 0.151 | 0.166 | 0.222 | 0.071 | 0.091 | 0.121 | 0.174 | 0.132 |
| **Avg.** | | 0.191 | 0.214 | 0.210 | 0.236 | 0.296 | 0.154 | 0.237 | 0.151 | 0.244 | 0.215 |

Table 9. Reading-order NED by page type.

Table 10. TEC-VQA accuracy for high-resource vs. low-resource language groups.

| Language group | Avg. accuracy (%) | Languages |
|---|---|---|
| High-resource | 45.78 | EN, ZH, ID, MS |
| Low-resource | 17.45 | LO, KM, MY, FIL PT, TH, VI |

*image" or equivalent in the question's language.*
*5. Do not make assumptions or provide information not visible in the image.*
*Question (in {LANG}): {QUESTION}*
*Answer (in {LANG}, be concise):*

Here {LANG} is replaced by the target language name (for example, "English", "Vietnamese"), and {QUESTION} is the TEC-VQA question text. The image is passed as the visual input channel supported by each MLLM. An illustrative example of the full prompt, including a pseudo document image and question, is shown in Figure 2.

Figure 2. Example prompt used for Multimodal Large Language Model (MLLM) TEC-VQA baselines. The document image and question are replaced with actual samples at inference time. (Pseudo example; not from the released dataset.)

For non-generative baselines that directly classify or retrieve answers from a candidate set, we do not use natural-language prompts and instead follow the original model configuration.

**Sampling and decoding settings.** All TEC-VQA baselines use a unified deterministic setup: no sampling (do_sample=False, temperature=0), and model-specific maximum output length (e.g., 32k for Qwen2.5-VL-72B, 4k for InternVL3.5-38B). Closed-source models run with their official default decoding settings.

## D. Annotation and Verification Protocols

In Section 3.2.2 and Section 4 of the main text, we described a hybrid pipeline that combines automatic processing with native-speaker verification for both Document Parsing and TEC-VQA. This part focuses on the four-step human verification protocol used to refine TEC-VQA question–answer (QA) pairs.

### D.1. TEC-VQA Human Verification Checklist

After automatic generation, every TEC-VQA QA pair is passed through a four-step human verification checklist, which corresponds to the four steps described in Section 3.2.2 of the main paper:

1. **Answerability and usefulness.** Annotators first decide whether the question is answerable using only the visible content in the image. Questions that require external knowledge, depend on hidden context, or are overly trivial (e.g., reading a single obvious word with no value for evaluation) are removed.
2. **Question clarity and answer normalization.** For retained QA pairs, annotators refine the wording of the question to be clear and unambiguous, and normalize the answer format. This includes standardizing number formats (decimal points, thousands separators), making units explicit where needed, and using consistent conventions for dates and currencies across the dataset.
3. **Strict alignment with visible text.** Annotators then check that all entities appearing in the question and answer (names, numbers, dates, monetary amounts) are present in the visible text of the image and that the answer can be derived solely from that text (possibly with simple reasoning or calculation). If the MLLM misreads a value, hallucinates content that does not exist in the image, or mixes languages, annotators correct the QA pair or discard it if no safe fix exists.
4. **Capability labeling.** Finally, annotators assign capability labels to each question from the following set: text recognition, numerical calculation, comparative analysis, logical reasoning, and spatial understanding. Multi-label annotations are allowed for questions that require several skills simultaneously (e.g., reading multiple prices and comparing the total).

Table 11 provides an illustrative example of how one QA pair evolves from the original MLLM output to the final, verified version under this checklist.

**Operational definitions of capability labels.** For consistency, annotators use short operational rules when assigning capability labels:

- **Text recognition:** the answer is obtained by directly reading and copying a span of visible text (e.g., a store name, product name, or single field).
- **Numerical calculation:** the answer requires explicit arithmetic (addition, subtraction, multiplication, division) or unit conversion using numbers in the image.
- **Comparative analysis:** the answer requires comparing multiple values or text snippets (e.g., choosing the largest price or earliest date).
- **Logical reasoning:** the answer requires combining multiple pieces of information with a condition or rule (e.g., applying a discount rule or free-item policy).
- **Spatial understanding:** the answer depends on layout or positional relations in the image (e.g., "the item listed at the top of the table").

**Document Parsing verification.** Document Parsing labels use the same tool and logging infrastructure but a simpler review protocol: annotators correct semantic categories, and reading order for each page, and a subset of pages is double-checked by a second annotator. Pages containing privacy-sensitive or politically sensitive content are removed before release.

## E. Limitations and License

### E.1. Limitations

SEA-Vision is a step toward comprehensive multilingual document and scene text understanding, but it still has several limitations.

First, coverage remains uneven. Some document types and languages are underrepresented, especially highly specialized formats and very low-resource languages, which may bias both training and evaluation. Second, all questions are restricted to those answerable from a single image; cross-document reasoning and the use of external knowledge are not evaluated. Third, despite the multi-stage quality control pipeline, there may be residual noise in OCR text, structural annotations, and question–answer pairs. Fourth, evaluation relies mainly on automatic metrics and does not address interpretability, reasoning transparency, or human preference. Finally, while the benchmark exposes large performance gaps for low-resource languages and complex reasoning skills, it does not yet close them. Extending SEA-Vision and combining it with complementary resources are natural directions for future work.

### E.2. Dataset License and Ethics

We plan to release SEA-Vision under the Creative Commons Attribution–NonCommercial 4.0 (CC BY-NC 4.0, Creative Commons Attribution–NonCommercial 4.0) license. This allows users to share and adapt the dataset for non-commercial purposes with proper attribution, while restricting direct commercial use.

All documents are collected and processed with privacy and legal considerations in mind. No personally identifi-

Table 11. Illustrative TEC-VQA example showing one question–answer pair at three stages: original MLLM generation, intermediate human checking, and the final version that satisfies all checklist items. The example is based on a receipt-style image; text here is illustrative.

| Stage | Question | Answer | Issues / actions |
|---|---|---|---|
| Original generation | What is the total? | 12.0 | Question is underspecified (no currency or context); answer format does not match the receipt ("12.80 SGD" printed on the image). |
| Intermediate check | What is the total price on the receipt (in dollars)? | 12.50 | Question is clearer, but the amount is misread ("12.80" on the image). Fails the strict alignment check; needs correction. |
| Final version | What is the total amount on the receipt (in SGD)? | 12.80 | Question specifies the context and currency; answer exactly matches the visible text. Capability labels: text recognition + numerical calculation. |

able information is included in the released data; many documents are synthetic or originate from public sources. For real documents (for example, public forms or published materials), we only retain content that is public or permitted for research use. The accompanying documentation will describe data sources, processing steps, and any usage constraints. Users are required not to attempt to reconstruct sensitive information or to deanonymize any content.


Figure 3. Representative document page samples for English (EN), Filipino (FIL), and Indonesian (ID) in SEA-Vision.

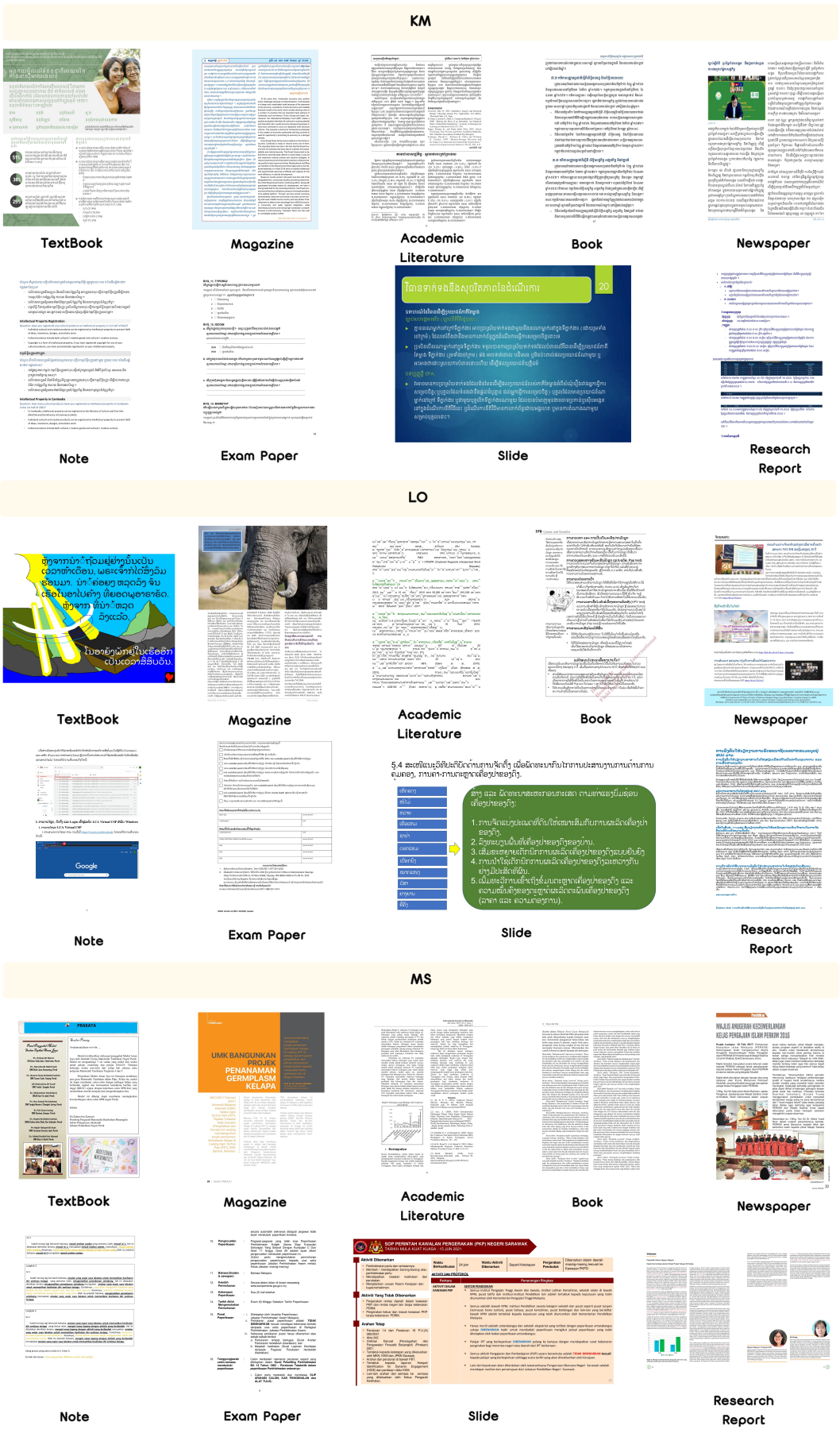



Figure 4. Representative document page samples for Khmer (KM), Lao (LO), and Malay (MS) in SEA-Vision.

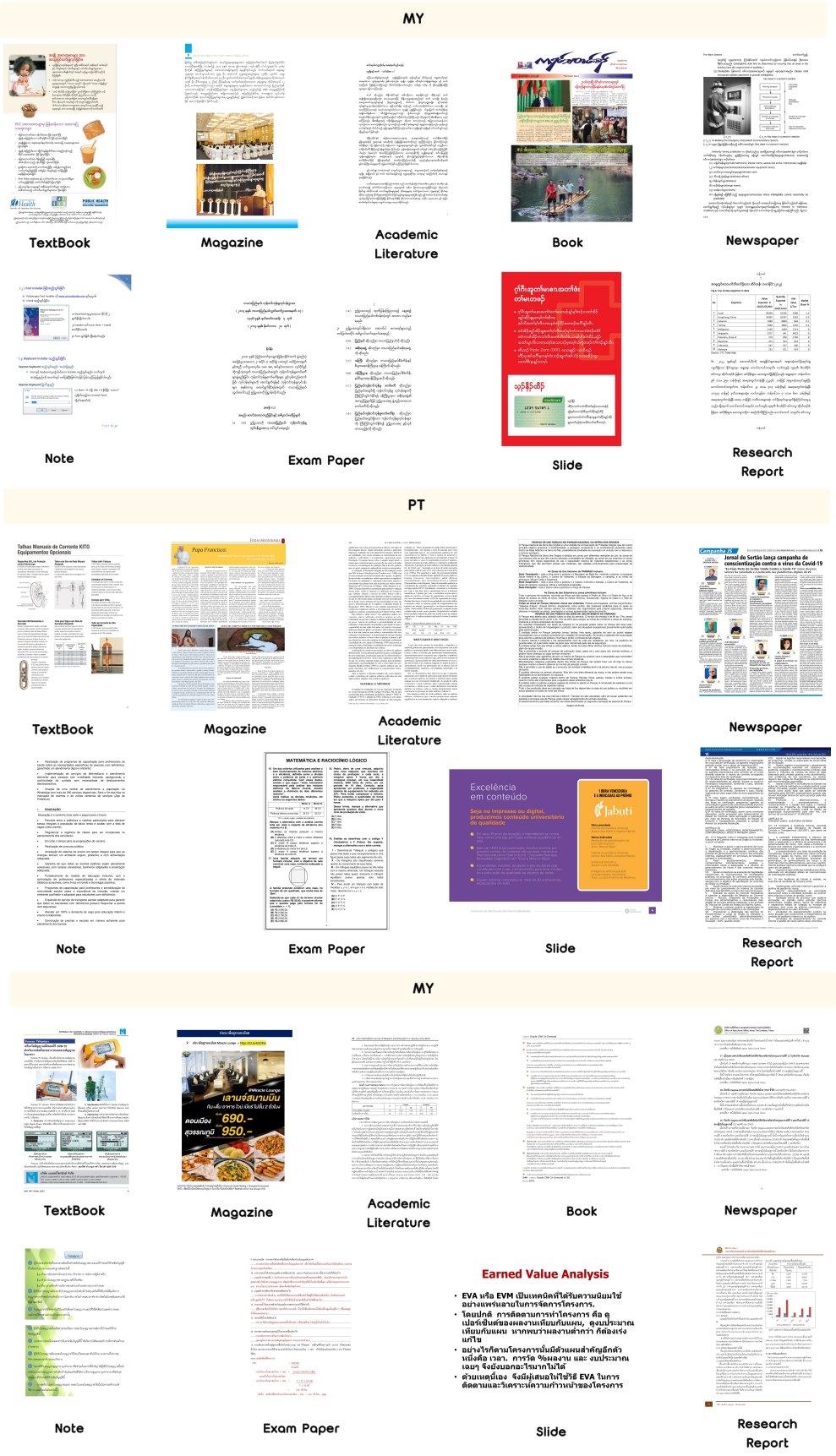



Figure 5. Representative document page samples for Burmese (MY), Portuguese (PT), and Thai (TH) in SEA-Vision.

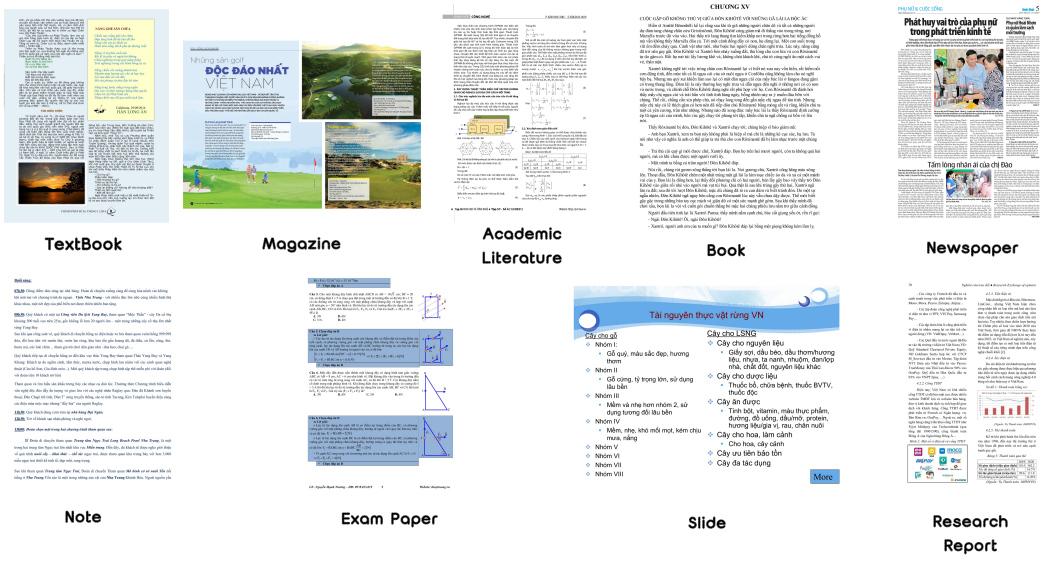



Figure 6. Representative document page samples for Vietnamese (VI) and Chinese (ZH) in SEA-Vision.

**Q:** Siapa pembangun/pemilik projek ini yang disenaraikan di papan iklan?
**A:** SCP TRADERS SQUARE SDN. BHD. (ID Syarikat: 100), Kumpulan SCP

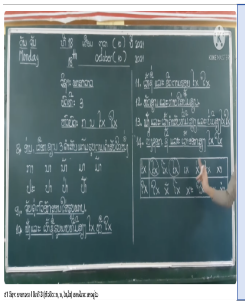
**Q:** Who is the developer/owner of this project listed on the billboard?
**A:** SCP TRADERS SQUARE SDN. BHD. (Company Number: 100), SCP Group

a. Malay-*public spaces*


**Q:** Có những mức giảm giá nào cho nhóm ăn uống (5 người trở lên) tại các nhà hàng được liệt kê trên tờ rơi bên trái?
**A:** Giảm giá 10%

**Q:** On the menu on the left, what discount can a group (5 people or more) enjoy when dining here?
**A:** 10% discount

b. Vietnamese-*consumer places*



**Q:**
ວ ັນທ ີ່ໃດທ ີ ່ຂຽນໄວ ້ໃນ ກະດ ານດ ໍ ?
**A:** ວ ັນຈ ັນທ ີ 18 ຕ ,ລ າ 2021

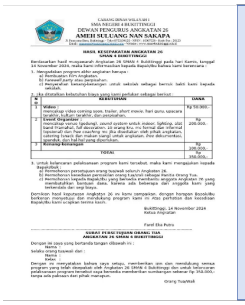
**Q:** What date is written on the blackboard?
**A:** Monday, October 18, 2021

c. Lao-work & study


**Q:**
น ้ำหน ักสุทธ ิของบรรจ ุภ ัณฑ ์ผล ิตภ ัณฑ ์ค ือเท าไ ?
**A:** 500 กร ัม

**Q:** What is the net weight of the product packaging?
**A:** 500g

d. Thai-*daily products*



**Q:** Siapa yang menandatangani dokumen ini sebagai ketua kelas? Tanggal berapa penandatanganannya?
**A:** Farruel Eka Putra tanggal 14 November 2024.

**Q:** Who signed this document in his capacity as the class chairperson? When was the signing date?
**A:** Farul Eka Putra, the class chairperson on November 14, 2024

e. Indonesian-*documents& IDs*


**Q:** Quantas reuniões estavam agendadas para 24 de outubro?
**A:** 2

**Q:** How many meetings were listed on October 24th?
**A:** 2

f. Portuguese-*digital displays*

Figure 7. Representative TEC-VQA examples across multiple Southeast Asian languages.