CodeMMR: Bridging Natural Language, Code, and Image for Unified Retrieval

Supplementary Material

A. Dataset Statistics

Data Schema. Each training instance consists of six keys: `qry`, `qry_img_path`, `pos_text`, `pos_img_path`, `neg_text`, and `neg_img_path`. The `qry` field contains the structured textual component (natural language and/or code), while `qry_img_path` specifies the associated image if available. Positive and negative examples are defined through their respective text and image paths, enabling contrastive multimodal learning. Evaluation instances follow a similar structure with four keys: `qry_text`, `qry_img_path`, `tgt_text`, and `tgt_img_path`. Different retrieval directions are instantiated by populating the appropriate modalities, such as image-to-code ($q_i \rightarrow r_c$), code-to-image ($q_c \rightarrow r_i$), or text-to-code ($q_t \rightarrow r_c$). The special `[image]` token in text fields explicitly marks multimodal inputs, ensuring consistent processing across all retrieval configurations. Comprehensive dataset statistics are reported in Table 1.

Dataset Token Length Analysis. Figure 7 presents the average token lengths of training and test samples across all datasets. We observe that the token-length distributions of the train and test splits are highly consistent, indicating no significant distributional drift between them. Most datasets have moderate average lengths around several hundred tokens, suggesting that they can be efficiently processed under typical model context limits. However, a few datasets exhibit exceptionally long sequences: Sketch2Code, MMSVG-Illustration, and MMSVG-Icon contain substantially longer textual representations (over 10k tokens on average in Sketch2Code and MMSVG-Illustration). Such long sequences pose challenges for multimodal retrieval models when the token budget is limited (e.g., 256 or 512 tokens), potentially leading to information truncation and degraded retrieval performance. Interestingly, despite some datasets such as WebSight, Web2Code, and Chart2Code also having average lengths slightly beyond the 256-token range, our retrieval model maintains strong performance on these sources. This indicates that our training strategy enables robust generalization even when facing inputs that exceed the nominal token limit. The model effectively captures the cross-modal alignment without overfitting to input length, suggesting that the learned representation scales gracefully with content complexity.

SVG Example. As shown in Figure 6, we illustrate an example of an SVG image and its corresponding XML-based code representation. The SVG format encodes visual shapes using textual geometric primitives such as paths, curves, and color attributes. Although the image and its underlying code describe the same content, their modalities differ drastically — the visual domain emphasizes spatial patterns and holistic structure, whereas the XML text focuses on low-level coordinate and syntax tokens. This discrepancy makes it extremely challenging for multimodal retrieval systems to establish direct correspondence between visual and textual cues. The mapping from complex XML sequences to perceptually coherent image features is highly non-linear, and minor code variations may yield visually indistinguishable results. Consequently, learning effective cross-modal alignment between SVG code and rendered images remains a difficult problem, limiting retrieval accuracy and representation consistency in multimodal tasks.

SVG Image	Code
	<svg xmlns=\"http://www.w3.org/2000/svg\" viewBox=\"0.0 0.0 200.0 200.0\" height=\"200.0px\" width=\"200.0px\"><path fill=\"currentColor\" fill-opacity=\"1.0\" filling=\"0\" d=\"M87.5 141.66659545898438 L33.333396911621094 141.66659545898438 L33.333396911621094 125.0 L87.5 125.0 A29.16659927368164 29.166599273 68164 0.0 1 1 60.18339920043945 164.41659545898438 L75.79160308837891 158.55839 53857422 A12.5 12.5 0.0 1 0 87.5 141.66659545898438 Z M41.666603088378906 91.6 666030883789 L154.16659545898438 91.66660308837891 A29.16659927368164 29.166599 27368164 0.0 1 1 126.84999847412111 131.08338928222656 L142.45840454101562 125. 22499084472656 A12.5 12.5 0.0 1 0 154.16659545898438 108.33339691162111 L41.666 603088378906 108.33339691162111 A25.0 25.0 0.0 0 1 41.666603088378906 58.333396 911621094 L112.5 58.333396911621094 A12.5 12.5 0.0 1 0 100.79159545898438 41.4 4160079956055 L85.18340301513672 35.59159851074219 A29.17500114440918 29.17500 114440918 0.0 0 1 141.66659545898438 45.833396911621094 A29.16659927368164 29. 16659927368164 0.0 0 1 112.5 75.0 L41.666603088378906 75.0 A8.333398818969727 8.333398818969727 0.0 1 0 41.666603088378906 91.66660308837891 Z\"></path></svg>

Figure 6. SVG example.

B. Additional Results

Multi-modal code RAG exhibits stronger robustness and better performance over fine-tuned VLM. We evaluate multiple vision-language models on the task of generating webpage code from a given screenshot on the WebCode2M-Mid dataset. Figure 8 shows the result of **LLaVA-7B**, which captures the basic list structure and hyperlink styling, but completely misses the header bar and overall page layout, rendering only a plain bulleted list without any container or background styling. Figure 9 presents **LLaVA-13B**, which correctly reproduces the header and the search item list within a styled container, but loses hyperlink formatting and applies an overly plain gray background, with no color differentiation for the section title. Figure 10 illustrates **LLaVA-13B + CodeMMR (ours)**, which integrates RAG with a code-aware multimodal retriever. The generated page closely replicates the dark header bar, the teal-colored "Related Searches" title, and the vertically spaced list items, achieving the highest visual fidelity among all evaluated models. Figure 11 shows **WebCoder** [7], a ViT-based model fine-tuned on WebCode2M, which recovers

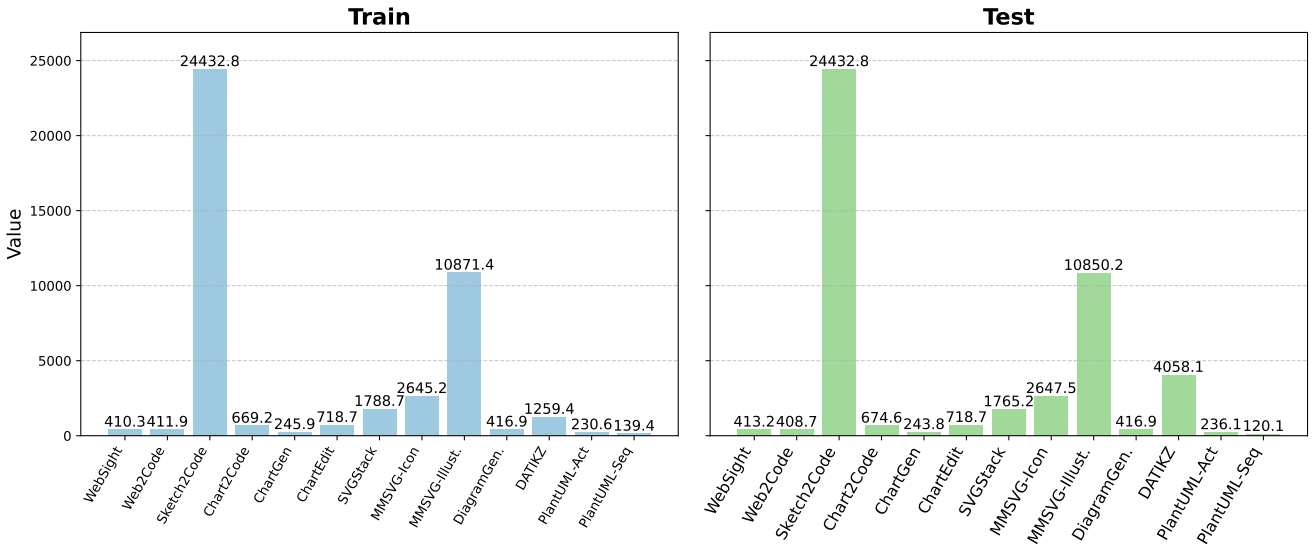

Figure 7. Average token length distribution across datasets in train/test splits.

the section title and list items but renders them in a horizontally misaligned layout, losing the vertical list structure and introducing irrelevant footer text. Figure 12 presents the **Ground Truth**, i.e., the original rendered webpage, featuring a dark navigation header, a teal "Related Searches" heading, and a clean vertically stacked list of search terms with proper spacing and footer links. Overall, our method LLaVA-13B + CodeMMR achieves the best reconstruction quality, closely matching the ground truth in layout, typography, and hierarchical organization, demonstrating the effectiveness of retrieval-enhanced multimodal reasoning for code generation from web screenshots.

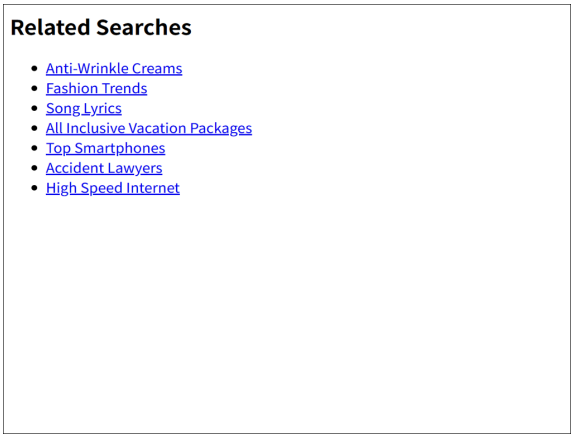

Figure 8. Qualitative comparison of web page code generation on WebCode2M-Mid: LLaVA-7B.

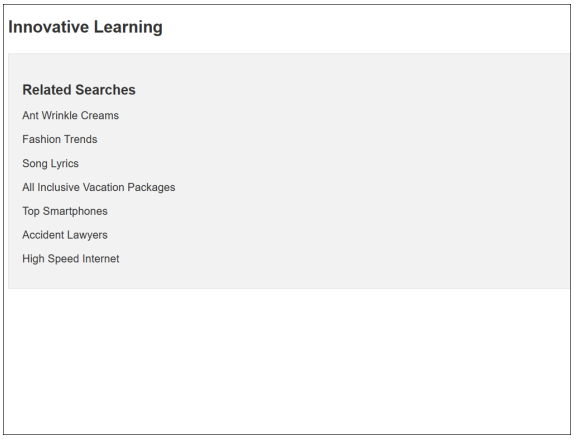

Figure 9. Qualitative comparison of web page code generation on WebCode2M-Mid: LLaVA-13B.

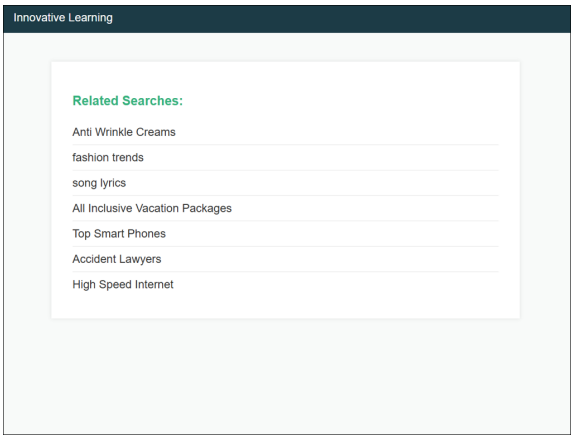

Figure 10. Qualitative comparison of web page code generation on WebCode2M-Mid: LLaVA-13B + CodeMMR.

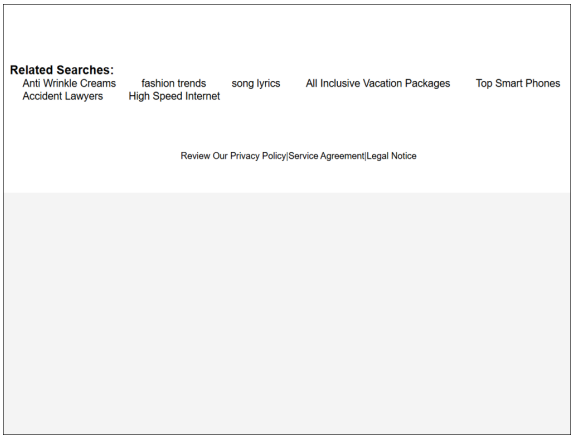

Figure 11. Qualitative comparison of web page code generation on WebCode2M-Mid: WebCoder.

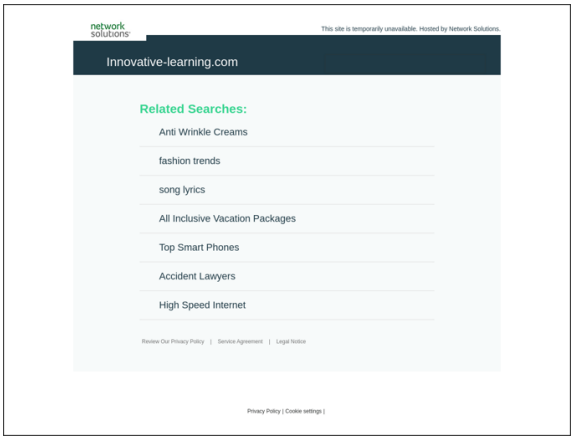

Figure 12. Qualitative comparison of web page code generation on WebCode2M-Mid: Ground Truth.