1. Appendix

1.1. Qualitative Analysis

1.1.1. Comparative Model Performance

Figure 2 and Figure **??** illustrate representative failure cases of baseline methods and how our text grounding approach overcomes these limitations. The OCR + LLM baseline struggles with stylized text where character boundaries are ambiguous or when text is heavily integrated with artistic elements. Nova Pro and Qwen can locate text regions effectively but cannot output precise bounding boxes. DeepSeek OCR struggles with differentiating between text and object elements within the same image as shown in Figure **??**. In contrast, our fine-tuned Florence-2 text grounding model successfully localizes text across these challenging scenarios, benefiting from its vision-language architecture that jointly reasons about visual appearance and semantic context.

1.1.2. Multilingual Text Detection

Figure 1 demonstrates our model's multilingual capabilities across diverse scripts.

Our vocabulary extension and multilingual finetuning strategy enables the Florence-2 model to effectively process Japanese, Chinese, Korean, Arabic, and Indic scripts without requiring script-specific detection pipelines. Critically, the model accepts word-level language specifications, allowing it to handle mixed-script images where different text regions may be in different languages. This represents a significant advantage over traditional OCR systems, which typically accept only image-level language cues and struggle with multilingual or mixed-script content. The model maintains consistent performance across languages, accurately localizing both title text and secondary text elements regardless of script complexity, character density, or text directionality. This unified approach eliminates the need for language-specific pre-processing or separate detection models for different writing systems.

1.2. Current Limitations

Despite strong overall performance, our text grounding approach exhibits specific failure modes that warrant further investigation.

Stylized Title Text with Artistic Effects. Figure 3 demonstrates challenges with highly stylized title text where letterforms incorporate artistic embellishments and visual effects. In the first example, the title "PROVA FINAL" features heavily stylized characters (especially the "A") that create ambiguity in character boundaries. The second example shows "THE HAUNTING" where the bounding boxes for each word are affected when characters are placed apart from each other. Furthermore, our current implementation uses axis-aligned bounding boxes, which cannot accurately capture slanted or curved text regions as shown in Figure 2. Future work should explore quadrilateral or polygon annotations to address these limitations.

Future Directions. These limitations highlight opportunities for future improvements, including specialized augmentation strategies for decorative text, explicit modeling of artistic typography effects, and potentially incorporating perceptual grouping principles to better handle text with complex visual treatments. Training with synthetic data that systematically varies decorative effects, color splits, and shadow styles may help the model develop more robust representations of stylized text boundaries.

Figure 1. Multilingual text grounding examples. Japanese,Chinese characters and Korean characters. The model successfully detects text across diverse scripts without language-specific tuning, with word-level language specification enabling robust mixed-script handling.

Figure 2. Comparative performance across baseline methods. Text grounding (ours) successfully detects stylized and artistic text where OCR + LLM, Nova Pro, and DeepSeek OCR fail or produce incomplete predictions.

Figure 3. Failure cases for the text grounding model. Left: Highly stylized character. Right: Far-placed characters and vertically attached text