7 APPENDIX

7.1 Extracting Items from Comments

We use ChatGPT (gpt-3.5-turbo-0125) to extract items from comments. We use the following prompts for books and music, respectively:

> In the following text, extract all books. Each book should be in the Author - Title format and separated by a newline. If an author is mentioned without the title, output the author only. If there are no mentions, output None.

> In the following text, extract all music tracks. Each track should be in the Artist - Title format and separated by a newline. If an artist is mentioned without the title, output the artist only. If there are no mentions, output None.

We then further process the responses to get a list of items. The following are examples of comments to extracted items.

- **Comment:** Book of A Thousand Days by Shannon Hale fits! It's one of my all-time favorites.
  → **Items:** Shannon Hale - Book of A Thousand Days
- **Comment:** For some reason, the song You Get What You Give by New Radicals popped in my head when looking at this picture.
  → **Items:** New Radicals - You Get What You Give
- **Comment:** Anything by Astro Man.
  → **Items:** Astro Man
- **Comment:** Lady of the Forest by Jennifer Roberson Rose Daughter and Spindle's End by Robin McKinley
  → **Items:** Jennifer Roberson - Lady of the Forest, Robin McKinley - Rose Daughter, Robin McKinley - Spindle's End

7.2 Statistics

We obtain 1,470 requests and 12,208 items for *books*; 796 requests and 38,204 items for *music*. Each request has an average of 3.11 (±3.14) images and 14.04 (±12.55) recommended items for *books*; 1.41 (±1.74) images and 65.92 (±101.00) items for *music*. Figure 4 shows the number of requests per year. There is a notable upward trend in the number of requests in both domains, with the counts for the initial two months of 2024 surpassing those of any full previous year. Figure 5 shows the number of images per request. Users tend to include multiple images for *books*, while most use just one image for *music*. Figure 4 plots the number of recommended items per request, showing long-tail distributions. It's not unusual to see requests receive recommendations for dozens of items, and specifically in *music*, some requests can accumulate more than a hundred items.

7.3 Prompts

We enlist all the prompts we use in our tasks.

7.3.1 *Title generation: standard prompting (books).*

> Given the request, provide recommendations. Enumerate 20 books (1., 2., ...) in the order of relevance. Each book should take the Author - Title format. Don't say anything else.

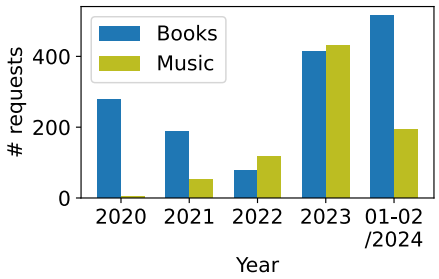

Figure 4: Requests per year

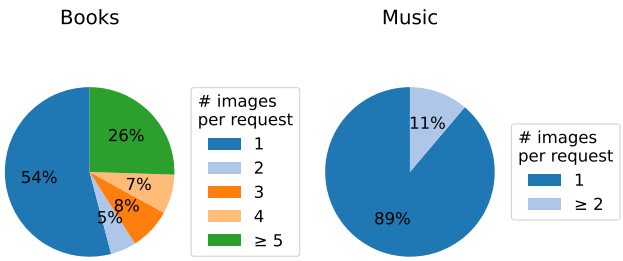

Figure 5: Images per request

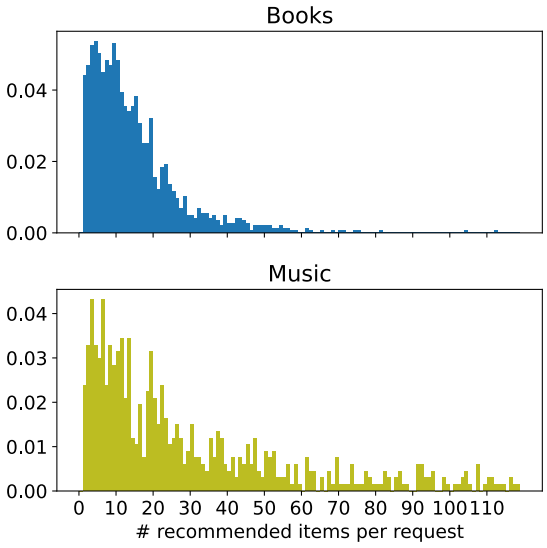

Figure 6: Recommended items per request

7.3.2 *Title generation: standard prompting (music).*

> Given the request, provide recommendations. Enumerate 20 music pieces (1., 2., ...) in the order of

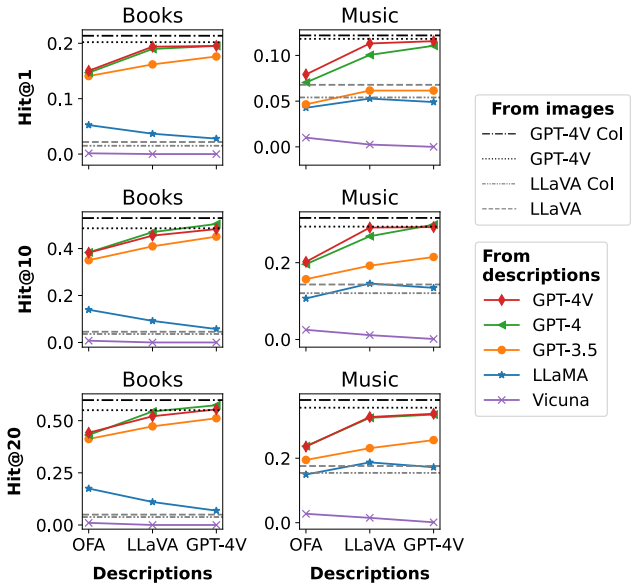

Figure 7: Results for title generation. Each row shows Hit@1, Hit@10, Hit@20 results, and each column shows *books* and *music* datasets, respectively. Vision models that take images as input are plotted in constant dotted lines; models that use only text descriptions (including GPT-4V without image input) are marked based on the type of description (x-axis).

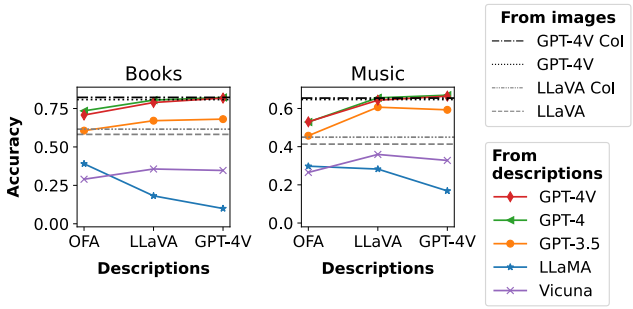

Figure 8: Results for multiple-choice selection. The plots are drawn in a similar style as in Figure 7.

Table 5: Chain-of-imagery (CoI) prompting results for *text-only* GPT-4 (from GPT-4V descriptions). Unlike in the case of GPT-4V (from images), CoI does not always improve results.

Dataset	Model	Hit@1	@10	@20	Acc (%)
Books	GPT-4 + CoI	.2034	.5102	.5946	81.70
	GPT-4	.1952	.5048	.5741	82.18
Music	GPT-4	.0867	.2663	.3077	66.45
	GPT-4	.1005	.2990	.3354	66.96

relevance. Each piece should take the Artist - Title format. Don't say anything else.

7.3.3 *Title generation: CoI prompting (books).*

Given the request, provide recommendations. Think step by step. First understand the given image(s) in detail, including its content, style, and vibe. Then, think of books that capture the essence of the image(s). Enumerate 20 books (1., 2., ...) in the order of relevance. Each book should take the Author - Title format. Don't say anything else.

7.3.4 *Title generation: CoI prompting (music).*

Given the request, provide recommendations. Think step by step. First understand the given image(s) in detail, including its content, style, and vibe. Then, think of music pieces that capture the essence of the image(s). Enumerate 20 music pieces (1., 2., ...) in the order of relevance. Each piece should take the Artist - Title format. Don't say anything else.

7.3.5 *Multiple-choice selection: standard prompting.*

Given the request, choose which recommendation is the most suitable. Choose a single item. Don't say anything else.

7.3.6 *Multiple-choice selection: CoI prompting.*

Given the request, choose which recommendation is the most suitable. Think step by step. First understand the given image(s) in detail, including its content, style, and vibe. Then, carefully inspect each choice. Recall its content and see if it captures the essence of the image(s). Select the item that best captures this essence. Choose a single item. Don't say anything else.

7.4 More results

We summarize all results for title generation in Figure 7, multiple-choice selection in Figure 8, and chain-of-imagery (CoI) prompting for text-only GPT-4 in Table 5. These results contain the information that support the takeaway messages of the paper:

- M1 Models struggle, particularly smaller ones.
- M2 Only larger models benefit from detailed descriptions.
- M3 Using descriptions can be better than using images.
- M4 Chain-of-imagery (CoI) prompting may help VLMs better harness their visual capabilities.

Please refer to the main paper for detailed explanations of each.