A. Prompts Used in the Dataset Construction Pipeline

A.1. Intention and Description Annotation

You are an excellent document analyst. I will give you a document image. Your task is to carefully (1) infer its intention, and (2) provide a detailed description of its content.
Instructions:
- For the intention, provide 1-2 high-level sentences about the document's purpose, goal, or the message it aims to convey. Please don't include visual details.
- For the description, write 2-3 sentences giving a clear, objective, and factual content description of the document. Focus on what is present in the document, without inferring beyond the visible elements.
- Please strictly follow this JSON format for your output, {"intention": " ", "description": " "}.
- Do not include ‵‵‵ or ‵‵‵json in your answer.
- Do not explain or justify your answer.

A.2. HTML/CSS Document Code Generation

You are a professional graphic designer. Your task is to create a high-quality document using HTML and CSS.
Instructions:
- Create the specified document as HTML/CSS.
- Do not use Javascript.
- Do not use any frameworks or libraries.
- Besides necessary structural tags (`<html>`, `<head>`, `<body>`, etc.), use only `<div>`, `<img>`, and `<p>` tags.
- All text must appear in a `<p>` tag.
- All images/icons/logos must appear in an `<img>` tag.
- The `<img>` tag must have both src and alt fields.
- For the src of all `<img>` tags, use only this source: "https://picsum.photos/W/H", by replacing "W" and "H" in the URL with the desired width and height. For example, "https://picsum.photos/200/300" for an image of width=200 and height=300. Make sure the width/height attributes of the `<img>` tag match the dimensions in the URL.
- In the alt text of all `<img>` tags, write a detailed description of what the image should be. This description will be used to search or generate an appropriate image, so the more detailed, the better!
- Do not use placeholder shapes for icons or logos.
- Do not include an image in "background: url(...)".
- Make the `<body>` tag a fixed size based on the document size.
- `<body>` and `<div>` tags must use either "display: flex" or "display: grid".
- Grids may only be used if there are at least 2 columns and 2 rows.
- Flex containers should always set "flex-wrap: nowrap". Flex direction should be "row" or "column". Do not use reverse.
- Do not use flex item reordering.
- Prefer to use "row-gap", "column-gap", "gap", "justify-content", and "align-items" to control spacing and alignment of elements.
- Prefer to use "flex-basis", "flex-grow", and "flex-shrink" to control the size of flex items.
- All elements should have "box-sizing: border-box". Do not use padding unless an element has a visible border, i.e. prefer margin to padding.
- Prefer specifying dimensions in percentages rather than pixels unless the dimension is small.
- No absolute positioning of elements.
Tips!!! Use these principles to create an aesthetically pleasing document:
* Proper content alignment is key.
* Use hierarchy to help focus your design.
* Leverage contrast to accentuate important design elements.
* Choose fonts and font sizes that are stylistically appropriate.
* Not all text needs to be the same font or size.
* Use repetition to create a cohesive look.
* Consider proximity when organizing your graphic elements.
* Make sure your design is balanced.
* Carefully select colors that are diverse, visually appealing, and appropriate for the theme.
* Leave negative space.

B. Supervised Fine-Tuning Input-Output Examples

I2D Input
{“width”: 3300, “height”: 2071, “category”: “planner”, “styles”: [“natural”], “intention”: “The document is intended to provide a structured template for outlining the details of a course, including the subject, days, teacher, and additional course details.”}
DD Input
{“width”: 3300, “height”: 2071, “screenshot”: “<image>”}
E2D Input
{“width”: 3300, “height”: 2071, “texts”: [“Dr. Evelyn Green”, “Wednesday”, “Materials”, “Additional Information”, “Duration”, “Prerequisites”, “Office Hours: Tue/Thu 1-3 PM”, “12 Weeks”, “Subject”, “evelyn.green@university.edu”, “2:00 PM - 4:00 PM”, “Course Details”, “10:00 AM - 12:00 PM”, “Botany & Ecology”, “Advanced”, “Advanced Botany Studies”, “Learning Outcomes”, “Intro to Biology”, “Course Planner Template - Natural Theme”, “Schedule”, “Upon completion of this course, students will be able to analyze plant physiological processes, evaluate ecosystem dynamics, conduct botanical research using scientific methods, and apply knowledge of plant biology to environmental conservation efforts.”, “10:00 AM - 12:00 PM”, “Greenhouse Lab B”, “Monday”, “Professor of Botany”, “Lab Manual & Field Kit”, “Exploring Plant Life and Ecosystems”, “Level”, “This comprehensive course explores the intricate relationships between plant life and their ecosystems. Students will examine plant physiology, taxonomy, and ecological roles through hands-on laboratory work and field studies. Topics include photosynthesis mechanisms, plant adaptations, biodiversity conservation, and sustainable gardening practices. The course emphasizes critical thinking and scientific methodology through research projects and collaborative learning experiences.”, “Projects & Exams”, “Fall Semester 2023”, “4.0”, “Course Description”, “Instructor”, “Assessment”, “Credits”, “Location”, “Office: Science Building, Room 304”, “Friday”], “images”: {“image_0”: “<image>”, “image_1”: “<image>”, “image_2”: “<image>”, “image_3”: “<image>”, “image_4”: “<image>”, “image_5”: “<image>”, “image_6”: “<image>”, “image_7”: “<image>”}}
Common Output

```
<!DOCTYPE html>
<html>
  <head>
    <style>
        * {
            box-sizing: border-box;
            margin: 0;
        }

        body {
            width: 3300px;
            height: 2071px;
            display: flex;
            flex-direction: column;
            ...
        }

        .header {
            height: 20%;
            display: flex;
            flex-direction: row;
            align-items: center;
            ...
        }

        ...
    </style>
  </head>
  <body>
    <div class="header">
      <div class="logo-container">
        <!-- I2D or DD img tag -->
        <img src="https://picsum.photos/200/200" alt="Natural green leaf logo representing growth
            and learning">
        <!-- E2D img tag -->
        <img src="image_6_200x200.png" alt="Natural green leaf logo representing growth and
            learning">
      </div>
      ...
    </div>
    ...
  </body>
</html>
```

C. More Baseline Implementation Details

GPT-4o and InternVL3-78B. We use these two general-purpose MLLMs in the zero-shot setting. The prompt is almost the same as the code generation prompt used in data construction (Section A.2), with the addition of task-specific instructions tailored to each generation task. The instructions are: (1) I2D: *Generate a high-quality document that accurately reflects the specified category, style, and intention.* (2) DD: *Convert the input document screenshot into corresponding HTML/CSS code that faithfully reproduces the visual layout and content.* (3) E2D: *Synthesize a coherent document by incorporating all provided elements exactly once, ensuring proper layout and visual harmony.*

FLUX.1-dev. We also use FLUX.1-dev as one of the baselines for I2D. The text prompt is *A flat, digital, single page {category} document in a top-down view, clearly presenting text and image content according to the intention: {intention}. Use {styles} styles.*

LaDeCo. LaDeCo serves as the baseline for E2D, employing a five-layer hierarchical framework consisting of background, underlay, image, text, and embellishment layers. However, as illustrated in Section B, our task formulation focuses on two element types: image and text. Hence, we adapt LaDeCo by utilizing only its image and text layers while leaving the remaining three layers empty (containing no elements) for fair comparison.

D. Dataset Statistics

Figure 8 visualizes the distributions of document categories and styles within DocHTML. The element count and document aspect ratio distributions are shown in Figure 9. From the figures, we can see that DocHTML covers a broad range of document categories, styles, layout complexities, and aspect ratios.

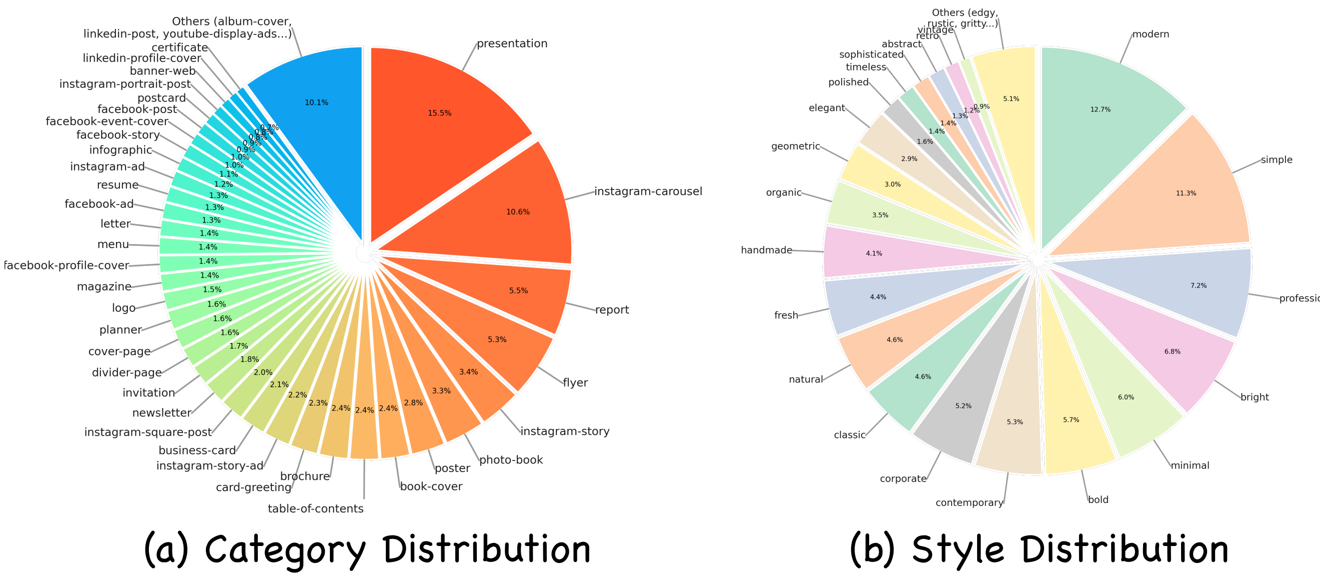


Figure 8. Distributions of document categories and styles in DocHTML.

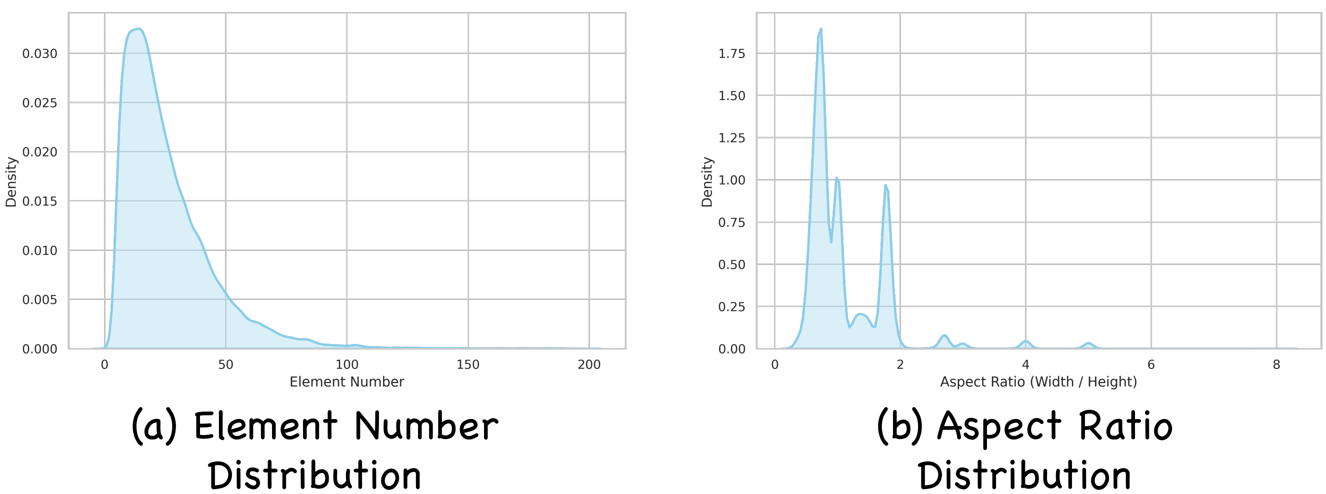
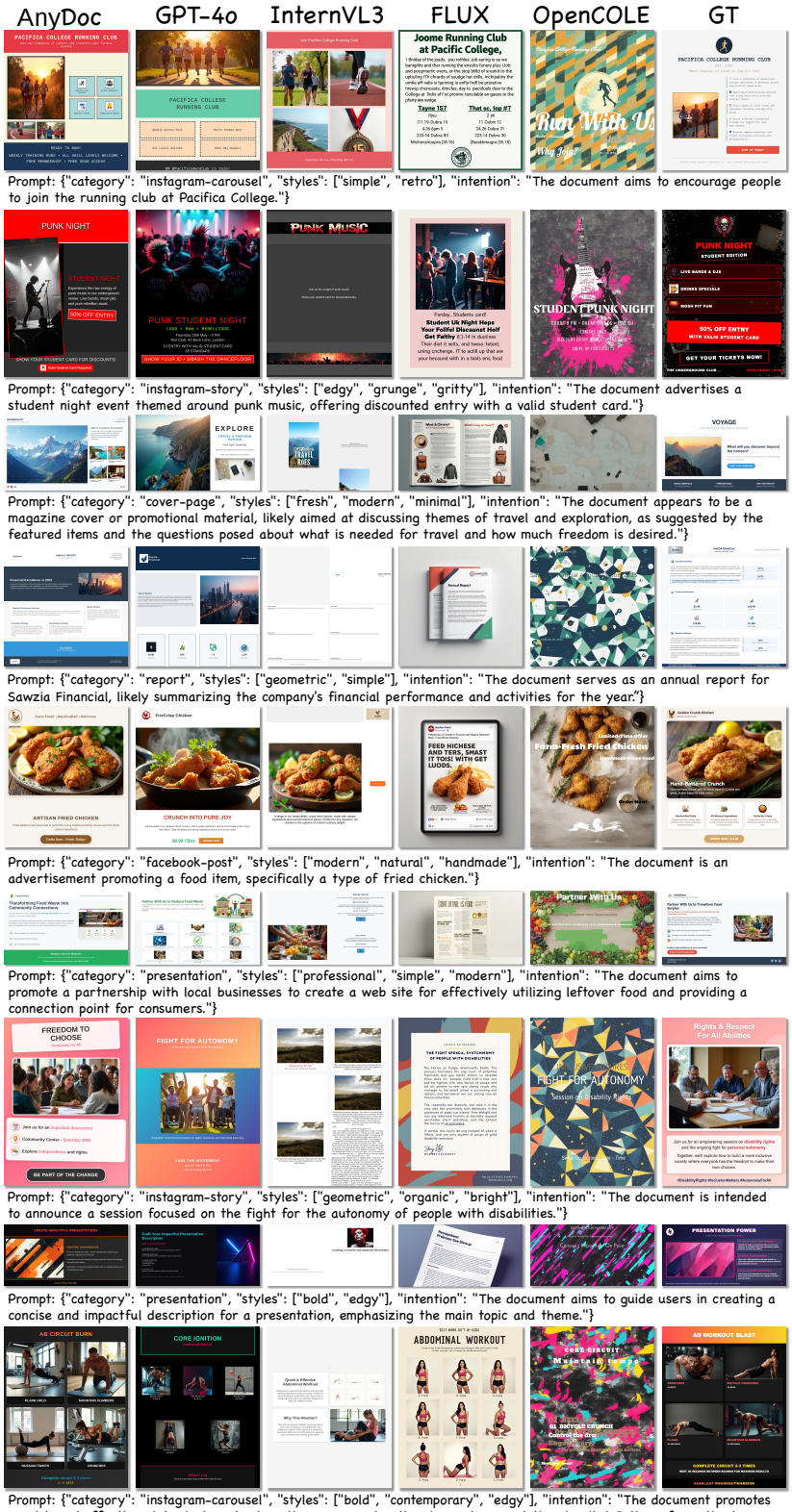


Figure 9. Distributions of document element count and document aspect ratio in DocHTML.

E. More Qualitative Results

In Figures 10, 11, 12, we show more qualitative comparison of the I2D, DD, and E2D tasks, respectively. AnyDoc exhibits remarkable document generation performance compared to the baseline models.



Prompt: {"category": "instagram-carousel", "styles": ["simple", "retro"], "intention": "The document aims to encourage people to join the running club at Pacifica College."}

Prompt: {"category": "instagram-story", "styles": ["edgy", "grunge", "gritty"], "intention": "The document advertises a student night event themed around punk music, offering discounted entry with a valid student card."}

Prompt: {"category": "cover-page", "styles": ["fresh", "modern", "minimal"], "intention": "The document appears to be a magazine cover or promotional material, likely aimed at discussing themes of travel and exploration, as suggested by the featured items and the questions posed about what is needed for travel and how much freedom is desired."}

Prompt: {"category": "report", "styles": ["geometric", "simple"], "intention": "The document serves as an annual report for Sawzia Financial, likely summarizing the company's financial performance and activities for the year."}

Prompt: {"category": "facebook-post", "styles": ["modern", "natural", "handmade"], "intention": "The document is an advertisement promoting a food item, specifically a type of fried chicken."}

Prompt: {"category": "presentation", "styles": ["professional", "simple", "modern"], "intention": "The document aims to promote a partnership with local businesses to create a web site for effectively utilizing leftover food and providing a connection point for consumers."}

Prompt: {"category": "instagram-story", "styles": ["geometric", "organic", "bright"], "intention": "The document is intended to announce a session focused on the fight for the autonomy of people with disabilities."}

Prompt: {"category": "presentation", "styles": ["bold", "edgy"], "intention": "The document aims to guide users in creating a concise and impactful description for a presentation, emphasizing the main topic and theme."}

Prompt: {"category": "instagram-carousel", "styles": ["bold", "contemporary", "edgy"], "intention": "The document promotes a quick and effective abdominal workout routine, encouraging the viewer to repeat the circuit 2-3 times for optimal results."}

Figure 10. More qualitative results on the intention-to-document task.

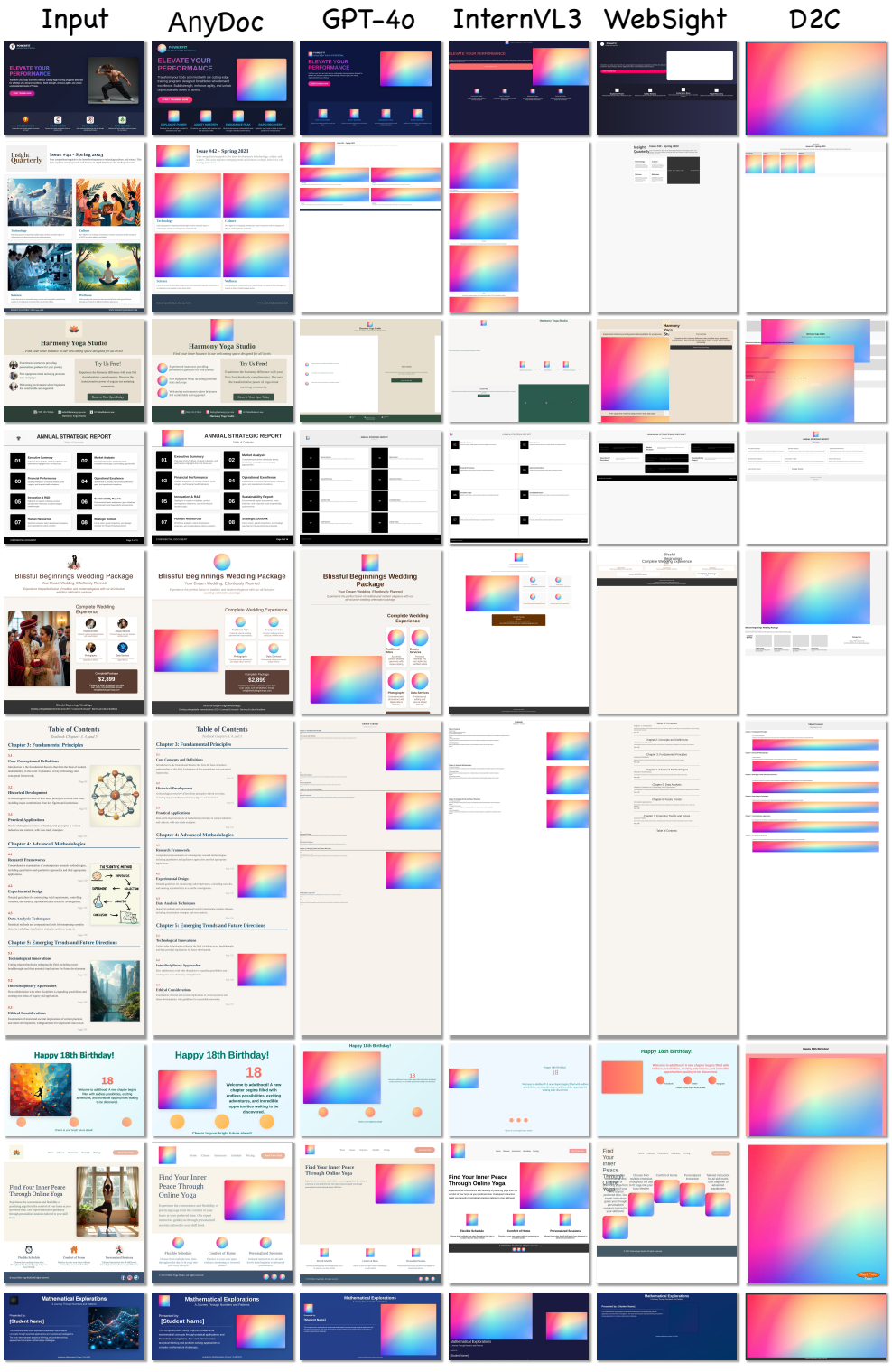


Figure 11. More qualitative results on the document derendering task.

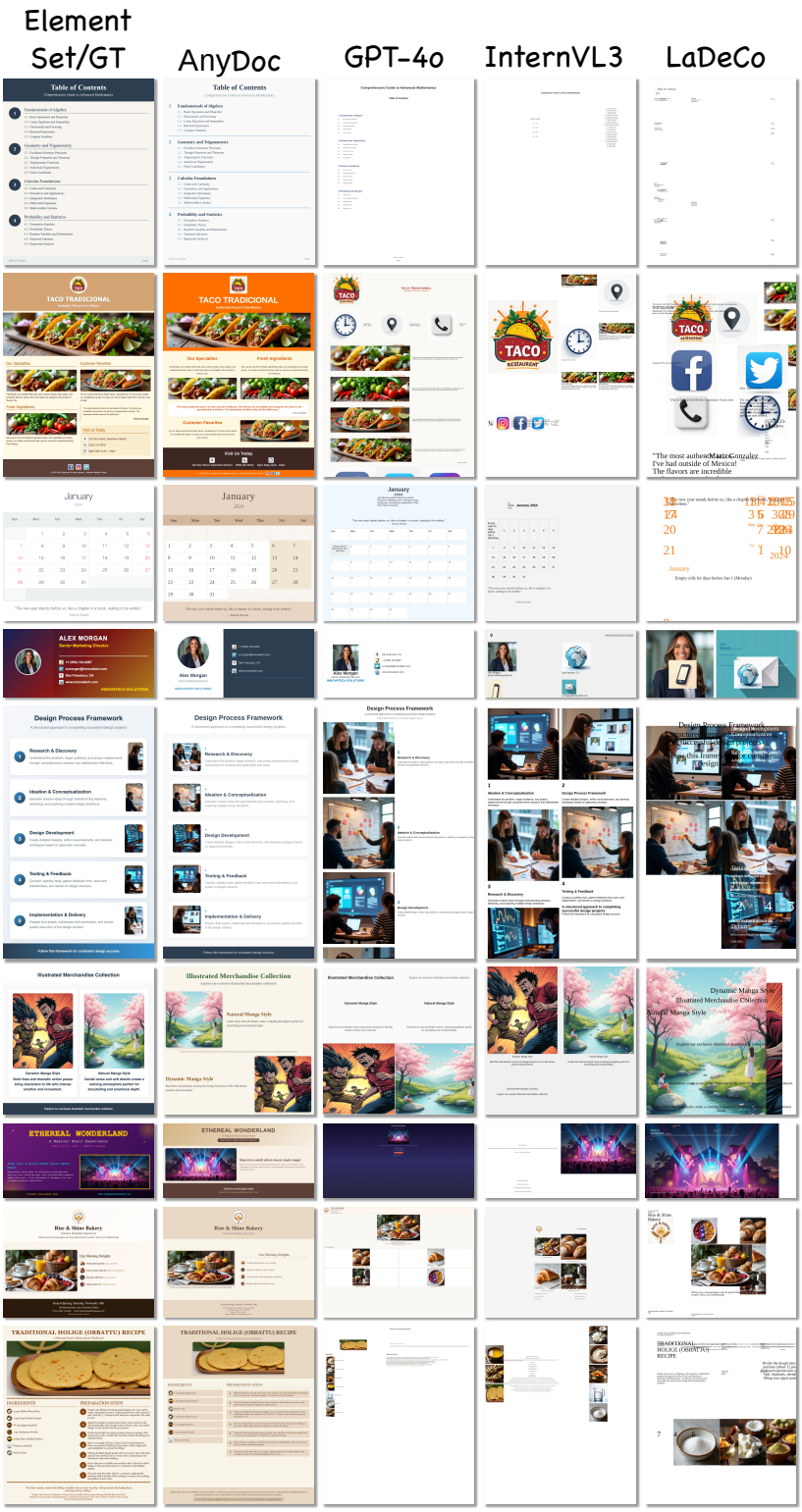


Figure 12. More qualitative results on the element-to-document task.